Leveraging PAC-Bayes Theory and Gibbs Distributions for Generalization Bounds with Complexity Measures

Paul Viallard[*,1] **Rémi Emonet**[†,◇,§] **Amaury Habrard**[†,◇,§]

Emilie Morvant[†] **Valentina Zantedeschi**[‡]

[*] Univ Rennes, Inria, CNRS IRISA - UMR 6074, F35000 Rennes, France
[†] Université Jean Monnet Saint-Etienne, CNRS, Institut d Optique Graduate School, Inria[◇], Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, FRANCE
§ Institut Universitaire de France (IUF)
‡ ServiceNow Research

Abstract

In statistical learning theory, a generalization bound usually involves a complexity measure imposed by the considered theoretical framework. This limits the scope of such bounds, as other forms of capacity measures or regularizations are used in algorithms. In this paper, we leverage the framework of disintegrated PAC-Bayes bounds to derive a *general* generalization bound instantiable with arbitrary complexity measures. One trick to prove such a result involves considering a commonly used family of distributions: the Gibbs distributions. Our bound stands in probability jointly over the hypothesis and the learning sample, which allows the complexity to be adapted to the generalization gap as it can be customized 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. This is often done by bounding a deviation, called the generalization gap, between these risks. An upper bound on this gap is usually a function of two main quantities: *(i)* the size of the training set, *(ii)* a complexity measure that captures how prone a model is to overfitting. The higher the complexity, the higher the number of examples needed to obtain a tight bound on the gap. One limitation is that existing frameworks are restricted to specific complexity measures, *e.g.*, the VC-dimension [Vapnik and Chervonenkis, 1971] or the Rademacher complexity [Bartlett and Mendelson, 2002] (known to be large [Nagarajan and Kolter, 2019]).

Recently, Lee et al. [2020, Proposition 1] related arbitrary complexity measures to their usage in generalization bounds. Indeed, if we interpret this bound, it says that the generalization gap is upper-bounded by a user-defined complexity measure with high probability if the complexity measure is close to the generalization gap. However, this bound is uncomputable since it relies on a measure of closeness between the measure and the gap. Hence, to our knowledge, there is no computable generalization bound able to capture, by construction, an arbitrary complexity measure that can serve as a good proxy for the generalization gap.

In this paper, we tackle this drawback by leveraging the framework of disintegrated PAC-Bayesian bounds (Theorem 1) to propose a novel general generalization bound instantiable with arbitrary complexity measures. To do so, we incorporate a user-defined parametric function characterizing the complexity in a probability distribution over the hypothesis set, expressed as a Gibbs distribution (also called Boltzmann distribution). This trick allows us to derive guarantees in terms of probabilistic bounds that depend on a model sampled from this user-parametrized Gibbs distribution. It is worth noticing that our result is general enough to ob-

[1]This research began when the author was affiliated with Laboratoire Hubert Curien and finished at Inria Paris.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

tain bounds on well-known complexity measures such as the VC dimension or the Rademacher complexity. We believe that our result provides new theoretical foundations for understanding the generalization abilities of machine learning models and for performing model selection in practice. As an illustration, we empirically show how some arbitrary complexity measures, studied by Jiang et al. [2019b], Dziugaite et al. [2020], Jiang et al. [2021], can be integrated into our framework. Moreover, inspired by Lee et al. [2020], we investigate how our bounds behave when provided with a complexity measure learned via a neural network.

Paper's Organization. Section 2 provides preliminary definitions and concepts. Then, Section 3 presents our framework. In Section 4, we provide a practical instantiation of our framework.

2 PRELIMINARIES

2.1 Notations and Setting

We stand in a supervised classification setting where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. An example $(\mathbf{x}, y) \in \mathcal{X}\times\mathcal{Y}$ is drawn from an unknown data distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$. 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 a sample by $\mathcal{D}^m$. Let $\mathcal{H}$ be a possibly infinite set of hypotheses $h : \mathcal{X}\to\mathcal{Y}$ that return a label from $\mathcal{Y}$ given an input from $\mathcal{X}$. Let $\mathcal{M}(\mathcal{H})$ be the set of strictly positive probability densities on $\mathcal{H}$ given a reference measure (*e.g.*, the Lebesgue measure). Given $\mathcal{S}$ and a loss function $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R}$, we aim to find $h \in \mathcal{H}$ that minimizes the true risk $\mathrm{R}_{\mathcal{D}}^{\ell}(h)=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\,\ell(h,(\mathbf{x},y))$. As $\mathcal{D}$ is unknown, $\mathrm{R}_{\mathcal{D}}^{\ell}(h)$ is in practice estimated with its empirical counterpart: the empirical risk $\mathrm{R}_{\mathcal{S}}^{\ell}(h)=\frac{1}{m}\sum_{i=1}^m \ell(h,(\mathbf{x}_i,y_i))$. We denote the generalization gap by $\phi:\mathbb{R}^2\to\mathbb{R}$, which quantifies how much the empirical risk is representative of the true risk; it is usually defined by $\phi(\mathrm{R}_{\mathcal{D}}^{\ell}(h),\mathrm{R}_{\mathcal{S}}^{\ell}(h))=|\mathrm{R}_{\mathcal{D}}^{\ell}(h)-\mathrm{R}_{\mathcal{S}}^{\ell}(h)|$.

In this paper, we leverage the PAC-Bayesian setting [Shawe-Taylor and Williamson, 1997, McAllester, 1998] to bound the generalization gap with a function involving an arbitrary measure of complexity (see Guedj [2019], Hellström et al. [2023], Alquier [2024] for recent surveys). In PAC-Bayes, we assume an *apriori* belief on the hypotheses in $\mathcal{H}$ modeled by a prior distribution $\pi \in \mathcal{M}(\mathcal{H})$ on $\mathcal{H}$. Instead of looking for the best $h \in \mathcal{H}$, we aim to learn, from $\mathcal{S}$ and π, a *posterior* distribution $\rho\in\mathcal{M}(\mathcal{H})$ on $\mathcal{H}$ to assign higher probability to the best hypotheses in $\mathcal{H}$ (the support of ρ is included in the one of π). A PAC-Bayesian generalization bound provides an upper bound in expectation over ρ, meaning it bounds the generalization gap expressed as $|\,\mathbb{E}_{h\sim\rho}[\mathrm{R}_{\mathcal{D}}^{\ell}(h)-\mathrm{R}_{\mathcal{S}}^{\ell}(h)]|$. The complexity depends here on the KL divergence between ρ and π defined as $\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{h\sim\rho}\ln\frac{\rho(h)}{\pi(h)}$. This complexity captures how much ρ and π deviate in expectation over all the hypotheses. To incorporate a custom complexity in a bound, we follow a slightly different framework called the disintegrated PAC-Bayesian bound (see below) in which the expectations on ρ are *disintegrated*: for a single h sampled from ρ, it upper-bounds the gap $\phi(\mathrm{R}_{\mathcal{D}}^{\ell}(h),\mathrm{R}_{\mathcal{S}}^{\ell}(h))=|\mathrm{R}_{\mathcal{D}}^{\ell}(h)-\mathrm{R}_{\mathcal{S}}^{\ell}(h)|$.

2.2 Disintegrated PAC-Bayesian Bounds

We recall now the framework of disintegrated PAC-Bayesian bounds (introduced by Catoni [2007, Th 1.2.7] and Blanchard and Fleuret [2007, Prop 3.1]) on which our contribution is based. As far as we know, despite their significance, they have received little attention in the literature and have only received renewed interest for deriving tight bounds in practice recently (*e.g.*, Rivasplata et al. [2020], Hellström and Durisi [2020], Viallard et al. [2024]). Such bounds provide guarantees for a hypothesis h sampled from a posterior distribution $\rho_{\mathcal{S}}$, where $\rho_{\mathcal{S}}$ depends on the learning sample $\mathcal{S}\sim\mathcal{D}^m$. In fact, these bounds stand with high probability (at least $1-\delta$) over the random choice of learning sample $\mathcal{S}\sim\mathcal{D}^m$ *and* a hypothesis h. This paper mainly focuses on the bound of Rivasplata et al. [2020, Th.1 *(i)*] recalled below in Theorem 1.

Theorem 1 (General Disintegrated Bound of Rivasplata et al. [2020])**.** *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any distribution $\pi \in \mathcal{M}(\mathcal{H})$, for any measurable function $\varphi : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m \to \mathbb{R}$, for any $\delta \in (0, 1]$, we have with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathcal{S}}$*

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

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

Remark that φ can be any (measurable) function. However, it is usually defined as $\varphi(h,\mathcal{S}) = m\,\phi(\mathrm{R}_{\mathcal{D}}^{\ell}(h),\mathrm{R}_{\mathcal{S}}^{\ell}(h))$, which is a deviation between the true risk $\mathrm{R}_{\mathcal{D}}^{\ell}(h)$ and the empirical risk $\mathrm{R}_{\mathcal{S}}^{\ell}(h)$. The bound depends on two terms: *(a)* the *disintegrated* KL divergence $\ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}$ defining how much π and $\rho_{\mathcal{S}}$ deviate for a single h, *(b)* the term $\ln\big[\frac{1}{\delta}\,\mathbb{E}_{\mathcal{V}}\,\mathbb{E}_g \exp\left(\varphi(g,\mathcal{V})\right)\big]$ which is constant *w.r.t.* $h\in\mathcal{H}$ and $\mathcal{S}\in(\mathcal{X}\times\mathcal{Y})^m$. Note that, to instantiate the bound with a given φ, the rightmost term *(b)* is usually upper-bounded. In fact, it is constant *w.r.t.* the hypothesis $g\sim\pi$ and the learning sample $\mathcal{V}\sim\mathcal{D}^m$. Then, to integrate the relevance of the prior belief and for the sake of simplicity, in the rest of the paper, we refer to as "*complexity measure*" the

right-hand side of the bound. This is in slight contrast with the standard definition of complexity (*e.g.*, in the case of the VC-dimension or the Rademacher complexity), where the term *(b)* is not included in the definition.

In the bound of Theorem 1, the disintegrated KL divergence suffers from drawbacks: the KL complexity term is imposed by the framework and can be subject to high variance in practice [Viallard et al., 2024]. Despite this shortcoming, it is important to notice that the disintegrated KL divergence has a clear advantage: it only depends on the sampled hypothesis $h \sim \rho_{\mathcal{S}}$ and the data sample $\mathcal{S}$, instead of the whole hypothesis class (as it is the case, for instance, with the Rényi divergence in the disintegrated PAC-Bayesian bounds of Viallard et al. [2024], or with the bounds based on the VC-dimension, or the Rademacher complexity). This might imply a better correlation between the generalization gap and some complexity measures. In the next section, we leverage the disintegrated KL divergence to derive our main contribution: a general bound that involves arbitrary complexity measures.

3 INTEGRATING MEASURES IN GENERALIZATION BOUNDS

In Sections 3.1 and 3.2, we give intuitions about our contribution and recall notions about the Gibbs distribution. Second, we formalize our result in Section 3.3.

3.1 The Framework

The idea to introduce our notion of complexity measure is to parametrize the complexity with an additional "customizable" function $\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$ that we call *parametric function*. Thanks to the function μ, we define the randomized complexity measure $\Phi^r_\mu(h,\mathcal{S},\delta)$ as a real-valued function parameterized by μ and an external randomness $r\sim\mathcal{R}$ which takes as argument a hypothesis $h \in \mathcal{H}$, a learning sample $\mathcal{S}\in(\mathcal{X}\times\mathcal{Y})^m$, and δ. As we will see in Section 3.3, the bound we derive in Theorem 3 depends on the complexity measure $\Phi^r_\mu(h,\mathcal{S},\delta)$ and takes the following form.

Definition 2. *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ be a loss function, $\phi : \mathbb{R}^2\to\mathbb{R}$ be the generalization gap, and $\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$ be a parametric function. A generalization bound with a complexity measure is defined such that if for any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any distribution $\mathcal{R}$ representing the randomness, for any hypothesis set $\mathcal{H}$, there exists a randomized real-valued function $\Phi^r_\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\times(0,1]\to\mathbb{R}$ such that for any $\delta\in(0,1]$, we have*

$$\mathop{\mathbb{P}}_{r\sim\mathcal{R},\mathcal{S}\sim\mathcal{D}^m,h\sim\rho_{\mathcal{S}}}\left[\phi(R^\ell_{\mathcal{D}}(h),R^\ell_{\mathcal{S}}(h)) \le \Phi^r_\mu(h,\mathcal{S},\delta)\right]\ge 1-\delta,$$

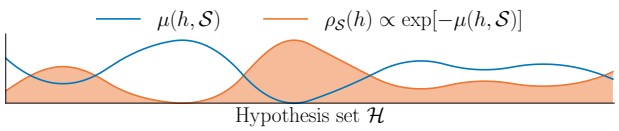

Figure 1: Illustration of the behavior of the Gibbs distribution $\rho_{\mathcal{S}}$ with a parametric function μ. The x-axis represents a (continuous) hypothesis set, and the y-axis the values of $\rho_{\mathcal{S}}$ and μ. The distribution $\rho_{\mathcal{S}}$ gives a higher probability to the hypotheses with a low μ value.

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

The main trick to obtain a bound that involves a parametrizable complexity measure is to consider a posterior distribution $\rho_{\mathcal{S}}$ that depends on μ. To do so, we propose to set $\rho_{\mathcal{S}}$ as the Gibbs distribution defined as

$$\rho_{\mathcal{S}}(h) \propto \exp\left[-\,\mu(h,\mathcal{S})\right]. \tag{1}$$

This formulation might look restrictive, but it can represent any probability density function provided that a relevant complexity measure is selected. For instance, let $\rho'_{\mathcal{S}}$ be a distribution on $\mathcal{H}$, *e.g.*, a Gaussian or a Laplace distribution, by setting $\mu(h,\mathcal{S}) = -\ln\rho'_{\mathcal{S}}(h)$ we can retrieve the distribution $\rho'_{\mathcal{S}}$. Moreover, this Gibbs distribution $\rho_{\mathcal{S}}$ is interesting from an optimization viewpoint: given a fixed learning sample $\mathcal{S}$, a hypothesis h is more likely to be sampled from it when $\mu(h,\mathcal{S})$ is low (see Figure 1 for an illustration). In fact, the function $h \mapsto \mu(h,\mathcal{S})$ can be seen as an objective function. For instance, to minimize the true risk $\mathrm{R}^\ell_{\mathcal{D}}(h)$, one can ideally set $\mu(h,\mathcal{S})=\alpha\mathrm{R}^\ell_{\mathcal{D}}(h)$ that is associated with a Gibbs distribution which samples hypotheses with small true risks and concentrates around the small risks when $\alpha\in\mathbb{R}^*_+$ increases. However, since the true risk is unknown, it must be replaced with a computable function μ. For instance, μ can be the empirical risk such as $\mu(h,\mathcal{S})=\alpha\mathrm{R}^\ell_{\mathcal{S}}(h)$.

3.2 Gibbs Distribution and Optimization

Given a differentiable parametric function defined by $\mu(h,\mathcal{S}) = \alpha\nu(h,\mathcal{S})$ (with α a concentration parameter), its associated Gibbs distribution can be related to the Stochastic Gradient Langevin Dynamics algorithm [SGLD, Welling and Teh, 2011] that learns the hypothesis $h\in\mathcal{H}$ by running iterations of the form

$$h_t \longleftarrow h_{t-1} - \eta\nabla\nu(h_t,\mathcal{S}) + \sqrt{\frac{2\eta}{\alpha}}\epsilon_t, \tag{2}$$

where $\epsilon_t\sim\mathcal{N}(\mathbf{0},\mathbf{I}_D)$, and h_t is the hypothesis learned at iteration $t\in\mathbb{N}$, and η is the learning rate, and α is the concentration parameter for the Gibbs distribution.

When α increases, the noise ϵ_t has less influence on the next iterate obtained from SGLD as $\sqrt{2\eta/\alpha}\epsilon_t$ decreases, and hence, minimizes better the function ν. Moreover, when the learning rate η tends to zero, SGLD becomes a continuous-time process called Langevin diffusion, defined as the stochastic differential equation in Equation (3). Indeed, Equation (2) can be seen as the Euler-Maruyama discretization [see, Raginsky et al., 2017] of Equation (3) defined for $t \geq 0$ as

$$dh_t = -\nabla\nu(h_t, \mathcal{S})dt + \sqrt{\frac{2}{\alpha}}B_t, \tag{3}$$

where B_t is the Brownian motion. Under some mild assumptions on the function ν, Chiang et al. [1987] show that the invariant distribution of the Langevin diffusion is the Gibbs distribution $\rho_{\mathcal{S}}$ with $\mu(h, \mathcal{S}) = \alpha\nu(h, \mathcal{S})$.

3.3 Bounds with Complexity Measures

We now introduce our main results, *i.e.*, generalization bounds with user-defined complexity measures. In Section 3.3.1, we present a general theorem that fulfills Definition 2. We specialize our result to uniform priors in Section 3.3.2 and informed priors in Section 3.3.3.

3.3.1 General Generalization Bound

We state below our theorem that introduces a bound on the generalization gap involving the parametric function μ, which stands for hypotheses sampled from the posterior distribution $\rho_{\mathcal{S}}(h) \propto \exp\left[-\mu(h, \mathcal{S})\right]$. Note that our bound is "general," meaning the generalization gap ϕ has to be further upper-bounded.

Theorem 3. *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ be a loss function and $\phi : \mathbb{R}^2\to\mathbb{R}$ be a generalization gap. For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\pi \in \mathcal{M}(\mathcal{H})$ on $\mathcal{H}$, for any $\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, for any $\delta \in (0, 1]$, we have with probability at least $1-\delta$ over $h' \sim \pi$, $\mathcal{S} \sim \mathcal{D}^m$, and $h \sim \rho_{\mathcal{S}}$*

$$\begin{aligned}\phi(R^\ell_{\mathcal{D}}(h), R^\ell_{\mathcal{S}}(h)) &\leq \mu(h', \mathcal{S}) - \mu(h, \mathcal{S}) + \ln\frac{\pi(h')}{\pi(h)} \\ &\quad + \ln\left[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(R^\ell_{\mathcal{D}}(g), R^\ell_{\mathcal{V}}(g))}\right] \\ &\triangleq \Phi^{h'}_\mu(h, \mathcal{S}, \delta),\end{aligned}$$

where $\rho_{\mathcal{S}}$ is the Gibbs distribution as in Equation (1).

The bound $\Phi^{h'}_\mu(h,\mathcal{S},\delta)$ of Theorem 3 depends on three terms: *(i)* the difference $\mu(h',\mathcal{S})-\mu(h,\mathcal{S})$, *(ii)* the log ratio $\ln(\pi(h')/\pi(h))$, *(iii)* a constant term $\ln[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{V}\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp[\phi(\mathrm{R}^\ell_{\mathcal{D}}(g), \mathrm{R}^\ell_{\mathcal{V}}(g))]]$. Compared to Theorem 1, we essentially upper-bound the disintegrated KL divergence $\ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}$ by the difference $\mu(h',\mathcal{S})-\mu(h,\mathcal{S})$ and the log ratio $\ln(\pi(h')/\pi(h))$. The advantage of these two terms is that they are easily computable, as long as we can compute $\mu(h', \mathcal{S})$, $\mu(h, \mathcal{S})$ and the density of π (up to its normalization constant). This is in contrast with the the result of Lee et al. [2020], that is essentially a bound that holds for all $\epsilon > 0$ and of the form,

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\Big[|\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathcal{S}}(h)| \leq \mu(h, \mathcal{S})+\epsilon\Big] \geq 1-\delta'(\epsilon),$$

where $\delta'(\epsilon)$ depends on n learning samples $\mathcal{S}_1, \ldots, \mathcal{S}_n$, on n hypotheses $h_1 \sim \rho_{\mathcal{S}_1}, \ldots, h_n \sim \rho_{\mathcal{S}_n}$, and on the unknown distribution $\mathcal{D}$. This result has no restriction on the form of the distribution $\rho_{\mathcal{S}}$, however, the dependence on $\mathcal{D}$ makes the term $\delta'(\epsilon)$ not computable (in contrast to our bound); see Appendix B.1 for more details. Note also that the term *(iii)* is usually negligible compared to *(i)* and *(ii)*, and it is upper-bounded when the generalization gap ϕ is instantiated. To get a bound that converges when m increases, it is sufficient to set ϕ as a function of m as it is done further. The tightness of the term *(ii)* depends on the instantiation of π; we propose two types of instantiation in Sections 3.3.2 and 3.3.3. Lastly, the term *(i)* depends on the choice of μ which has a big influence on the sampled hypothesis $h\sim\rho_{\mathcal{S}}$ and so on the gap $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h), \mathrm{R}^\ell_{\mathcal{S}}(h))$. For instance, when $\mu(h, \mathcal{S})=0$, the difference $\mu(h', \mathcal{S})-\mu(h, \mathcal{S})=0$, but, in this case, the posterior distribution $\rho_{\mathcal{S}}$ is uniform which does not permit to sample a hypothesis minimizing the true risk $\mathrm{R}^\ell_{\mathcal{D}}(h)$. There is hence a trade-off to find between minimizing this difference and sampling a hypothesis minimizing the gap $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h), \mathrm{R}^\ell_{\mathcal{S}}(h))$ and the true risk $\mathrm{R}^\ell_{\mathcal{D}}(h)$. In Section 4, we see how to instantiate the parametric function μ. Note that, when instantiated correctly, it also allows to get uniform-convergence-based and algorithm-dependent bounds; see Appendix C.

3.3.2 Practical Bound with Uniform Priors

The remaining challenge to get a practical bound is to upper-bound $\ln[\frac{4}{\delta^2}\,\mathbb{E}_{\mathcal{V}\sim\mathcal{D}^m}\,\mathbb{E}_{g\sim\pi}\exp[\phi(\mathrm{R}^\ell_{\mathcal{D}}(g), \mathrm{R}^\ell_{\mathcal{V}}(g))]]$ and $\ln(\pi(h')/\pi(h))$. As an illustration, we restrict ourselves in the rest of the paper to the case where the loss is bounded, *i.e.*, we consider a loss function $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y}) \to [0, 1]$. Under this assumption, we provide in the next corollary an instantiation of Theorem 3 for the generalization gap $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h), \mathrm{R}^\ell_{\mathcal{S}}(h))=m\,\mathrm{kl}[\mathrm{R}^\ell_{\mathcal{S}}(h)\|\mathrm{R}^\ell_{\mathcal{D}}(h)]$ 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 $q \in [0, 1]$ and with a uniform distribution π on a bounded set $\mathcal{H}$.

Corollary 4. *For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any bounded hypothesis set $\mathcal{H}$, given the uniform prior π on $\mathcal{H}$, for any loss $\ell : \mathcal{H} \times (\mathcal{X}\times\mathcal{Y}) \to [0, 1]$, for any $\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, 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*

$$\mathrm{kl}\Big[R^{\ell}_{\mathcal{S}}(h)\|R^{\ell}_{\mathcal{D}}(h)\Big] \leq \frac{\mu(h',\mathcal{S})-\mu(h,\mathcal{S})+\ln\frac{8\sqrt{m}}{\delta^2}}{m}, \quad (4)$$

with $\rho_{\mathcal{S}}$ *defined in Equation* (1).

Interestingly, Corollary 4 gives a computable bound on $\mathrm{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h)\|\mathrm{R}^{\ell}_{\mathcal{D}}(h)]$. From Equation (4), we obtain the following generalization bounds on the true risk $\mathrm{R}^{\ell}_{\mathcal{D}}(h)$:

$$\mathrm{R}^{\ell}_{\mathcal{D}}(h) \leq \overline{\mathrm{kl}}\left[\mathrm{R}^{\ell}_{\mathcal{S}}(h) \;\middle|\; \frac{\mu(h',\mathcal{S})-\mu(h,\mathcal{S})+\ln\frac{8\sqrt{m}}{\delta^2}}{m}\right], \quad (5)$$

with $\overline{\mathrm{kl}}[q|\tau]=\max\{p\in(0,1) \mid \mathrm{kl}(q\|p)\leq\tau\}$. We use these bounds in Section 4 to illustrate the generalization guarantees for different parametric functions μ. For some trivial cases, the convergence rate can be arbitrary, *e.g.*, when $\mu(h,\mathcal{S}) = m\mathrm{R}^{\ell}_{\mathcal{S}}(h)$. For example, for a large empirical risk $\mathrm{R}^{\ell}_{\mathcal{S}}(h')$ (which is common when h' is sampled from a uniform prior on $\mathcal{H}$), the right-hand side of Equation (4) simplifies to $\Phi^{h'}_{\mu}(h,\mathcal{S},\delta)=[(\mathrm{R}^{\ell}_{\mathcal{S}}(h')-\mathrm{R}^{\ell}_{\mathcal{S}}(h))+\frac{1}{m}\ln(2\sqrt{m}/\delta)]_+$ and is large, for all m. In order for the bound to be meaningful, we have to set μ such that the distribution $\rho_{\mathcal{S}}$ allows to sample h minimizing the empirical risk $\mathrm{R}^{\ell}_{\mathcal{S}}(h)$ and the generalization gap, and we want the complexity measure $\Phi^{h'}_{\mu}(h,\mathcal{S},\delta)$ to be tight (with $h'\sim\pi$).

3.3.3 Practical Bound with Informed Priors

While it is common to consider uninformed priors when we have no *apriori* belief, informative priors can be necessary and useful to get better results. For that purpose, one solution is to consider distribution-dependent priors, heavily used in PAC-Bayes [see *e.g.*, Parrado-Hernández et al., 2012, Dziugaite et al., 2021, Pérez-Ortiz et al., 2021]. We use a strategy similar to that for the posterior $\rho_{\mathcal{S}}$ by defining the prior π as follows

$$\pi(h) \propto \exp\left[-\omega(h)\right], \quad (6)$$

where $\omega : \mathcal{H}\to\mathbb{R}$ can depend on the distribution $\mathcal{D}$. Hence, the prior can depend on an additional learning sample $\mathcal{S}'\in(\mathcal{X}\times\mathcal{Y})^{m'}$ sampled from $\mathcal{D}$. We prove the following corollary with the prior of Equation (6).

Corollary 5. *For any* $\mathcal{D}$ *on* $\mathcal{X}\times\mathcal{Y}$, *for any hypothesis set* $\mathcal{H}$, *for any loss* $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,1]$, *for any* $\mu:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, *for any* $\omega:\mathcal{H}\to\mathbb{R}$, *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*

$$\mathrm{kl}\Big[R^{\ell}_{\mathcal{S}}(h)\|R^{\ell}_{\mathcal{D}}(h)\Big] \leq \frac{1}{m}\Big[[\mu(h',\mathcal{S})-\omega(h')] - [\mu(h,\mathcal{S})-\omega(h)] + \ln\frac{8\sqrt{m}}{\delta^2}\Big], \quad (7)$$

with $\rho_{\mathcal{S}}$ *and* π *resp. defined in Equations* (1) *and* (6).

4 USING COMPLEXITY MEASURES IN PRACTICE

Section 4.1 presents our experimental setting. In Section 4.2, we first compare Corollaries 4 and 5 to two additional bounds with $\mu(h,\mathcal{S}) = \alpha\mathrm{R}^{\ell'}_{\mathcal{S}}(h)$ (where ℓ' is a differentiable loss). In Section 4.3 we study the behavior of our bounds when the parametric function μ is defined as a regularized empirical risk. Finally, in Section 4.4 we assess the tightness of our bounds when the complexity term is learned with a neural network.

4.1 General Experimental Setting[2]

In this section, we investigate the tightness of Corollaries 4 and 5's bounds on the MNIST [LeCun et al., 1998] and FashionMNIST [Xiao et al., 2017] datasets. Specifically, we consider the bounds on the true risk and the empirical risk endowed with the 01-loss $\ell(h,(\mathbf{x},y))=\mathrm{I}[h(\mathbf{x})\neq y]$ where $\mathrm{I}[a]=1$ if a is true and 0 otherwise.

Model. Inspired by the setting of Viallard et al. [2024], we train an All Convolutional Network [Springenberg et al., 2015] that is fitted for the two datasets MNIST and FashionMNIST. This network comprises 4 convolutional layers of 10 channels and a kernel of size 5×5 followed by a Leaky ReLU activation function (where the padding and the stride are set to 1 except for the second layer where the stride is set to 2). Finally, the network ends with an average pooling of size 8×8 followed by a Softmax activation function. The weights are initialized with the Xavier Glorot uniform initializer [Glorot and Bengio, 2010], and the biases are initialized uniformly between $-1/\sqrt{250}$ and $+1/\sqrt{250}$ for all biases instead for the first layer they are initialized uniformly in $[-1/5,+1/5]$.

Datasets. We keep the original test set $\mathcal{T}$ to estimate the true risk that we refer to as test risk $\mathrm{R}_{\mathcal{T}}(h)$. To evaluate Corollary 4's bounds and to sample $h\sim\rho_{\mathcal{S}}$, we keep the original learning set $\mathcal{S}$ to evaluate. To evaluate Corollary 5's bound, and to sample $h\sim\rho_{\mathcal{S}}$ and $h'\sim\pi$, the original learning set is split into 2 sets $\mathcal{S}$ and $\mathcal{S}'$ respectively of size m and m'; When $\frac{m'}{m+m'}=0$, the prior distribution is the uniform (non-data-dependent) and we retrieve Corollary 4.

Sampling and Bound Computation. To compute Corollaries 4 and 5's bounds, we aim to sample $h\sim\rho_{\mathcal{S}}$ and $h'\sim\pi$ by performing SGLD[3] (described in Equation (2)) and to evaluate these hypotheses with $\overline{\mathrm{kl}}$. Note that using SGLD is efficient for sampling since it does not require computing the normalization

[2]The source code is available at `https://github.com/paulviallard/AISTATS24-Complexity-Measures`.

[3]Dziugaite and Roy [2018] also perform SGLD to sample from a Gibbs distribution.

constants of the two distributions. To tune the learning rate, *(1)* we compute the mean loss over the learning sample (without training), *(2)* we start with a learning rate of size 0.1, and we decrease it (by a factor of 0.1) and reinitialize the model after each epoch if the mean loss is not decreasing (to retrain from scratch), *(3)* if the learning rate attains 10^{-10}, we set the learning rate to its starting value 0.1 and start learning from scratch. Once the initial learning rate that decreases the mean loss is set, we perform SGLD for 10 epochs and with a mini-batch of size 64. After each epoch, we decrease the learning rate by a factor of 0.5. Whenever we have to sample a risk value with SGLD, we replace the 01-loss by the differentiable bounded cross entropy of Dziugaite and Roy [2018] $\ell'(h,(\mathbf{x},y)) = -\frac{1}{4}\ln(e^{-4}+(1-2e^{-4})h(\mathbf{x})[y])$, where $h[y]$ is the probability assigned to the label y by h. The advantage of Dziugaite and 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. For all experiments, we perform 5 runs to obtain a mean and a standard deviation, which involves sampling from $\rho_{\mathcal{S}}$ and π for each evaluation of the bounds and risks.

4.2 Experiments on the Empirical Risk

In this section, we compare the tightness of our bounds of Corollaries 4 and 5 with bounds that share similarities with the literature. More precisely, this comparison is done for the Gibbs distribution $\rho_{\mathcal{S}}$ defined with the parametric function $\mu(h,\mathcal{S}) = \alpha \mathrm{R}^{\ell'}_{\mathcal{S}}(h)$. Indeed, this Gibbs distribution was already studied in the classical and disintegrated PAC-Bayesian theory but led to uncomputable bounds. We adapt these bounds to make them computable so that we can report them as baselines (more details are given in Appendix B). More precisely, we compare our bounds to the following one (similar to Lever et al. [2013]): with probability at least $1-\delta$, we have

$$\mathrm{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h)\|\mathrm{R}^{\ell}_{\mathcal{D}}(h)] \leq \frac{1}{m}\left[\frac{\alpha^2}{8m} + \sqrt{\frac{\alpha^2}{2m}\ln\frac{6\sqrt{m}}{\delta}} + \ln\frac{6\sqrt{m}}{\delta}\right]. \quad (8)$$

We also adapt the proof technique of Dziugaite and Roy [2018] to obtain with probability at least $1-\delta$

$$\mathrm{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h)\|\mathrm{R}^{\ell}_{\mathcal{D}}(h)] \leq \frac{1}{m}\Big[\alpha\Big[\mathrm{R}^{\ell'}_{\mathcal{S}}(h') - \mathrm{R}^{\ell'}_{\mathcal{S}}(h)\Big] + \alpha'\Big[\mathrm{R}^{\ell'}_{\mathcal{S}}(h) - \mathrm{R}^{\ell'}_{\mathcal{S}}(h')\Big] + 2\alpha' + \ln\frac{8\sqrt{m}}{\delta^2}\Big], \quad (9)$$

where $h' \sim \pi$ and $\pi(h) \propto \exp[-\alpha' \mathrm{R}^{\ell'}_{\mathcal{S}}(h)]$. For Corollary 5, we set the prior π with the parametric function $\omega(h) = \alpha \mathrm{R}^{\ell'}_{\mathcal{S}'}(h)$ and with a learning sample $\mathcal{S}'$ satisfying the split ratio $\frac{m'}{m+m'} = 0.5$. For all the bounds, we take α (and α') uniformly spaced on the logarithmic scale between $\sqrt{m}$ and m. For each parameter α and bound, we select the prior minimizing the bound and average its value over 5 runs. We report in Figure 2 the evolution of the different bounds and the test risks *w.r.t.* to α.

Analysis. As expected, the standard deviations of all the bounds are small only for large α, as this parameter controls the concentration of the Gibbs distributions. A larger α tends to imply lower test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$. However, the bounds become large as α increases except for our bound of Corollary 5. This is an expected behavior in the case of Equation (8) since the bound increases when α increases. For Corollary 4, the bound is large when the difference $\alpha[\mathrm{R}^{\ell'}_{\mathcal{S}}(h') - \mathrm{R}^{\ell'}_{\mathcal{S}}(h)]$ is large. This is effectively the case because $\mathrm{R}^{\ell'}_{\mathcal{S}}(h')$ is large since h' is sampled from a uniform distribution and $\mathrm{R}^{\ell'}_{\mathcal{S}}(h)$ is small because h is sampled from $\rho_{\mathcal{S}}(h) \propto \exp[-\alpha \mathrm{R}^{\ell'}_{\mathcal{S}}(h)]$. The same phenomenon arises with Equation (9) since $\mathrm{R}^{\ell'}_{\mathcal{S}}(h')$ is large when α' is small, *i.e.*, the concentration is not sufficient to minimize the empirical risk. The tightness of Corollary 5 comes from the fact that both empirical risks for $h' \sim \pi$ and $h \sim \rho_{\mathcal{S}}$ are small, and so is the bound when the risks $\mathrm{R}^{\ell'}_{\mathcal{S}}(h')$ and $\mathrm{R}^{\ell'}_{\mathcal{S}'}(h)$ are small as well. Moreover, note that, for small α, the test risks and the bound values are higher compared to the others. This is due to the fact that we use half of the data ($\frac{m'}{m+m'}=0.5$) for learning an informed prior. Indeed, the value of α is twice as small as for the other bounds, which makes the bound values and the test risks higher as the Gibbs distribution is less concentrated.

4.3 Experiments on Regularized Risks

In order to tighten the bounds in Corollaries 4 and 5, one might assume that selecting a hypothesis with a small trade-off between its empirical risk and a norm is a reasonable solution. Against all odds, we will see in this section that regularizing the empirical risk with a parametric function does not help to tighten the bounds. To define the norms used as regularizers, we assume that the model h, composed of L layers, is parameterized by weights (and biases) $\mathbf{w} \in \mathbb{R}^d$; we denote by $h_{\mathbf{w}^2}$ the hypothesis h that has its weights replaced by $\mathbf{w}^2$. We define by $\mathbf{w}_i$ the weights and biases on the i-th layer. Moreover, we denote the parameters obtained at initialization by $\mathbf{v} \in \mathbb{R}^d$. Thanks to this additional notation, we can now define 6 parametric functions μ associated with 6 Gibbs distributions (and bounds). We consider regularized empirical risks with the optimizable norms studied by Jiang et al. [2019b, Sec.C] and defined as follows:

- $\textsc{DistFro}^{\boldsymbol{R}}_{\beta}(h,\mathcal{S}) = \alpha(\beta \mathrm{R}^{\ell'}_{\mathcal{S}}(h) + \bar{\beta}\textsc{DistFro}(h,\mathcal{S}))$,
- $\textsc{DistL}_2{}^{\boldsymbol{R}}_{\beta}(h,\mathcal{S}) = \alpha(\beta \mathrm{R}^{\ell'}_{\mathcal{S}}(h) + \bar{\beta}\textsc{DistL}_2(h,\mathcal{S}))$,
- $\textsc{ParNorm}^{\boldsymbol{R}}_{\beta}(h,\mathcal{S}) = \alpha(\beta \mathrm{R}^{\ell'}_{\mathcal{S}}(h) + \bar{\beta}\textsc{ParNorm}(h,\mathcal{S}))$,

- $\textsc{PathNorm}_\beta^{\boldsymbol{R}}(h,\mathcal{S}) = \alpha(\beta \mathrm{R}_{\mathcal{S}}^{\ell'}(h) + \bar{\beta}\textsc{PathNorm}(h,\mathcal{S}))$,
- $\textsc{SumFro}_\beta^{\boldsymbol{R}}(h,\mathcal{S}) = \alpha(\beta \mathrm{R}_{\mathcal{S}}^{\ell'}(h) + \bar{\beta}\textsc{SumFro}(h,\mathcal{S}))$,
- $\textsc{Gap}_\beta^{\boldsymbol{R}}(h,\mathcal{S}) = \alpha(\beta \mathrm{R}_{\mathcal{S}}^{\ell'}(h) + \bar{\beta}\textsc{Gap}(h,\mathcal{S}))$,

where $\bar{\beta} = 1-\beta$ and with

- $\textsc{DistFro}(h,\mathcal{S}) = \sum_{i=1}^{L} \|\mathbf{w}_i - \mathbf{v}_i\|_2$,
- $\textsc{DistL}_2(h,\mathcal{S}) = \|\mathbf{w}-\mathbf{v}\|_2$,
- $\textsc{ParNorm}(h,\mathcal{S}) = \sum_{i=1}^{L} \|\mathbf{w}_i\|_2^2$,
- $\textsc{PathNorm}(h,\mathcal{S}) = \sum_{y\in\mathcal{Y}} h_{\mathbf{w}^2}(\mathbf{1})[y]$,
- $\textsc{SumFro}(h,\mathcal{S}) = L[\prod_{i=1}^{L} \|\mathbf{w}_i\|_2^2]^{\frac{1}{L}}$,
- $\textsc{Gap}(h,\mathcal{S}) = |\mathrm{R}_{\mathcal{T}}^{\ell'}(h) - \mathrm{R}_{\mathcal{S}}^{\ell'}(h)|$.

Remark that Gap is not a function that can be computed in practice, however, it can be interpreted as an ideal case when we want the norm to be representative of the gap and check if it correlates with it as done by Jiang et al. [2019a]. Note that during SGLD, we do not evaluate the gap on the whole learning samples $\mathcal{S}$ and $\mathcal{T}$ but on a mini-batch of size 64. Moreover, by taking into account the Dziugaite and Roy [2018]'s bounded cross-entropy instead of the classical one allows the parametric function not to focus too much on the risk since we want to take into account the norm. We consider the bounds of Corollaries 4 and 5 with a split ratio of $\frac{m'}{m+m'}=0.5$ and $\frac{m'}{m+m'}=0.0$ while we set the parametric function ω associated with π to $\omega(h_{\mathbf{w}}) = \mu(h_{\mathbf{w}}, \mathcal{S}')$ (the same function as for $\rho_{\mathcal{S}}$). We report in Figure 3 the evolution of the test risks $\mathrm{R}_{\mathcal{T}}(h)$ and the bound values for the different parametric functions as a function of the trade-off parameter β.

Analysis. The main striking result is that the Gap behaves differently than the norms. Its test risk rapidly decreases (until $\beta = 0.3$) while the associated bound remains tight. In contrast, the norms' curves show two regimes depending on the split ratio $\frac{m'}{m+m'}$. For instance, when $\frac{m'}{m+m'} = 0.0$, the test risks and the bounds decrease when $\beta \geq 0.7$ but their gap increases. When $\frac{m'}{m+m'} = 0.5$, the test risks decrease for $\beta \geq 0.7$ and the bounds stay tight. Note that the bounds and test risks for ParNorm stay high because SGLD fails to minimize the regularized risk. This experiment suggests that the norms are not good predictors of the generalization gap, as the norms' bounds are not close to the ideal one, given by Gap.

4.4 Experiments on Neural Complexities

In light of the previous results, we now study how our bounds behave when computed with a better predictor of the generalization gap. Indeed, while Sections 4.2 and 4.3 focus on hypotheses that minimize (regularized) empirical risks, we are ideally interested in concentrating the probability measure associated with $\rho_{\mathcal{S}}$ on the hypotheses with small generalization gaps. To do so, the parametric function for $\rho_{\mathcal{S}}$ can depend on an estimation of the gap (this latter being not available in practice). In this section, we consider the bound of Corollary 4 (without data-dependent priors) and study the following parametric functions μ

$$\mu(h,\mathcal{S}) = f^{\boldsymbol{D}}(h,\mathcal{S}) = \alpha |f(h,\mathcal{S}) - f(h_{\mathrm{SGD}},\mathcal{S})|,$$

where $f \in \{\textsc{DistFro}, \textsc{DistL}_2, \textsc{ParNorm}, \textsc{PathNorm}, \textsc{SumFro}, \textsc{Gap}\}$, and $\alpha = m$, and h_{SGD} is obtained by Stochastic Gradient Descent (SGD). This particular choice of μ allows us to sample hypotheses close to the value of the parametric function f evaluated on h_{SGD}. We additionally assess a parametric function $\textsc{Neural}^{\boldsymbol{D}}$, consisting of a neural network learned to predict the generalization gap. More precisely, we learn the function $\textsc{Neural}(h,\mathcal{S})$, which becomes the output of a feed-forward neural network (learned from $\mathcal{S}$), taking the parameters $\mathbf{w}$ of the model h and outputting a positive real that must represent the generalization gap. $\textsc{Neural}^{\boldsymbol{D}}$ is thus the function comparing the output of the feed-forward neural network associated with $h \sim \rho_{\mathcal{S}}$ and h_{SGD}. Note that learning a neural network for predicting the gap was previously proposed by [Lee et al., 2020]; we refer the reader to Appendix D.3 for a discussion and a detailed presentation of the learning setting. To obtain the model, we first run SGD on a random number of epochs uniformly sampled between 1 and 10, and with the same parameters as SGLD (Section 4.1). To sample $h \sim \rho_{\mathcal{S}}$, we start from h_{SGD} as the initialization, and we run SGLD for 10 epochs (unless the learning rate attains 10^{-10}). Finally, we consider a second setting, with the parametric functions are noted DistFro, $\textsc{DistL}_2$, ParNorm, PathNorm, SumFro, Gap (scaled by $\alpha = m$) and Neural (without $\boldsymbol{D}$), where the parametric function $\mu(h,\mathcal{S})$ is evaluated *w.r.t.* the initial value of h instead of h_{SGD}. Note that for Gap and $\textsc{Gap}^{\boldsymbol{D}}$, we skip the SGLD phase to have a bound on the model obtained from SGD directly; these two parametric functions correspond to our ideal cases. We plot in Figure 4 the mean and the standard deviation of the bound values and test risks (averaged over 5 runs) for the considered parametric functions. We provide additional experiments on Neural and $\textsc{Neural}^{\boldsymbol{D}}$ in Appendix D.3.

Analysis. As a first general remark, the bounds with complexity measures based on norms behave differently than the ones based on Neural, $\textsc{Neural}^{\boldsymbol{D}}$, Gap, and $\textsc{Gap}^{\boldsymbol{D}}$. Indeed, the mean bound values for the complexity measures based on the norms are all vacuous (*i.e.*, they are equal to 1), while their test risks are high for DistFro, $\textsc{DistL}_2$, ParNorm, PathNorm, and SumFro, which is expected since we want to sample a hypothesis with a low norm (and not far from the initialization). Similarly, for $\textsc{DistFro}^{\boldsymbol{D}}$, $\textsc{DistL}_2{}^{\boldsymbol{D}}$,

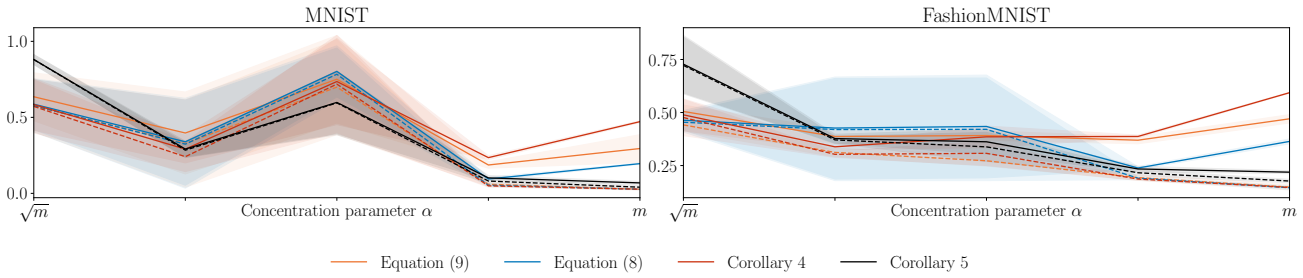

Figure 2: Evolution of the bounds (the plain lines) and the test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$ (the dashed lines) *w.r.t.* the concentration parameter α. The lines correspond to the mean, while the bands are the standard deviations.

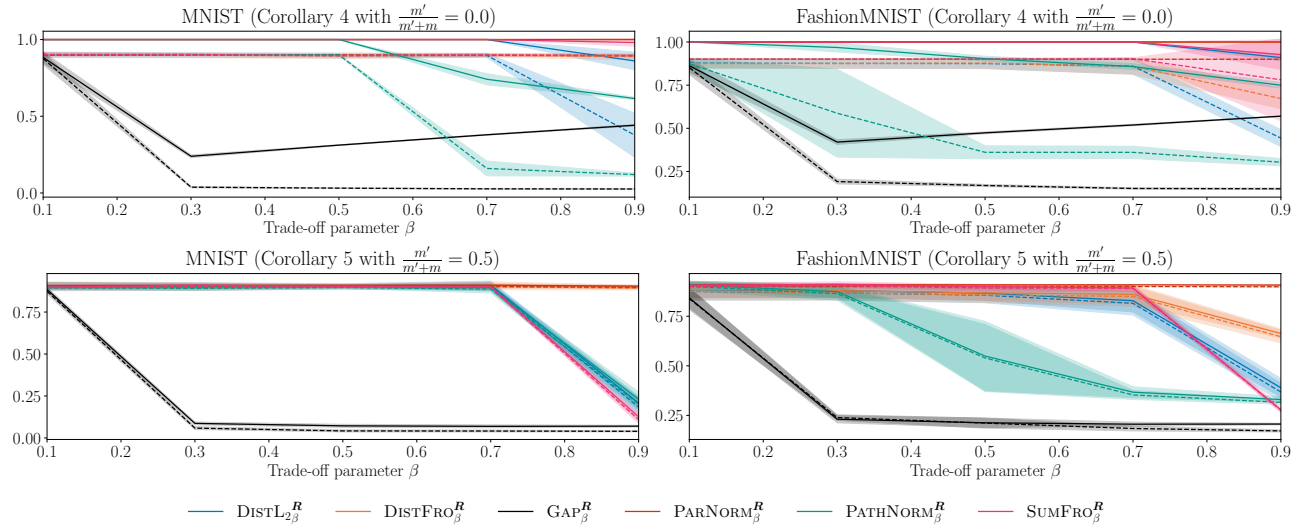

Figure 3: Evolution of the bounds (the plain lines) and the test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$ (the dashed lines) *w.r.t.* the trade-off parameter β for $\alpha = m$. The lines correspond to the mean, while the bands are the standard deviations.

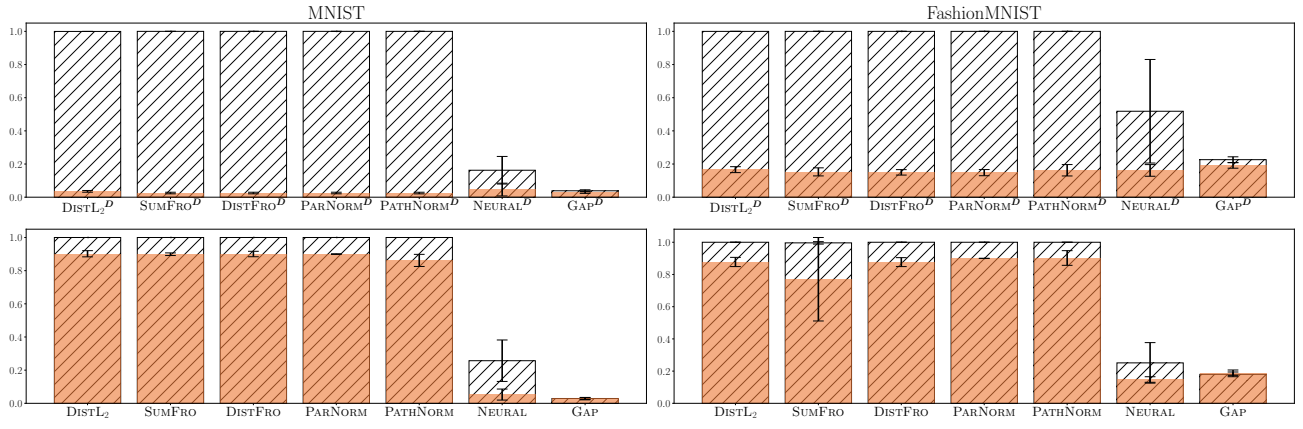

Figure 4: Bar plot of the bound value associated with Corollary 4 and the different parametric functions. The mean bound values of the sampled hypotheses $h \sim \rho_{\mathcal{S}}$ are shown with the hatched bars, and the mean test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$ are plotted in the colored bars. Moreover, the standard deviations are plotted in black.

ParNorm$^{\boldsymbol{D}}$, PathNorm$^{\boldsymbol{D}}$, and SumFro$^{\boldsymbol{D}}$, we want to sample a hypothesis with a norm that is close to the one of h_{SGD} and thus a hypothesis with a test risk $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$ close to $\mathrm{R}^{\ell}_{\mathcal{T}}(h_{\text{SGD}})$. In these cases, the bounds are vacuous because the parametric functions evaluated on h are close to zero, and the ones evaluated on h' are high. This highlights a drawback of the empirical studies of Jiang et al. [2019b], Dziugaite et al. [2020]: they study the correlation between the norms and the generalization gaps on *trained* neural networks. However, considering a norm as a good proxy for the generalization gap is impossible in this case. Indeed, rescaling the weights of the networks (by a scalar) gives the exact same predictions and thus keeps the same generalization gap while changing the norm; this is due to the use of non-negative homogeneous activation functions, such as the standard (Leaky) RELU (see *e.g.*, [Neyshabur et al., 2015, Dinh et al., 2017]. In contrast, the two parametric functions Neural and Neural$^{\boldsymbol{D}}$ give tight bounds and are close to the ideal bounds of Gap and Gap$^{\boldsymbol{D}}$.[4] This clearly illustrates that learning a parametric function (and so a complexity measure) can help to obtain tighter generalization bounds. Note that the bounds with Neural and Neural$^{\boldsymbol{D}}$ are tight even without a data-dependent prior, which is usually needed to obtain tight bounds for neural networks (see *e.g.*, [Dziugaite and Roy, 2017, Dziugaite et al., 2021, Pérez-Ortiz et al., 2021, Viallard et al., 2024]). This is an encouraging result and a step toward eliminating the need for data-dependent priors in PAC-Bayes to obtain tight bounds for neural networks.

5 CONCLUSION

In contrast to classical statistical learning theory frameworks, for which a complexity measure is imposed, we provide a generic and novel generalization bound where the user can choose any parametric function acting as a complexity. This measure incorporates a data and model-dependent function, which can be devised to favor desired properties for the hypotheses. In particular, we show that when such a function is learned to be representative of the generalization gap, our bounds are tight even without data-dependent priors. To the best of our knowledge, our framework is one of the few general enough to bring theoretical guarantees for learned complexity measures and for ones used in practice, *e.g.*, based on some weight norms. Last but not least, we believe this work paves the way for new research directions on bridging the gap between statistical learning theory and practice. Indeed, our framework could provide meaningful insights into the generalization of deep models by plugging in new complexity measures, e.g., given by: *(i)* learning an interpretable model based on features such as training trajectory and network configuration, or *(ii)* new handcrafted parametric functions that are simple but predictive of generalization.

[4] For Gap, the bounds are sometimes lower than the test risks. This is normal if the gap of h_{SGD} is much higher than 0 because sampling h_{SGD} is unlikely in this context.

Acknowledgements

This work was partially funded by the French ANR projects APRIORI ANR-18-CE23-0015, ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), and FAMOUS ANR-23-CE23-0019.

References

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*, 2024.

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 *Advances in Neural Information Processing System (NeurIPS)*, 2021.

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

Gilles Blanchard and François Fleuret. Occam's Hammer. In *Annual Conference on Learning Theory (COLT)*, 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 Journal on Selected Areas in Information Theory*, 2020.

Olivier Catoni. Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *arXiv*, 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.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. In *International Conference on Machine Learning (ICML)*, 2017.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in Adaptive Data Analysis and Holdout Reuse. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing System (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 *Advances in Neural Information Processing System (NeurIPS)*, 2020.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel M. Roy. On the role of data in PAC-Bayes Bounds. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian Learning of Linear Classifiers. In *International Conference on Machine Learning (ICML)*, 2009.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

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)*, 2017.

Benjamin Guedj. A Primer on PAC-Bayesian Learning. *arXiv*, abs/1901.05353, 2019.

Fredrik Hellström and Giuseppe Durisi. Generalization Bounds via Information Density and Conditional Information Density. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization Bounds: Perspectives from Information Theory and PAC-Bayes. *arXiv*, abs/2309.04381, 2023.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, 2015.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the Generalization Gap in Deep Networks with Margin Distributions. In *International Conference on Learning Representation (ICLR)*, 2019a.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. In *International Conference on Learning Representation (ICLR)*, 2019b.

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.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representation (ICLR)*, 2015.

Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-Dependent Analysis of Gibbs-ERM Principle. In *Annual Conference on Learning Theory (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/`.

Yoonho Lee, Juho Lee, Sung Ju Hwang, Eunho Yang, and Seungjin Choi. Neural Complexity Measures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks. In *Advances in Neural Information Processing System (NeurIPS)*, 2019.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 2013.

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

David McAllester. Some PAC-Bayesian Theorems. In *Annual Conference on Computational Learning Theory (COLT)*, 1998.

Frank McSherry and Kunal Talwar. Mechanism Design via Differential Privacy. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2007.

Ilya Mironov. Rényi Differential Privacy. In *IEEE Computer Security Foundations Symposium*, 2017.

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

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing System (NeurIPS)*, 2019.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-Normalized Optimization in Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

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. *Journal of Machine Learning Research*, 2021.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *IEEE Information Theory Workshop*, 2016.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Annual Conference on Learning Theory (COLT)*, 2017.

David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian Processes by Minimizing PAC-Bayesian Generalization Bounds. In *Advances in Neural Information Processing System (NeurIPS)*, 2018.

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 C. Williamson. A PAC Analysis of a Bayesian Estimator. In *Annual Conference on Computational Learning Theory (COLT)*, 1997.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. In *International Conference on Learning Representations (ICLR) – Workshop Track*, 2015.

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 practical disintegration of PAC-Bayesian bounds. *Machine Learning*, 2024.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *International Conference on Machine Learning (ICML)*, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv*, abs/1708.07747, 2017.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing System (NeurIPS)*, 2017.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

The appendix is organized as follows:

(i) Appendix A is dedicated to the proof of Theorem 3 (in Appendix A.1), and to the proof of Corollaries 4 and 5 (in Appendix A.2),

(ii) Appendix B is dedicated to the theoretical results related to the comparison with other theoretical results of the literature,

(iii) In Appendix C, we explain how to obtain uniform-convergence and algorithmic-dependent bounds by setting appropriately the parametric function,

(iv) Additional details on the experiments are provided in Appendix D.

A PROOF OF THE MAIN RESULTS

This section is dedicated to the proof of the results. More precisely, in Appendix A.1, we provide the proof of Theorem 3, whereas in Appendix A.2, we prove Corollaries 4 and 5.

A.1 Proof of Theorem 3

Theorem 3. *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ be a loss function and $\phi:\mathbb{R}^2\to\mathbb{R}$ be a generalization gap. For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\pi\in\mathcal{M}(\mathcal{H})$ on $\mathcal{H}$, for any $\mu:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, for any $\delta\in(0,1]$, we have with probability at least $1-\delta$ over $h'\sim\pi$, $\mathcal{S}\sim\mathcal{D}^m$, and $h\sim\rho_{\mathcal{S}}$*

$$\phi(R^\ell_{\mathcal{D}}(h), R^\ell_{\mathcal{S}}(h)) \le \mu(h',\mathcal{S}) - \mu(h,\mathcal{S}) + \ln\frac{\pi(h')}{\pi(h)} + \ln\left[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(R^\ell_{\mathcal{D}}(g),R^\ell_{\mathcal{V}}(g))}\right] \triangleq \Phi^{h'}_{\mu}(h,\mathcal{S},\delta),$$

where $\rho_{\mathcal{S}}$ is the Gibbs distribution as in Equation (1)*.*

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

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

We apply Theorem 1 with $\frac{\delta}{2}$ instead of δ and with the function $\varphi(h,\mathcal{S})=\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h))$ to obtain

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

We develop the term $\ln\left[\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}\right]$ in Theorem 1. We have

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

Hence, we obtain the following inequality

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\rho_{\mathcal{S}}}\left[\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) \le \ln\left[\frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(\mathrm{R}^\ell_{\mathcal{D}}(g),\mathrm{R}^\ell_{\mathcal{V}}(g))}\right] - \mu(h,\mathcal{S}) - \ln\left(\mathbb{E}_{g\sim\pi}\frac{\pi(h)}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right)\right] \ge 1-\frac{\delta}{2}. \tag{10}$$

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

$$\forall h\in\mathcal{H},\quad \forall\mathcal{S}\in(\mathcal{X}\times\mathcal{Y})^m,\quad \mathop{\mathbb{P}}_{h'\sim\pi}\left[\frac{\pi(h)}{\pi(h')}e^{-\mu(h',\mathcal{S})}\le\frac{2}{\delta}\mathop{\mathbb{E}}_{g\sim\pi}\frac{\pi(h)}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right]\ge 1-\frac{\delta}{2}$$
$$\iff \mathop{\mathbb{P}}_{h'\sim\pi}\left[-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\frac{\pi(h)}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right)\le\ln\frac{2}{\delta}-\ln\left(\frac{\pi(h)}{\pi(h')}e^{-\mu(h',\mathcal{S})}\right)\right]\ge 1-\frac{\delta}{2}.$$

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

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

Hence, we obtain the following inequality

$$\mathop{\mathbb{P}}_{h'\sim\pi}\left[-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\frac{\pi(h)}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right)\le\ln\frac{2}{\delta}+\ln\frac{\pi(h')}{\pi(h)}+\mu(h',\mathcal{S})\right]\ge 1-\frac{\delta}{2}. \tag{11}$$

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

A.2 Proof of Corollaries 4 and 5

In order to prove Corollaries 4 and 5, we first provide another corollary that will be necessary.

Corollary 6. *For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any loss $\ell:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,1]$, for any prior distribution $\pi\in\mathcal{M}(\mathcal{H})$ on $\mathcal{H}$, for any $\mu:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, 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*

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

Proof. We instantiate Theorem 3 with $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h))=m\,\mathrm{kl}[\mathrm{R}^\ell_{\mathcal{S}}(h)\|\mathrm{R}^\ell_{\mathcal{D}}(h)]$. By rearranging the term, we have

$$\mathrm{kl}[\mathrm{R}^\ell_{\mathcal{S}}(h)\|\mathrm{R}^\ell_{\mathcal{D}}(h)]\le\frac{1}{m}\left[\mu(h',\mathcal{S})-\mu(h,\mathcal{S})+\ln\frac{\pi(h')}{\pi(h)}+\ln\left[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}e^{m\,\mathrm{kl}[\mathrm{R}^\ell_{\mathcal{S}}(g)\|\mathrm{R}^\ell_{\mathcal{D}}(g)]}\right]\right].$$

We also have to upper-bound $\mathbb{E}_{\mathcal{V}\sim\mathcal{D}^m}\mathbb{E}_{g\sim\pi}\exp\left(m\,\mathrm{kl}\left[\mathrm{R}^\ell_{\mathcal{V}}(g)\|\mathrm{R}^\ell_{\mathcal{D}}(g)\right]\right)$. Indeed, we have

$$\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}e^{m\,\mathrm{kl}\left[\mathrm{R}^\ell_{\mathcal{V}}(g)\|\mathrm{R}^\ell_{\mathcal{D}}(g)\right]}=\mathop{\mathbb{E}}_{g\sim\pi}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}e^{m\,\mathrm{kl}\left[\mathrm{R}^\ell_{\mathcal{V}}(g)\|\mathrm{R}^\ell_{\mathcal{D}}(g)\right]} \tag{12}$$
$$\text{and}\quad \mathop{\mathbb{E}}_{g\sim\pi}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}e^{m\,\mathrm{kl}\left[\mathrm{R}^\ell_{\mathcal{V}}(g)\|\mathrm{R}^\ell_{\mathcal{D}}(g)\right]}\le 2\sqrt{m}, \tag{13}$$

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]. □

Thanks to Corollary 6, we are now able to prove Corollary 4.

Corollary 4. *For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any bounded hypothesis set $\mathcal{H}$, given the uniform prior π on $\mathcal{H}$, for any loss $\ell:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,1]$, for any $\mu:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, 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*

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

with $\rho_{\mathcal{S}}$ defined in Equation (1).

Proof. We instantiate Corollary 6 with π being a uniform distribution. Moreover, we have $\ln \frac{\pi(h')}{\pi(h)} = 0$. □

Similarly than for Corollary 4, we prove Corollary 5 thanks to Corollary 6.

Corollary 5. *For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any loss $\ell : \mathcal{H} \times (\mathcal{X}\times\mathcal{Y}) \to [0,1]$, for any $\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m \to \mathbb{R}$, for any $\omega : \mathcal{H}\to\mathbb{R}$, 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*

$$\mathrm{kl}\Big[R^{\ell}_{\mathcal{S}}(h)\|R^{\ell}_{\mathcal{D}}(h)\Big] \le \frac{1}{m}\Big[[\mu(h',\mathcal{S})-\omega(h')] - [\mu(h,\mathcal{S})-\omega(h)] + \ln\tfrac{8\sqrt{m}}{\delta^2}\Big], \tag{7}$$

with $\rho_{\mathcal{S}}$ and π resp. defined in Equations (1) *and* (6).

Proof. We instantiate Corollary 6 with $\pi(h) \propto \exp(-\omega(h))$. Moreover, we have $\ln \frac{\pi(h')}{\pi(h)} = \omega(h) - \omega(h')$. □

B COMPARISON WITH THE BOUNDS OF THE LITERATURE

In this section, we first provide the bound of Lee et al. [2020] in Appendix B.1. Additionally, we discuss three bounds that are in the (classical or disintegrated) PAC-Bayesian literature and that consider Gibbs distributions. More precisely, we discuss in Appendix B.2 a disintegrated bound of Catoni [2007, Section 1.2.4] that was proven for a specific Gibbs distribution (*i.e.*, with a fixed parametric function μ). Moreover, in Appendices B.3 and B.4, we provide two disintegrated PAC-Bayesian bounds for $\mu(h,\mathcal{S}) = \alpha \mathrm{R}^{\ell'}_{\mathcal{S}}(h)$ inspired by two works from the classical PAC-Bayesian literature. Moreover, in Appendix B.5, we discuss the related works.

B.1 About Lee et al. [2020]'s Bound

For the sake of completeness, we provide a (refined) proof of the Lee et al. [2020]'s bound.

Theorem 7. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any distribution $\pi\in\mathcal{M}(\mathcal{H})$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $\mathcal{S}_1\sim\mathcal{D}^m,\dots,\mathcal{S}_n\sim\mathcal{D}^m$ and $h_1\sim\rho_{\mathcal{S}_1},\dots h_n\sim\rho_{\mathcal{S}_n}$ we have for all $\epsilon>0$*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\Big[|R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)| \le \mu(h,\mathcal{S}) + \epsilon\Big] \ge 1 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\Big[|R_{\mathcal{D}}(h_i) - R_{\mathcal{S}_i}(h_i)| - \mu(h_i,\mathcal{S}_i) > \epsilon\Big] - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \triangleq 1-\delta'(\epsilon),$$

where $\mathrm{I}[a] = 1$ *if a is true and* 0 *otherwise.*

Proof. First of all, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\Big[|\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)| \le \mu(h,\mathcal{S}) + \epsilon\Big] = \mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\Big[|\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)| - \mu(h,\mathcal{S}) \le \epsilon\Big] \triangleq F(\epsilon),$$

where $F(\cdot)$ is the cumulative distribution function of $|\mathrm{R}_{\mathcal{D}}(h) - \mathrm{R}_{\mathcal{S}}(h)| - \mu(h,\mathcal{S})$ where $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathcal{S}}$. Then, from the Dvoretzky–Kiefer–Wolfowitz inequality, we have with probability at least $1-\delta$ over $\mathcal{S}_1\sim\mathcal{D}^m,\dots,\mathcal{S}_n\sim\mathcal{D}^m$ and $h_1\sim\rho_{\mathcal{S}_1},\dots h_n\sim\rho_{\mathcal{S}_n}$

$$F(\epsilon) \ge \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\Big[|\mathrm{R}_{\mathcal{D}}(h_i) - \mathrm{R}_{\mathcal{S}_i}(h_i)| - \mu(h_i,\mathcal{S}_i) \le \epsilon\Big] - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

Moreover, remark that we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\Big[|\mathrm{R}_{\mathcal{D}}(h_i) - \mathrm{R}_{\mathcal{S}_i}(h_i)| - \mu(h_i,\mathcal{S}_i) \le \epsilon\Big] &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \mathrm{I}\Big[|\mathrm{R}_{\mathcal{D}}(h_i) - \mathrm{R}_{\mathcal{S}_i}(h_i)| - \mu(h_i,\mathcal{S}_i) > \epsilon\Big]\right)\\ &= 1 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\Big[|\mathrm{R}_{\mathcal{D}}(h_i) - \mathrm{R}_{\mathcal{S}_i}(h_i)| - \mu(h_i,\mathcal{S}_i) > \epsilon\Big].\end{aligned}$$

Finally, combining the equations gives the desired result. □

Note that we improve their result by replacing $\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathcal{S}}(h)-\mu(h,\mathcal{S})$ with $|\mathrm{R}_{\mathcal{D}}(h)-\mathrm{R}_{\mathcal{S}}(h)|-\mu(h,\mathcal{S})$ in the empirical cumulative distribution function, which improves the constant in the statistical term $\sqrt{\ln\frac{2}{\delta}/2n}$. This does not change the interpretation of their results. Indeed, while the term $\mu(h,\mathcal{S})+\epsilon$ is computable, the probability term $1-\delta'(\epsilon)$ is not since $\mathrm{R}_{\mathcal{D}}(h_i)$ is unknown (as it depends on the data distribution $\mathcal{D}$). This makes the overall bound uncomputable as the probability $\delta'(\epsilon)$ by which it stands is unknown.

B.2 About Catoni [2007]'s Bound

Catoni [2007, Theorem 1.2.7] proved the following disintegrated PAC-Bayesian bound; we give a proof for the sake of completeness.

Lemma 8. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any distribution $\pi\in\mathcal{M}(\mathcal{H})$, for any $\delta\in(0,1]$, we have with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathcal{S}}$*

$$R^{\ell}_{\mathcal{D}}(h)\leq\frac{1}{1-e^{-c}}\left\{1-\exp\left(-cR^{\ell}_{\mathcal{S}}(h)-\frac{1}{m}\left[\ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}+\ln\frac{1}{\delta}\right]\right)\right\}.$$

Proof. We apply Theorem 1 with $\varphi(h,\mathcal{S})=m\left[-\ln(1-\mathrm{R}^{\ell}_{\mathcal{D}}(h)\,[1-e^{-c}])-c\mathrm{R}^{\ell}_{\mathcal{S}}(h)\right]$. By rearranging the terms, we obtain

$$\mathrm{R}^{\ell}_{\mathcal{D}}(h)\leq\frac{1}{1-e^{-c}}\left\{1-\exp\left(-c\mathrm{R}^{\ell}_{\mathcal{S}}(h)-\frac{1}{m}\left[\ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}+\ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}e^{m\left[-\ln(1-\mathrm{R}^{\ell}_{\mathcal{D}}(h)[1-e^{-c}])-c\mathrm{R}^{\ell}_{\mathcal{S}}(h)\right]}\right)\right]\right)\right\}. \tag{14}$$

Moreover, from Fubini's theorem, Maurer [2004, Lemma 3], and Germain et al. [2009, Corollary 2.2], we have

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi}e^{m\left[-\ln(1-\mathrm{R}^{\ell}_{\mathcal{D}}(h)[1-e^{-c}])-c\mathrm{R}^{\ell}_{\mathcal{S}}(h)\right]}\leq 1. \tag{15}$$

Finally, by merging Equations (14) and (15), we have the stated result. □

Compared to the bounds that we provided in this paper, this one depends on a parameter $c>0$ that is fixed before seeing the learning sample $\mathcal{S}\sim\mathcal{D}^m$ and the hypothesis $h\sim\rho_{\mathcal{S}}$. Catoni applied Lemma 8 for a particular Gibbs distribution. In the following, we provide a more general corollary. To obtain Catoni's corollary, we have to fix $\mu(h,\mathcal{S})=cm\mathrm{R}^{\ell}_{\mathcal{S}}(h)-\ln\pi(h)$.

Corollary 9. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any loss $\ell:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,1]$, for any prior distribution $\pi\in\mathcal{M}(\mathcal{H})$ on $\mathcal{H}$, for any parametric function $\mu:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$, $h\sim\rho_{\mathcal{S}}$ we have*

$$R^{\ell}_{\mathcal{D}}(h)\leq\frac{1}{1-e^{-c}}\left\{1-\exp\left(-cR^{\ell}_{\mathcal{S}}(h)-\frac{1}{m}\left[-\mu(h,\mathcal{S})-\ln\pi(h)-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\frac{1}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right)+\ln\frac{1}{\delta}\right]\right)\right\},$$

where $\rho_{\mathcal{S}}$ is the Gibbs distribution (see Equation (1)).

Proof. Starting from Lemma 8, we develop the disintegrated KL divergence $\ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}$ as in Theorem 3. We have

$$\begin{aligned}\ln\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}&=-\mu(h,\mathcal{S})-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\frac{\pi(h)}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right)\\&=-\mu(h,\mathcal{S})-\ln\pi(h)-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\frac{1}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right),\end{aligned}$$

which leads to the desired result. □

In its current form, the generalization bound presented in Corollary 9 is not computable because of the expectation $\mathbb{E}_{g\sim\pi}\frac{1}{\pi(g)}\exp[-\mu(g,\mathcal{S})]$. In order to obtain a term that is computable, we can do the same trick as in Theorem 3. This gives the following bound.

Corollary 10. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any loss $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y}) \to [0,1]$, for any prior distribution $\pi \in \mathcal{M}(\mathcal{H})$ on $\mathcal{H}$, for any parametric function $\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m \to \mathbb{R}$, 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*

$$R^{\ell}_{\mathcal{D}}(h) \le \frac{1}{1-e^{-c}}\left\{1-\exp\left(-cR^{\ell}_{\mathcal{S}}(h) - \frac{1}{m}\left[\mu(h',\mathcal{S}) - \mu(h,\mathcal{S}) + \ln\frac{\pi(h')}{\pi(h)} + \ln\frac{4}{\delta^2}\right]\right)\right\},$$

where $\rho_{\mathcal{S}}$ is the Gibbs distribution (see Equation (1)).

Proof. We first consider Corollary 9 with $\delta/2$ instead of δ. This gives the following bound

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\mathrm{R}^{\ell}_{\mathcal{D}}(h) \le \frac{1}{1-e^{-c}}\left\{1-\exp\left(-c\mathrm{R}^{\ell}_{\mathcal{S}}(h) - \frac{1}{m}\left[-\mu(h,\mathcal{S}) - \ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\frac{\pi(h)}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right) + \ln\frac{2}{\delta}\right]\right)\right\}\right] \ge 1-\frac{\delta}{2}. \tag{16}$$

Then, we use Equation (11), which tells us that

$$\mathop{\mathbb{P}}_{h'\sim\pi}\left[-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi}\frac{\pi(h)}{\pi(g)}e^{-\mu(g,\mathcal{S})}\right) \le \ln\frac{2}{\delta} + \ln\frac{\pi(h')}{\pi(h)} + \mu(h',\mathcal{S})\right] \ge 1-\frac{\delta}{2}.$$

Finally, combining Equation (16) with Equation (11) gives us the desired result. □

As we can remark, the bound of Corollary 10 depends on the same terms as Corollary 6. Hence, in order to compare Corollary 10 and Corollary 6, we prove the following proposition.

Proposition 11. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any loss $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y}) \to [0,1]$, for any prior distribution $\pi \in \mathcal{M}(\mathcal{H})$ on $\mathcal{H}$, for any parametric function $\mu : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m \to \mathbb{R}$, 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*

$$\begin{aligned} R^{\ell}_{\mathcal{D}}(h) &\le \inf_{c>0}\left\{\frac{1}{1-e^{-c}}\left\{1-\exp\left(-cR^{\ell}_{\mathcal{S}}(h) - \frac{1}{m}\left[\mu(h',\mathcal{S}) - \mu(h,\mathcal{S}) + \ln\frac{\pi(h')}{\pi(h)} + \ln\frac{8\sqrt{m}}{\delta^2}\right]\right)\right\}\right\} \\ &= \underbrace{\overline{\mathrm{kl}}\left[R^{\ell}_{\mathcal{S}}(h) \;\middle|\; \frac{1}{m}\left(\mu(h',\mathcal{S}) - \mu(h,\mathcal{S}) + \ln\frac{\pi(h')}{\pi(h)} + \ln\frac{8\sqrt{m}}{\delta^2}\right)\right]}_{\textit{Bound of Corollary 6}}, \end{aligned}$$

where $\rho_{\mathcal{S}}$ is the Gibbs distribution (see Equation (1)).

Proof. We apply the same proof of Letarte et al. [2019]'s Theorem 3 where in our case their "$\mathcal{L}_{\mathcal{D}}(G_\theta)$", "$\mathcal{L}_{\mathcal{S}}(G_\theta)$" are respectively $\mathrm{R}^{\ell}_{\mathcal{D}}(h)$ and $\mathrm{R}^{\ell}_{\mathcal{S}}(h)$ and "ξ" is defined by $\xi \triangleq \frac{1}{m}\left(\mu(h',\mathcal{S}) - \mu(h,\mathcal{S}) + \ln\frac{\pi(h')}{\pi(h)} + \ln\frac{8\sqrt{m}}{\delta^2}\right)$. □

In other words, the bound of Corollary 6 is a Catoni-like bound where the parameter $c > 0$ is optimized. At first sight, the bound in Corollary 10 might appear slightly tighter than Corollary 6 (in light of Proposition 11). Indeed, Corollary 6's bound has an additional cost of $\frac{\ln(2\sqrt{m})}{m}$, which is negligible for a large number of examples m. However, the parameter $c > 0$ in Corollary 10 cannot be optimized since the bound holds for a fixed parameter. In order to optimize the bound, the bound must hold for a set of parameters c. This can be done through the union bound (that adds an additional cost to the bound). Hence, in order for Corollary 10 to be tighter, the additional cost cannot be larger than $\frac{\ln(2\sqrt{m})}{m}$, which is challenging for large m. Hence, for the experiments, we did not consider the bound of Corollary 10, which is only as tight as Corollary 6 or larger.

B.3 About Equation (8)

Lever et al. [2013, Lemma 5] proved a (classical) PAC-Bayesian bound on the expected risk with $\mu(h,\mathcal{S}) = \alpha\mathrm{R}^{\ell'}_{\mathcal{S}}(h)$, *i.e.*, they proved a bound on $\mathrm{kl}[\mathbb{E}_{h\sim\rho_{\mathcal{S}}}\mathrm{R}^{\ell}_{\mathcal{S}}(h) \,\|\, \mathbb{E}_{h\sim\rho_{\mathcal{S}}}\mathrm{R}^{\ell}_{\mathcal{D}}(h)]$. For the sake of comparison, we prove the following disintegrated bound that is similar to the one of Lever et al. [2013]. We consider this bound as a baseline for the experiments in Section 4.

Theorem 12. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any losses $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y}) \to [0,1]$ and $\ell' : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y}) \to [0,1]$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$, $h\sim\rho_{\mathcal{S}}$ we have*

$$\mathrm{kl}[R^{\ell}_{\mathcal{S}}(h)\|R^{\ell}_{\mathcal{D}}(h)] \leq \frac{1}{m}\left[\frac{\alpha^2}{8m}+\sqrt{\frac{\alpha^2}{2m}\ln\frac{6\sqrt{m}}{\delta}}+\ln\frac{6\sqrt{m}}{\delta}\right],$$

where the posterior $\rho_{\mathcal{S}}$ and the prior π are defined respectively by $\rho_{\mathcal{S}}(h) \propto e^{-\alpha R^{\ell'}_{\mathcal{S}}(h)}$ and $\pi(h) \propto e^{-\alpha R^{\ell'}_{\mathcal{D}}(h)}$.

Compared to the other bounds, Theorem 12 does not depend on a parametric function μ. Instead, it depends only on the concentration parameter $\alpha \in \mathbb{R}$ and the number of examples m. To obtain such a bound, the disintegrated KL divergence $\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}$ is upper-bounded. Hence, to prove Theorem 12, we first prove the following lemma (that is also inspired by Lever et al.'s Lemma 4).

Lemma 13 (Disintegrated version of Lever et al.'s Lemma 4)**.** *Given the posterior $\rho_{\mathcal{S}}$ and the prior π defined as $\rho_{\mathcal{S}}(h) \propto e^{-\mu(h,\mathcal{S})}$ and $\pi(h) \propto e^{-\omega(h)}$, we have the following upper-bound:*

$$\forall h \in \mathcal{H}, \quad \ln_+\frac{\rho_{\mathcal{S}}(h)}{\pi(h)} \leq [\omega(h)-\mu(h,\mathcal{S})]_+ + \left[\mathop{\mathbb{E}}_{h'\sim\pi}\mu(h',\mathcal{S})-\omega(h')\right]_+,$$

where $[\cdot]_+ \triangleq \max(\cdot,0)$ and $\ln_+(\cdot) \triangleq [\ln(\cdot)]_+$.

Proof. First of all, we denote as $Z_{\rho_{\mathcal{S}}} = \int_{\mathcal{H}} \exp\left[-\mu(g,\mathcal{S})\right] d\lambda(g)$ and $Z_\pi = \int_{\mathcal{H}} \exp\left[-\omega(g)\right] d\lambda(g)$, the normalization constant of the Gibbs distributions $\rho_{\mathcal{S}}$ and π respectively while λ is the reference measure on $\mathcal{H}$. Then, we have

$$\begin{aligned}
\ln_+\frac{\rho_{\mathcal{S}}(h)}{\pi(h)} &= \ln_+\frac{Z_\pi e^{-\mu(h,\mathcal{S})}}{Z_{\rho_{\mathcal{S}}}e^{-\omega(h)}} \\
&\leq [\omega(h)-\mu(h,\mathcal{S})]_+ + \ln_+\frac{Z_\pi}{Z_{\rho_{\mathcal{S}}}} \qquad (17)\\
&= [\omega(h)-\mu(h,\mathcal{S})]_+ + \max\left(\ln\frac{Z_\pi}{Z_{\rho_{\mathcal{S}}}}, 0\right)\\
&= [\omega(h)-\mu(h,\mathcal{S})]_+ + \max\left(-\ln\left(\frac{1}{Z_\pi}\int_{\mathcal{H}} e^{-\mu(g,\mathcal{S})}d\lambda(g)\right), 0\right)\\
&= [\omega(h)-\mu(h,\mathcal{S})]_+ + \max\left(-\ln\left(\frac{1}{Z_\pi}\int_{\mathcal{H}} e^{\omega(g)}e^{-\omega(g)}e^{-\mu(g,\mathcal{S})}d\lambda(g)\right), 0\right)\\
&= [\omega(h)-\mu(h,\mathcal{S})]_+ + \max\left(-\ln\left(\int_{\mathcal{H}} \pi(g)e^{\omega(g)-\mu(g,\mathcal{S})}d\lambda(g)\right), 0\right)\\
&= [\omega(h)-\mu(h,\mathcal{S})]_+ + \max\left(-\ln\left(\mathop{\mathbb{E}}_{h'\sim\pi} e^{\omega(h')-\mu(h',\mathcal{S})}\right), 0\right)\\
&\leq [\omega(h)-\mu(h,\mathcal{S})]_+ + \max\left(-\mathop{\mathbb{E}}_{h'\sim\pi}\left[\omega(h')-\mu(h',\mathcal{S})\right], 0\right) \qquad (18)\\
&= [\omega(h)-\mu(h,\mathcal{S})]_+ + \left[\mathop{\mathbb{E}}_{h'\sim\pi}\mu(h',\mathcal{S})-\omega(h')\right]_+,
\end{aligned}$$

where Equation (17) is obtained thanks to the inequality $[a+b]_+ \leq [a]_+ + [b]_+$ while Equation (18) holds thanks to Jensen's inequality and because $[\cdot]_+$ is monotonically increasing. □

Moreover, in order to prove Theorem 12, we need the following lemma, which is an application of Theorem 1 given by Rivasplata et al. [2020].

Lemma 14. *For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any loss $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y}) \to [0,1]$, for any prior distribution $\pi \in \mathcal{M}(\mathcal{H})$ on $\mathcal{H}$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$, $h\sim\rho_{\mathcal{S}}$ we have*

$$\mathrm{kl}[R^{\ell}_{\mathcal{S}}(h)\|R^{\ell}_{\mathcal{D}}(h)] \leq \frac{1}{m}\left[\ln_+\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right], \qquad (19)$$

$$\textit{and} \quad \left|R^{\ell}_{\mathcal{S}}(h)-R^{\ell}_{\mathcal{D}}(h)\right| \leq \sqrt{\frac{1}{2m}\left[\ln_+\frac{\rho_{\mathcal{S}}(h)}{\pi(h)}+\ln\frac{2\sqrt{m}}{\delta}\right]}, \qquad (20)$$

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

Proof. We apply Theorem 1 with $\varphi(h, \mathcal{S}) = m \operatorname{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h) \| \mathrm{R}^{\ell}_{\mathcal{D}}(h)]$ to obtain

$$\operatorname{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h) \| \mathrm{R}^{\ell}_{\mathcal{D}}(h)] \leq \frac{1}{m} \left[\ln \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \left[\frac{1}{\delta} \mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{g \sim \pi} e^{m \operatorname{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(g) \| \mathrm{R}^{\ell}_{\mathcal{D}}(g)]} \right] \right].$$

From Fubini's theorem and Maurer [2004], we have

$$\mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{g \sim \pi} e^{m \operatorname{kl}\left[\mathrm{R}^{\ell}_{\mathcal{V}}(g) \| \mathrm{R}^{\ell}_{\mathcal{D}}(g)\right]} = \mathop{\mathbb{E}}_{g \sim \pi} \mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} e^{m \operatorname{kl}\left[\mathrm{R}^{\ell}_{\mathcal{V}}(g) \| \mathrm{R}^{\ell}_{\mathcal{D}}(g)\right]} \leq 2\sqrt{m}. \tag{21}$$

By definition of $\ln_+(\cdot)$, we have $\ln \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} \leq \ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)}$, which is Equation (19). Finally, thanks to Pinsker's inequality, we have $2(\mathrm{R}^{\ell}_{\mathcal{S}}(h) - \mathrm{R}^{\ell}_{\mathcal{D}}(h))^2 \leq \operatorname{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h) \| \mathrm{R}^{\ell}_{\mathcal{D}}(h)]$ and we obtain Equation (20). □

Thanks to Lemmas 13 and 14, we are now able to prove Theorem 12.

Proof of Theorem 12. Starting from Lemma 14 (and Equation (19)) with probability at least $1 - \delta/3$ instead of $1 - \delta$, we have

$$\operatorname{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h) \| \mathrm{R}^{\ell}_{\mathcal{D}}(h)] \leq \frac{1}{m} \left[\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \frac{6\sqrt{m}}{\delta} \right]. \tag{22}$$

From Lemma 13, we have

$$\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} \leq \alpha \left[\mathrm{R}^{\ell'}_{\mathcal{D}}(h) - \mathrm{R}^{\ell'}_{\mathcal{S}}(h) \right]_+ + \alpha \left[\mathop{\mathbb{E}}_{h' \sim \pi} \mathrm{R}^{\ell'}_{\mathcal{S}}(h') - \mathrm{R}^{\ell'}_{\mathcal{D}}(h') \right]_+. \tag{23}$$

From [Maurer, 2004, Equation (4)] and Pinsker's inequality, we have with probability at least $1 - \delta/3$ over $\mathcal{S} \sim \mathcal{D}^m$

$$\alpha \left[\mathop{\mathbb{E}}_{h' \sim \pi} \mathrm{R}^{\ell'}_{\mathcal{S}}(h') - \mathrm{R}^{\ell'}_{\mathcal{D}}(h') \right]_+ \leq \alpha \left| \mathop{\mathbb{E}}_{h' \sim \pi} \mathrm{R}^{\ell'}_{\mathcal{S}}(h') - \mathrm{R}^{\ell'}_{\mathcal{D}}(h') \right| \leq \sqrt{\frac{\alpha^2}{2m} \ln \frac{6\sqrt{m}}{\delta}}. \tag{24}$$

Moreover, from Lemma 14 (and Equation (20)), we can obtain with probability at least $1 - \delta/3$ over $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathcal{S}}$

$$\alpha \left[\mathrm{R}^{\ell'}_{\mathcal{D}}(h) - \mathrm{R}^{\ell'}_{\mathcal{S}}(h) \right]_+ \leq \alpha \left| \mathrm{R}^{\ell'}_{\mathcal{S}}(h) - \mathrm{R}^{\ell'}_{\mathcal{D}}(h) \right| \leq \sqrt{\frac{\alpha^2}{2m} \left[\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \frac{6\sqrt{m}}{\delta} \right]}. \tag{25}$$

From combining Equations (23) and (24) with a union bound, we have with probability at least $1 - 2\delta/3$ over $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \rho_{\mathcal{S}}$

$$\begin{aligned}
&\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} \leq \sqrt{\frac{\alpha^2}{2m} \left[\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \frac{6\sqrt{m}}{\delta} \right]} + \sqrt{\frac{\alpha^2}{2m} \ln \frac{6\sqrt{m}}{\delta}} \\
\iff &\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \frac{6\sqrt{m}}{\delta} - \ln \frac{6\sqrt{m}}{\delta} \leq \sqrt{\frac{\alpha^2}{2m} \left[\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \frac{6\sqrt{m}}{\delta} \right]} + \sqrt{\frac{\alpha^2}{2m} \ln \frac{6\sqrt{m}}{\delta}} \\
\iff &\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \frac{6\sqrt{m}}{\delta} - \ln \frac{6\sqrt{m}}{\delta} - \sqrt{\frac{\alpha^2}{2m} \left[\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln \frac{6\sqrt{m}}{\delta} \right]} - \sqrt{\frac{\alpha^2}{2m} \ln \frac{6\sqrt{m}}{\delta}} \leq 0.
\end{aligned}$$

We obtain the upper-bound on $\ln_+ \frac{\rho_{\mathcal{S}}(h)}{\pi(h)}$ by solving the quadratic (in)equation

$$ax^2 + bx + c \leq 0 \quad \text{such that} \quad x \in \mathbb{R}^+ \quad \text{with} \quad a = 1, \quad b = -\sqrt{\frac{\alpha^2}{2m}}, \quad \text{and} \quad c = -\ln \frac{6\sqrt{m}}{\delta} - \sqrt{\frac{\alpha^2}{2m} \ln \frac{6\sqrt{m}}{\delta}}.$$

Hence solving the quadratic (in)equation gives

$$x \in \left[0, \sqrt{\frac{\alpha^2}{8m} + \ln\frac{6\sqrt{m}}{\delta} + \sqrt{\frac{\alpha^2}{2m}\ln\frac{6\sqrt{m}}{\delta}}} - \sqrt{\frac{\alpha^2}{8m}}\right].$$

Hence, we can deduce that

$$\begin{aligned}\sqrt{\ln_+\frac{\rho_{\mathcal{S}}(h)}{\pi(h)} + \ln\frac{6\sqrt{m}}{\delta}} &\leq \sqrt{\frac{\alpha^2}{8m} + \ln\frac{6\sqrt{m}}{\delta} + \sqrt{\frac{\alpha^2}{2m}\ln\frac{6\sqrt{m}}{\delta}}} - \sqrt{\frac{\alpha^2}{8m}} \\ &\leq \sqrt{\frac{\alpha^2}{8m} + \ln\frac{6\sqrt{m}}{\delta} + \sqrt{\frac{\alpha^2}{2m}\ln\frac{6\sqrt{m}}{\delta}}}\end{aligned}$$

and

$$\ln_+\frac{\rho_{\mathcal{S}}(h)}{\pi(h)} \leq \frac{\alpha^2}{8m} + \sqrt{\frac{\alpha^2}{2m}\ln\frac{6\sqrt{m}}{\delta}}. \tag{26}$$

Combining Equations (22) and (26) gives the desired result. □

Note that the proof technique differs from the one of Lever et al. [2013] because we have to use two disintegrated PAC-Bayesian bounds and one classical PAC-Bayesian bound instead of only one classical PAC-Bayesian bound. Indeed, since the disintegrated bounds are valid only for one posterior distribution, we have to use one bound to obtain Equation (22) and one, in Equation (25), that serves to upper-bound the disintegrated KL divergence. The classical PAC-Bayesian bound in Equation (24) serves to upper-bound the second term for the disintegrated KL divergence.

B.4 About Equation (9)

More recently, Dziugaite and Roy [2018, Theorem 4.2] proved a (classical) PAC-Bayesian bound on the expected risk with $\mu(h,\mathcal{S}) = \alpha \mathrm{R}^{\ell}_{\mathcal{S}}(h)$ and considers data-dependent priors obtained from a ϵ-differentially private mechanism. However, their proof relies on the *approximate max-information* [Dwork et al., 2015] that we cannot straightforwardly adapt to the disintegrated setting. Instead, our proof is based on the definition of ϵ-differential privacy (given by Mironov [2017, Section III]).

Definition 15. *A randomized mechanism π is ϵ-differentially private if and only if for any learning samples $\mathcal{T}'$ and $\mathcal{T}$ differing from one example we have*

$$D_\infty(\pi_{\mathcal{T}'} \| \pi_{\mathcal{T}}) \triangleq \ln\left(\operatorname*{esssup}_{h\sim\pi_{\mathcal{T}'}} \frac{\pi_{\mathcal{T}'}(h)}{\pi_{\mathcal{T}}(h)}\right) \leq \epsilon.$$

Put into words, a randomized mechanism (*i.e.*, the sampling from the prior π) is ϵ-differentially private if the ratio between the densities obtained from the two learning samples $\mathcal{T}'$ and $\mathcal{T}$ (differing from one point) is bounded by ϵ. Intuitively, the two densities must be close when the learning samples $\mathcal{T}$ and $\mathcal{T}'$ differ from only one example.

From the definition, we are able to prove the following bound.

Lemma 16. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any ϵ-differentially private randomized mechanism π, for any measurable function $\varphi : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m \to \mathbb{R}$, for any $\delta\in(0,1]$, we have with probability at least $1-\delta$ over $\mathcal{S}'\sim\mathcal{D}^m$, $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathcal{S}}$*

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

Proof. First of all, note that we can apply Theorem 1 with the data-dependent prior $\pi_{\mathcal{S}'}$ depending on the ghost sample $\mathcal{S}'$. Indeed, we have with probability at least $1-\delta$ over $\mathcal{S}'\sim\mathcal{D}^m$, $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\rho_{\mathcal{S}}$

$$\varphi(h,\mathcal{S}) \leq \ln\frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}'}(h)} + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi_{\mathcal{S}'}} e^{\varphi(g,\mathcal{V})}\right]. \tag{27}$$

Let's denote by $\mathcal{S}^{(i)}$ the learning sample $\mathcal{S}$ such that the examples from index 1 to i have been replaced by the examples coming from the learning sample $\mathcal{S}'$. By convention, we have thus $\mathcal{S}^{(0)} = \mathcal{S}$ and $\mathcal{S}^{(m)} = \mathcal{S}'$. We can hence upper-bound the disintegrated KL divergence by

$$\begin{aligned}
\ln \frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}'}(h)} &= \ln \frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}^{(m)}}(h)} \\
&\leq \ln \frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}^{(m-1)}}(h)} + \epsilon \\
&\cdots \\
&\leq \ln \frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}^{(0)}}(h)} + m\epsilon \\
&= \ln \frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(h)} + m\epsilon. \qquad (28)
\end{aligned}$$

By combining Equations (27) and (28), we obtain the stated result. □

Lemma 16 can be interpreted as a special case of Theorem 1 where π is a ϵ-differentially private randomized mechanism. Note that the bound is also in probability over $\mathcal{S}' \sim \mathcal{D}^m$, which is a ghost sample (that we do not have in practice). However, it is not problematic since Equation (9) does not depend explicitly on $\mathcal{S}' \sim \mathcal{D}^m$. In order to prove further Equation (9), we now specialize Lemma 16 to obtain a bound with a parametric function μ and a ϵ-differentially private randomized mechanism π.

Theorem 17. *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ be a loss function and $\phi:\mathbb{R}^2\to\mathbb{R}$ be a generalization gap. For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any ϵ-differentially private randomized mechanism π, for any parametric function $\mu:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m\to\mathbb{R}$, for any $\delta\in(0,1]$, we have with probability at least $1-\delta$ over $\mathcal{S}' \sim \mathcal{D}^m$, $\mathcal{S} \sim \mathcal{D}^m$, $h' \sim \pi_{\mathcal{S}}$ and $h \sim \rho_{\mathcal{S}}$*

$$\phi(R^{\ell}_{\mathcal{D}}(h), R^{\ell}_{\mathcal{S}}(h)) \leq \mu(h', \mathcal{S}) - \mu(h, \mathcal{S}) + \ln \frac{\pi_{\mathcal{S}}(h')}{\pi_{\mathcal{S}}(h)} + m\epsilon + \ln\left[\frac{4}{\delta^2} \mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{g\sim\pi_{\mathcal{S}'}} e^{\phi(R^{\ell}_{\mathcal{D}}(g), R^{\ell}_{\mathcal{V}}(g))}\right],$$

where $\rho_{\mathcal{S}}$ is the Gibbs distribution (see Equation (1)).

Proof. Starting from Lemma 16, we follow the same steps as for Theorem 3 to obtain the result. Indeed, we first develop the term $\ln \frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(h)}$ to have

$$\ln \frac{\rho_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(h)} = -\mu(h, \mathcal{S}) - \ln\left(\mathop{\mathbb{E}}_{g\sim\pi_{\mathcal{S}}} \frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(g)} e^{-\mu(g,\mathcal{S})}\right).$$

Hence, we obtain the following inequality

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\rho_{\mathcal{S}}}\left[\phi(\mathrm{R}^{\ell}_{\mathcal{D}}(h), \mathrm{R}^{\ell}_{\mathcal{S}}(h)) \leq \ln\left[\frac{2}{\delta} \mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{g\sim\pi_{\mathcal{S}'}} e^{\phi(\mathrm{R}^{\ell}_{\mathcal{D}}(g), \mathrm{R}^{\ell}_{\mathcal{V}}(g))}\right] - \mu(h,\mathcal{S}) - \ln\left(\mathbb{E}_{g\sim\pi_{\mathcal{S}}} \frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(g)} e^{-\mu(g,\mathcal{S})}\right)\right] \geq 1 - \frac{\delta}{2}. \qquad (29)$$

We can now upper-bound the term $-\ln\left(\mathbb{E}_{g\sim\pi_{\mathcal{S}}} \frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(g)} e^{-\mu(g,\mathcal{S})}\right)$. To do so, since $\frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(h')} e^{-\mu(h',\mathcal{S})} > 0$, we apply Markov's inequality to obtain

$$\begin{aligned}
&\forall h \in \mathcal{H}, \quad \forall \mathcal{S} \in (\mathcal{X}\times\mathcal{Y})^m, \quad \mathop{\mathbb{P}}_{h'\sim\pi_{\mathcal{S}}}\left[\frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(h')} e^{-\mu(h',\mathcal{S})} \leq \frac{2}{\delta} \mathop{\mathbb{E}}_{g\sim\pi_{\mathcal{S}}} \frac{\pi_{\mathcal{S}}(h)}{\pi(g)} e^{-\mu(g,\mathcal{S})}\right] \geq 1 - \frac{\delta}{2} \\
\iff &\mathop{\mathbb{P}}_{h'\sim\pi_{\mathcal{S}}}\left[-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi_{\mathcal{S}}} \frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(g)} e^{-\mu(g,\mathcal{S})}\right) \leq \ln\frac{2}{\delta} - \ln\left(\frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(h')} e^{-\mu(h',\mathcal{S})}\right)\right] \geq 1-\delta.
\end{aligned}$$

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

$$-\ln\left(\frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(h')} e^{-\mu(h',\mathcal{S})}\right) = \ln\frac{\pi_{\mathcal{S}}(h')}{\pi_{\mathcal{S}}(h)} + \mu(h',\mathcal{S}).$$

Hence, we obtain the following inequality

$$\mathop{\mathbb{P}}_{h'\sim\pi_{\mathcal{S}}}\left[-\ln\left(\mathop{\mathbb{E}}_{g\sim\pi_{\mathcal{S}}}\frac{\pi_{\mathcal{S}}(h)}{\pi_{\mathcal{S}}(g)}e^{-\,\mu(g,\mathcal{S})}\right)\leq\ln\frac{2}{\delta}+\ln\frac{\pi_{\mathcal{S}}(h')}{\pi_{\mathcal{S}}(h)}+\mu(h',\mathcal{S})\right]\geq 1-\frac{\delta}{2}. \tag{30}$$

By using a union bound on Equations (29) and (30) and rearranging the terms, we obtain the claimed result. □

We are now able to prove the bound stated in Equation (9).

Corollary 18. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, for any losses $\ell:\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ and $\ell':\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$, for any $\alpha,\alpha'\geq 0$, for any $\delta\in(0,1]$, we have with probability at least $1-\delta$ over $\mathcal{S}'\sim\mathcal{D}^m$, $\mathcal{S}\sim\mathcal{D}^m$, $h'\sim\pi_{\mathcal{S}}$ and $h\sim\rho_{\mathcal{S}}$*

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

where the posterior $\rho_{\mathcal{S}}$ and the prior π are defined respectively by $\rho_{\mathcal{S}}(h)\propto e^{-\alpha R^{\ell'}_{\mathcal{S}}(h)}$ and $\pi(h)\propto e^{-\alpha' R^{\ell'}_{\mathcal{S}}(h)}$.

Proof. We instantiate Theorem 17 with $\phi(\mathrm{R}^{\ell}_{\mathcal{D}}(h),\mathrm{R}^{\ell}_{\mathcal{S}}(h))=m\,\mathrm{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h)\|\mathrm{R}^{\ell}_{\mathcal{D}}(h)]$. Additionally, from Fubini's theorem and Maurer [2004] we have $\mathbb{E}_{\mathcal{V}\sim\mathcal{D}^m}\mathbb{E}_{g\sim\pi}\exp(m\,\mathrm{kl}[\mathrm{R}^{\ell}_{\mathcal{V}}(g)\|\mathrm{R}^{\ell}_{\mathcal{D}}(g)])\leq 2\sqrt{m}$. Hence, we can deduce that

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

Let the posterior distribution $\rho_{\mathcal{S}}$ and the prior distribution π defined respectively by $\rho_{\mathcal{S}}\propto\exp(-\alpha\mathrm{R}^{\ell'}_{\mathcal{S}}(h))$ (*i.e.*, $\mu(h,\mathcal{S})=\alpha\mathrm{R}^{\ell'}_{\mathcal{S}}(h)$) and $\pi_{\mathcal{S}}\propto\exp(-\alpha'\mathrm{R}^{\ell'}_{\mathcal{S}}(h))$. From these definitions, we obtain

$$\mathrm{kl}[\mathrm{R}^{\ell}_{\mathcal{S}}(h)\|\mathrm{R}^{\ell}_{\mathcal{D}}(h)]\leq\frac{1}{m}\left[\alpha\Big[\mathrm{R}^{\ell'}_{\mathcal{S}}(h')-\mathrm{R}^{\ell'}_{\mathcal{S}}(h)\Big]+\alpha'\Big[\mathrm{R}^{\ell'}_{\mathcal{S}}(h)-\mathrm{R}^{\ell'}_{\mathcal{S}}(h')\Big]+m\epsilon+\ln\frac{8\sqrt{m}}{\delta^2}\right]. \tag{31}$$

From McSherry and Talwar [2007, Theorem 6], we can deduce that the randomized mechanism $\pi_{\mathcal{S}}$ (*i.e.*, the prior) gives $\epsilon=2\alpha'\frac{1}{m}$-differential privacy. Hence, by simplifying the Equation (31), we have the desired result. □

Even though Corollary 18 does not use the approximate max-information as done by Dziugaite and Roy [2018], we are still able to provide a bound with a prior that gives a hypothesis h' from an ϵ-differentially private randomize mechanism. The main advantage of these bounds compared to the others is that the prior can depend on the learning sample $\mathcal{S}$. This is why this bound is a good candidate for a baseline in Section 4.

B.5 Related Works

The Gibbs distribution has been used in information-theoretic generalization bounds[5] that upper-bound the expected generalization gap $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\rho_{\mathcal{S}}}\,\mathrm{R}^{\ell}_{\mathcal{D}}(h)-\mathrm{R}^{\ell}_{\mathcal{S}}(h)$. For instance, Raginsky et al. [Theorem 4, 2016] provided bounds for $\mu(h,\mathcal{S})=\alpha\mathrm{R}^{\ell}_{\mathcal{S}}(h)$ with losses bounded between 0 and 1, while Kuzborskij et al. [Theorem 1, 2019] with sub-Gaussian losses. Aminian et al. [Theorem 1, 2021] proved a closed-form solution of the expected generalization gap with the Gibbs distribution with $\mu(h,\mathcal{S})=\alpha\mathrm{R}^{\ell}_{\mathcal{S}}(h)$ (where the loss is non-negative); they also considered regularized empirical risks. Xu and Raginsky [2017], Kuzborskij et al. [2019] upper-bound the expected true risk $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\rho_{\mathcal{S}}}\,\mathrm{R}^{\ell}_{\mathcal{D}}(h)$ by excess risk bounds, *i.e.*, bounds *w.r.t.* the minimal true risk over the hypothesis set. In the PAC-Bayesian literature, Alquier et al. [2016] develop PAC-Bayesian bounds on the true risk $\mathbb{E}_{h\sim\rho_{\mathcal{S}}}\,\mathrm{R}^{\ell}_{\mathcal{D}}(h)$ with $\mu(h,\mathcal{S})=\alpha\mathrm{R}^{\ell}_{\mathcal{S}}(h)$. However, all these bounds consider a (regularised) empirical risk scaled by α for the parametric function, while we are interested in user-defined parametric functions μ. Moreover, these bounds are in expectation over $h\sim\rho_{\mathcal{S}}$, while we are interested in the risk of a *single* hypothesis h sampled from $\rho_{\mathcal{S}}$. Hence, to the best of our knowledge, our contribution is the first to derive probabilistic bounds for a single hypothesis sampled from a Gibbs distribution with general parametric functions μ.

[5] See Xu and Raginsky [2017], Goyal et al. [2017], Bu et al. [2020] for some examples of information-theoretic bounds.

C OBTAINING UNIFORM-CONVERGENCE AND ALGORITHMIC-DEPENDENT BOUNDS

In this section, we theoretically compare generalization bounds with arbitrary complexity measures and the literature's bounds. To do so, we prove in Corollaries 21 and 23 that, from an appropriate parametric function μ, we can obtain two types of generalization bounds: the uniform-convergence-based and the algorithmic-dependent generalization bounds. Hence, Corollaries 21 and 23 do not present new uniform-convergence bounds but show how to obtain existing bounds by integrating a specific complexity measure. In other words, we show that Theorem 3 is general enough to obtain one bound belonging to one of these frameworks. As we see in Appendices C.2 and C.3, this is done by *(i)* assuming that we can effectively find an upper bound of the generalization gap and *(ii)* fixing the appropriate function μ. In order to present our results in Corollaries 21 and 23, we first recall the definitions of the literature's bounds.

C.1 Types of Generalization Bounds in the Literature

The uniform-convergence-based bounds were the first type to be introduced, notably in Vapnik and Chervonenkis [1971] using the VC-dimension. Other bounds were later developed based on the Gaussian/Rademacher complexity [Bartlett and Mendelson, 2002]. The definition of this type of bounds is the following.

Definition 19 (Uniform Convergence Bound)**.** *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ be a loss function and $\phi : \mathbb{R}^2\to\mathbb{R}$ be a generalization gap. A uniform convergence bound is defined such that if for any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any hypothesis set $\mathcal{H}$, there exists a function $\Phi_{\mathtt{u}} : (0,1]\to\mathbb{R}$, such that for any $\delta \in (0,1]$ we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall h\in\mathcal{H},\ \phi(R^\ell_{\mathcal{D}}(h), R^\ell_{\mathcal{S}}(h)) \le \Phi_{\mathtt{u}}\big(\delta\big)\right] = \mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} \phi(R^\ell_{\mathcal{D}}(h), R^\ell_{\mathcal{S}}(h)) \le \Phi_{\mathtt{u}}\big(\delta\big)\right] \ge 1-\delta. \tag{32}$$

This definition encompasses different complexity measures, such as $\Phi_{\mathtt{u}}(\delta)=\mathtt{rad}(\mathcal{H})+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$ for the Rademacher complexity $\mathtt{rad}(\mathcal{H})$, or $\Phi_{\mathtt{u}}(\delta)=\sqrt{\frac{1}{m}2\mathtt{vc}(\mathcal{H})\ln\frac{em}{\mathtt{vc}(\mathcal{H})}}+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$ for the VC-dimension $\mathtt{vc}(\mathcal{H})$ [see Theorem 3.3 and Corollary 3.19 of Mohri et al., 2012] where the generalization gap is defined by $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) = \mathrm{R}^\ell_{\mathcal{D}}(h)-\mathrm{R}^\ell_{\mathcal{S}}(h)$ and ℓ is the 01-loss. This definition also highlights the worst-case nature of the uniform-convergence bounds: given a confidence δ, the generalization gap $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h))$ is upper-bounded by a complexity measure $\Phi_{\mathtt{u}}(\delta)$ constant for all $(h,\mathcal{S})\in\mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m$. The upper bound $\Phi_{\mathtt{u}}(\delta)$ can generally be improved by considering algorithmic-dependent bounds [Bousquet and Elisseeff, 2002, Xu and Mannor, 2012]. This kind of bounds upper-bound the generalization gap for the hypothesis $h_{\mathcal{S}}$ learned by an algorithm from a learning sample $\mathcal{S}$. The definition of such bounds is recalled below.

Definition 20 (Algorithmic-dependent Generalization Bound)**.** *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ be a loss function and $\phi:\mathbb{R}^2\to\mathbb{R}$ be a generalization gap. An algorithmic-dependent generalization bound is defined such that if for any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, there exists a function $\Phi_{\mathtt{a}} : (0,1]\to\mathbb{R}$, such that for any $\delta \in (0,1]$ we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\phi(R^\ell_{\mathcal{D}}(h_{\mathcal{S}}), R^\ell_{\mathcal{S}}(h_{\mathcal{S}})) \le \Phi_{\mathtt{a}}(\delta)\right] \ge 1-\delta, \tag{33}$$

where $h_{\mathcal{S}} \in \mathcal{H}$ is the hypothesis learned from an algorithm with $\mathcal{S}\sim\mathcal{D}^m$.

For example, when $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h_{\mathcal{S}}),\mathrm{R}^\ell_{\mathcal{S}}(h_{\mathcal{S}})) = \mathrm{R}^\ell_{\mathcal{D}}(h_{\mathcal{S}})-\mathrm{R}^\ell_{\mathcal{S}}(h_{\mathcal{S}})$, the upper bound $\Phi_{\mathtt{a}}(\delta) = 2\beta+(4m\beta+1)\sqrt{\frac{\ln 1/\delta}{2m}}$ where β is the uniform stability parameter [see, Bousquet and Elisseeff, 2002] and a bounded loss $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,1]$. Similarly to the uniform-convergence-based bounds, the upper bound $\Phi_{\mathtt{a}}(\delta)$ is a constant *w.r.t.* the hypothesis $h_{\mathcal{S}}$ and the learning sample $\mathcal{S}$.

C.2 Obtaining Uniform-convergence Bounds

Since the parametric function μ in Theorem 3 depends on the learning sample $\mathcal{S}$ and the hypothesis h, we can obtain from a specific μ a uniform-convergence-based bound (Equation (32)). Indeed, from Theorem 3, we obtain the following uniform-convergence-based bound.

Corollary 21. *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}$ be a loss function, $\phi:\mathbb{R}^2\to\mathbb{R}$ be the generalization gap and assume that there exists a function $\Phi_{\mathtt{u}} : (0,1] \to \mathbb{R}$ fulfilling Definition 19. Applying Theorem 3 with the parametric function μ defined by*

$$\forall(h,\mathcal{S}) \in \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m, \quad \mu(h,\mathcal{S}) = -\phi(R^\ell_{\mathcal{D}}(h), R^\ell_{\mathcal{S}}(h)) - \Phi_{\mathtt{u}}(\tfrac{\delta}{2}) - \ln\pi(h)$$

gives the following bound

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h'\sim\pi}\left[\sup_{f\in\mathcal{H}}\phi(R^\ell_{\mathcal{D}}(f),R^\ell_{\mathcal{S}}(f)) \le \underbrace{\Phi_{\mathtt{u}}(\tfrac{\delta}{2}) + \ln\left[\frac{16}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(R^\ell_{\mathcal{D}}(g),R^\ell_{\mathcal{V}}(g))-\phi(R^\ell_{\mathcal{D}}(h'),R^\ell_{\mathcal{S}}(h'))}\right]}_{\triangleq\ \Phi'_{\mathtt{u}}(\delta)}\right] \ge 1-\delta. \tag{34}$$

Proof. Given the definition of $\rho_{\mathcal{S}}$ (with the parametric function μ defined above), we deduce from Theorem 3 that

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi,h\sim\rho_{\mathcal{S}}}\Bigg[\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) \le & \underbrace{-\phi(\mathrm{R}^\ell_{\mathcal{D}}(h'),\mathrm{R}^\ell_{\mathcal{S}}(h')) - \Phi_{\mathtt{u}}(\tfrac{\delta}{2}) - \ln\pi(h')}_{\mu(h',\mathcal{S})}\\ & + \underbrace{\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) + \Phi_{\mathtt{u}}(\tfrac{\delta}{2}) + \ln\pi(h)}_{-\mu(h,\mathcal{S})}\\ & + \ln\frac{\pi(h')}{\pi(h)} + \ln\left[\frac{16}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(\mathrm{R}^\ell_{\mathcal{D}}(g),\mathrm{R}^\ell_{\mathcal{V}}(g))}\right]\Bigg] \ge 1-\frac{\delta}{2}.\end{aligned}$$

Moreover, thanks to Definition 19, with probability at least $1-\frac{\delta}{2}$ over the random choice of $\mathcal{S}$, we have $-\Phi_{\mathtt{u}}(\frac{\delta}{2}) \le -\sup_{f\in\mathcal{H}}\phi(\mathrm{R}^\ell_{\mathcal{D}}(f),\mathrm{R}^\ell_{\mathcal{S}}(f))$. Hence, with the union bound, we have that

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi,h\sim\rho_{\mathcal{S}}}\Bigg[\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) \le & -\phi(\mathrm{R}^\ell_{\mathcal{D}}(h'),\mathrm{R}^\ell_{\mathcal{S}}(h')) - \sup_{f\in\mathcal{H}}\phi(\mathrm{R}^\ell_{\mathcal{D}}(f),\mathrm{R}^\ell_{\mathcal{S}}(f)) - \ln\pi(h')\\ & + \phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) + \Phi_{\mathtt{u}}(\tfrac{\delta}{2}) + \ln\pi(h)\\ & + \ln\frac{\pi(h')}{\pi(h)} + \ln\left[\frac{16}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(\mathrm{R}^\ell_{\mathcal{D}}(g),\mathrm{R}^\ell_{\mathcal{V}}(g))}\right]\Bigg] \ge 1-\delta.\end{aligned}$$

Therefore, by rearranging the terms, we obtain the desired result. □

Corollary 21 underlines the fact that our framework is general enough to allow us to obtain a uniform-convergence-based bound. Indeed, if we are able to find (with high probability) an upper bound of the worst-case generalization gap $\sup_{f\in\mathcal{H}}\phi(\mathrm{R}^\ell_{\mathcal{D}}(f),\mathrm{R}^\ell_{\mathcal{S}}(f))$ denoted by $\Phi_{\mathtt{u}}(\delta)$, then our framework allows us to obtain a bound depending on $\Phi_{\mathtt{u}}(\delta)$. For instance, when we consider the bound $\Phi_{\mathtt{u}}(\delta) = \mathtt{rad}(\mathcal{H}) + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$ depending on the Rademacher complexity $\mathtt{rad}(\mathcal{H})$, we are able to obtain a bound depending on $\Phi_{\mathtt{u}}(\delta)$ thanks to our framework; it is shown in the following corollary.

Corollary 22. *Let $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})\to[0,1]$ be a loss function. By applying Theorem 3 with the parametric function μ defined by*

$$\forall(h,\mathcal{S}) \in \mathcal{H}\times(\mathcal{X}\times\mathcal{Y})^m, \quad \mu(h,\mathcal{S}) = -\sqrt{m}[R^\ell_{\mathcal{D}}(h) - R^\ell_{\mathcal{S}}(h)] - \sqrt{m}\left[\mathtt{rad}(\mathcal{H}) + \sqrt{\tfrac{1}{2m}\ln\tfrac{2}{\delta}}\right] - \ln\pi(h),$$

we can deduce the following bound

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\sup_{f\in\mathcal{H}} R^\ell_{\mathcal{D}}(f) - R^\ell_{\mathcal{S}}(f) \le \mathtt{rad}(\mathcal{H}) + \sqrt{\frac{1}{2m}\ln\frac{4}{\delta}} + \frac{\ln\frac{128}{\delta^3}+2}{\sqrt{m}}\right] \ge 1-\delta.$$

Proof. We apply Corollary 21 with the generalization gap $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h), \mathrm{R}^\ell_{\mathcal{S}}(h)) = \sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(h) - \mathrm{R}^\ell_{\mathcal{S}}(h)]$ and with $\Phi_{\mathtt{u}}(\delta) = \mathtt{rad}(\mathcal{H}) + \sqrt{\frac{1}{2m} \ln \frac{1}{\delta}}$ (see Theorem 3.3 of Mohri et al. [2012]). To obtain with probability at least $1 - \delta/2$ over $\mathcal{S} \sim \mathcal{D}^m$ and $h' \sim \pi$ we have

$$\sup_{f \in \mathcal{H}} \mathrm{R}^\ell_{\mathcal{D}}(f) - \mathrm{R}^\ell_{\mathcal{S}}(f) \leq \Phi_{\mathtt{u}}(\tfrac{\delta}{4}) + \frac{1}{\sqrt{m}} \ln \left[\frac{64}{\delta^2} \mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{g \sim \pi} e^{\sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(g) - \mathrm{R}^\ell_{\mathcal{V}}(g)] - \sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(h') - \mathrm{R}^\ell_{\mathcal{S}}(h')]} \right].$$

Moreover, thanks to Markov's inequality and the union bound, we obtain with probability at least $1 - \delta$ over $\mathcal{S} \sim \mathcal{D}^m$ we have

$$\sup_{f \in \mathcal{H}} \mathrm{R}^\ell_{\mathcal{D}}(f) - \mathrm{R}^\ell_{\mathcal{S}}(f) \leq \Phi_{\mathtt{u}}(\tfrac{\delta}{4}) + \frac{1}{\sqrt{m}} \ln \left[\frac{128}{\delta^3} \mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{h' \sim \pi} \mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{g \sim \pi} e^{\sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(g) - \mathrm{R}^\ell_{\mathcal{V}}(g)] - \sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(h') - \mathrm{R}^\ell_{\mathcal{S}}(h')]} \right].$$

Then, we upper-bound the term $\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \pi} \mathbb{E}_{\mathcal{V} \sim \mathcal{D}^m} \mathbb{E}_{g \sim \pi} e^{[\mathrm{R}^\ell_{\mathcal{D}}(g) - \mathrm{R}^\ell_{\mathcal{V}}(g)] - [\mathrm{R}^\ell_{\mathcal{D}}(h') - \mathrm{R}^\ell_{\mathcal{S}}(h')]}$. To do so, we first use Fubini's theorem to have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{h' \sim \pi} \mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{g \sim \pi} e^{\sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(g) - \mathrm{R}^\ell_{\mathcal{V}}(g)] - \sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(h') - \mathrm{R}^\ell_{\mathcal{S}}(h')]} \\
&= \mathop{\mathbb{E}}_{h' \sim \pi} \mathop{\mathbb{E}}_{g \sim \pi} \mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} e^{\sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(g) - \mathrm{R}^\ell_{\mathcal{V}}(g)] - \sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(h') - \mathrm{R}^\ell_{\mathcal{S}}(h')]}.
\end{aligned}$$

Moreover, we upper-bound the term $\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{\mathcal{V} \sim \mathcal{D}^m} e^{\sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(g) - \mathrm{R}^\ell_{\mathcal{V}}(g)] - \sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(h') - \mathrm{R}^\ell_{\mathcal{S}}(h')]}$ thanks to Hoeffding's lemma, to obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{\mathcal{S} \sim \mathcal{D}^m} \exp\left(\sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(g) - \mathrm{R}^\ell_{\mathcal{V}}(g)] - \sqrt{m}[\mathrm{R}^\ell_{\mathcal{D}}(h') - \mathrm{R}^\ell_{\mathcal{S}}(h')] \right) \\
&= \prod_{i=1}^m \left[\mathop{\mathbb{E}}_{(\mathbf{x}'_i, y'_i) \sim \mathcal{D}} \mathop{\mathbb{E}}_{(\mathbf{x}_i, y_i) \sim \mathcal{D}} \exp\left(\frac{1}{\sqrt{m}} \left[\mathop{\mathbb{E}}_{(\mathbf{x}, z) \sim \mathcal{D}} \ell(g, (\mathbf{x}, y)) - \ell(g, (\mathbf{x}'_i, y'_i)) - \mathop{\mathbb{E}}_{(\mathbf{x}, z) \sim \mathcal{D}} \ell(h', (\mathbf{x}, y)) + \ell(h', (\mathbf{x}_i, y_i)) \right] \right) \right] \\
&\leq \prod_{i=1}^m \left[\exp\left(\frac{2}{m} \right) \right] \\
&= \exp(2).
\end{aligned}$$

Hence, by rearranging the terms, we obtain the stated result. □

The bound that we can obtain in Corollary 22 is greater than the bound of Mohri et al. [2012]'s Theorem 3.3. This is normal since we use the bound in the parametric function μ. However, the higher the number of examples m, the closer our bound will be to the original bound of Mohri et al. [2012]. Obtaining new uniform-convergence bound (without relying on previously known bounds) by setting a specific parametric function μ is highly non-trivial and is thus an exciting line of research that can be explored in the future.

C.3 Obtaining Algorithmic-dependent Bounds

Similarly, we can obtain an algorithmic-dependent generalization bound (Definition 20) by using the same technique as in Corollary 21. Indeed, we can obtain the following result.

Corollary 23. *Let $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ be a loss function, $\phi : \mathbb{R}^2 \to \mathbb{R}$ be the generalization gap and assume that there exists a function $\Phi_{\mathtt{a}} : (0, 1] \to \mathbb{R}$ fulfilling Definition 20. Applying Theorem 3 with the parametric function μ defined by*

$$\forall (h, \mathcal{S}) \in \mathcal{H} \times (\mathcal{X} \times \mathcal{Y})^m, \quad \mu(h, \mathcal{S}) = -\phi(R^\ell_{\mathcal{D}}(h), R^\ell_{\mathcal{S}}(h)) - \Phi_{\mathtt{a}}(\tfrac{\delta}{2}) - \ln \pi(h)$$

gives the following bound

$$\mathop{\mathbb{P}}_{\mathcal{S} \sim \mathcal{D}^m, h' \sim \pi} \left[\phi(R^\ell_{\mathcal{D}}(h_{\mathcal{S}}), R^\ell_{\mathcal{S}}(h_{\mathcal{S}})) \leq \underbrace{\Phi_{\mathtt{a}}(\tfrac{\delta}{2}) + \ln \left[\frac{16}{\delta^2} \mathop{\mathbb{E}}_{\mathcal{V} \sim \mathcal{D}^m} \mathop{\mathbb{E}}_{g \sim \pi} e^{\phi(R^\ell_{\mathcal{D}}(g), R^\ell_{\mathcal{V}}(g)) - \phi(R^\ell_{\mathcal{D}}(h'), R^\ell_{\mathcal{S}}(h'))} \right]}_{\triangleq\ \Phi'_{\mathtt{a}}(\delta)} \right] \geq 1 - \delta. \tag{35}$$

Proof. The proof is similar to the one of Corollary 21. Given the definition of $\rho_{\mathcal{S}}$ (with the parametric function μ defined above), we deduce from Theorem 3 that

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi,h\sim\rho_{\mathcal{S}}}\left[\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) \leq \underbrace{-\phi(\mathrm{R}^\ell_{\mathcal{D}}(h'),\mathrm{R}^\ell_{\mathcal{S}}(h')) - \Phi_{\mathtt{a}}(\tfrac{\delta}{2}) - \ln\pi(h')}_{\mu(h',\mathcal{S})} + \underbrace{\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) + \Phi_{\mathtt{a}}(\tfrac{\delta}{2}) + \ln\pi(h)}_{-\mu(h,\mathcal{S})} + \ln\frac{\pi(h')}{\pi(h)} + \ln\left[\frac{16}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(\mathrm{R}^\ell_{\mathcal{D}}(g),\mathrm{R}^\ell_{\mathcal{V}}(g))}\right]\right] \geq 1-\frac{\delta}{2}.$$

Moreover, thanks to Definition 20, with probability at least $1-\frac{\delta}{2}$ over the random choice of $\mathcal{S}$, we have $-\Phi_{\mathtt{a}}(\frac{\delta}{2}) \leq -\phi(\mathrm{R}^\ell_{\mathcal{D}}(h_{\mathcal{S}}),\mathrm{R}^\ell_{\mathcal{S}}(h_{\mathcal{S}}))$. Hence, with the union bound, we have that

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\ h'\sim\pi,h\sim\rho_{\mathcal{S}}}\left[\phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) \leq -\phi(\mathrm{R}^\ell_{\mathcal{D}}(h'),\mathrm{R}^\ell_{\mathcal{S}}(h')) - \phi(\mathrm{R}^\ell_{\mathcal{D}}(h_{\mathcal{S}}),\mathrm{R}^\ell_{\mathcal{S}}(h_{\mathcal{S}})) - \ln\pi(h') + \phi(\mathrm{R}^\ell_{\mathcal{D}}(h),\mathrm{R}^\ell_{\mathcal{S}}(h)) + \Phi_{\mathtt{a}}(\tfrac{\delta}{2}) + \ln\pi(h) + \ln\frac{\pi(h')}{\pi(h)} + \ln\left[\frac{16}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{V}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{g\sim\pi} e^{\phi(\mathrm{R}^\ell_{\mathcal{D}}(g),\mathrm{R}^\ell_{\mathcal{V}}(g))}\right]\right] \geq 1-\delta.$$

Therefore, by rearranging the terms, we obtain the desired result. □

Hence, our framework is also general enough to retrieve algorithmic-dependent bounds. More precisely, the generalization gap $\phi(\mathrm{R}^\ell_{\mathcal{D}}(h_{\mathcal{S}}),\mathrm{R}^\ell_{\mathcal{S}}(h_{\mathcal{S}}))$ associated with the hypothesis $h_{\mathcal{S}}$ is upper-bounded by a constant $\Phi'_{\mathtt{a}}(\delta)$. As for Corollaries 21 and 22, the drawback of Corollary 23 is that we have to rely on a previously known bound to obtain our result. Therefore, further investigations must be done to derive entirely new algorithmic-depend bounds by setting a specific parametric function μ.

D ADDITIONAL INFORMATION ON THE EXPERIMENTS

In this section, we first provide more experiments about Section 4.3 by varying α. Appendix D.3 presents how NEURAL is obtained and more experiments on this parametric function (along with NEURAL$^{\boldsymbol{D}}$).

D.1 About Computing the Bounds (with $\overline{\mathrm{kl}}$)

The evaluated bounds that we consider have all the same structure: with high probability, we have $\mathrm{kl}(q\|p) \leq \tau$, where q is the empirical risk, p is the true risk, and τ is the bound. As shown, *e.g.*, in Equation (5), we can evaluate the bound on the true risk p by computing

$$\overline{\mathrm{kl}}[q|\tau] = \max\left\{p\in(0,1)\ \middle|\ \mathrm{kl}(q\|p)\leq\tau\right\}.$$

We use the bisection method to solve this optimization problem, as proposed by Reeb et al. [2018]. This method consists of refining the interval $[p_{\min},p_{\max}]$ in which p belongs. To do so, we first initialize $p_{\min}=q$ and $p_{\max}=1$. Then, for each iteration, we first set $p_{\mathrm{tmp}}=\frac{1}{2}(p_{\max}-p_{\min})$, and then we change the values of $p_{\min}$ and $p_{\max}$ depending on the value of the temporary parameter p_{tmp}. Indeed, if $\mathrm{kl}(q\|p_{\mathrm{tmp}})>\tau$ (*resp.*, $\mathrm{kl}(q\|p_{\mathrm{tmp}})<\tau$), we update $p_{\max}=p_{\mathrm{tmp}}$ (*resp.*, $p_{\min}=p_{\mathrm{tmp}}$). Moreover, if we have $\mathrm{kl}(q\|p_{\mathrm{tmp}})=\tau$, a small interval (with $p_{\max}-p_{\min}<\epsilon$), or if we attain the maximum number of iterations, we return p_{tmp} as $\overline{\mathrm{kl}}[q|\tau]$. In the experiments, we set $\epsilon=10^{-9}$ and the maximum number of iterations to 1000.

D.2 About Section 4.3

In the experiments introduced in Section 4.3, we fix $\alpha=m$. We propose additional experiments in Figures 5 and 6 where α varies between $\sqrt{m}$ and m. As we can remark in Figures 5 and 6, the concentration parameter α

plays an important role: the higher α, the lower the test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$. Moreover, the bounds are tighter than those in Figure 3; however, the test risks remain high, so the bounds are not sufficiently concentrated. In this setting, there is a trade-off between concentrating the distribution (by increasing α) and having a tight bound.

D.3 About Section 4.4

In this section, we first introduce in Appendix D.3.1 the setting to learn the parametric functions NEURAL and NEURAL$^{\boldsymbol{D}}$ with neural networks, and we show additional experiments in Appendix D.3.2.

D.3.1 Training the Neural Parametric Functions

NEURAL's dataset. In order to learn the neural networks associated with NEURAL, we have to train first neural networks (to have the weights as input) and save their corresponding generalization gap (that is further used as a label). To do so, for MNIST and FashionMNIST, we train models with a size of the validation set that varies in order to obtain models with diverse generalization gaps. Starting from the original training set of MNIST or FashionMNIST, we split the dataset into a training set and a validation set (to compute the gaps); we denote by m_{val} the size of the validation set and m_{train} the size of the training set. After fixing the split, we launch training and save the model with its corresponding gap after each epoch. We launch 1000 trainings with the split ratio $\frac{m_{\text{val}}}{m_{\text{val}}+m_{\text{train}}} \in \{0.99, 0.97, 0.95, 0.93\}$, 120 trainings with $\frac{m_{\text{val}}}{m_{\text{val}}+m_{\text{train}}} = 0.90$ and 110 trainings with $\frac{m_{\text{val}}}{m_{\text{val}}+m_{\text{train}}} \in \{0.80, 0.70, 0.60, 0.50, 0.40, 0.30, 0.20, 0.10\}$. In each training, we learn the model with the same architecture as in Section 4.1, and we optimize in the same way as for the sampling from $\rho_{\mathcal{S}}$, except that we run SGD instead of SGLD (*i.e.*, we remove the Gaussian noise) for at least 8000 iterations (we finish the epoch after reaching the number of iterations). We show in Figure 7 the histogram of the obtained generalization gaps. Note that our method of creating the dataset is similar to Lee et al. [2020], except that the size of our validation set varies more, and we save the parameters directly (instead of the predictions).

NEURAL's model. As summarized in Section 4.4, the parametric function NEURAL is a neural network that is learned with the dataset created previously. This model is a feed-forward neural network with 3 hidden layers of width 1024. The input is the weights and biases $\mathbf{w}$ of the network h, whereas the output is a scalar representing $\text{NEURAL}(h, \mathcal{S})$, *i.e.*, the value of the parametric function learned from the neural network. After the input, we normalize the parameters $\mathbf{w}$ with its ℓ_2 norm, and we use a batch normalization layer [Ioffe and Szegedy, 2015] (with a momentum of 0.1 and $\epsilon = 0.0$). Moreover, the activation functions are leaky ReLU, and the output is squared to obtain a positive output. The weights are initialized with the Xavier Glorot uniform initializer [Glorot and Bengio, 2010]. The biases are initialized with a uniform distribution between $-1/\sqrt{1024}$ and $+1/\sqrt{1024}$ for all biases, except for the first layer, they are initialized uniformly in $[-1/\sqrt{d}, +1/\sqrt{d}]$, where d is the number of parameters of the models in the dataset. The model is learned from Adam optimizer [Kingma and Ba, 2015] for 100 epochs by minimizing the mean absolute error. The model is selected by early stopping: the NEURAL's dataset is split between a training set (of size m_{train}) and a validation set (of size m_{val}). Moreover, in order to handle the fact that the NEURAL's dataset is unbalanced, we rebalance it by *(i)* putting the gaps into 50 bins, *(ii)* merging neighboring bins if they represent less than 1% of the dataset and *(iii)* sampling a bin with a probability proportional to the inverse of the number of examples in the bin (the examples in the bin are sampled uniformly). The merging procedure is done as follows: we merge the bin with its neighbor (that contains higher gaps) when it has less than 1% of the dataset; we perform this until stabilization. We learn different networks (that give several parametric functions NEURAL), with a batch size of 64, 128, or 256; Adam's learning rate is either 0.001 or 0.0001 (the other parameters remain the default ones); and the ratio $\frac{m_{\text{val}}}{m_{\text{train}}+m_{\text{val}}}$ is either 0.1, 0.3 or 0.5. Note that the work of Lee et al. [2020] differs from ours by the fact that *(i)* we have a simpler architecture, and *(ii)* we take the parameters as input in order to have a differentiable parametric function (for sampling from $\rho_{\mathcal{S}}$).

D.3.2 Additional Experiments

In Figures 8 to 11, we show the bar plots of the parametric functions NEURAL and NEURAL$^{\boldsymbol{D}}$ learned with the different hyperparameters. The figures highlight that hyperparameter tuning is extremely important. Indeed, the performance of the parametric functions can change drastically between two sets of hyperparameters. We believe that understanding the role of these hyperparameters is an exciting future work that might improve the performance of such parametric functions.

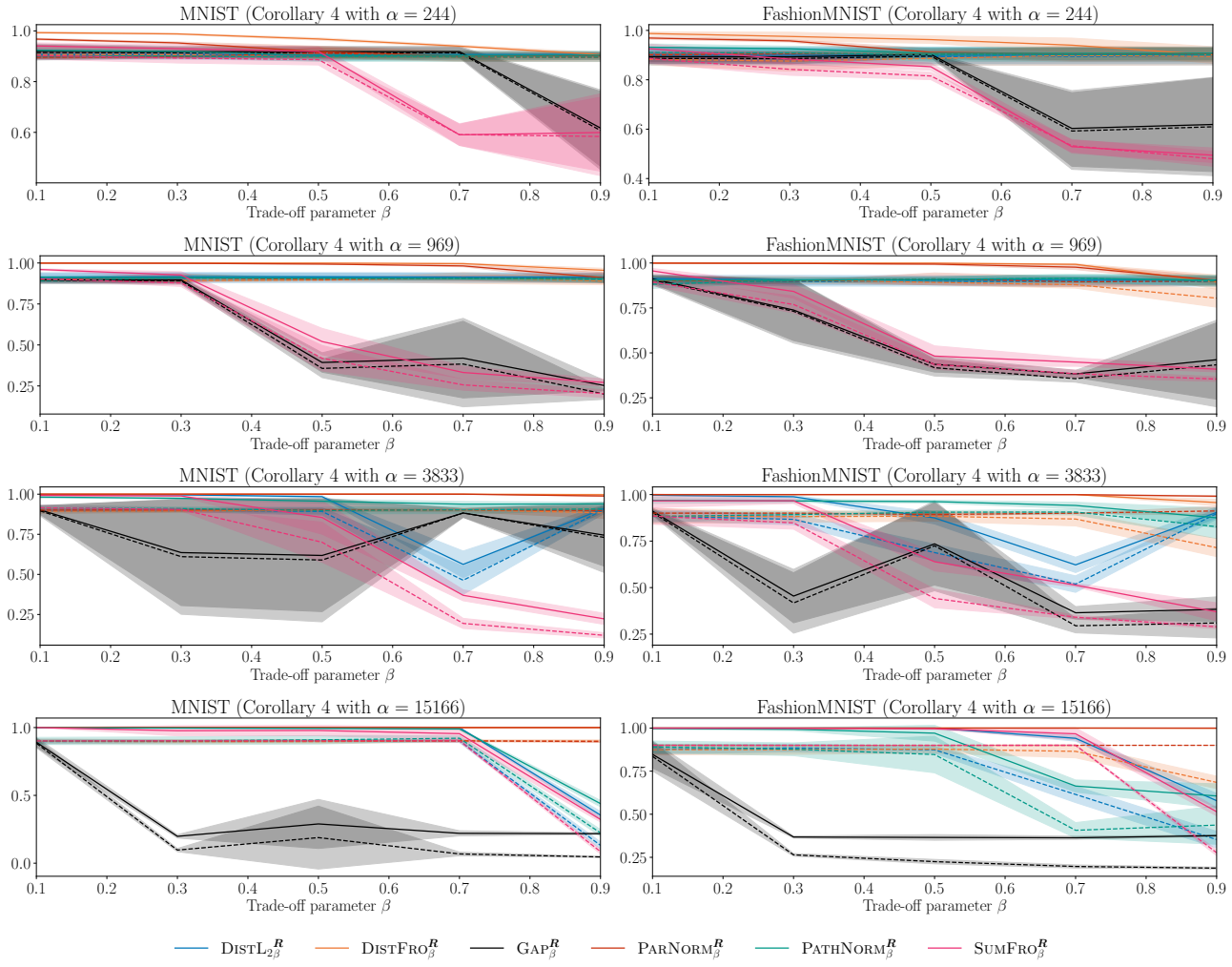

Figure 5: Evolution of the bounds (the plain lines) and the test risks $\mathrm{R}_{\mathcal{T}}^{\ell}(h)$ (the dashed lines) *w.r.t.* the trade-off parameter β for varying α and $\frac{m'}{m'+m} = 0.0$. The lines correspond to the mean, while the bands are the standard deviations.

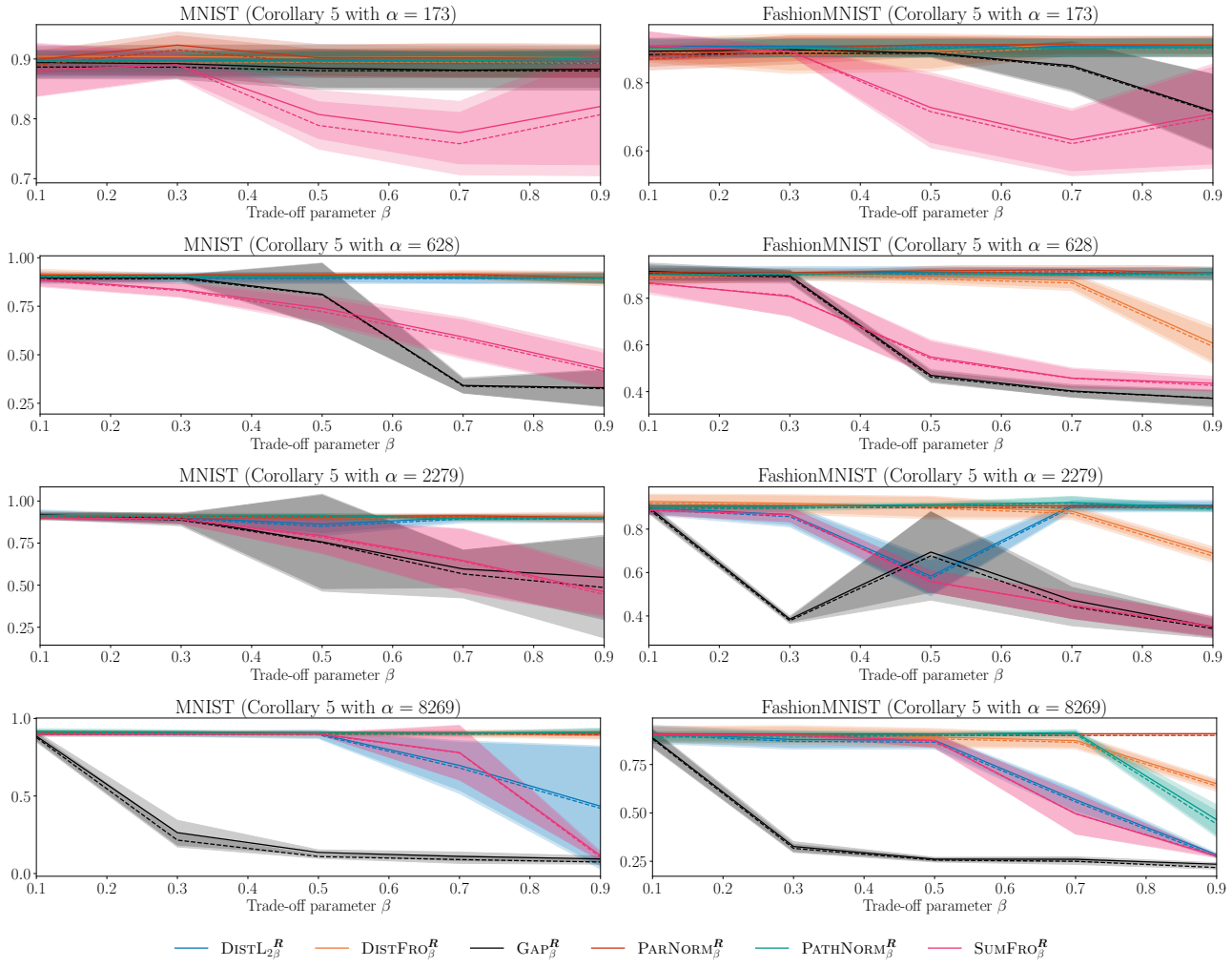

Figure 6: Evolution of the bounds (the plain lines) and the test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$ (the dashed lines) *w.r.t.* the trade-off parameter β for varying α and $\frac{m'}{m'+m} = 0.5$. The lines correspond to the mean, while the bands are the standard deviations.

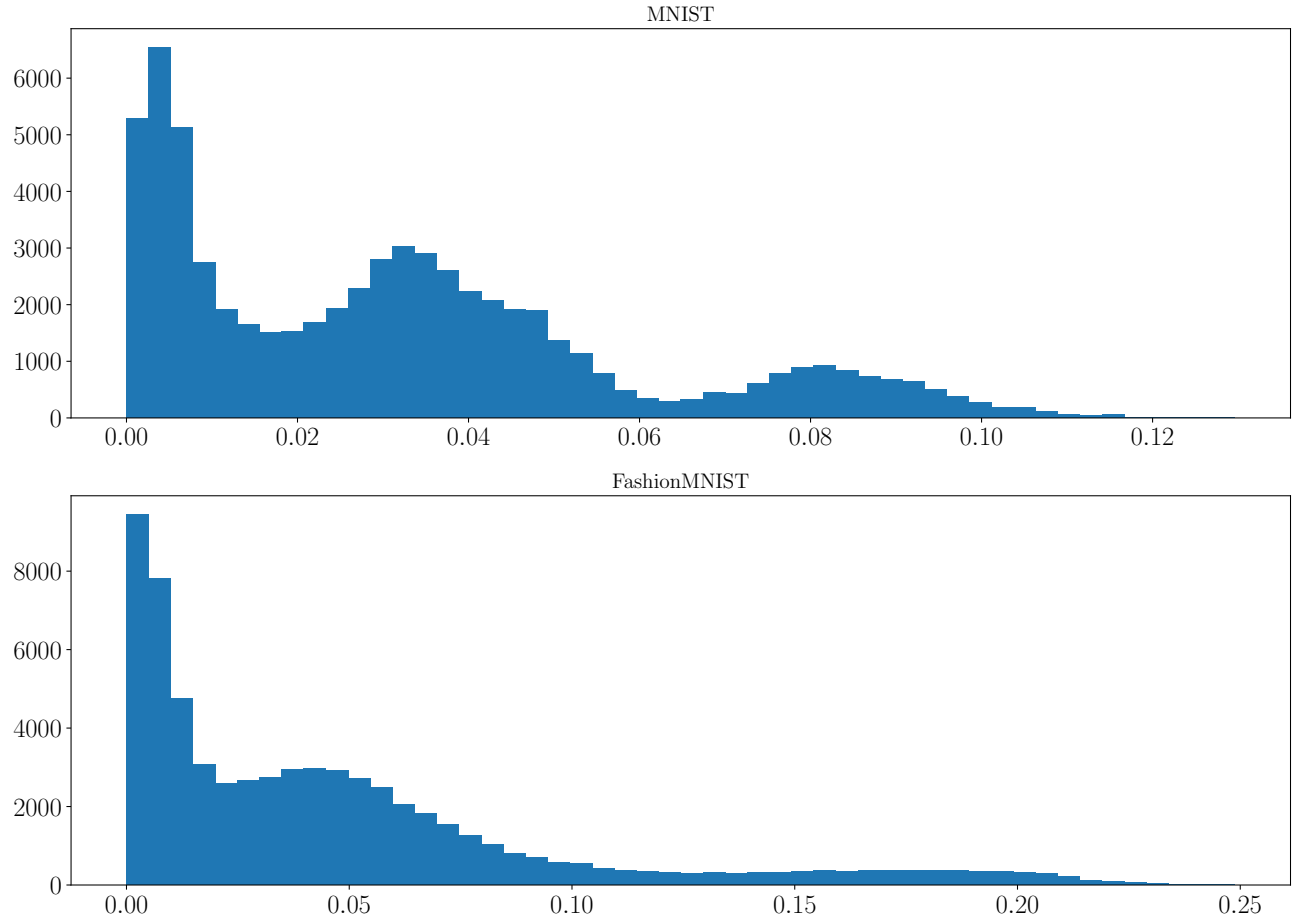

Figure 7: Histograms of the generalization gaps associated with the neural networks in the dataset to learn NEURAL.

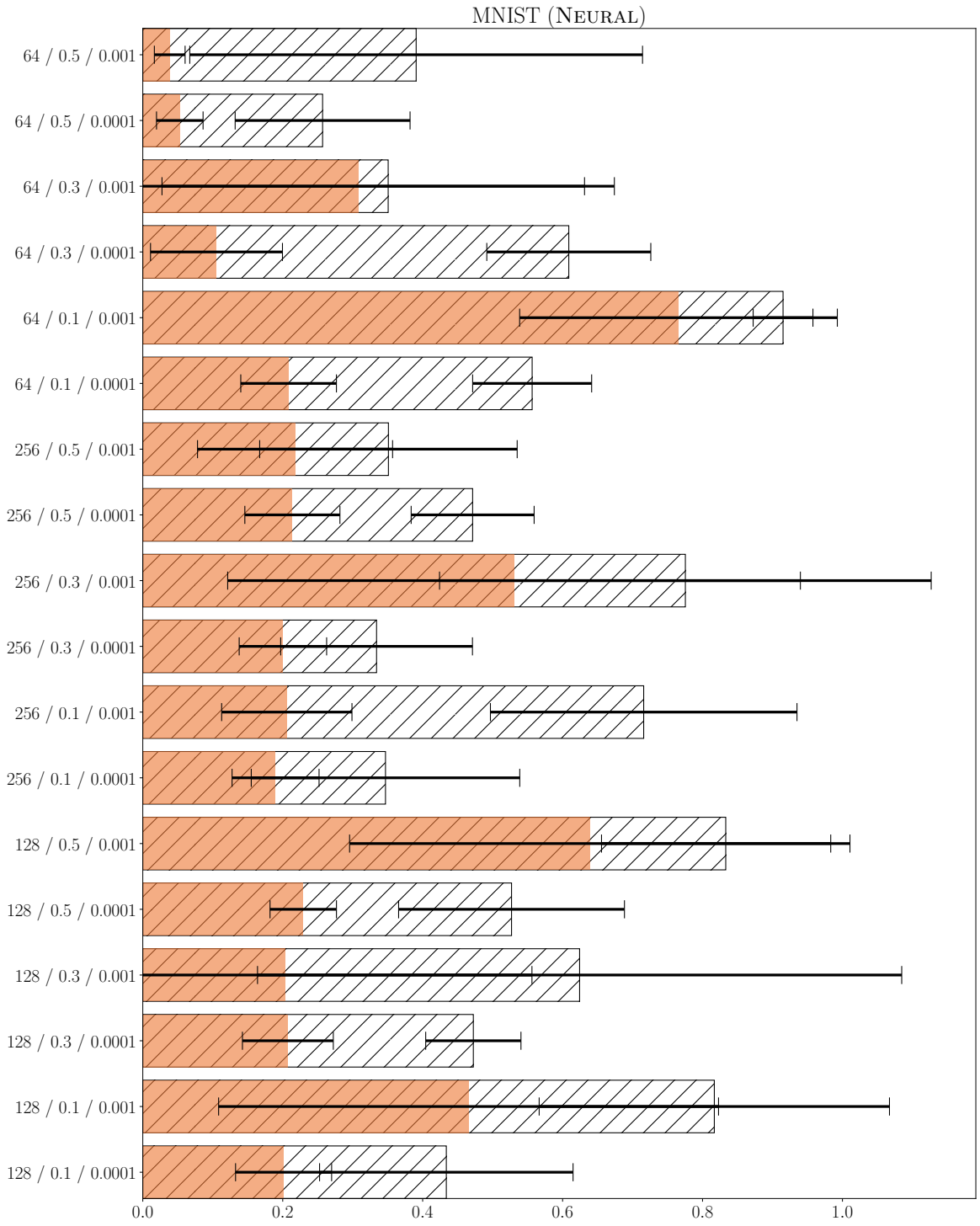

Figure 8: Bar plot of the bound value associated with Corollary 4 and MNIST for the parametric function NEURAL learned with different hyperparameters. On the y-axis, the bar labels "A / B / C" represent the three hyperparameters that vary: "A" is the batch size, "B" is the size of the validation set compared to the original dataset (of neural networks), and "C" is the learning rate of the Adam optimizer. The mean bound values of the sampled hypotheses $h \sim \rho_{\mathcal{S}}$ are shown with the hatched bars, and the mean test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$ are plotted in the colored bars. Moreover, the standard deviations are plotted in black.

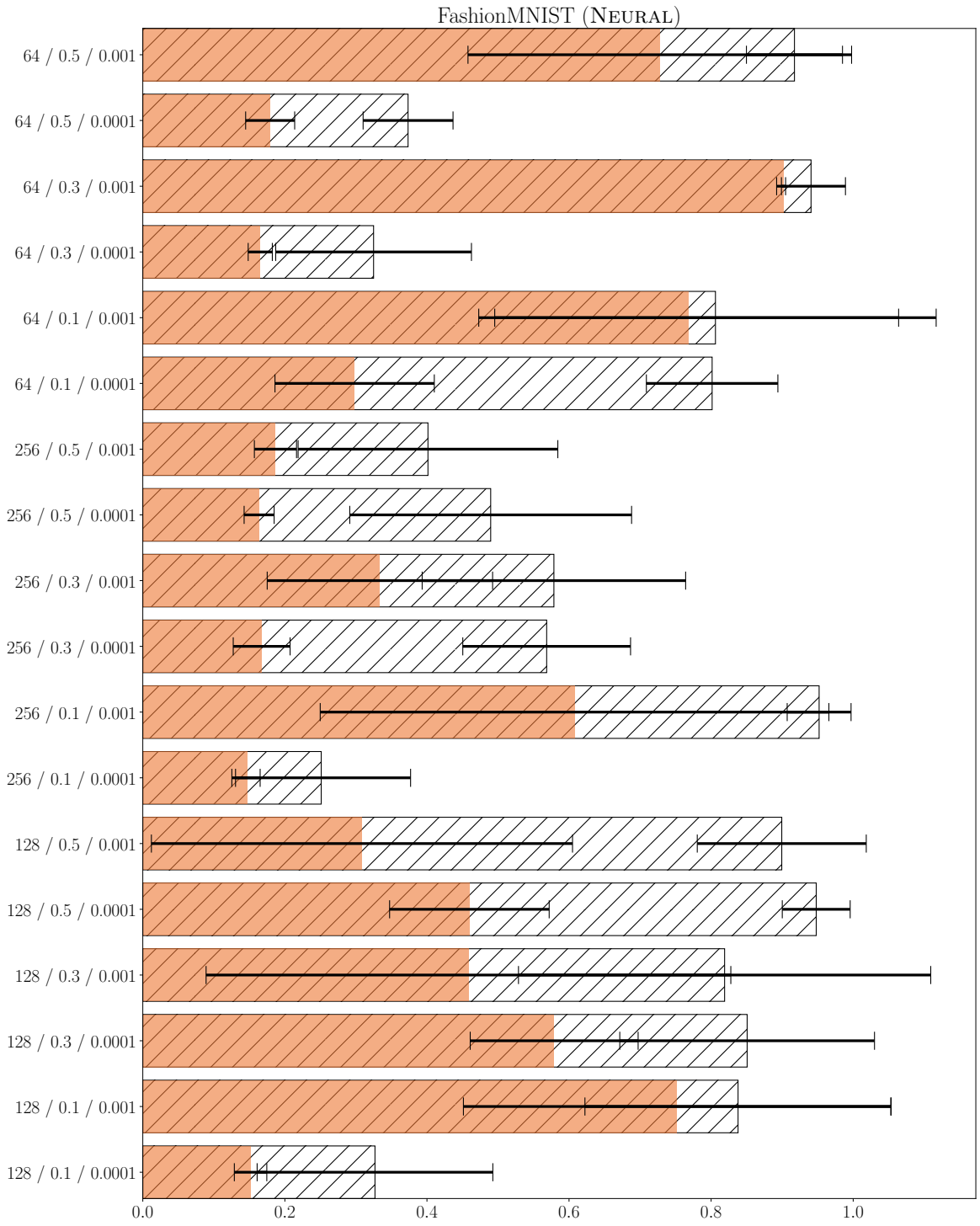

Figure 9: Bar plot of the bound value associated with Corollary 4 and FashionMNIST for the parametric function NEURAL learned with different hyperparameters. On the y-axis, the bar labels "A / B / C" represent the three hyperparameters that vary: "A" is the batch size, "B" is the size of the validation set compared to the original dataset (of neural networks), and "C" is the learning rate of the Adam optimizer. The mean bound values of the sampled hypotheses $h \sim \rho_S$ are shown with the hatched bars, and the mean test risks $R_T^\ell(h)$ are plotted in the colored bars. Moreover, the standard deviations are plotted in black.

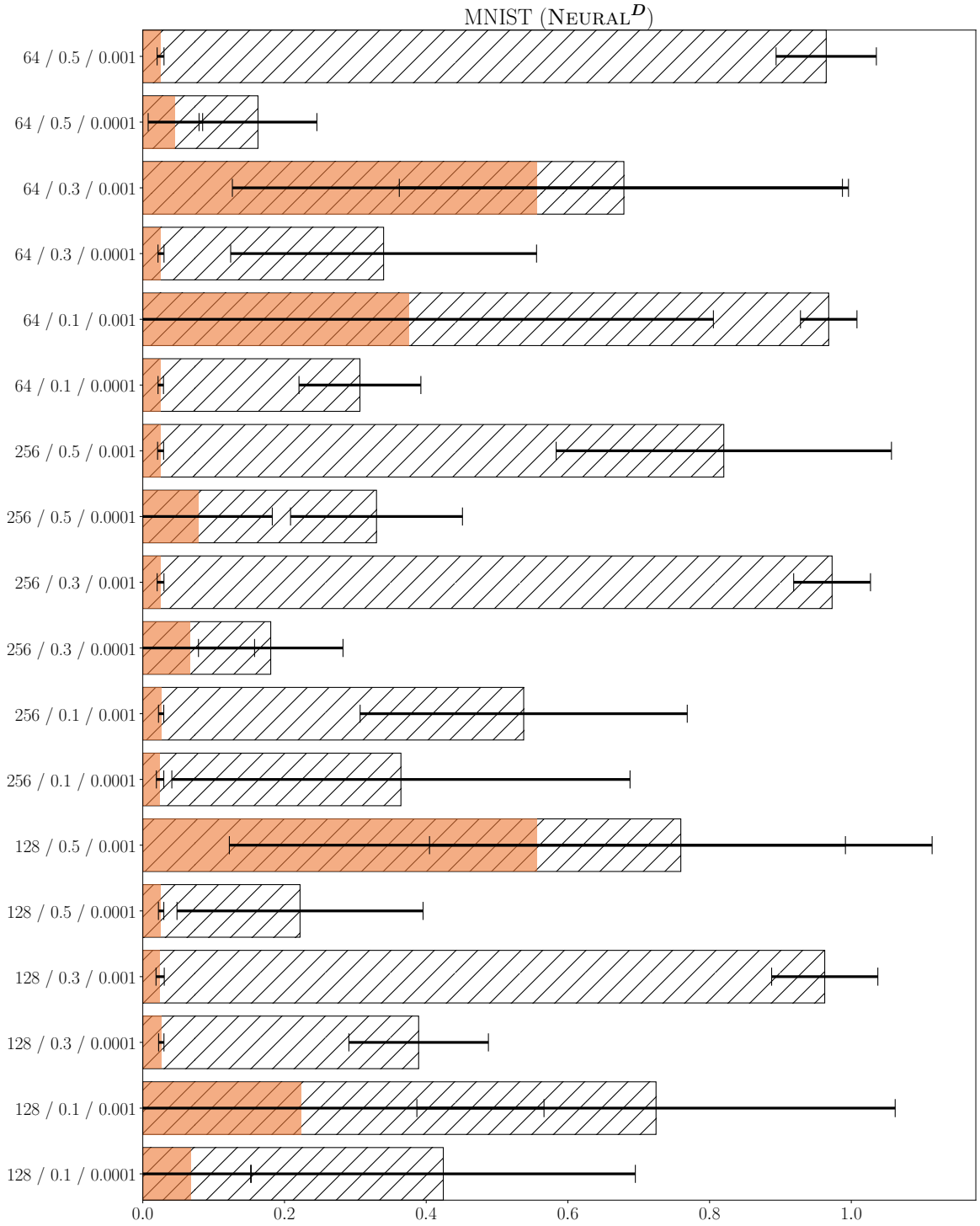

Figure 10: Bar plot of the bound value associated with Corollary 4 and MNIST for the parametric function NEURAL^D learned with different hyperparameters. On the y-axis, the bar labels "A / B / C" represent the three hyperparameters that vary: "A" is the batch size, "B" is the size of the validation set compared to the original dataset (of neural networks), and "C" is the learning rate of the Adam optimizer. The mean bound values of the sampled hypotheses $h \sim \rho_S$ are shown with the hatched bars, and the mean test risks $R^{\ell}_{\mathcal{T}}(h)$ are plotted in the colored bars. Moreover, the standard deviations are plotted in black.

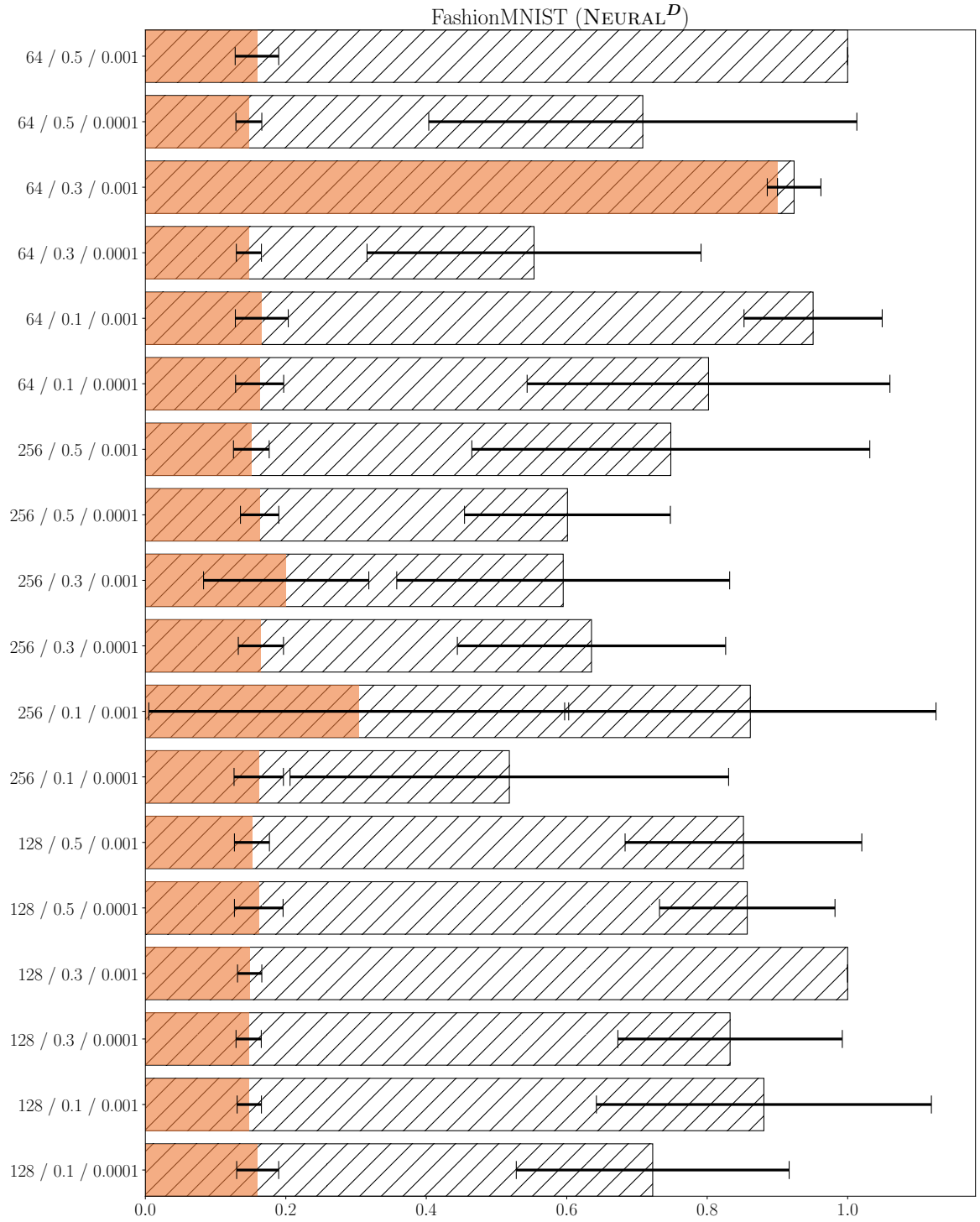

Figure 11: Bar plot of the bound value associated with Corollary 4 and FashionMNIST for the parametric function NEURAL^D learned with different hyperparameters. On the y-axis, the bar labels "A / B / C" represent the three hyperparameters that vary: "A" is the batch size, "B" is the size of the validation set compared to the original dataset (of neural networks), and "C" is the learning rate of the Adam optimizer. The mean bound values of the sampled hypotheses $h \sim \rho_{\mathcal{S}}$ are shown with the hatched bars, and the mean test risks $\mathrm{R}^{\ell}_{\mathcal{T}}(h)$ are plotted in the colored bars. Moreover, the standard deviations are plotted in black.