# 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 determined by the considered theoretical framework. This limits the scope of such analyses, as other forms of capacity measures or regularization are used in practical algorithms. In this paper, we leverage the framework of disintegrated PAC-Bayesian bounds and combine it with Gibbs distributions to derive generalization bounds involving a complexity measure that can be defined by the user. Our bounds stand 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 is 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. 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.

In this paper, we tackle this drawback by leveraging the framework of disintegrated PAC-Bayesian bound (Theorem 2.1) to propose a novel generalization bound with arbitrary complexity measures. To do so, we make use of the Gibbs probability distributions (Equation (2)) that depend on a user-defined parametric function characterizing the complexity. It allows us to derive guarantees in terms of probabilistic bounds that depend on a model sampled from a Gibbs distribution mentioned above. It is worth noticing that our result allows retrieving the uniform convergence and algorithm-dependent bounds.

We believe that our novel result provides theoretical foundations for the many regularizations used in practice to perform model selection. For instance, our result allows integrating complexity measures studied empirically in a recent line of work on over-parametrized models (Jiang et al., 2019; Dziugaite et al., 2020; Jiang et al., 2021). In our experimental evaluation, we show how these measures can be easily integrated into our framework in practice. We notably provide a stochastic version of the Metropolis Adjusted Langevin algorithm to compute empirical estimates of our bounds.

**Organization of the paper.** In Section 2, we provide some preliminary 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 supervised classification learning setting where $\mathbb{X}$ denotes the input space and $\mathbb{Y}$ is 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 $\pi$, 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 $\rho$ being included in the support of $\pi$). The classical PAC-Bayesian generalization bounds provide upper bounds in expectation over $\rho$, meaning that they bound the generalization gap expressed as $\left|\,\mathbb{E}_{h\sim\rho}[\mathrm{R}_\mathcal{D}(h)-\mathrm{R}_\mathcal{S}(h)]\,\right|$, and where the complexity term depends on the KL divergence between $\rho$ and $\pi$ defined as $\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)}$. This standard complexity hence captures how much the prior and the posterior distribution deviate in expectation over all the hypotheses. To incorporate custom complexities in the bounds, we follow a slightly different framework recalled below (the disintegrated PAC-Bayesian bounds) in which the expectations on $\rho$ 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 $\rho$ 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)[1]. As far as we know, despite their significance, 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 provide guarantees for a hypothesis $h$ sampled from a posterior distribution $\rho_\mathcal{S}$. They take the form of a bound that stands with high probability (at least $1-\delta$) over the random choice of training set $\mathcal{S} \sim \mathcal{D}^m$ and hypothesis $h$. This paper mainly focuses on a particular bound, namely, the one of Rivasplata et al. (2020, Theorem 1 *(i)*) 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*

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

*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

[1] Disintegrated PAC-Bayesian bounds have also been introduced as a "single-draw case" by Hellström & Durisi (2020).

for a single $h$, and *(ii)* the term $\ln\left[\frac{1}{\delta}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp\left(\varphi(g,\mathcal{S}')\right)\right]$ 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 $\delta$ 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 data sample $\mathcal{S}$, 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 $\delta$ 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 custom complexity measures in the bound; we call "*parametric function*" the function $\mu()$. This function must, in fact, serves to obtain a complexity measure $\Phi_\mu(h,\mathcal{S},\delta)$ that is representative of the generalization gap (which is unknown). For instance, 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 entirely new since, for example, uniform stability bounds allow one to consider such norms (see, *e.g.*, Kakade et al., 2008). This example is just for illustration purposes. Our framework is compatible with broader families 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*

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

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

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

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 $\mathcal{S}$. In the ideal case, since we want to minimize the generalization gap $\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h))$, one can define the function $\mu(h,\mathcal{S})=\alpha\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h))-\alpha\mathrm{R}_{\mathcal{S}}(h)$ to obtain a Gibbs distribution that samples hypotheses with small gaps. However, since the generalization gaps are unknown, they must be replaced with a computable function $\mu()$. For instance, the function $\mu()$ can serve as a "regularizing term" (when $\mu()$ is a norm), 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. Equation (2) might look restrictive, but it can actually represent

any probability density function. 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 $\rho'_{\mathcal{S}}$. The Gibbs distribution is well-known and studied in learning theory. In the following, we discuss the principal theoretical works based on it and highlight the differences with our framework.

### 3.2 Related works using the Gibbs Distribution

This section highlights two lines of work that are related to our setting: *(i)* the link between the Gibbs distribution and optimization and *(i)* the usage of the Gibbs distribution in generalization bounds.

**Relationship between optimization and the Gibbs distribution.** Given an objective function $f : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$, the information risk minimization principle (Zhang, 2006) is related to the Gibbs distribution, *i.e.*, by taking

$$\rho_{\mathcal{S}} = \operatorname*{argmin}_{\rho \in \mathbb{M}(\mathbb{H})} \left\{ \mathop{\mathbb{E}}_{h \sim \rho} f(h, \mathcal{S}) + \frac{\mathrm{KL}(\rho \| \pi)}{\alpha} \right\} \quad \text{where} \quad \rho_{\mathcal{S}}(h) \propto \exp\left[ -\alpha f(h, \mathcal{S}) + \ln \pi(h) \right].$$

Note that in our case, we have $f(h, \mathcal{S}) = \mathrm{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\, \mu(h, \mathcal{S}) - \frac{1}{\alpha} \ln \pi(h)$. This distribution is also linked to the Stochastic Gradient Langevin Dynamics (SGLD) algorithm (Welling & Teh, 2011) that learns the hypothesis $h \in \mathbb{H}$ by running several iterations of the form

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

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

$$dh_t = -\nabla f(h_t, \mathcal{S}) dt + \sqrt{2\alpha} B_t, \tag{4}$$

where $B_t$ is the Brownian motion. 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 $\exp(-\alpha f(h_t, \mathcal{S}))$.

**Gibbs distributions in generalization bounds.** The Gibbs distribution is introduced in the PAC-Bayesian theory by Catoni (2004; 2007). Alquier et al. (2016, Theorems 4.2 & 4.3) further develop PAC-Bayesian generalization bounds based on the Gibbs distribution of Equation (2) with $\mu(h, \mathcal{S}) = 0$ as posterior. The Gibbs distribution has also been considered in information-theoretic generalization bounds (see *e.g.*, Xu & Raginsky, 2017; Goyal et al., 2017; Bu et al., 2020) that upper-bound the expected generalization gap $\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \rho_{\mathcal{S}}} \mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)$. For instance, Kuzborskij et al. (Theorem 1, 2019) provides generalization bounds for $f$ being the empirical risk (with sub-Gaussian losses). Aminian et al. (Theorem 1, 2021) prove a closed-form solution of the expected generalization gap with the Gibbs distribution defined with a non-negative $f$. The expected true risk $\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \rho_{\mathcal{S}}} \mathrm{R}_{\mathcal{D}}(h)$ has also been upper bounded by excess risk bounds (Xu & Raginsky, 2017; Kuzborskij et al., 2019), *i.e.*, bounds *w.r.t.* the minimal true risk over the hypothesis set. However, all these bounds consider expected risks while we are interested in the risk of a *single* hypothesis $h$ sampled from $\rho_{\mathcal{S}}$. Hence, to the best of our knowledge, we are the first to derive probabilistic bounds for a single hypothesis sampled from a Gibbs distribution (see Corollary 3.1, Theorem 3.1).

### 3.3 Our Main Result: Generalization Bound with Complexity Measures

We now state our main result: a bound on the generalization gap involving a custom $\mu$, standing for hypotheses sampled from the 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*

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

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

This theorem is general since it depends only on the functions $\phi()$ (expressing the generalization gap) and $\mu()$ (expressing the complexity) chosen by the user. Moreover, we show that this theorem allows obtaining uniform-convergence-based and algorithm-dependent bounds with the integration of complexity measures. We defer the proof of this result to Appendix D.

Given $\phi()$ and $\mu()$, we note a point that can be surprising at 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 likely to be drawn since the density is higher when the objective is low. 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 expected result when considering regularized learning.

In general, the bound may appear loose as there is no explicit dependence on the size of the data sample $m$. 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 $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) = 2m[\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)]^2$ where $\mathrm{kl}(q\|p) \triangleq q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$ for $p \in (0, 1)$ and $q \in [0, 1]$. Then, the tightness of the bound depends on $m$, apart from $\phi()$, $\mu()$ and $\alpha$.

The remaining challenge is to upper-bound $\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m} \mathbb{E}_{g\sim\pi} \exp[\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))]$ and $\ln \frac{\pi(h')}{\pi(h)}$ 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 $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) = 2m[\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)]^2$; and for $\pi$ is a uniform distribution on a bounded set $\mathbb{H}$.

**Corollary 3.1** (Practical Generalization Bound with Complexity Measures)**.** *For any* $\mathcal{D}$ *on* $\mathbb{X} \times \mathbb{Y}$*, for any bounded hypothesis set* $\mathbb{H}$*, given the uniform prior* $\pi$ *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*

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

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

*where* $[a]_+ = \max(0, a)$*, and* $\rho_{\mathcal{S}}$ *is the Gibbs distribution defined in 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 $\mathrm{R}_{\mathcal{D}}(h)$. We obtain 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 $\overline{\mathrm{kl}}(q|\tau) = \max\{p \in (0, 1) \mid \mathrm{kl}(q\|p) \leq \tau\}$. These bounds are used in Section 4 to illustrate the generalization guarantees for different values of $\mu()$ and $\alpha$. In general, Equation (7) provides a tighter bound on the true risk than Equation (8). This can be proven with Pinsker's inequality (Appendix G) and is shown in our experiments. Notice that the r.h.s. of Equations (5) and (6) enjoys asymptotic convergence for $m \to \infty$. However, for some trivial cases, the convergence rate can be arbitrarily degraded by increasing $[\alpha \mathrm{R}_{\mathcal{S}}(h') + \mu(h', \mathcal{S})] - [\alpha \mathrm{R}_{\mathcal{S}}(h) + \mu(h, \mathcal{S})]$. For example,

for a large empirical risk $\mathrm{R}_{\mathcal{S}}(h')$ (which is common when $h'$ is sampled from a uniform prior on $\mathbb{H}$), and for $\alpha{=}m$ and $\mu(h,\mathcal{S}){=}0$, the r.h.s. for $\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h)) = \mathrm{kl}[\mathrm{R}_{\mathcal{S}}(h)\|\mathrm{R}_{\mathcal{D}}(h)]$ simplifies to $\Phi_\mu(h,\mathcal{S},\delta) = [[\mathrm{R}_{\mathcal{S}}(h'){-}\mathrm{R}_{\mathcal{S}}(h)] + \frac{1}{m}\ln\frac{2\sqrt{m}}{\delta}]_+$ and is large, no matter $m$. In order for the bound to be meaningful, we have then to set $\alpha$ and $\mu()$ such that *(i)* the distribution $\rho_{\mathcal{S}}$ allows us 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. For example, for $\alpha{=}\sqrt{m}$ 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: $[\frac{1}{\sqrt{m}}\,[\mathrm{R}_{\mathcal{S}}(h'){-}\mathrm{R}_{\mathcal{S}}(h)] + \frac{1}{m}\ln\frac{2\sqrt{m}}{\delta}]_+$. Lastly, in the ideal case with $\mu(h,\mathcal{S}){=}\frac{\alpha}{m}\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h)){-}\alpha\mathrm{R}_{\mathcal{S}}(h)$ and $\alpha{=}\sqrt{m}$, the upper-bound of $\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h)) = m\,\mathrm{kl}[\mathrm{R}_{\mathcal{S}}(h)\|\mathrm{R}_{\mathcal{D}}(h)]$ becomes $[\frac{1}{\sqrt{m}}\,(\mathrm{kl}[\mathrm{R}_{\mathcal{S}}(h')\|\mathrm{R}_{\mathcal{D}}(h')]{-}\,\mathrm{kl}[\mathrm{R}_{\mathcal{S}}(h)\|\mathrm{R}_{\mathcal{D}}(h)]) + \frac{1}{m}\ln\frac{2\sqrt{m}}{\delta}]_+$ which is tight when the gaps of $h$ and $h'$ are small; the tightness arise with high probability since the density $\rho_{\mathcal{S}}(h){\propto}\exp(-\frac{\alpha}{m}\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h)))$ is concentrated around the small gaps. This also highlights that the choice of the parametric function $\mu()$ is key to obtaining a tight generalization bound.

In our previous analysis, we considered a uniform distribution for the prior $\pi$ for illustration purposes. It is nevertheless clear that if the prior is good, *i.e.*, it associates a higher probability to hypotheses having a low objective function, then the bounds become tighter. The most favorable case is when both the prior $\pi$ and the posterior $\rho_{\mathcal{S}}$ associate high probabilities to these hypotheses. While the posterior $\rho_{\mathcal{S}}$ is generally learned from data, the choice of the prior $\pi$ matters to get tight bounds. When no prior knowledge of 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). In the context of our practical evaluation hereafter, we consider only uniform distributions for the prior $\pi$, as we think it helps us assess generalization better. Indeed, a hypothesis $h$ sampled from the uniform distribution $\pi$ has a high chance of underfitting. Hence, if the hypothesis $h \sim \rho_{\mathcal{S}}$ has a tight bound, it must be that this hypothesis generalizes well. On the other hand, when using data-dependent priors, we cannot tell if the bound is tight because the hypothesis generalizes well or because the posterior is close to the prior.

## 4 Using Arbitrary Complexities in Practice

The bound of Corollary 3.1 is not directly applicable 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 reports an overview of results on the tightness of the bound. We report more results on the influence of $\alpha$ 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 a tractable distribution, we propose in Algorithm 1 a stochastic version of the Metropolis Adjusted Langevin Algorithm (MALA, Besag (1994))[2]. Its objective is to generate samples from $\rho_{\mathcal{S}}$ by iteratively refining the tractable distribution that we define as

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

where $\mathrm{R}^{\ell}_{\mathcal{U}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{U}}\,\ell(h_{\mathbf{w}},(\mathbf{x},y))$ 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) minimizing $\mathrm{R}^{\ell}_{\mathcal{S}}(\mathbf{w}){+}\frac{1}{\alpha}\,\mu(\mathbf{w},\mathcal{S})$ (which is approximated by $\mathrm{R}^{\ell}_{\mathcal{U}}(\mathbf{w}){+}\frac{1}{\alpha}\,\mu(\mathbf{w},\mathcal{U})$ for each mini-batch $\mathcal{U}$).

[2]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 $\beta$, parameter $\alpha$
3: **for** $t \leftarrow 1 \dots T$ **do**
4: $\quad \mathcal{U} \leftarrow$ Sample (without replacement) a mini-batch from $\mathcal{S}$
5: $\quad \mathbf{w}' \leftarrow$ Sample from the distribution $\mathrm{P}^{\mathbf{w}}_{\mathcal{U}}$
6: $\quad \tau \leftarrow \min\left(1, \frac{\rho_{\mathcal{U}}(\mathbf{w}')\mathrm{P}^{\mathbf{w}'}_{\mathcal{U}}(\mathbf{w})}{\rho_{\mathcal{U}}(\mathbf{w})\mathrm{P}^{\mathbf{w}}_{\mathcal{U}}(\mathbf{w}')}\right)$
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** $\mathbf{w}$

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 $\mathrm{P}^{\mathbf{w}}_{\mathcal{U}}$; 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 $\tau = \min\big(1, \frac{\rho_{\mathcal{U}}(\mathbf{w}')\mathrm{P}^{\mathbf{w}'}_{\mathcal{U}}(\mathbf{w})}{\rho_{\mathcal{U}}(\mathbf{w})\mathrm{P}^{\mathbf{w}}_{\mathcal{U}}(\mathbf{w}')}\big)$ is larger than a control value $u$ sampled from the uniform distribution $\mathrm{Uni}(0,1)$ on $[0,1]$. Note that, to compute $\tau$, it is not necessary to know the normalization constants of the two distributions appearing in $\tau$ since they cancel out. In other words, only the function (without the normalizations) associated to the distributions are required to compute $\tau$. Under the mild assumption that $\rho_{\mathcal{S}}$ is absolute continuous *w.r.t.* $\mathrm{P}^{\mathbf{w}}_{\mathcal{S}}$ (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 $\mathrm{P}^{\mathbf{w}}_{\mathcal{S}}$ 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 Appendix E. We consider $L{\in}\{9, 12, 15\}$ and $H{\in}\{128, 256\}$. Furthermore, we initialize the network with the weights $\mathbf{w}^0{\in}\mathbb{R}^D$ obtained by 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)[3]. These 6 functions are actually independent of the learning sample $\mathcal{S}$ ($\mathcal{S}$ is dropped below for convenience) and defined as follows:

$$\textsc{Dist\_Fro}(h_{\mathbf{w}}) = \sum_{i=1}^{L} \|\mathbf{w}_i{-}\mathbf{w}_i^0\|_2, \quad \text{and} \quad \textsc{Dist\_L2}(h_{\mathbf{w}}) = \|\mathbf{w}{-}\mathbf{w}^0\|_2,$$

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

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

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

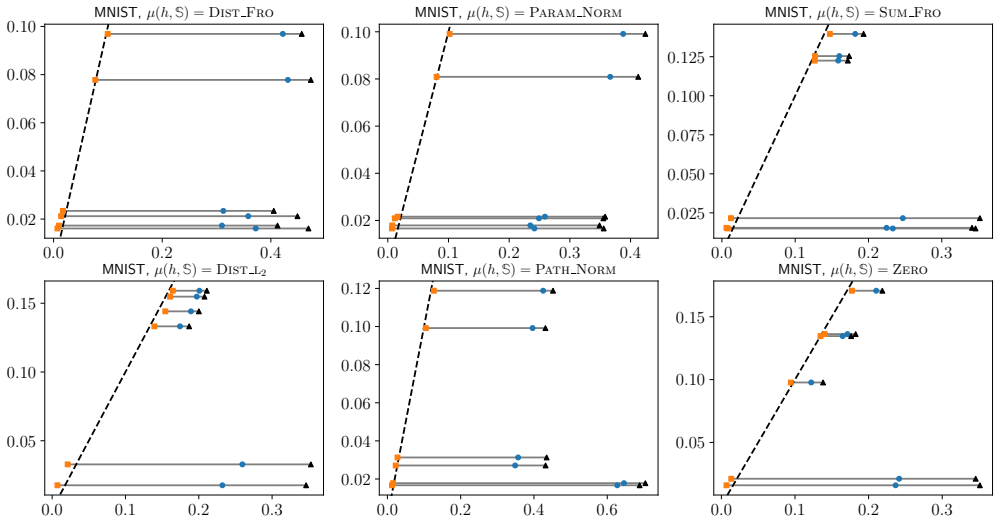


Figure 1: Scatter plot given a parametric function $\mu(h, \mathcal{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given $\alpha$, width $H$ and depth $L$. Each segment has a corresponding orange square and a blue circle. The orange squares corresponds to the empirical 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 $\alpha$ taken among 5 values uniformly spaced between $[\sqrt{m}, m]$. Note that, as mentioned above, these 6 parametric functions are independent of 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 a 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 $\mathrm{R}^{\ell}_{\mathcal{S}}(\mathbf{w}) + \frac{1}{\alpha}\,\mu(\mathbf{w}, \mathcal{S})$ 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.* $\alpha$) 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 $\pi$. 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. Indeed, the more the objective function $\mathrm{R}_{\mathcal{S}}(h) + \frac{1}{\alpha}\,\mu(h, \mathcal{S})$ is representative of the gap of $h$, the tighter the bound. On the other hand, we can also note that for some measures such as DIST_FRO, PARAM_NORM, and ZERO (mainly for FashionMNIST), the bounds remain similar whatever the hypothesis which illustrates that these

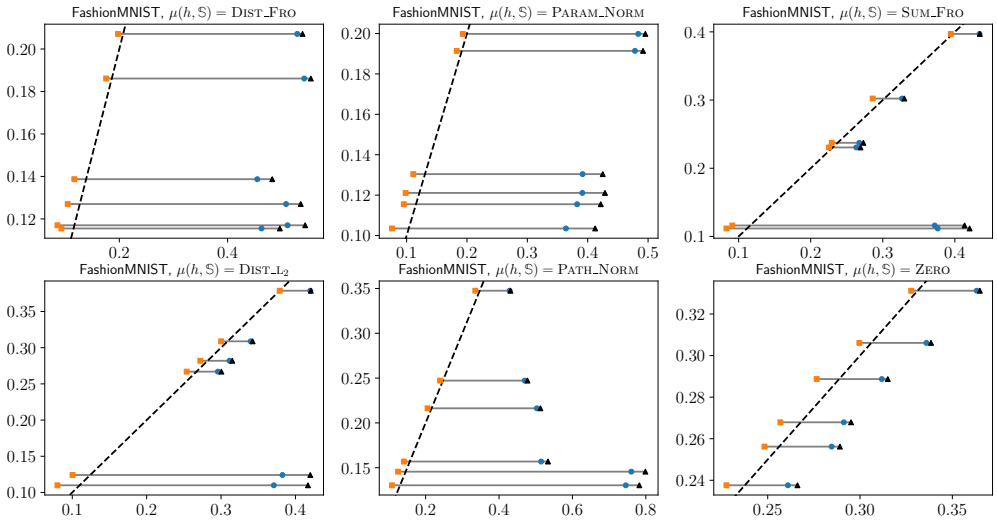


Figure 2: Scatter plot given a parametric function $\mu(h, \mathcal{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given $\alpha$, width $H$ and depth $L$. Each segment has a corresponding orange square and a blue circle. The orange squares correspond to the empirical 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.

latter measures do not really help to capture some information about the generalization gap. 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 $\alpha$ and the depth/width of the network. As expected, the bounds tend to increase when $\alpha$ becomes large for smaller $\alpha$ (*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_$\mathrm{L}_2$. 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.*, based on some norms or a validation set. Such a framework may be adapted to other settings, such as transfer learning, offering new research directions. However, one limitation of this work is clearly that the hypothesis is obtained from a distribution difficult to use, namely, the Gibbs distribution, which uses a specific sampling algorithm, *e.g.*, our algorithm stochastic MALA. It would be interesting to study the performance of such a sampling theoretically. Alternately, 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. Another interesting perspective is to study SGD-based algorithms, either by analyzing models learned by SGD through our framework or by developing SGD alternatives to optimize our bounds. In conclusion, we believe that this work paves the way for new research directions that try to bridge statistical learning theory and practice.

## REPRODUCIBILITY STATEMENT

In order to ensure the reproducibility 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 an immediate potential negative social impact from these contributions. We followed classic ethical guidelines in machine learning which in our case mainly consists in bringing information about reproducibility 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.

Gholamali Aminian, Yuheng Bu, Laura Toni, Miguel R. D. Rodrigues, and Gregory W. Wornell. An exact characterization of the generalization error for the gibbs algorithm. In *NeurIPS*, 2021.

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.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE J. Sel. Areas Inf. Theory*, 2020.

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.

Anil Goyal, Emilie Morvant, Pascal Germain, and Massih-Reza Amini. PAC-Bayesian Analysis for a Two-Step Hierarchical Multiview Learning Approach. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 10535 of *Lecture Notes in Computer Science*, pp. 205–221. Springer, 2017.

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.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE J. Sel. Areas Inf. Theory*, 2020.

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.

Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-dependent analysis of gibbs-erm principle. In *COLT*, 2019.

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.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.

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.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *NIPS*, 2017.

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

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inf. Theory*, 2006.

**Supplementary Material**

In this supplementary material, we present the proofs deferred in Appendix and we give details on the experiments. We prove in Appendix A the proof of Theorem 2.1 and in Appendix B the proof of Theorem 3.1. Appendix C is dedicated to Corollary 3.1. Appendix D is dedicated to the theoretical results related to the comparison with other theoretical results of the literature. Additional details on the experiments are provided in Appendix E. Appendices F and G recall respectively Maurer (2004)'s result and the Pinsker inequality.

## A Proof of Theorem 2.1

**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*

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

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

*Proof.* Note that $\exp\left[\varphi(h,\mathcal{S}) - \ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right]$ is a non-negative random variable. Thus, we can apply Markov's inequality to obtain

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\exp\left(\varphi(h,\mathcal{S}) - \ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right) \leq \frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\rho_{\mathcal{S}}}\exp\left(\varphi(g,\mathcal{S}') - \ln\frac{\rho_{\mathcal{S}'}(g)}{\pi(g)}\right)\right] \geq 1-\delta.$$

Hence, by rearranging the terms, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\exp\left(\varphi(h,\mathcal{S}) - \ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right) \leq \frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\rho_{\mathcal{S}}}\frac{\pi(g)}{\rho_{\mathcal{S}'}(g)}e^{\varphi(g,\mathcal{S}')}\right] \geq 1-\delta$$

$$\iff \mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\exp\left(\varphi(h,\mathcal{S}) - \ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right) \leq \frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}e^{\varphi(g,\mathcal{S}')}\right] \geq 1-\delta.$$

Since both sides of the inequality are strictly positive, we can apply the logarithm, *i.e.*, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\varphi(h,\mathcal{S}) \leq \ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}e^{\varphi(g,\mathcal{S}')}\right)\right] \geq 1-\delta,$$

which is the desired result. □

## B Proof of Theorem 3.1

**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*

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

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

*Proof.* First of all, we denote as $Z = \int_{\mathbb{H}} \exp\left[-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})\right] d\lambda(g)$, the normalization constant of the Gibbs distribution $\rho_{\mathcal{S}}$ and $\lambda$ the reference measure on $\mathbb{H}$. In other words, we have

$$\rho_{\mathcal{S}}(h) = \frac{1}{Z} \exp\left[-\alpha \mathrm{R}_{\mathcal{S}}(h) - \mu(h, \mathcal{S})\right] \propto \exp\left[-\alpha \mathrm{R}_{\mathcal{S}}(h) - \mu(h, \mathcal{S})\right].$$

We apply Theorem 2.1 with $\frac{\delta}{2}$ instead of $\delta$ and with the function $\varphi(h, \mathcal{S}) = \phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h))$ to obtain

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) \leq \ln\left[\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right] + \ln\left[\frac{2}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))}\right]\right] \geq 1 - \frac{\delta}{2}.$$

We develop the term $\ln\left[\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right]$ in Theorem 2.1. We have

$$\begin{aligned}
\ln\left[\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right] &= \ln\left(\frac{\exp\left[-\alpha \mathrm{R}_{\mathcal{S}}(h) - \mu(h, \mathcal{S})\right]}{Z} \frac{1}{\pi(h)}\right) \\
&= \ln\left(\exp\left[-\alpha \mathrm{R}_{\mathcal{S}}(h) - \mu(h, \mathcal{S})\right]\right) \\
&\quad - \ln\left(\pi(h) \int_{\mathbb{H}} \exp\left[-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})\right] d\lambda(g)\right) \\
&= -\alpha \mathrm{R}_{\mathcal{S}}(h) - \mu(h, \mathcal{S}) \\
&\quad - \ln\left(\pi(h) \int_{\mathbb{H}} \frac{\pi(g)}{\pi(g)} \exp\left[-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})\right] d\lambda(g)\right) \\
&= -\alpha \mathrm{R}_{\mathcal{S}}(h) - \mu(h, \mathcal{S}) - \ln\left(\mathop{\mathbb{E}}_{g\sim\pi} \frac{\pi(h)}{\pi(g)} e^{-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})}\right).
\end{aligned}$$

Hence, we obtain the following inequality

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\Bigg[\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) \leq \ln\left[\frac{2}{\delta} \mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))}\right] \\
- \alpha \mathrm{R}_{\mathcal{S}}(h) - \mu(h, \mathcal{S}) - \ln\left(\mathbb{E}_{g\sim\pi} \tfrac{\pi(h)}{\pi(g)} e^{-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})}\right)\Bigg] \geq 1 - \frac{\delta}{2}. \qquad (10)
\end{aligned}$$

We can now upper-bound the term $-\ln\left(\mathbb{E}_{g\sim\pi} \frac{\pi(h)}{\pi(g)} e^{-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})}\right)$. To do so, since $\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathrm{R}_{\mathcal{S}}(h') - \mu(h', \mathcal{S})} > 0$ for all $h \in \mathbb{H}$ and $\mathcal{S} \in (\mathbb{X} \times \mathbb{Y})^m$, we apply Markov's inequality to obtain for any $h \in \mathbb{H}$ and $\mathcal{S} \in (\mathbb{X} \times \mathbb{Y})^m$ with probability at least $1 - \frac{\delta}{2}$ over $h' \sim \pi$

$$\begin{aligned}
&\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathrm{R}_{\mathcal{S}}(h') - \mu(h', \mathcal{S})} \leq \frac{2}{\delta} \mathop{\mathbb{E}}_{g\sim\pi}\left(\frac{\pi(h)}{\pi(g)} e^{-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})}\right) \\
\iff &-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\left[\frac{\pi(h)}{\pi(g)} e^{-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})}\right]\right) \leq \ln\frac{2}{\delta} - \ln\left(\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathrm{R}_{\mathcal{S}}(h') - \mu(h', \mathcal{S})}\right).
\end{aligned}$$

Moreover, by simplifying the right-hand side of the inequality, we have

$$-\ln\left(\frac{\pi(h)}{\pi(h')} e^{-\alpha \mathrm{R}_{\mathcal{S}}(h') - \mu(h', \mathcal{S})}\right) = \ln\frac{\pi(h')}{\pi(h)} + \alpha \mathrm{R}_{\mathcal{S}}(h') + \mu(h', \mathcal{S}).$$

Hence, we obtain the following inequality

$$\mathop{\mathbb{P}}_{h'\sim\pi}\left[-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\left[\frac{\pi(h)}{\pi(g)} e^{-\alpha \mathrm{R}_{\mathcal{S}}(g) - \mu(g, \mathcal{S})}\right]\right) \leq \ln\frac{2}{\delta} + \ln\frac{\pi(h')}{\pi(h)} + \alpha \mathrm{R}_{\mathcal{S}}(h') + \mu(h', \mathcal{S})\right] \geq 1 - \frac{\delta}{2}. \qquad (11)$$

By using an union bound on Equations (10) and (11) and rearranging the terms, we obtain the claimed result. □

## C PROOF OF COROLLARY 3.1

**Corollary 3.1** (Practical Generalization Bound with Complexity Measures)**.** *For any $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any bounded hypothesis set $\mathbb{H}$, given the uniform prior $\pi$ 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*

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

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

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

*Proof.* Since $\pi$ is the uniform distribution we have: $\ln\frac{\pi(h')}{\pi(h)}=0$. We instantiate Theorem 3.1 with $\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h)) = m\,\mathrm{kl}\left[\mathrm{R}_{\mathcal{S}}(h)\|\mathrm{R}_{\mathcal{D}}(h)\right]$. It remains to upper-bound $\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{g\sim\pi}\exp\left(m\,\mathrm{kl}\left[\mathrm{R}_{\mathcal{S}'}(g)\|\mathrm{R}_{\mathcal{D}}(g)\right]\right)$. We have

$$\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{m\,\mathrm{kl}[\mathrm{R}_{\mathcal{S}'}(g)\|\mathrm{R}_{\mathcal{D}}(g)]} = \mathop{\mathbb{E}}_{g\sim\pi}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} e^{m\,\mathrm{kl}[\mathrm{R}_{\mathcal{S}'}(g)\|\mathrm{R}_{\mathcal{D}}(g)]} \tag{12}$$

$$\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{m\,\mathrm{kl}[\mathrm{R}_{\mathcal{S}'}(g)\|\mathrm{R}_{\mathcal{D}}(g)]} \leq 2\sqrt{m}, \tag{13}$$

where Equation (12) is due to Fubini's theorem (*i.e.*, we can exchange the two expectations), and Equation (13) is due to Maurer (2004) (see Lemma F.1). By rearranging the terms, with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$, $h\sim\rho_{\mathcal{S}}$, and $h'\sim\pi$ we have

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

Hence, by definition of $[a]_+$, we can deduce Equation (5). From Pinsker's inequality (Theorem G.1), we have

$$2(\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathcal{S}}(h))^2 \leq \mathrm{kl}\left[\mathrm{R}_{\mathcal{S}}(h)\|\mathrm{R}_{\mathcal{D}}(h)\right].$$

Hence, thanks to this inequality and by rearranging the terms, we obtain Equation (6). □

## D COMPARISON WITH THE GENERALIZATION BOUNDS OF THE LITERATURE

In this section, we theoretically compare generalization bounds with arbitrary complexity measures compared to the literature's bounds. To do so, we prove in Corollary D.1 that the upper bound on the generalization gap $\phi(\mathrm{R}_{\mathcal{D}}(h),\mathrm{R}_{\mathcal{S}}(h))$ can be identical, under a mild assumption, as those in the literature. In order to present our result in Corollary D.1, we first give the definitions of the literature's bounds.

We first give the definition of the uniform-convergence-based bounds that are the first type to be introduced, notably in Vapnik & Chervonenkis (1971) using the VC-dimension. Other bounds were later developed based on the Gaussian/Rademacher complexity (Bartlett & Mendelson, 2002) instead. The definition of this bound is the following.

**Definition D.1** (Uniform Convergence Bound)**.** *Let $\phi:[0,1]^2\to\mathbb{R}$ be a generalization gap. A uniform convergence bound 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 function $\Phi_{\mathtt{u}}:(0,1]\to\mathbb{R}$, such that for any $\delta\in(0,1]$ we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\sup_{h\in\mathbb{H}}\phi(R_{\mathcal{D}}(h),R_{\mathcal{S}}(h))\leq\Phi_{\mathtt{u}}(\delta)\right]\geq 1-\delta, \tag{14}$$

*where usually $\phi(R_{\mathcal{D}}(h),R_{\mathcal{S}}(h))=R_{\mathcal{D}}(h)-R_{\mathcal{S}}(h)$.*

This definition encompasses different complexity measures, such as $\Phi_{\mathtt{u}}(\delta)=\mathtt{rad}(\mathbb{H})+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$ for the Rademacher complexity $\mathtt{rad}(\mathbb{H})$, or $\Phi_{\mathtt{u}}(\delta)=\sqrt{\frac{1}{m}2\mathtt{vc}(\mathbb{H})\ln\frac{em}{\mathtt{vc}(\mathbb{H})}}+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$ for the VC-dimension $\mathtt{vc}(\mathbb{H})$ (see Theorem 3.3 and Corollary 3.19 of Mohri et al., 2012).

This definition also highlights the worst-case nature of the uniform-convergence bounds: given a confidence $\delta$, the generalization gap $\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h))$ is upper-bounded by a complexity measure $\Phi_{\mathtt{u}}(\delta)$ constant for all $(h, \mathcal{S}) \in \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m$. The upper bound $\Phi_{\mathtt{u}}(\delta)$ can generally be improved by considering algorithmic-dependent bounds (Bousquet & Elisseeff, 2002; Xu & Mannor, 2012). Indeed, only the generalization gap associated with the hypothesis $h_{\mathcal{S}}$ output by a learning algorithm given $\mathcal{S}$ is upper-bounded. The definition of such bounds is recalled in the following.

**Definition D.2** (Algorithmic-dependent Generalization Bound)**.** *Let $\phi : [0,1]^2 \to \mathbb{R}$ a generalization gap. An algorithmic-dependent generalization bound is defined such that if for any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, there exists a function $\Phi_{\mathtt{a}} : (0,1] \to \mathbb{R}$, such that for any $\delta \in (0,1]$ we have*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\phi(R_{\mathcal{D}}(h_{\mathcal{S}}), R_{\mathcal{S}}(h_{\mathcal{S}})) \leq \Phi_{\mathtt{a}}(\delta)\right] \geq 1-\delta, \tag{15}$$

*where usually $\phi(R_{\mathcal{D}}(h), R_{\mathcal{S}}(h)) = R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)$ and $h_{\mathcal{S}}$ is the hypothesis learned from an algorithm with $\mathcal{S} \sim \mathcal{D}^m$.*

For example, when $\phi(\mathrm{R}_{\mathcal{D}}(h_{\mathcal{S}}) = \mathrm{R}_{\mathcal{D}}(h_{\mathcal{S}}) - \mathrm{R}_{\mathcal{S}}(h_{\mathcal{S}})$, the upper bound $\Phi_{\mathtt{a}}(\delta) = 2\beta + (4m\beta)\sqrt{\frac{\ln 1/\delta}{2m}}$ where $\beta$ is the uniform stability parameter (see Bousquet & Elisseeff, 2002). Similarly to the uniform-convergence-based bounds, these bounds are still not as permissive as our framework. Indeed, the upper bound $\Phi_{\mathtt{a}}(\delta)$ is a constant *w.r.t.* the hypothesis $h_{\mathcal{S}}$ and the learning sample $\mathcal{S}$ (like in the uniform convergence bounds). Hence, since our bound can depend on the learning sample $\mathcal{S}$ and the hypothesis $h$, we can retrieve the upper bounds $\Phi_{\mathtt{u}}(\delta)$ and $\Phi_{\mathtt{a}}(\delta)$ (Equations (14) and (15)) since our upper bound in Theorem 3.1 can depend on the learning sample $\mathcal{S}$ and the hypothesis $h$. Indeed, from Theorem 3.1, we can obtain the following generalization bounds.

**Corollary D.1.** *Let $\phi : [0,1]^2 \to \mathbb{R}$ be the generalization gap and assume that there exists a function $\Phi_{\mathtt{u}} : (0,1] \to \mathbb{R}$ fulfilling Definition D.1 and a function $\Phi_{\mathtt{a}} : (0,1] \to \mathbb{R}$ fulfilling Definition D.2. For any distribution $\mathcal{D}$ on $\mathbb{X} \times \mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$, for any $\delta \in (0,1]$, there is a function $\mu : \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m \to \mathbb{R}$ such that*

*if $\Phi_{\mathtt{u}}(\delta) \geq \ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp\left(\phi(R_{\mathcal{D}}(g), R_{\mathcal{S}'}(g))\right)\right]$, then*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\phi(R_{\mathcal{D}}(h), R_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{u}}(\delta)\right] \geq 1-\delta, \tag{16}$$

*if $\Phi_{\mathtt{a}}(\delta) \geq \ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp\left(\phi(R_{\mathcal{D}}(g), R_{\mathcal{S}'}(g))\right)\right]$, then*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\phi(R_{\mathcal{D}}(h), R_{\mathcal{S}}(h)) \leq \Phi_{\mathtt{a}}(\delta)\right] \geq 1-\delta. \tag{17}$$

*Proof.* Let the parametric function $\mu()$ defined as

$$\forall (h, \mathcal{S}) \in \mathbb{H}\times(\mathbb{X}\times\mathbb{Y})^m, \quad \mu(h, \mathcal{S}) = -\alpha \mathrm{R}_{\mathcal{S}}(h) - \ln \pi(h) + \Phi_{\mathtt{u}}(\delta).$$

Given the definition of $\rho_{\mathcal{S}}$ (with the parametric function $\mu()$ defined above), we can deduce from Theorem 3.1 that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi, h\sim\rho_{\mathcal{S}}}\left[\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) \leq \ln\left(\frac{4}{\delta^2}\,\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathop{\mathbb{E}}_{g\sim\pi}\exp\left[\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))\right]\right)\right]$$

$$= \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m,\ h\sim\rho_{\mathcal{S}}}\left[\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) \leq \ln\left(\frac{4}{\delta^2}\,\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathop{\mathbb{E}}_{g\sim\pi}\exp\left[\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))\right]\right)\right] \geq 1-\delta.$$

Note that the equality holds since $h' \sim \pi$ does not appear in the bound. If the assumption $\Phi_{\mathtt{u}}(\delta) \geq \ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp\left(\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))\right)\right]$ is satisfied, then, we can deduce Equation (16). Similarly, if the assumption $\Phi_{\mathtt{a}}(\delta) \geq \ln\left[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp\left(\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))\right)\right]$ is satisfied then we have Equation (17). □

Note that, to prove Corollary D.1, we require an additional assumption: a lower bound on $\Phi_{\mathtt{u}}(\delta)$ and $\Phi_{\mathtt{a}}(\delta)$. Actually, this lower bound will be low since the (exponentiated) gap $\exp\left[\phi(\mathrm{R}_{\mathcal{D}}(g), \mathrm{R}_{\mathcal{S}'}(g))\right]$

is *averaged* over the learning sample $\mathcal{S} \sim \mathcal{D}^m$ and the hypothesis $g$ sampled from the prior distribution $\pi$. Hence, under this assumption, our framework is general enough to retrieve uniform-convergence-based or algorithmic-dependent bounds when the hypothesis $h$ is sampled from $\rho_{\mathcal{S}}$. Moreover, in the case of the uniform-convergence bounds, the sampling $h \sim \rho_{\mathcal{S}}$ involved in Equation (16) is not necessary: the bound $\Phi_{\mathrm{u}}(\delta)$ holds for all hypothesis $h \in \mathbb{H}$ with high probability. In other words,

$$\text{if } \underbrace{\mathrm{I}\left[\sup_{h\in\mathbb{H}} \phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathrm{u}}(\delta)\right]}_{\text{from Definition D.1}} = 1, \text{then } \underbrace{\mathop{\mathbb{E}}_{h\in\rho_{\mathcal{S}}} \mathrm{I}\left[\phi(\mathrm{R}_{\mathcal{D}}(h), \mathrm{R}_{\mathcal{S}}(h)) \leq \Phi_{\mathrm{u}}(\delta)\right]}_{\text{from Corollary D.1}} = 1,$$

which arises with high probability.

# E Additional information on the experiments

In this section, we introduce additional figures concerning the tightness, the influence of $\alpha$, and the influence of the number of parameters. Additionally, we provide more experiments with data-dependent complexity measures that we present in Appendix E.2.

## E.1 Details on the neural network model

As said before, 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 $\mathbf{w}_i$. More precisely, the modules of this model can be described as follows. A module takes two parameters as arguments: the number of input channels $H_{\text{in}}$ and the number of output channels $H_{\text{out}}$ and applies consecutively three convolutional layers (each followed by a leaky ReLU activation function). The first layer is composed of a $3\times3$ kernel (where the stride *resp.* padding is set to 2 *resp.* 1) with $H_{\text{in}}$ channels as input and $H_{\text{out}}$ as output. The two other layers have a $1\times1$ kernel with $H_{\text{out}}$ channels as input and output. Then, the network is constructed as follows: *(a)* we have a module where $H_{\text{out}} = H$ and $H_{\text{in}}$ is the number of channels in the input *(b)* we have $(L/3) - 1$ modules with $H_{\text{in}} = H_{\text{out}} = H$ and *(c)* we have a convolutional layer with a $1\times1$ kernel with $\mathrm{card}(\mathbb{Y})$ channels as output followed by a leaky ReLU activation and an average pooling layer. In the experiments, 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's interval).

## E.2 Data-dependent complexity measures $\Phi_\mu(h, \mathcal{S}, \delta)$

As we have pointed out in the paper, the parametric function $\mu()$ depends on the learning sample $\mathcal{S}$. We illustrate this dependence with other parametric functions defined as

$$\begin{aligned}
\textsc{Dist_Fro}-\textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}) &= \textsc{Dist_Fro}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}),\\
\textsc{Dist_L}_2-\textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}) &= \textsc{Dist_L}_2(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}),\\
\textsc{Param_Norm}-\textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}) &= \textsc{Param_Norm}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}),\\
\textsc{Path_Norm}-\textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}) &= \textsc{Path_Norm}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}),\\
\textsc{Sum_Fro}-\textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}) &= \textsc{Sum_Fro}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}),\\
\textsc{Zero}-\textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}) &= \textsc{Zero}-\textsc{Aug}(h_{\mathbf{w}}) + \textsc{Aug}(h_{\mathbf{w}}, \mathcal{S}),
\end{aligned}$$

where

$$\textsc{Aug}(h, \mathcal{S}) = -\frac{1}{2}\mathrm{R}_{\mathcal{S}}(h) + \frac{1}{2}\mathrm{R}_{\widehat{\mathcal{S}}}(h),$$

and $\widehat{\mathcal{S}}$ is a data-augmented learning sample. More precisely, we apply to each example $(\mathbf{x}, y) \in \mathcal{S}$ *(a)* a random rotation (with a maximum angle set to $20^\circ$) and *(b)* a random translation (with a maximum of 3 translated pixels per dimension).

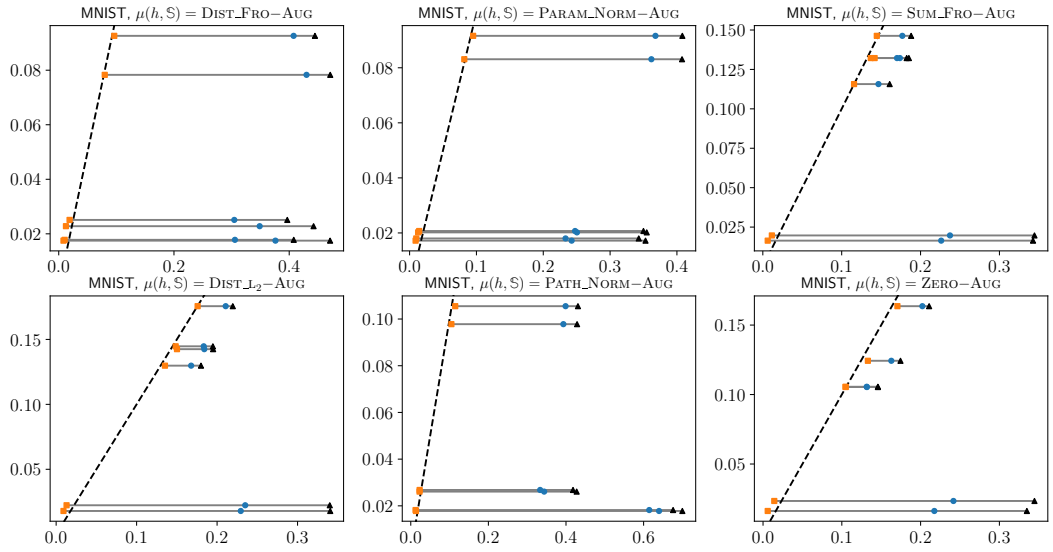


Figure 3: Scatter plot given a parametric function $\mu(h, \mathcal{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given $\alpha$, width $H$, and depth $L$. Each segment has a corresponding orange square and a blue circle. The orange squares corresponds to the empirical 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.

### E.3 Tightness of the Bounds

Figures 3 and 4 report the tightness of the bounds for the data-dependent complexity measures introduced in Appendix E.2. The results are comparable to the ones presented in Section 4.2. We can nevertheless observe some small gains in terms of test risk for Sum_Fro−Aug and Dist_$\mathrm{L}_2$−Aug sometimes coupled with some slight bounds improvement. However, for most of the cases, the fact that the bounds remain the same is consistent with the limited improvements due to the data augmentation.

### E.4 Influence of the Parameter $\alpha$

We analyze the influence of the parameter $\alpha$ in Equation (7). To do so, we plot an overview of the evolution for the bounds and the test risks $\mathrm{R}_{\mathcal{T}}(h)$. Figures 5 to 8 shows the influence of the parameter $\alpha$ for all parametric functions $\mu()$, data-independent and data-dependent introduced in Section 4.2 and appendix E.2 respectively. For each parameter $\alpha$, we plot the mean, the standard deviation, the minimum, and the maximum for the different parameters (depth and width). In general, the bound increases when the $\alpha$ tends to $m$ but the test risks $\mathrm{R}_{\mathcal{T}}(h)$ are less prone to variations. Indeed, the higher the parameter $\alpha$, the more concentrated around the minimizers the hypothesis will be sampled. On the contrary, for a small $\alpha$ (*e.g.*, $\alpha = \sqrt{m}$), the Gibbs distribution defined in Equation (2) is less concentrated making the test risks potentially high with a tighter generalization bound. We recall that the two datasets have $m = 60000$ instances in training and $10000$ in the test set.

### E.5 Influence of the Depth/Width

Figures 9 to 12 shows the influence of the depth and the width for all parametric functions on the evolution of Equation (7). Interestingly, the evolution of the bounds highly depends on the chosen parametric function $\mu()$. For instance, the bound increases with Path_Norm when the depth and the width increase. This is in contrast with Param_Norm, which decreases when the number of parameters increases. This shows the interest of our framework: considering a user-specified

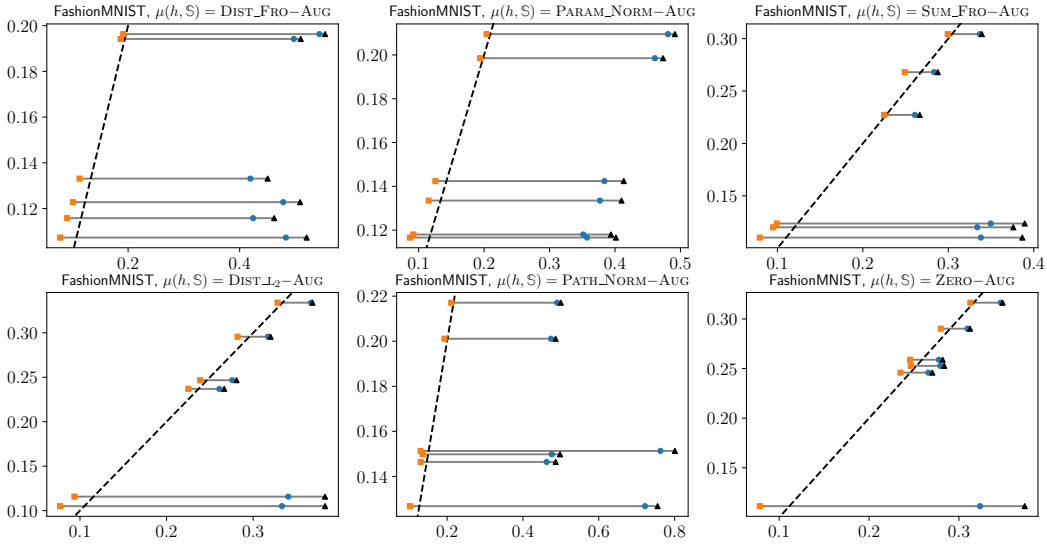


Figure 4: Scatter plot given a parametric function $\mu(h, \mathcal{S})$, where each segment represents a neural network $h_{\mathbf{w}}$ learned with a given $\alpha$, width $H$ and depth $L$. Each segment has a corresponding orange square and a blue circle. The orange squares corresponds to the empirical 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.

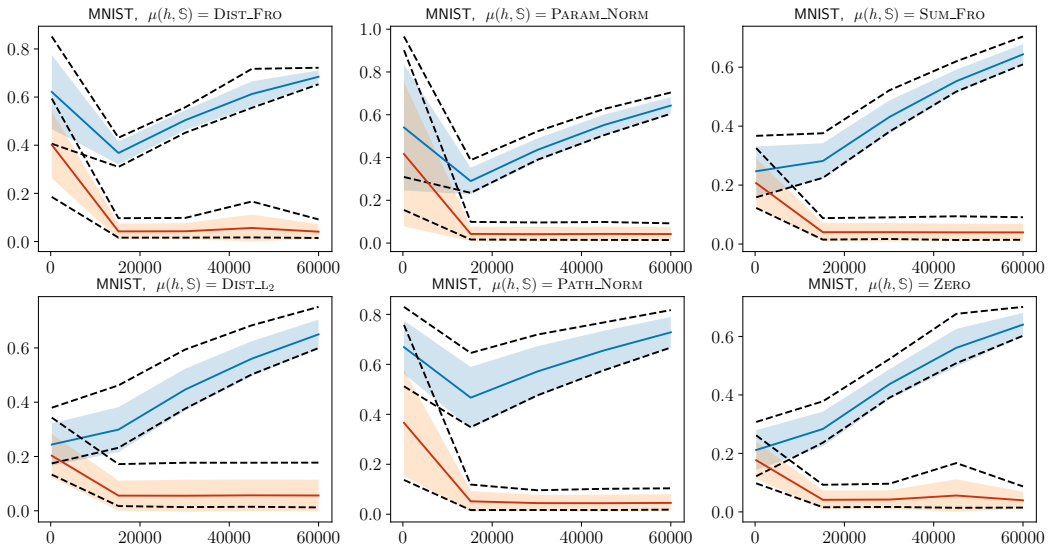


Figure 5: Influence of the parameter $\alpha$ in the x-axis. The bound values are represented in blue, and the test risk is in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

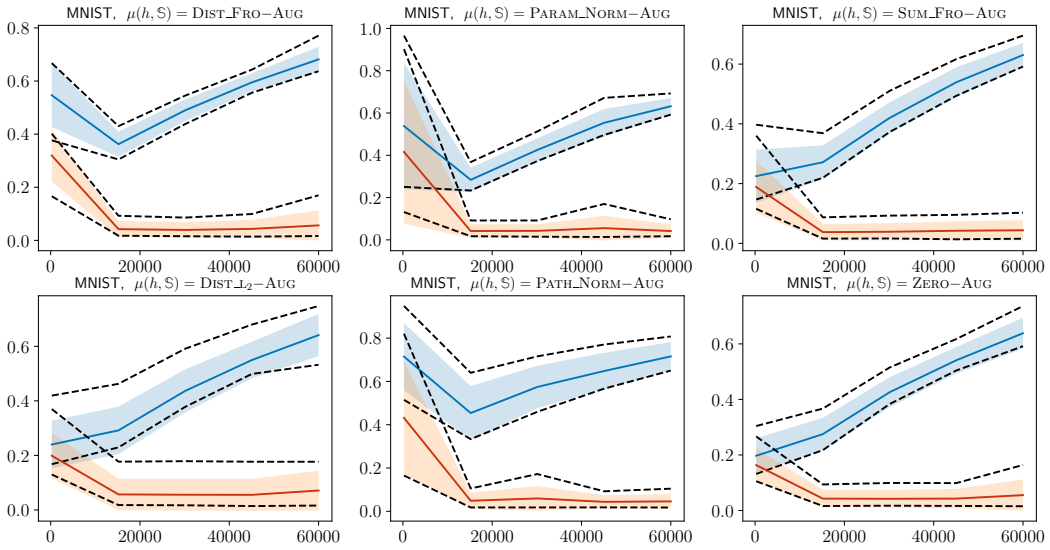


Figure 6: Influence of the parameter $\alpha$ in the x-axis. The bound values are represented in blue, and the test risk is in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

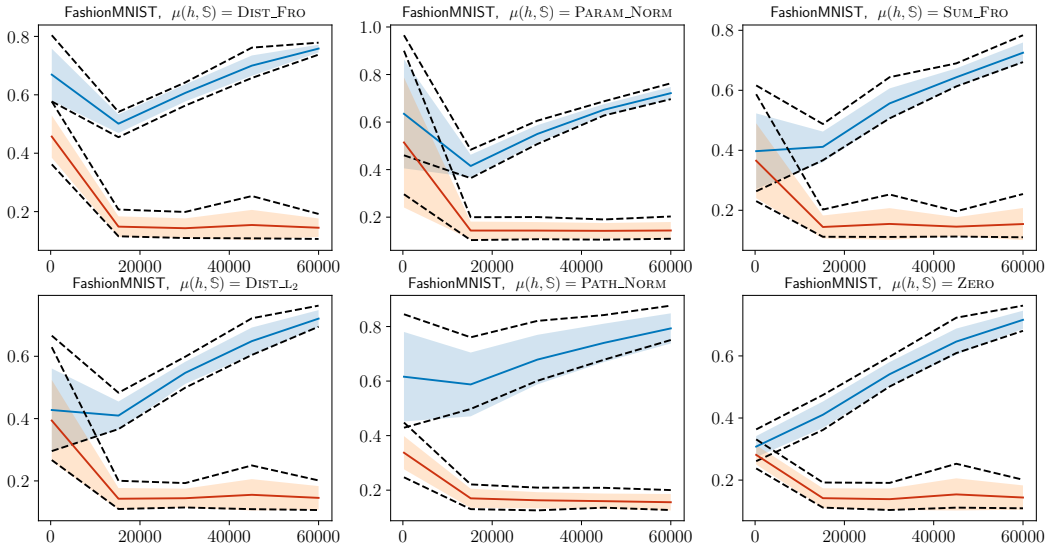


Figure 7: Influence of the parameter $\alpha$ in the x-axis. The bound values are represented in blue, and the test risk is in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

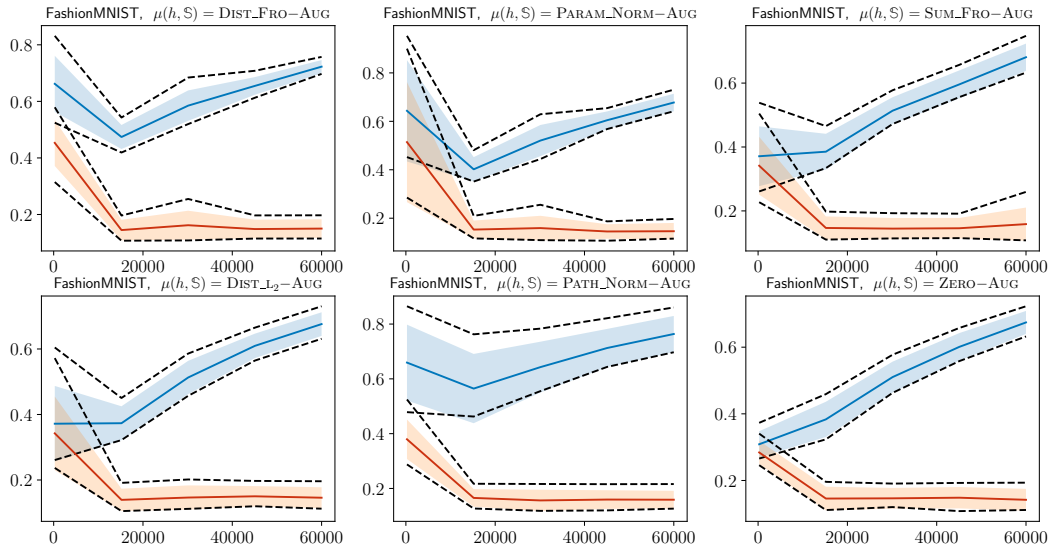


Figure 8: Influence of the parameter $\alpha$ in the x-axis. The bound values are represented in blue, and the test risk is in red. The two (solid) lines are the mean values computed on the depths and widths; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

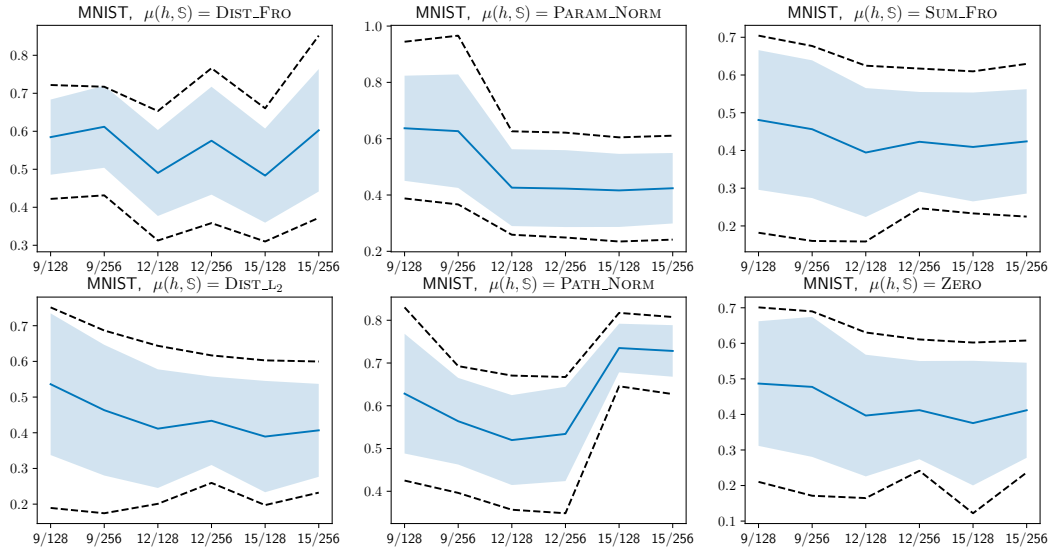


Figure 9: Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of $\alpha$; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

complexity measure $\Phi_\mu()$ can help to understand the generalization of over-parameterized models and, in particular, the effect of using some particular regularizations.

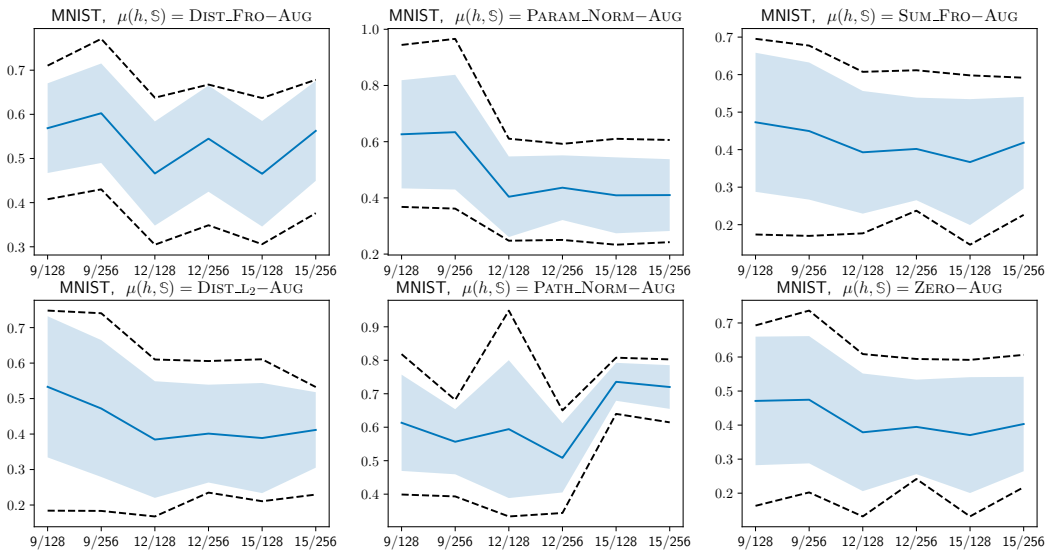


Figure 10: Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of $\alpha$; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

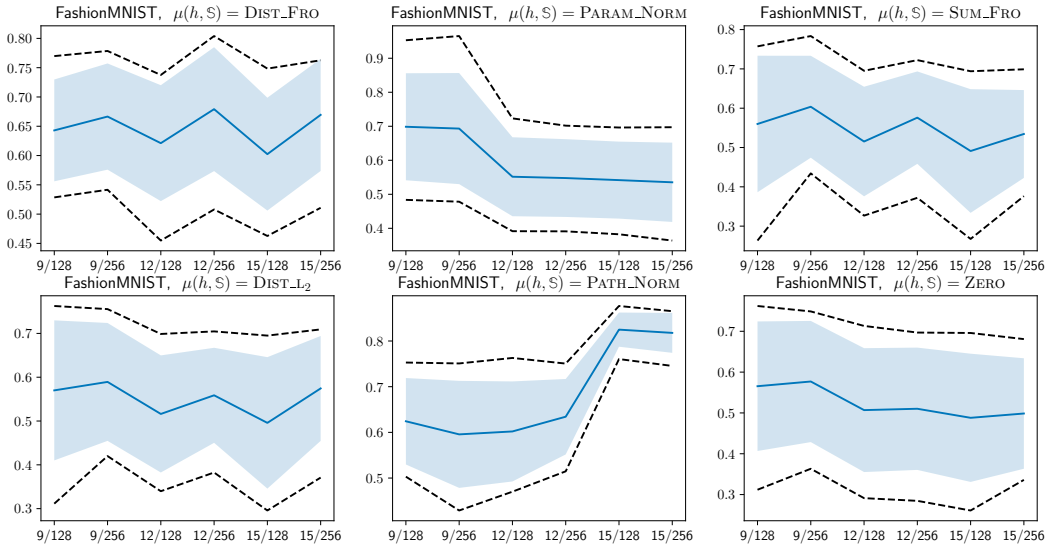


Figure 11: Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of $\alpha$; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

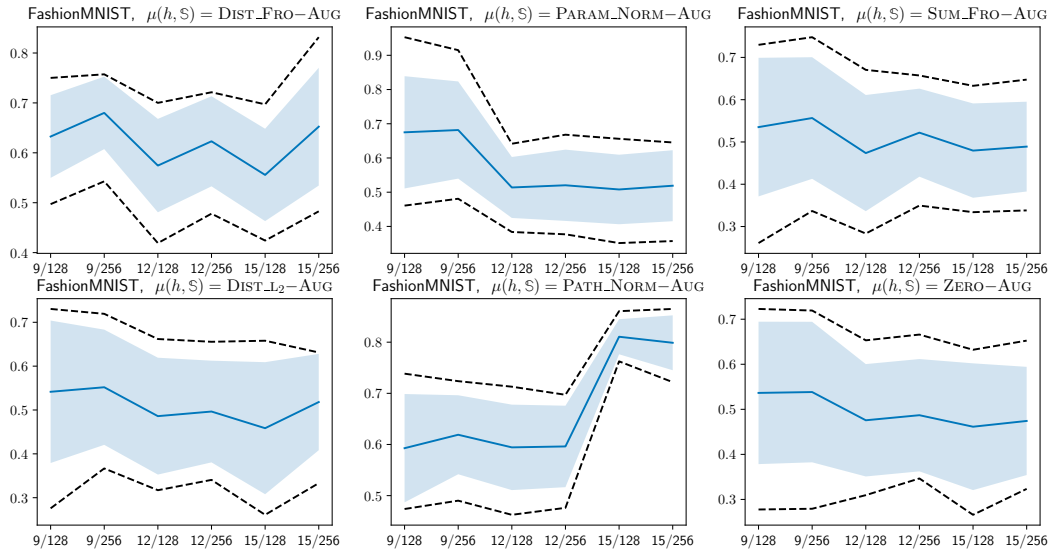


Figure 12: Influence of the depth and the width in the x-axis as "depth/width". The (solid) lines are the mean values computed on the different values of $\alpha$; the shadows are the standard deviation. The dashed lines are the minimum and the maximum values.

## F MAURER (2004)'S RESULT

**Lemma F.1** (Maurer (2004)). *For any distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, for any hypothesis set $\mathbb{H}$, for any loss $\ell : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y}) \to [0, 1]$, we have*

$$\forall h \in \mathbb{H}, \quad \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m} \exp\left[m \operatorname{kl}(R^\ell_{\mathcal{S}}(h) \| R^\ell_{\mathcal{D}}(h))\right] \leq 2\sqrt{m}.$$

## G PINSKER'S INEQUALITY

**Theorem G.1** (Pinsker's inequality). *For any $p \in (0, 1)$ and $q \in [0, 1]$, we have*

$$2(q - p)^2 \leq \operatorname{kl}(q\|p),$$

*where* $\operatorname{kl}(q\|p) \triangleq q \ln \frac{q}{p} + (1 - q) \ln \frac{1-q}{1-p}$.