GENERALIZATION BOUNDS WITH ARBITRARY COMPLEXITY MEASURES

Anonymous authors
Paper under double-blind review

ABSTRACT

In statistical learning theory, generalization bounds usually involve a complexity measure that is constrained by the considered theoretical framework. This limits the scope of such analysis, as in practical algorithms, other forms of regularization are used. Indeed, the empirical work of Jiang et al. (2019) shows that *(i)* common complexity measures (such as the VC-dimension) do not correlate with the generalization gap and that *(ii)* there exist *arbitrary* complexity measures that are better correlated with the generalization gap, but come without generalization guarantees. In this paper, we bridge the gap between this line of empirical works and generalization bounds of statistical learning theory. To do so, we leverage the framework of disintegrated PAC-Bayes bounds to derive a generalization bound that involves an *arbitrary* complexity measure. Our bound stands in probability jointly over the hypotheses and the learning sample, which allows us to tighten the complexity for a given generalization gap since it can be set to fit both the hypothesis class and the task.

1 INTRODUCTION

Statistical learning theory offers various theoretical frameworks to assess generalization by studying whether the empirical risk is representative of the true risk thanks to an upper bounding strategy of the generalization gap. The generalization gap represents a deviation between the true risk and the empirical risk. An upper bound on this gap is generally a function of two main quantities: *(i)* the size of the training sample and *(ii)* a complexity measure that captures how prone a model is to overfitting. One potential limitation is that existing frameworks are restricted to particular complexity measures, among them the VC-dimension (Vapnik & Chervonenkis, 1971) or the Rademacher complexity (Bartlett & Mendelson, 2002) for which some generalization bounds can be derived.

In order to study generalization capabilities, a recent line of works, mainly in the context of neural network learning, is dedicated to the empirical study of different complexity measures to find those that correlate the most with the generalization gap (Jiang et al., 2019; Dziugaite et al., 2020; Jiang et al., 2021). For example, given an arbitrary complexity measure, Jiang et al. propose to estimate a level of correlation between the generalization gap and the complexity measure for over-parametrized models empirically. They are able to rank the measure and the gap to obtain Kendall's rank coefficient (Kendall, 1938) to evaluate how a measure reflects the generalization capability of a particular model. However, this correlation coefficient was criticized by Dziugaite et al. (2020) because they found that if the measures empirically correlate well *on average*, they are not robust to changes in the fixed parameters (such as the depth or the width of the model).

On the one hand, while these results are extremely important to understand generalization, notably for over-parametrized models, they remain limited in the sense that they provide only empirical assessments. On the other hand, the bounds in the literature are restrictive because they cannot easily integrate user-defined complexity measures. In other words, to the best of our knowledge, there is no generalization bound able to take into account by construction some arbitrary complexity measures that can serve as good proxies for the generalization gap. This paper aims to bridge this gap by providing a general framework for deriving generalization bounds with arbitrary complexity measures. We believe that this direction is of important interest in advancing the understanding of generalization, as the generalization gap can be then *provably* upper-bounded with a term that depends on a user-specified complexity measure. For this purpose, our contribution leverages the framework of

disintegrated PAC-Bayesian bound (Theorem 2.1). This framework allows us to derive guarantees in terms of probabilistic bounds that depend on a model distribution involving a parametric function characterizing the complexity measure, the learning sample and the aforementioned parametric function. Our novel results provide theoretical foundations for the many regularizations used in practice to perform model selection (*e.g.*, L2 regularization) and are also related to uniform convergence and algorithm dependent bounds.

This paper is organized as follows. In Section 2, we provide some preliminaries definitions and concepts. Then, we present our main contribution in Section 3. In Section 4, we provide a practical instantiation of our framework, before concluding in Section 5.

2 Preliminaries

2.1 Setting

We consider the classic supervised learning setting where $\mathbb{X}$ denotes the input space and $\mathbb{Y}$ the label space. We consider that an example $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$ is sampled from an unknown data distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$. A learning sample $\mathcal{S}=\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ contains m examples drawn *i.i.d.* from $\mathcal{D}$; we denote the distribution of such an m-sample by $\mathcal{D}^m$. Let $\mathbb{H}$ be a potentially infinite set of functions $h : \mathbb{X}\to\mathbb{Y}$, called hypotheses (or models), that associate a label from $\mathbb{Y}$ given an input from $\mathbb{X}$. Let $\mathbb{M}(\mathbb{H})$ be the set of probability densities over $\mathbb{H}$ given a reference measure (*e.g.*, the Lebesgue measure); we denote by $\mathbb{M}^*(\mathbb{H}) \subseteq \mathbb{M}(\mathbb{H})$ the set of strictly positive probability densities.
Given a learning sample $\mathcal{S}$, we aim to find $h \in \mathbb{H}$ that minimizes the so-called true risk $\mathrm{R}_{\mathcal{D}}(h)=\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\mathrm{I}\left[h(\mathbf{x}) \neq y\right]$, where $\mathrm{I}[a]=1$ if a is true, and 0 otherwise. In practice, as the data distribution $\mathcal{D}$ is unknown, we estimate the true risk with its empirical counterpart: the empirical risk $\mathrm{R}_{\mathcal{S}}(h) = \frac{1}{m}\sum_{i=1}^m \mathrm{I}\left[h(\mathbf{x}_i) \neq y_i\right]$. We hereafter denote the generalization gap by $\phi : [0, 1]^2 \to \mathbb{R}$, which is usually defined by $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) = |\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathcal{S}}(h)|$ that quantifies how much the empirical risk is representative of the true risk.

In this paper, we leverage the PAC-Bayesian framework Shawe-Taylor & Williamson (1997); McAllester (1998); Guedj (2019); Alquier (2021) to upper-bound the generalization gap with a function that depends on an *arbitrary* measure of complexity. In PAC-Bayes, we consider an *apriori* belief on the hypotheses in $\mathbb{H}$ that is modeled by a prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$. We aim to learn, from $\mathcal{S}$ and π, a *posterior* distribution $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$ to assign higher probability to the best hypotheses in $\mathbb{H}$ (the support of ρ being included in the support of π). The classical PAC-Bayesian generalization bounds provide upper bounds in expectation over ρ, meaning that they bound the generalization gap expressed as $|\,\mathbb{E}_{h\sim\rho}[\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathcal{S}}(h)]\,|$, and where the complexity term depends on the KL divergence between ρ and π defined as $\mathbb{E}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)}$. This standard complexity hence captures how much the prior and the posterior distribution deviate in expectation over all the hypotheses. To incorporate arbitrary complexities in the bounds, we follow a slightly different framework recalled below (the disintegrated PAC-Bayesian bounds) in which the expectations on ρ are "disintegrated": the gap $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h))=|\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathcal{S}}(h)|$ of a single h sampled from ρ is considered in the bounds.

2.2 Disintegrated PAC-Bayesian Bounds

The disintegrated PAC-Bayesian bounds have been introduced by Catoni (2007, Th 1.2.7) and Blanchard & Fleuret (2007, Prop 3.1). As far as we know, despite their interest, they have been little used in the literature and received only recently renewed interest for deriving tight bounds in practice (*e.g.*, Rivasplata et al. (2020); Viallard et al. (2021)). Such bounds depend on a posterior distribution $\rho_{\mathcal{S}}$ obtained after obtaining the learning sample $\mathcal{S}$ and have the following form.

Definition 2.1 (Disintegrated PAC-Bayesian Generalization Bound)**.** *Let $\phi : [0, 1]^2\to[0, 1]$ be a generalization gap. A disintegrated PAC-Bayesian bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, there exists a function $\Phi : \mathbb{M}(\mathbb{H})\times\mathbb{M}^*(\mathbb{H})\times(0, 1]\to\mathbb{R}$ such that for any $\delta \in (0, 1]$ we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\phi(R_{\mathcal{D}}(h), R_{\mathcal{S}}(h)) \leq \Phi\big(\rho_{\mathcal{S}}, \pi, \delta\big)\right] \geq 1 - \delta,$$

where $\rho_{\mathcal{S}}$ is a posterior distribution $\rho_{\mathcal{S}} \in \mathbb{M}(\mathbb{H})$, and, e.g., $\phi(R_{\mathcal{D}}(h), R_{\mathcal{S}}(h)) = |R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)|$.

Put into words, given a training set $\mathcal{S} \sim \mathcal{D}^m$, we can learn the distribution $\rho_{\mathcal{S}}$ and then sample the hypothesis h from $\rho_{\mathcal{S}}$ to get a bound with high probability (at least $1-\delta$) over the random choice of $\mathcal{S}$ and h. This paper mainly focuses on a particular bound, namely, the one of Rivasplata et al. (2020) recalled below.

Theorem 2.1 (General Disintegrated Bound of Rivasplata et al. (2020))**.** *For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any distribution $\pi \in \mathbb{M}^*(\mathbb{H})$, for any measurable function $\varphi : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0,1]$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\varphi(h,\mathcal{S}) \le \underbrace{\ln\left[\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right] + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\pi}\exp\left(\varphi(h',\mathcal{S}')\right)\right]}_{\Phi(\rho_{\mathcal{S}},\pi,\delta)}\right] \ge 1-\delta,$$

where $\rho_{\mathcal{S}}$ is a posterior distribution such that $\rho_{\mathcal{S}} \in \mathbb{M}(\mathbb{H})$.

In this case, the function $\varphi(h,\mathcal{S}) = m\,\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h))$ is a deviation between the true risk $\mathrm{R}_{\mathcal{D}}(h)$ and the empirical risk $\mathrm{R}_{\mathcal{S}}(h)$. Moreover, the function $\Phi(\rho_{\mathcal{S}},\pi,\delta)$ is constituted of 2 terms: *(i)* the *disintegrated* KL divergence $\ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}$ defining how much the prior and posterior distributions deviate for a single h, and *(ii)* the term $\ln\big[\frac{1}{\delta}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\pi}\exp\left(\varphi(h',\mathcal{S}')\right)\big]$ which is constant *w.r.t.* $h \in \mathbb{H}$ and $\mathcal{S} \in (\mathbb{X}\times\mathbb{Y})^m$ and usually upper-bounded to instantiate the bound. In the following, we refer to the whole right-hand side of the bound, $\Phi()$, as the *complexity measure* for the sake of simplicity. Note that this is in slight contrast with the standard definitions of complexity, where the term *(ii)* (related to δ and the sample size m) is not included. This additional term is, in fact, constant *w.r.t.* the hypothesis $h\sim\rho_{\mathcal{S}}$ and the learning sample $\mathcal{S}\sim\mathcal{D}^m$.

In the bound of Theorem 2.1, the complexity term $\Phi()$ depends on the disintegrated KL divergence and suffers from drawbacks: the KL complexity term is imposed by the framework and can be subject to high variance in practice (Viallard et al., 2021). However, it is important to notice that this disintegrated KL divergence has a clear advantage: it only depends on the hypothesis h and $\mathcal{S}$, instead of the whole hypothesis class (as it is often the case for instance with the KL divergence in PAC-Bayesian bounds, or the VC-dimension). This might imply a better correlation between the generalization gap and some complexity measures. In the next section, we leverage this disintegrated KL divergence to derive our main contribution: a general bound that involves arbitrary complexity measures.

3 Integrating Arbitrary Complexities in Generalization Bounds

We first begin with a short presentation of our result to give some preliminary intuitions and to introduce the notion of Gibbs distribution which is a key element in the exposition of our contribution. We then formalize our theoretical result in Section 3.3.

3.1 An Introduction to our Results

Let $\Phi_\mu(h,\mathcal{S},\delta)$ be a real-valued function that takes a hypothesis $h \in \mathbb{H}$, a learning sample $\mathcal{S} \in (\mathbb{X}\times\mathbb{Y})^m$, and the parameter δ as arguments and that is dependent on an additional function $\mu : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$. The idea is to use this function $\mu()$ to parametrize the complexity measure with respect to the data sample $\mathcal{S}$ and the model h, in order to introduce arbitrary complexity measures in the bound; we call "*parametric function*" the function $\mu()$. As an example, when $\mathbb{H}$ is a set of hypotheses $h_{\mathbf{w}}$ parameterized by some weights $\mathbf{w} \in \mathbb{R}^d$, we can fix $\mu(h_{\mathbf{w}},\mathcal{S}) = \|\mathbf{w}\|$, for some norm $\|\cdot\|$. This means that $\mu(h_{\mathbf{w}},\mathcal{S})$ can be set to the regularization term of the chosen objective function so that the complexity, hence the bound, will depend on it. This is not really new since, for example, uniform stability bounds allow one to consider such norms (see, *e.g.*, Kakade et al., 2008). Here we want to illustrate that we simply incorporate some norms of parameters, but as the reader may guess, our framework allows one to consider a broader family of complexity measures, as we will see later. Given such a parametric function $\mu()$, the bound we derive in Theorem 3.1 takes the following form.

Definition 3.1 (Generalization Bound with Complexity Measures)**.** *Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap, $\mu : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$ be a parametric function. A generalization bound with arbitrary complexity measures is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$,*

there exists a real-valued function $\Phi_\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \times (0,1] \to \mathbb{R}$ *such that for any* $\delta \in (0,1]$*, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h\sim\rho_\mathcal{S}} \left[\phi(R_\mathcal{D}(h), R_\mathcal{S}(h)) \leq \Phi_\mu(h, \mathcal{S}, \delta) \right] \geq 1-\delta. \tag{1}$$

The main trick to obtain such a result is to consider a particular posterior distribution $\rho_\mathcal{S}$: we incorporate the function $\mu()$ by fixing the distribution $\rho_\mathcal{S}$ as a Gibbs distribution defined by

$$\rho_\mathcal{S}(h) \propto \exp\left[-\alpha \mathrm{R}_\mathcal{S}(h) - \mu(h, \mathcal{S})\right], \quad \text{where } \alpha \in \mathbb{R}^+. \tag{2}$$

This Gibbs distribution $\rho_\mathcal{S}$ is interesting from an optimization viewpoint: a hypothesis h is more likely to be sampled from it when the objective function $h \mapsto \mathrm{R}_\mathcal{S}(h) + \frac{1}{\alpha}\,\mu(h, \mathcal{S})$ is low for a given $\mathcal{S}$. For example, when $\mu(h, \mathcal{S}) = 0$, a hypothesis h is more likely to be sampled when its empirical risk $\mathrm{R}_\mathcal{S}(h)$ is low. Conversely, when $\mu()$ is non-constant, the function serves as a "regularizing term", so that a hypothesis is more likely to be sampled when the trade-off $\mathrm{R}_\mathcal{S}(h) + \frac{1}{\alpha}\,\mu(h, \mathcal{S})$ is low. In both cases, the higher α, the more the density of $\rho_\mathcal{S}$ is concentrated around those hypotheses that minimize the empirical risk. Moreover, it seems that Equation (2) is restrictive, but it can actually represent any probability density functions. Indeed, let $\rho'_\mathcal{S}$ be a distribution on $\mathbb{H}$, *e.g.*, a Gaussian or a Laplace distribution, by setting $\mu(h, \mathcal{S}) = -\alpha \mathrm{R}_\mathcal{S}(h) - \ln \rho'_\mathcal{S}(h)$ we can retrieve the distribution $\rho'_\mathcal{S}$. Actually, the Gibbs distribution is well-known and studied in learning theory, as discussed below.

3.2 About the Gibbs Distribution

In this section, we would like to highlight two main lines of works that are related to our setting: *(i)* the usage of the Gibbs distribution in the *classical* PAC-Bayesian theory and *(ii)* the link between this distribution and optimization.

Gibbs distribution in the PAC-Bayesian theory. The Gibbs distribution has started to be studied in the PAC-Bayesian theory by Catoni (2004; 2007). Moreover, Alquier et al. (2016, Theorems 4.2 & 4.3) developed PAC-Bayesian generalization bounds with the Gibbs distribution of Equation (2) with $\mu(h, \mathcal{S}) = 0$ as posterior. However, their theorems analyze the expected true risk $\mathbb{E}_{h\sim\rho_\mathcal{S}}\, \mathrm{R}_\mathcal{D}(h)$ while we are interested in a *single* hypothesis h sampled from $\rho_\mathcal{S}$. Moreover, their bounds involve the non-computable and hypothesis-set-dependent KL divergence between the Gibbs distribution and a prior distribution. Hence, the computation of the KL divergence must be upper-bounded to allow one to instantiate their bounds in practice. We will see further that the bounds of Theorem 3.1 and Corollary 3.1 do not have this issue since they only require to know the expression of the density (up to a normalization constant) for $h \sim \rho_\mathcal{S}$ and $h' \sim \pi$.

Relationship between optimization and the Gibbs distribution. Given an objective function $f : \mathbb{R}^D \to \mathbb{R}$, the Stochastic Gradient Langevin Dynamics (SGLD) algorithm (Welling & Teh, 2011) learns some parameter $\mathbf{w} \in \mathbb{R}^D$ by running several iterations of the form

$$\mathbf{w}_t \longleftarrow \mathbf{w}_{t-1} - \beta \nabla f(\mathbf{w}) + \sqrt{\frac{2\beta}{\alpha}}\epsilon_t, \quad \text{with} \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D), \tag{3}$$

where $\mathbf{w}_t$ is the weight vector learned at the iteration $t \in \mathbb{N}$, the parameter β is the learning rate, and α is the concentration parameter of the Gibbs distribution. This algorithm has an interesting feature: when the learning rate β tends to zero, the SGLD algorithm becomes a continuous-time process called Langevin diffusion, defined as the following stochastic differential equation in Equation (4). Indeed, Equation (3) can be seen as the Euler-Maruyama discretization (see, *e.g.*, Raginsky et al., 2017) of Equation (4) defined for $t \geq 0$ as

$$d\mathbf{w}(t) = -\nabla f(\mathbf{w}(t))dt + \sqrt{2\alpha}\mathbf{B}(t), \tag{4}$$

where $\mathbf{B}(t)$ is the Brownian motion in $\mathbb{R}^D$. Under some mild assumptions on the function $f()$, Chiang et al. (1987) show that the invariant distribution of the Langevin diffusion is the Gibbs distribution proportional to $\exp(\alpha f(\mathbf{w}))$.

3.3 Our Main Result: Generalization Bound with Complexity Measures

Thanks to the definition of $\rho_\mathcal{S}$, we now state our main result: a bound on the generalization gap involving our parametric measure μ with respect to a posterior $\rho_\mathcal{S}(h) \propto \exp\left[-\alpha \mathrm{R}_\mathcal{S}(h) - \mu(h, \mathcal{S})\right]$.

Theorem 3.1 (Generalization Bound with Complexity Measures)**.** *Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap. For any $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0,1]$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi,\ h\sim\rho_{\mathcal{S}}} \Bigg[\phi(R_{\mathcal{D}}(h), R_{\mathcal{S}}(h)) \leq \Big[\alpha R_{\mathcal{S}}(h') + \mu(h',\mathcal{S})\Big] - \Big[\alpha R_{\mathcal{S}}(h) + \mu(h,\mathcal{S})\Big] + \ln \frac{\pi(h')}{\pi(h)} + \ln \left(\frac{4}{\delta^2} \mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h'\sim\pi} \exp\left[\phi(R_{\mathcal{D}}(h'), R_{\mathcal{S}'}(h'))\right] \right) \Bigg] \geq 1-\delta,$$

where $\rho_{\mathcal{S}}$ is the Gibbs distribution defined by Equation (2).

This theorem is general since it depends only on the functions $\phi()$ and $\mu()$ that the user must fix. Moreover, we can show that this theorem generalizes uniform-convergence-based and algorithm-dependent bounds with the integration of some complexity measures. The justification of this claim is deferred to Appendix D, and to the best of our knowledge, this result is novel.

Given $\phi()$ and $\mu()$, we note an element that can be surprising at the first reading: it appears indeed possible to sample hypotheses with a high objective $\mathrm{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\mu(h,\mathcal{S})$ value and to obtain a tight generalization bound. However, by definition of the Gibbs distribution $\rho_{\mathcal{S}}$, such a sampled hypothesis $h \sim \rho_{\mathcal{S}}$ is less probable since the density is higher when the objective is low. On the other hand, it is highly probable to sample a hypothesis h from $\rho_{\mathcal{S}}$ with a low objective $\mathrm{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\mu(h,\mathcal{S})$ value. In other words, when $\mu(h,\mathcal{S})$ acts as a regularizer, the bound holds more likely for the hypotheses achieving a low regularized empirical risk, which is a rather reasonable expected result when considering regularized learning.

Concerning the tightness of the bound, it may appear loose. However to get a bound that converges when m increases, it is sufficient to fix $\phi()$ as a function of m such as $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) = m\,\mathrm{kl}[\mathrm{R}_{\mathcal{S}}(h) \| \mathrm{R}_{\mathcal{D}}(h)]$ or $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) = 2m[\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)]^2$. Then, the tightness of the bound depends on the generalization gap $\phi()$, the parametric function $\mu()$ and α. The remaining challenge is to upper-bound $\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m} \mathbb{E}_{h'\sim\pi} \exp[\phi(\mathrm{R}_{\mathcal{D}}(h'), \mathrm{R}_{\mathcal{S}'}(h'))]$ and $\mathbb{E}_{h'\sim\pi} \ln \frac{\pi(h')}{\pi(h)}$ to get a practical bound. As an illustration we provide in the next corollary an instantiation of Theorem 3.1 for two generalization gaps: $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) = m\,\mathrm{kl}[\mathrm{R}_{\mathcal{S}}(h) \| \mathrm{R}_{\mathcal{D}}(h)]$ and $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) = 2m[\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)]^2$ when π is a uniform distribution on a bounded set $\mathbb{H}$.

Corollary 3.1 (Practical Generalization Bound with Complexity Measures)**.** *For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any bounded hypothesis set $\mathbb{H}$, given the uniform prior distribution π on $\mathbb{H}$, for any $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^m$, $h' \sim \pi$, $h \sim \rho_{\mathcal{S}}$ we have*

$$\mathrm{kl}\left[R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)\right] \leq \frac{1}{m} \left[\Big[\alpha R_{\mathcal{S}}(h') + \mu(h',\mathcal{S})\Big] - \Big[\alpha R_{\mathcal{S}}(h) + \mu(h,\mathcal{S})\Big] + \frac{8\sqrt{m}}{\delta^2} \right]_+, \tag{5}$$

$$\text{and } \Big| R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h) \Big| \leq \sqrt{\frac{1}{2m} \left[\Big[\alpha R_{\mathcal{S}}(h') + \mu(h',\mathcal{S})\Big] - \Big[\alpha R_{\mathcal{S}}(h) + \mu(h,\mathcal{S})\Big] + \frac{8\sqrt{m}}{\delta^2} \right]_+}, \tag{6}$$

where $[a]_+ = \max(0,a)$, and $\rho_{\mathcal{S}}$ is the Gibbs distribution defined by Equation (2).

Interestingly, Corollary 3.1 gives a bound on $\mathrm{kl}\left[\mathrm{R}_{\mathcal{S}}(h) \| \mathrm{R}_{\mathcal{D}}(h)\right]$ and $|\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)|$ where all terms except $\mathrm{R}_{\mathcal{D}}(h)$ are computable. To compute Equations (5) and (6) we can rearrange the terms to obtain a generalization bound on the true risk $\mathrm{R}_{\mathcal{D}}(h)$. Indeed, we have respectively

$$\mathrm{R}_{\mathcal{D}}(h) \leq \overline{\mathrm{kl}} \left(\mathrm{R}_{\mathcal{S}}(h) \;\middle|\; \frac{1}{m} \left[[\alpha \mathrm{R}_{\mathcal{S}}(h') + \mu(h',\mathcal{S})] - [\alpha \mathrm{R}_{\mathcal{S}}(h) + \mu(h,\mathcal{S})] + \frac{8\sqrt{m}}{\delta^2} \right]_+ \right), \tag{7}$$

$$\text{and } \mathrm{R}_{\mathcal{D}}(h) \leq \mathrm{R}_{\mathcal{S}}(h) + \sqrt{\frac{1}{2m} \left[[\alpha \mathrm{R}_{\mathcal{S}}(h') + \mu(h',\mathcal{S})] - [\alpha \mathrm{R}_{\mathcal{S}}(h) + \mu(h,\mathcal{S})] + \frac{8\sqrt{m}}{\delta^2} \right]_+}, \tag{8}$$

where $\overline{\mathrm{kl}}(q|\tau) = \max\left\{ p \in (0,1) \;\middle|\; \mathrm{kl}(q\|p) \leq \tau \right\}$. These bounds are used in Section 4 to illustrate the generalization guarantees with different values of the parametric function $\mu()$ and α. Moreover,

as illustrated in the experiments, Equation (7) is a tighter bound on the true risk $\mathrm{R}_{\mathcal{D}}(h)$ than Equation (8) which can be proven with Pinsker's inequality (Appendix G).

Nevertheless, the complexity measure $\Phi_\mu(h,\mathcal{S},\delta)$ of Equations (5) and (6) implies asymptotic convergence for $m\to+\infty$. However, as mentioned previously, the complexity also depends on α and $\mu()$: the convergence rate can then be degraded by $[\alpha\mathrm{R}_{\mathcal{S}}(h')+\mu(h',\mathcal{S})]-[\alpha\mathrm{R}_{\mathcal{S}}(h)+\mu(h,\mathcal{S})]$ which may be potentially large. For example, for all $(h,\mathcal{S})\in\mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m$, if the empirical risk $\mathrm{R}_{\mathcal{S}}(h')$ is large (which is often the case when h' is sampled from a uniform prior on $\mathbb{H}$), $\alpha{=}m$, and $\mu(h,\mathcal{S}){=}0$, then the complexity measure, simplified into $\Phi_\mu(h,\mathcal{S},\delta) = \left[[\mathrm{R}_{\mathcal{S}}(h')-\mathrm{R}_{\mathcal{S}}(h)]+\frac{1}{m}\ln\frac{2\sqrt{m}}{\delta}\right]_+$ for $\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h)) = \mathrm{kl}[\mathrm{R}_{\mathcal{S}}(h)\|\mathrm{R}_{\mathcal{D}}(h)]$, is large. We thus have to set α and $\mu()$ such that *(i)* the distribution $\rho_{\mathcal{S}}$ allows to sample a hypothesis h associated with a low objective function $h\mapsto\mathrm{R}_{\mathcal{S}}(h)+\frac{1}{\alpha}\mu(h,\mathcal{S})$ and *(ii)* the complexity measure $\Phi_\mu(h,\mathcal{S},\delta)$ is tight, resulting in a meaningful bound. For example, with $\alpha{=}\sqrt{m}$ and $\mu(h,\mathcal{S}){=}0$, the distribution $\rho_{\mathcal{S}}$ is less concentrated around the minimizers of the empirical risk, but the complexity measure is tighter compared to the previous example: $\left[\frac{1}{\sqrt{m}}[\mathrm{R}_{\mathcal{S}}(h')-\mathrm{R}_{\mathcal{S}}(h)]+\frac{1}{m}\ln\frac{2\sqrt{m}}{\delta}\right]_+$.

In our previous analysis, we considered a uniform distribution for the prior π for illustration purpose. It is nevertheless clear that if the prior is good, *i.e.*, it associates a higher propability to hypotheses having a low objective function, then the bounds become tighter. The most favorable case being when both the prior π and the posterior $\rho_{\mathcal{S}}$ associate high probabilities to these hypotheses. While the posterior $\rho_{\mathcal{S}}$ is generally learned from the learning data, the choice of the prior π matters to get tight bounds. When no good prior knowledge on the problem is available, to obtain better bounds, one solution is to consider data-dependent priors that have been heavily used in the PAC-Bayesian literature (see, *e.g.*, Parrado-Hernández et al., 2012; Dziugaite et al., 2021; Pérez-Ortiz et al., 2021). However, in the context of our practical evaluation hereafter, we propose to consider the uniform distribution for the prior π. Actually, we think that the uniform distribution helps to understand the generalization phenomenon better. Indeed, an hypothesis h sampled from the uniform distribution π has a high chance of underfitting. Hence, if the hypothesis $h\sim\rho_{\mathcal{S}}$ has its associated bound value that is tight, we can easily interpret that this hypothesis generalizes well. On the other hand, if a generalization bound with a data-dependent prior π is tight, it means that a posterior $\rho_{\mathcal{S}}$ (not very far from the prior) tends to generalize well. In this case, we do not know if the generalization capability is essentially due to the choice of a good prior π or if it comes mainly from the learned posterior $\rho_{\mathcal{S}}$.

4 Using Arbitrary Complexities in Practice

The bound of Corollary 3.1 is not directly usable in practice: the remaining challenge is to sample h from the Gibbs distribution $\rho_{\mathcal{S}}$ defined in Equation (2); we address the sampling issue in Section 4.1. Then, we make use of the proposed solution to assess our bound in practice. Section 4.2 introduces our experimental setting and Section 4.3 gives an overview of the tightness of the bound. Additionally, we give an overview of the influence of α and the other parameters in Appendix E.

4.1 Sampling from the Gibbs Distribution

Sampling from the Gibbs distribution of Equation (2) is a hard task: naively, it requires to evaluate the function $h\mapsto-\alpha\mathrm{R}_{\mathcal{S}}(h)-\mu(h,\mathcal{S})$ for all $h\in\mathbb{H}$, which is intractable when $\mathbb{H}$ is infinite or even large. In an empirical study of our bound, we tackle this issue for over-parameterized models, which we later consider in Section 4.2. Let us consider a set $\mathbb{H}$ of hypotheses $h_{\mathbf{w}}$ parameterized by $\mathbf{w}\in\mathbb{R}^D$, and a tractable distribution denoted $\mathrm{P}^{\mathbf{w}}_{\mathcal{U}}$ (*e.g.*, a Gaussian distribution) such that its density approximates the density of $\rho_{\mathcal{S}}$. In this setting, to learn such an auxiliary distribution, we propose in Algorithm 1 a stochastic version of the Metropolis Adjusted Langevin Algorithm (MALA, Besag (1994))[1]. Its objective is to generate samples from $\rho_{\mathcal{S}}$ by iteratively refining the auxiliary distribution that we define as

$$\mathrm{P}^{\mathbf{w}}_{\mathcal{U}} = \mathcal{N}\left(\mathbf{w}-\beta\nabla\left[\mathrm{R}^{\ell}_{\mathcal{U}}(\mathbf{w})+\frac{1}{\alpha}\mu(\mathbf{w},\mathcal{U})\right],\ \tfrac{2\beta}{\alpha}\mathbf{I}\right), \tag{9}$$

[1] See Chib & Greenberg (1995) for an introduction on Metropolis-Hastings Algo on which MALA is based.

Algorithm 1 Stochastic MALA

1: **Input:** Learning set $\mathcal{S}$, weights $\mathbf{w}$, function $\mu()$, loss function $\ell()$
2: **Hyperparameters:** Number of iterations T, learning rate β, parameter α
3: **for** $t \leftarrow 1 \ldots T$ **do**
4: $\quad \mathcal{U} \leftarrow$ Sample (without replacement) a mini-batch from $\mathcal{S}$
5: $\quad \mathbf{w}' \leftarrow$ Sample from the distribution $\mathrm{P}_{\mathcal{U}}^{\mathbf{w}}$
6: $\quad \tau \leftarrow \min\left(1, \frac{\rho_{\mathcal{U}}(\mathbf{w}')\mathrm{P}_{\mathcal{U}}^{\mathbf{w}'}(\mathbf{w})}{\rho_{\mathcal{U}}(\mathbf{w})\mathrm{P}_{\mathcal{U}}^{\mathbf{w}}(\mathbf{w}')}\right)$
7: $\quad u \leftarrow$ Sample from the distribution $\mathrm{Uni}(0,1)$
8: $\quad$ **if** $u \leq \tau$ **then**
9: $\quad\quad \mathbf{w} \leftarrow \mathbf{w}'$
10: **return** $\mathbf{w}$

where $\mathrm{R}_{\mathcal{U}}^{\ell}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{U}}\,\ell(h_{\mathbf{w}},(\mathbf{x},y))$ is the empirical risk on the mini-batch $\mathcal{U} \subseteq \mathcal{S}$, and $\ell : \mathbb{H} \times (\mathbb{X}\times\mathbb{Y}) \to [0,1]$ is a loss function. Concretely, we initialize the parameters $\mathbf{w}$ of the model as the output of an optimization algorithm (Vanilla SGD in our case). Then, we refine them as follows: at each iteration, given the current weights $\mathbf{w}$ and a mini-batch $\mathcal{U} \subseteq \mathcal{S}$ (Line 4), we sample a candidate vector $\mathbf{w}'$ (Line 5) according to the distribution $\mathrm{P}_{\mathcal{U}}^{\mathbf{w}}$; then (Line 6 to 9) we decide to reject or accept the new candidate to become our current weights $\mathbf{w}$, depending on its ratio $\tau = \min\left(1, \frac{\rho_{\mathcal{U}}(\mathbf{w}')\mathrm{P}_{\mathcal{U}}^{\mathbf{w}'}(\mathbf{w})}{\rho_{\mathcal{U}}(\mathbf{w})\mathrm{P}_{\mathcal{U}}^{\mathbf{w}}(\mathbf{w}')}\right)$ being larger than a control value u sampled from the uniform distribution $\mathrm{Uni}(0,1)$ on $[0,1]$. Under the mild assumption that $\rho_{\mathcal{S}}$ is absolute continuous *w.r.t.* $\mathrm{P}_{\mathcal{S}}^{\mathbf{w}}$ (see Chib & Greenberg, 1995, for details), when the number of iterations tends to infinity and when $\mathcal{U}=\mathcal{S}$, the returned $\mathbf{w}$ is sampled according to $\rho_{\mathcal{S}}$ (Smith & Roberts, 1993). Note that this assumption requires that the tractable distribution $\mathrm{P}_{\mathcal{S}}^{\mathbf{w}}$ has a strictly positive density when the density of $\rho_{\mathcal{S}}$ is strictly positive as well (see Chib & Greenberg, 1995).

4.2 Experimental Setting

In this section, we investigate the tightness of our bounds of Equations (7) and (8) on the MNIST (LeCun et al., 1998) and FashionMNIST (Xiao et al., 2017) datasets. We keep the original learning set as $\mathcal{S}$ and the original test set $\mathcal{T}$ to estimate the true risk that we refer to as test risk $\mathrm{R}_{\mathcal{T}}(h)$.

Model. We use a "Convolutional Network in Network" (Lin et al., 2013) similarly to Jiang et al. (2019) and Dziugaite et al. (2020) , that consists of several modules of 3 convolutional layers each followed by a leaky ReLU activation function (its negative slope is set to 10^{-2}). The depth of the network L is the number of convolutional layers, and the width H is the number of channels of each convolution. In addition, for each layer i, we denote its weights by $\mathbf{w}_i$. For full details of the architecture, we refer the reader to the Appendix. We consider $L\in\{9,12,15\}$ and $H\in\{128,256\}$. Furthermore, we initialize the network with the weights $\mathbf{w}^0\in\mathbb{R}$ obtained the uniform Kaiming He initializer He et al. (2015). The set $\mathbb{H}$ corresponds to the hypotheses $h_{\mathbf{w}}$ that can be obtained from this initialization (and we clamp the weights during the optimization in the initialization interval).

Arbitrary complexity measures. We study 6 different complexity measures parametrized by different functions $\mu(h_{\mathbf{w}},\mathcal{S})$ from Jiang et al. (2019, Sec. C)[2]. These 6 functions are actually independent from the learning sample $\mathcal{S}$ ($\mathcal{S}$ is dropped below for convenience) and defined as follows:

$$\textsc{Dist_Fro}(h_{\mathbf{w}}) = \sum_{i=1}^{L} \|\mathbf{w}_i - \mathbf{w}_i^0\|_2, \quad \text{and} \quad \textsc{Dist_L2}(h_{\mathbf{w}}) = \|\mathbf{w} - \mathbf{w}^0\|_2,$$

$$\text{and} \quad \textsc{Param_Norm}(h_{\mathbf{w}}) = \sum_{i=1}^{L} \|\mathbf{w}_i\|_2^2, \quad \text{and} \quad \textsc{Path_Norm}(h_{\mathbf{w}}) = \sum_{i=1}^{\mathrm{card}(\mathbb{Y})} h_{\mathbf{w}^2}(\mathbf{1})[i],$$

$$\text{and} \quad \textsc{Sum_Fro}(h_{\mathbf{w}}) = L\left(\prod_{i=1}^{L} \|\mathbf{w}_i\|_2^2\right)^{\frac{1}{L}}, \quad \text{and} \quad \textsc{Zero}(h_{\mathbf{w}}) = 0.$$

[2]Note we consider a subset of the functions studied by Jiang et al.; we select those that are optimizable.

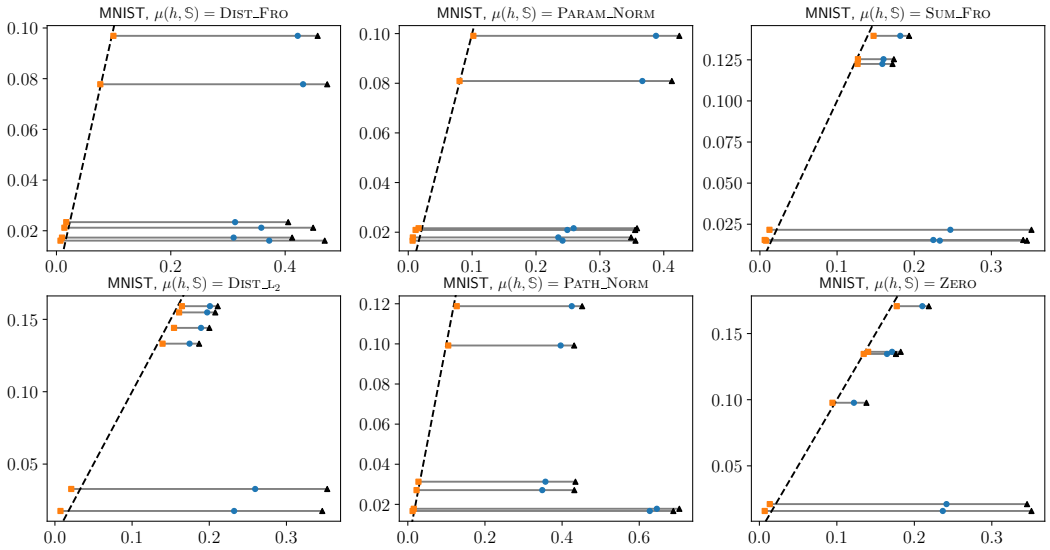

Figure 1: Scatter plot given a parametric function $\mu(h, \mathcal{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given α, width H and depth L. Each segment has a corresponding orange square and a blue circle. The orange squares corresponds to the empirical risk $\mathrm{R}_{\mathcal{S}}(h)$ (x-axis) and test risk $\mathrm{R}_{\mathcal{T}}(h)$ (y-axis). The blue circle *resp.* the black triangle represents Equation (7) *resp.* Equation (8) in the x-axis and the test risk $\mathrm{R}_{\mathcal{T}}(h)$ in the y-axis. The dashed line is the identity function.

We define the considered measures with α taken among 5 values uniformly spaced between $[\sqrt{m}, m]$. Note that as mentioned above these 6 parametric functions are independent from the sample $\mathcal{S}$, we have also analyzed other parametric functions that depend on $\mathcal{S}$. The results obtained are similar, we decided to defer these results in Appendix E.

Bound optimization. To compute our bound in Equations (7) and (8), we aim to sample an hypothesis $h \sim \rho_{\mathcal{S}}$ via Algorithm 1. We set the loss function to the bounded cross entropy from Dziugaite & Roy (2018): $\ell(h, (\mathbf{x}, y)) = -\frac{1}{4}\ln(e^{-4} + (1-2e^{-4})h[y])$, where $h[y]$ is the probability assigned to label y by h. The advantage of Dziugaite & Roy (2018)'s cross-entropy is that it lies in $\ell(h, (\mathbf{x}, y)) \in [0, 1]$, whereas the classical cross-entropy is unbounded. Indeed, taking into account the classical cross-entropy when optimizing the objective function would lead to focusing too much on the risk minimization, while we want to take into account $\frac{1}{\alpha}\,\mu(\mathbf{w}, \mathcal{U})$. We initialize the weights $\mathbf{w} \in \mathbb{R}^D$ to the solution found by optimizing the objective function with a Vanilla SGD (with 10 epochs, a learning rate of 10^{-1}, and a batch size of 64). Given these initial parameters $\mathbf{w}$, we execute Algorithm 1 for 1 epoch with a mini-batch of size 64, where $\beta = 10^{-4}$.

4.3 Tightness of the Bounds

For each parametric function $\mu()$, we report in Figures 1 and 2, the test risks $\mathrm{R}_{\mathcal{T}}(h)$ and the values of the tightest bound on $\mathrm{R}_{\mathcal{D}}(h)$ (*w.r.t.* α) associated to Equations (7) and (8) for different parameters (depth L, width H). First of all, we observe that the bounds correctly upper-bound the test risks and that some measures lead to tighter bounds, such as Sum_Fro or Dist_L2. We also remark that certain empirical risks are high, in particular for these latter measures. This is due to the sampling of the hypothesis h from the distribution $\rho_{\mathcal{S}}$: the hypothesis does not necessarily minimizes the objective function $h \mapsto \mathrm{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\,\mu(h, \mathcal{S})$. We nevertheless observe that the bounds' values are higher when the empirical risk $\mathrm{R}_{\mathcal{S}}(h)$ is low. This can be explained by the fact that $[\alpha \mathrm{R}_{\mathcal{S}}(h') + \mu(h', \mathcal{S})] - [\alpha \mathrm{R}_{\mathcal{S}}(h) + \mu(h, \mathcal{S})]$ is large in this case due notably to the non informative prior π. Interestingly, when the empirical risks are a bit worse or close to the true risks, the bounds become tighter for certain parametric functions such as Dist_L2 and Sum_Fro which then appear to capture more information on the generalization capabilities. On the other hand, we can also note that for some measures such as Dist_Fro, Param_Norm and Zero (mainly for FashionMNIST)

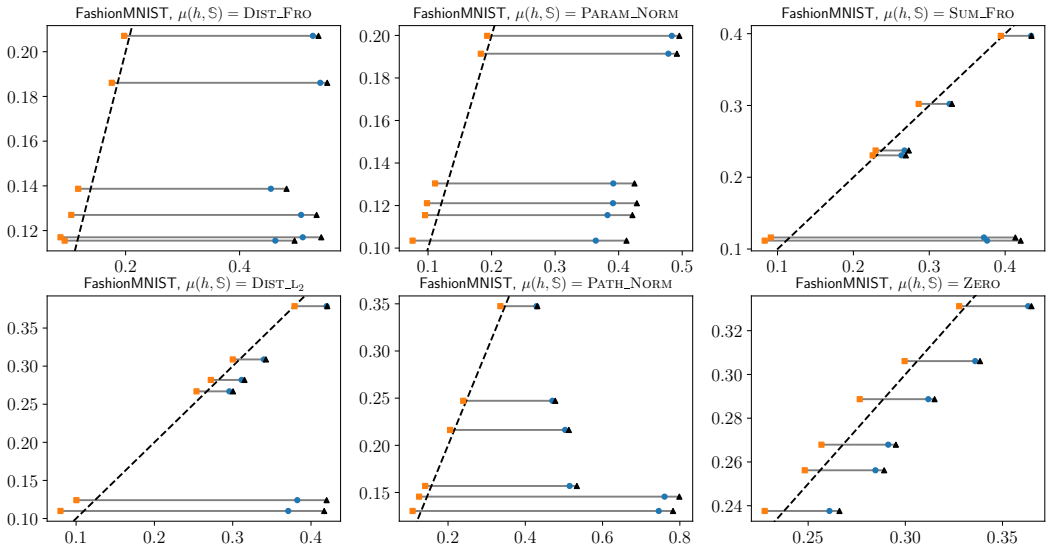

Figure 2: Scatter plot given a parametric function $\mu(h, \mathcal{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given α, width H and depth L. Each segment has a corresponding orange square and a blue circle. The orange squares correspond to the empirical risk $\mathrm{R}_{\mathcal{S}}(h)$ (x-axis) and test risk $\mathrm{R}_{\mathcal{T}}(h)$ (y-axis). The blue circle *resp.* the black triangle represents Equation (7) *resp.* Equation (8) in the x-axis and the test risk $\mathrm{R}_{\mathcal{T}}(h)$ in the y-axis. The dashed line is the identity function.

the bounds remain similar whatever the hypothesis which illustrates that these latter measures do not really help to capture some information about the evolution of the generalization capabilities with the different hypotheses. This confirms that there is an interest in using a parametric function that captures information on the model during the training phase to assess its generalization capability.

In Appendix E, we provide additional results on the influence of the parameter α and the depth/width of the network. As expected, the bounds tend to increase when α becomes large, for smaller α (*e.g.*, close to $\sqrt{m}$), the bounds are improved but to the price of potentially higher risks. In contrast, about the depth/width impact, some measures are less sensitive to the increase of such parameters such as Param_Norm and to a lesser extend Sum_Fro and Dist_L_2. This illustrates our framework's interest in studying the impact of some regularization when learning (over-)parameterized models.

5 Conclusion

In this paper, we provide a novel generalization bound that is able to incorporate arbitrary complexity measures, unlike classical learning theory frameworks (for which the framework imposes the complexity). These measures incorporate a data and model-dependent function, which can favor tightening the complexity for the generalization gap. To the best of our knowledge, our framework is one of the few able to be general enough to bring theoretical guarantees for most of the arbitrary complexity measures used in practice, *e.g.*, for regularization purposes. One limitation of this work is clearly that the hypothesis is obtained from a distribution difficult to use, namely the Gibbs distribution, which requires a specific sampling algorithm. Hopefully, the generality of this framework allows one to avoid the sampling if we consider uniform-convergence-type bounds, for example, as in Corollary D.1. Improving the framework in this direction is an interesting future work. In particular investigating the use of other distributions for sampling the hypothesis could be a possible direction. Another one could be to consider other specific ways to define informative data-dependent priors in order to obtain better bounds. For instance, the parametric function can be leveraged in order to include informative prior. In conclusion, we believe that this work paves the way to new research directions that try to bridge statistical learning theory and practice.

REPRODUCTIBILITY STATEMENT

In order to ensure the reproductibility of our results, we complete the presentation of the experimental setup of the main text in Section 4.2 with a more complete description of the setting, models and parameter used in Appendix E where some additional results are also provided. We also include the code of our method as an additional zip file in the supplementary material in order to facilitate the reproduction of the experiments.
Regarding the theoretical contributions, we provide in Appendices A to C the proofs of the results presented in the main paper, namely Theorems 2.1 and 3.1 and corollary 3.1. We also provide in Appendix D some additional results and discussion about the comparison of our framework with the uniform convergence and algorithm-dependent generalization bounds.

ETHIC STATEMENT

The contributions of this paper are essentially fundamental and theoretical, we do not see immediate potential negative social impact from these contributions. We followed classic ethical guidelines in machine learning which in our case mainly consists in bringing informations about reproductibility issues which is addressed in the previous paragraph.

REFERENCES

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *CoRR*, abs/2110.11216, 2021.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 2016.

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 2002.

Julian Besag. Comments on "Representations of knowledge in complex systems" by U. Grenander and MI Miller. *Journal of the Royal Statistical Society, Series B.*, 1994.

Gilles Blanchard and François Fleuret. Occam's hammer. In *Annual Conference on Learning Theory (COLT)*, volume 4539 of *Lecture Notes in Computer Science*, pp. 112–126. Springer, 2007.

Olivier Bousquet and André Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*, 2002.

Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*. Springer Science & Business Media, 2004.

Olivier Catoni. Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *CoRR*, abs/0712.0248, 2007.

Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 1987.

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 1995.

Gintare Karolina Dziugaite and Daniel Roy. Data-dependent PAC-Bayes priors via differential privacy. In *NeurIPS*, 2018.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In *NeurIPS*, 2020.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes. In *AISTATS*, 2021.

Benjamin Guedj. A primer on pac-bayesian learning. *CoRR*, abs/1901.05353, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE ICCV*, 2015.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *ICLR*, 2019.

Yiding Jiang, Parth Natekar, Manik Sharma, Sumukh K. Aithal, Dhruva Kashyap, Natarajan Subramanyam, Carlos Lassance, Daniel M. Roy, Gintare Karolina Dziugaite, Suriya Gunasekar, Isabelle Guyon, Pierre Foret, Scott Yak, Hossein Mobahi, Behnam Neyshabur, and Samy Bengio. Methods and Analysis of The First Competition in Predicting Generalization of Deep Learning. In *NeurIPS 2020 Competition and Demonstration Track*, 2021.

Sham Kakade, Karthik Sridharan, and Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In *NIPS*, 2008.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 1938.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. THE MNIST DATASET of handwritten digits, 1998. URL http://yann.lecun.com/exdb/mnist/.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2013.

Andreas Maurer. A Note on the PAC Bayesian Theorem. *CoRR*, cs.LG/0411099, 2004.

David McAllester. Some PAC-Bayesian Theorems. In *COLT*, 1998.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *NeurIPS*, 2019.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes Bounds with Data Dependent Priors. *Journal of Machine Learning Research*, 2012.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *J. Mach. Learn. Res.*, 2021.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *COLT*, 2017.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes Analysis Beyond the Usual Bounds. In *Advances in Neural Information Processing System (NeurIPS)*, 2020.

John Shawe-Taylor and Robert Williamson. A PAC Analysis of a Bayesian Estimator. In *COLT*, 1997.

Adrian FM Smith and Gareth O Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1993.

Vladimir Vapnik and Alexey Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*, 1971.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the disintegration of pac-bayesian bounds. *arXiv preprint arXiv:2102.08649*, 2021.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *ICML*, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, 2017.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*, 2012.