A General Framework for the Disintegration of PAC-Bayesian Bounds

Paul Viallard, Pascal Germain, Amaury Habrard, Emilie Morvant

A General Framework for the Disintegration of PAC-Bayesian Bounds

Paul Viallard[1], Pascal Germain[2], Amaury Habrard[1], and Emilie Morvant[1]

[1]Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France
[2]Département d'informatique et de génie logiciel, Université Laval, Québec, Canada

Abstract

PAC-Bayesian bounds are known to be tight and informative when studying the generalization ability of randomized classifiers. However, when applied to some family of deterministic models such as neural networks, they require a loose and costly derandomization step. As an alternative to this step, we introduce new PAC-Bayesian generalization bounds that have the originality to provide *disintegrated* bounds, *i.e.*, they give guarantees over one *single* hypothesis instead of the usual *averaged* analysis. Our bounds are easily optimizable and can be used to design learning algorithms. We illustrate the interest of our result on neural networks and show a significant practical improvement over the state-of-the-art framework.

Introduction

PAC-Bayesian theory (Shawe-Taylor and Williamson, 1997; McAllester, 1998) provides a powerful framework for analyzing the generalization ability of machine learning models such as linear classifiers (Germain et al., 2009), SVM (Ambroladze et al., 2006) or neural networks (Dziugaite and Roy, 2017). PAC-Bayesian analyses usually take the form of bounds on the average risk of a randomized classifier with respect to a learned posterior distribution given a chosen prior distribution defined over a set of hypotheses. While such bounds are very effective for analyzing stochastic classifiers, some machine learning methods need nevertheless guarantees on deterministic models. In this case, a *derandomization step* of the bound is required. Different forms of derandomization have been introduced in the literature for specific settings. Among them, Langford and Shawe-Taylor (2002) propose a derandomization for Gaussian posteriors over linear classifiers: thanks to the Gaussian symmetry, a bound on the risk of the *maximum a posteriori* (deterministic) classifier is obtainable from the bound on the average risk of the randomized classifier. Also relying on Gaussian posteriors, Letarte et al. (2019) derived a PAC-Bayes bound for a very specific deterministic network architecture using sign functions as activations. Another line of works derandomizes neural networks (Neyshabur et al., 2018; Nagarajan and Kolter, 2019a). While being technically different, it starts from PAC-Bayesian guarantees on the randomized classifier and use an "output perturbation" bound to convert a guarantee from a random classifier to the mean classifier. These works highlight the need of a general framework for the derandomization of classic PAC-Bayesian bounds.

In this paper, we focus on another kind of derandomization, sometimes referred to as *disintegration of the PAC-Bayesian bound*, and first proposed by Catoni (2007, Th.1.2.7) and Blanchard and Fleuret (2007): Instead of bounding the *average risk of a randomized* classifier with respect to the posterior distribution, the *desintegrated PAC-Bayesian bounds* upper-bound the *risk of a sampled* (unique) classifier from the posterior distribution. Despite their interest in derandomizing PAC-Bayesian bounds, this kind of bounds have only received little study in the literature; especially we can cite the recent work of Rivasplata et al. (2020, Th.1*(i)*) who derived a general disintegrated PAC-Bayesian theorem. Moreover, these bounds have never been used in practice. Driven by machine learning practical purposes, our objective is twofold: To derive new tight *disintegrated* PAC-Bayesian bounds *(i)* that directly derandomize any type of classifiers without any other additional step and with (almost) no impact on the guarantee, *(ii)* that can be easily optimized to learn classifiers with strong guarantees. Our main contribution has a practical objective: Providing a new general framework based on the Rényi divergence allowing efficient learning. We also derive an information-theoretic bound giving interesting new insights on disintegration procedures. Note that for the sake of readability we deferred to the technical appendix the proofs of our theoretical results.

Setting and basics

General notations. We tackle supervised classification tasks with $\mathcal{X}$ the *input space*, $\mathcal{Y}$ the *label set*, and $\mathcal{D}$ an unknown *data distribution* on $\mathcal{X}\times\mathcal{Y}=\mathcal{Z}$. An *example* is denoted by $z=(\mathbf{x},y)\in\mathcal{Z}$, and the *learning sample* $\mathcal{S}=\{z_i\}_{i=1}^m$ is constituted by m examples drawn *i.i.d.* from $\mathcal{D}$; The distribution of such a m-sample being $\mathcal{D}^m$. We consider a *hypothesis set* $\mathcal{H}$ of functions $h:\mathcal{X}\to\mathcal{Y}$. The learner aims to find $h\in\mathcal{H}$ that assigns a label y to an input $\mathbf{x}$ as accurately as possible. Given an example z and a hypothesis h, we assess the quality of the prediction of h with a *loss* function $\ell:\mathcal{H}\times\mathcal{Z}\to[0,1]$ evaluating to which extent the prediction is accurate. Given the loss ℓ, the *true risk* $R_{\mathcal{D}}(h)$ of a hypothesis h on the distribution $\mathcal{D}$ and its empirical counterpart $R_{\mathcal{S}}(h)$ estimated on $\mathcal{S}$ are defined as

$$R_{\mathcal{D}}(h) \triangleq \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \ell(h,z)\,, \quad \text{and} \quad R_{\mathcal{S}}(h) \triangleq \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)\,.$$

Then, we want to find the hypothesis from $\mathcal{H}$ that minimizes $R_{\mathcal{D}}(h)$, that is, however, not computable since $\mathcal{D}$ is unknown. In practice we can consider the Empirical Risk Minimization (ERM) principle that looks for a hypothesis minimizing $R_{\mathcal{S}}(h)$. Generalization guarantees over unseen data from $\mathcal{D}$ can be obtained by quantifying how much the empirical risk $R_{\mathcal{S}}(h)$ is a good estimate of $R_{\mathcal{D}}(h)$. The statistical machine learning theory (*e.g.*, Vapnik, 2000) studies the conditions of consistency and convergence of ERM towards the true value. This kind of result is called *generalization bounds*, often referred to as PAC (Probably Approximately Correct) bounds (Valiant, 1984), and take the form:

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\left|R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)\right| \le \epsilon\big(\tfrac{1}{\delta}, \tfrac{1}{m}, \mathcal{H}\big)\right] \ge 1-\delta.$$

Put into words, with high probability $(1-\delta)$ on the random choice of $\mathcal{S}$, we obtain good generalization properties when the deviation between the true risk $R_{\mathcal{D}}(h)$ and its empirical estimate $R_{\mathcal{S}}(h)$ is low, *i.e.*, $\epsilon\big(\frac{1}{\delta}, \frac{1}{m}, \mathcal{H}\big)$ should be as small as possible. The function ϵ depends on 3 quantities: *(i)* the number of examples m for statistical precision, *(ii)* the hypothesis set $\mathcal{H}$ for assessing how its specificities influence generalization, *(iii)* the confidence parameter δ. We now recall 3 classic bounds with a focus on the PAC-Bayesian theory at the heart of our contribution.

Uniform convergence bound. A first classical type of generalization bounds is referred to as *Uniform Convergence* bounds based on a measure of complexity of the set $\mathcal{H}$ and stand for all the hypotheses of $\mathcal{H}$. Among the most renowned complexity measure, we can cite the VC-dimension or the Rademacher complexity. This type of bound takes the form

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}} |R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)| \le \epsilon\big(\tfrac{1}{\delta}, \tfrac{1}{m}, \mathcal{H}\big)\right] \ge 1-\delta.$$

This bound holds for all $h\in\mathcal{H}$, including the best, but also the worst. This *worst-case* analysis makes hard to obtain a non-vacuous bound with $\epsilon(\frac{1}{\delta}, \frac{1}{m}, \mathcal{H})<1$. The ability of such bounds to explain the generalization of deep learning has been recently challenged (Nagarajan and Kolter, 2019b).

Algorithmic-Dependent Bound. Other bounds depend on the learning algorithm (Bousquet and Elisseeff, 2002; Xu and Mannor, 2012), and directly involve some particularities of the learning algorithm L. This allows obtaining bounds that stand for a single hypothesis $h_{L(\mathcal{S})}$, the one learned with L from the learning sample $\mathcal{S}$. The form of such bounds is

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\left|R_{\mathcal{D}}(h_{L(\mathcal{S})}) - R_{\mathcal{S}}(h_{L(\mathcal{S})})\right| \le \epsilon\big(\tfrac{1}{\delta}, \tfrac{1}{m}, L\big)\right] \ge 1-\delta.$$

PAC-Bayesian Bound. This paper leverages on PAC-Bayes bounds that stand in the PAC framework but borrow inspiration from the Bayesian probabilistic view that deals with randomness and uncertainty in machine learning (McAllester, 1998). That being said, considering the set $\mathcal{M}(\mathcal{H})$ of probability measures on $\mathcal{H}$, a PAC-Bayesian generalization bound is a bound in expectation over the hypothesis set $\mathcal{H}$ and involves a *prior* distribution $\mathcal{P}\in\mathcal{M}(\mathcal{H})$ on $\mathcal{H}$ and a *posterior* distribution $\mathcal{Q}\in\mathcal{M}(\mathcal{H})$ on $\mathcal{H}$ learned from $\mathcal{S}$. The form of such a bound is

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\forall\mathcal{Q},\ \mathbb{E}_{h\sim\mathcal{Q}}|R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)| \le \epsilon\big(\tfrac{1}{\delta}, \tfrac{1}{m}, \mathcal{Q}\big)\right] \ge 1-\delta.$$

We recall below the classical PAC-Bayesian theorem in a slightly different form from the usual one: From Theorem 1, $\phi(h,\mathcal{S}) = \exp(\phi(h,\mathcal{S}))$ gives the usual form of the bounds.

Theorem 1 (General PAC-Bayes bounds)**.** *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any $\phi:\mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}^+$, for any $\delta>0$ we have*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\Big(\forall\mathcal{Q},\ \mathbb{E}_{h\sim\mathcal{Q}}\ln(\phi(h,\mathcal{S})) \le \qquad \text{(Germain et al., 2009)}$$
$$\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\Big[\tfrac{1}{\delta}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathbb{E}_{h\sim\mathcal{P}}\phi(h,\mathcal{S})\Big]\Big) \ge 1-\delta, \tag{0.1}$$

and
$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\Big(\forall\mathcal{Q},\ \tfrac{\alpha}{\alpha-1}\ln\Big[\mathbb{E}_{h\sim\mathcal{Q}}\phi(h,\mathcal{S})\Big] \le \qquad \text{(Bégin et al., 2016)}$$
$$D_\alpha(\mathcal{Q}\|\mathcal{P}) + \ln\Big[\tfrac{1}{\delta}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathbb{E}_{h\sim\mathcal{P}}\phi(h,\mathcal{S})^{\frac{\alpha}{\alpha-1}}\Big]\Big) \ge 1-\delta, \tag{0.2}$$

with $D_\alpha(\mathcal{Q}\|\mathcal{P}) \triangleq \frac{1}{\alpha-1}\ln\Big[\mathbb{E}_{h\sim\mathcal{P}}\Big[\frac{\mathcal{Q}(h)}{\mathcal{P}(h)}\Big]^\alpha\Big]$ *the Rényi divergence* $(\alpha>1)$ *and* $\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) \triangleq \mathbb{E}_{h\sim\mathcal{Q}}\ln\frac{\mathcal{Q}(h)}{\mathcal{P}(h)}$ *the KL-divergence.*

Equation (0.2) is more general than Equation (0.1) since the Rényi divergence can be seen as a generalization of the KL-divergence: When α tends to 1, then $D_\alpha(\mathcal{Q}\|\mathcal{P})$ tends to $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})$. The advantage of these bounds is that they are general since they can be seen as the starting point for deriving different forms of bounds by, first, instantiating ϕ to capture a deviation between the true and empirical risks, second, upper-bounding the term $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathbb{E}_{h\sim\mathcal{P}}\phi(h,\mathcal{S})$. For instance, with $\phi(h,\mathcal{S}) = e^{m2[R_{\mathcal{D}}(h)-R_{\mathcal{S}}(h)]^2}$ we retrieve from Equation (0.1) the form of McAllester (1998):

$$\mathbb{E}_{h\sim\mathcal{Q}} R_{\mathcal{D}}(h) \le \mathbb{E}_{h\sim\mathcal{Q}} R_{\mathcal{S}}(h) + \sqrt{\frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{2\sqrt{m}}{\delta}}{2m}}.$$

This bound illustrates the trade-off between the average empirical risk and the bound. Another example, leading to a slightly tighter but less interpretable bound, is the Seeger (2002); Maurer (2004)'s bound that we retrieve with $\phi(h,\mathcal{S}) = e^{m\,\mathrm{kl}[R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h)]}$ and $\mathrm{kl}[q\|p] = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$:

$$\mathbb{E}_{h\sim\mathcal{Q}}\mathrm{kl}(R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h)) \le \frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{2\sqrt{m}}{\delta}}{m}. \tag{0.3}$$

Such PAC-Bayesian bounds are known to be tight, but they stand for a randomized classifier by nature (due to the expectation on $\mathcal{H}$). A key issue for usual machine learning tasks is then the derandomization of the bounds to obtain a guarantee for a deterministic classifier. In some cases, this derandomization is a result of the structure of the hypotheses such as for randomized linear classifier that can be directly expressed as one deterministic linear classifier (Germain et al., 2009). However, in other cases the derandomization is much more complex and specific to a class of hypothesis, such as for neural networks (*e.g.*, Neyshabur et al. (2018), Nagarajan and Kolter (2019b, Ap. J)).

In the following, we state our contribution: A general derandomization framework for desintegrating PAC-Bayesian bounds into a bound for a single hypothesis from $\mathcal{H}$.

Disintegrated PAC-Bayesian theorems

Form of a disintegrated PAC-Bayes bound. First, we recall another kind of bound introduced by Blanchard and Fleuret (2007) and Catoni (2007, Th.1.2.7) and referred to as the disintegrated PAC-Bayesian bound; Its form is

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,\, h\sim\mathcal{Q}_\mathcal{S}}\left[|R_\mathcal{D}(h) - R_\mathcal{S}(h)| \le \epsilon\big(\tfrac{1}{\delta}, \tfrac{1}{m}, \mathcal{Q}_\mathcal{S}\big)\right] \ge 1-\delta.$$

This kind of generalization bound allows one to derandomize the usual PAC-Bayes bounds as follows. Instead of considering a bound that stands for all the posterior distributions on $\mathcal{H}$ as usually done in PAC-Bayes (Theorem 1), we propose to consider only the posterior distribution $\mathcal{Q}_\mathcal{S}$ obtained through a deterministic algorithm taking the learning sample $\mathcal{S}$ as input. More formally, we consider a *deterministic* algorithm $A : \mathcal{Z}^m \to \mathcal{M}(\mathcal{H})$ chosen *a priori* which *(i)* takes a learning sample $\mathcal{S} \in \mathcal{Z}^m$ as input and *(ii)* outputs a *data-dependent* distribution $\mathcal{Q}_\mathcal{S} \triangleq A(\mathcal{S})$ from the set $\mathcal{M}(\mathcal{H})$ of all possible probability measures on $\mathcal{H}$. Then, the above bound stands for a unique hypothesis $h\sim\mathcal{Q}_\mathcal{S}$ instead of a randomized classifier: The individual risks are no longer averaged with respect to $\mathcal{Q}_\mathcal{S}$; This is the *PAC-Bayesian bound disintegration*. The dependence in probability on $\mathcal{Q}_\mathcal{S}$ means that the bound is valid with probability at least $1-\delta$ over the random choice of the learning sample $\mathcal{S}\sim\mathcal{D}^m$ and the hypothesis $h\sim\mathcal{Q}_\mathcal{S}$. Under this principle, we present in the following two new general disintegrated PAC-Bayesian bounds in Theorems 2 and 4. Importantly, they have the advantage to be instantiable to specific settings. Doing so, one obtains an easily optimizable bound, leading to a self-bounding algorithm with theoretical guarantees. In this paper, we provide such an instantiation for neural networks.

Disintegrated PAC-Bayesian bounds with the Rényi divergence. In the same spirit as Equation (0.2) our first result stated below can be seen as a general bound depending on the Rényi divergence $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ of order $\alpha > 1$.

Theorem 2 (Disintegrated PAC-Bayes Bound)**.** *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}^+$, for any $\alpha > 1$, for any $\delta > 0$, for any algorithm $A : \mathcal{Z}^m \to \mathcal{M}(\mathcal{H})$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\Bigg(\frac{\alpha}{\alpha-1}\ln\left(\phi(h,\mathcal{S})\right) \;\le\; \frac{2\alpha-1}{\alpha-1}\ln\frac{2}{\delta} + D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) + \ln\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]\Bigg) \ge 1-\delta,$$

where $\mathcal{Q}_\mathcal{S} \triangleq A(\mathcal{S})$ is output by the deterministic algorithm A.

Proof sketch (see Appendix for details). First, note that $\mathcal{Q}_\mathcal{S}$ is obtained with algorithm A from $\mathcal{S}$. Then, applying Markov's inequality on $\phi(h,\mathcal{S})$ with the random variable (r.v.) h and using Hölder's inequality to introduce $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$, we have with probability at least $1-\frac{\delta}{2}$ on $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$

$$\begin{aligned}\tfrac{\alpha}{\alpha-1}\ln\left(\phi(h,\mathcal{S})\right) &\le \tfrac{\alpha}{\alpha-1}\ln\Big(\tfrac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\Big)\\ &\le D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) + \tfrac{\alpha}{\alpha-1}\ln\tfrac{2}{\delta} + \ln\Big(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\Big).\end{aligned}$$

By applying again Markov's inequality on $\phi(h,\mathcal{S})$ with the r.v. $\mathcal{S}$, we have with probability at least $1-\frac{\delta}{2}$ on $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$

$$\ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right) \le \ln\left(\tfrac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right).$$

By combining the two bounds with a union bound argument, we obtain the desired result. □

Note that Hölder's inequality is used differently in the classic PAC-Bayes bound's proof. Indeed, in Bégin et al. (2016, Th. 8), the change of measure (based on Hölder's inequality) is key to obtain a bound that holds for all posteriors $\mathcal{Q}$ with high probability, while our bound holds for a unique posterior $\mathcal{Q}_\mathcal{S}$ dependent on the sample $\mathcal{S}$. In fact, we use Hölder's inequality to make appear a prior $\mathcal{P}$ independent from $\mathcal{S}$, a crucial point for our bound instantiated in Corollary 6. Compared to Equation (0.2), our bound requires an additional term $\ln 2 + \frac{\alpha}{\alpha-1}\ln\frac{2}{\delta}$. However, by taking $\phi(h,\mathcal{S}) = m\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))$ and $\alpha = 2$, the term $\ln\frac{8}{\delta^2}$ is multiplied by $\frac{1}{m}$, allowing us to obtain a reasonable overhead to "derandomize" a bound into a disintegrated one. For instance, if $m = 5,000$ (reasonable sample size) and $\delta = .05$, we have $\frac{1}{m}\ln\frac{8}{\delta^2} \approx .002$.

We instanciate below Theorem 2 for $\alpha \to 1^+$ and $\alpha \to +\infty$.

Corollary 3. *Under the assumptions of Theorem 2, when $\alpha \to 1^+$, with probability at least $1-\delta$ we have*

$$\ln\phi(h,\mathcal{S}) \le \ln\tfrac{2}{\delta} + \ln\left[\operatorname{esssup}_{\mathcal{S}'\in\mathcal{Z}, h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right].$$

When $\alpha \to +\infty$, with probability at least $1-\delta$ we have

$$\ln\phi(h,\mathcal{S}) \le \ln\operatorname*{esssup}_{h'\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')} + \ln\Big[\tfrac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\Big],$$

where esssup *is the essential supremum.*

This corollary illustrates that the parameter α controls the trade-off between the Rényi divergence $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ and $\ln\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]$. Indeed, when $\alpha\to 1^+$, the Rényi divergence vanishes while the other term converges toward $\ln\left[\operatorname{esssup}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right]$, roughly speaking the maximal value possible for the second term. On the other hand, when $\alpha\to+\infty$, the Rényi divergence increases and converges toward $\ln\operatorname{esssup}_{h'\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}$ and the other term decreases toward $\ln[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')]$.

For the sake of comparison, we recall the bound of Rivasplata et al. (2020, Th.1*(i)*), that is more general than the bounds of Blanchard and Fleuret (2007) and Catoni (2007):

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\Bigg(\ln(\phi(h,\mathcal{S})) \leq \ln\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)} + \ln\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\Bigg) \geq 1-\delta. \quad (0.4)$$

Our bound has a worse dependence in δ, which is the price we pay for a general result with a divergence. Indeed, the main difference with Equation (0.4) is that our bound involves the Rényi divergence $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ between the prior $\mathcal{P}$ and the posterior $\mathcal{Q}_\mathcal{S}$. In contrast, the term $\ln\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}$ (also involved in the bounds of Catoni (2007); Blanchard and Fleuret (2007)) can be seen as a "disintegrated[1] KL-divergence" depending only on the sampled $h\sim\mathcal{Q}_\mathcal{S}$. Overall, from the properties of the Rényi divergence (van Erven and Harremos, 2014), our bound is expected to be looser. However, as we show in our experiment, the divergence term of Equation (0.4) makes it harder to optimize since it can be subject to a high variance. Optimizing our bound with $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ makes the procedure more stable and efficient which is thus more interesting in practice, allowing us to get better empirical results.

We now provide a parametrized version of our bound enlarging its practical scope. In the PAC-Bayes literature, parametrized bounds have been introduced (*e.g.*, Catoni (2007); Thiemann et al. (2017)) to control the trade-off between the empirical risk and the divergence along with the additional term. We follow a similar approach to introduce a version of a disintegrated Rényi divergence-based bound that has the advantage to be parametrizable.

Theorem 4 (Parametrizable Disintegrated PAC-Bayes Bound)**.** *Under the same assumptions than Theorem 2, we have*

$$\mathop{\mathbb{P}}_{\substack{\mathcal{S}\sim\mathcal{D}^m\\ h\sim\mathcal{Q}_\mathcal{S}}}\Bigg(\forall\lambda>0,\ \ln(\phi(h,\mathcal{S})) \leq \ln\left[\frac{\lambda}{2}e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}+\frac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\Bigg) \geq 1-\delta.$$

Of note, $e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}$ is closely related to the χ^2-distance: $e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}-1=\chi^2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$. An asset of Theorem 4 is the parameter λ controlling the trade-off between the exponentiated Rényi divergence $e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}$ and $\frac{1}{\delta^3}\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2$. Our bound is valid for all $\lambda>0$, then from a practical view we can learn the parameter to minimize the bound (and control the possible numerical instability due to $e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}$. Despite this algorithmic advantage, for a given $\mathcal{P}$ and $\mathcal{Q}_\mathcal{S}$, Proposition 5 shows the optimal bound of Theorem 4 is the bound of Theorem 2.

Proposition 5. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any $\delta>0$ and for any algorithm $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, let*

$$\lambda^*=\mathop{\mathrm{argmin}}_{\lambda>0}\ \ln\left[\frac{\lambda}{2}e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}+\frac{\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}8\phi(h',\mathcal{S}')^2}{2\lambda\delta^3}\right],$$

$$\text{then}\quad \overbrace{2\ln\left[\frac{\lambda^*}{2}e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}+\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{8\phi(h',\mathcal{S}')^2}{2\lambda^*\delta^3}\right]}^{\text{Theorem 4}} = \underbrace{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\ln\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{8\phi(h',\mathcal{S}')^2}{\delta^3}\right]}_{\text{Theorem 2 with } \alpha=2.},$$

where $\lambda^*=\sqrt{\frac{\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}8\phi(h',\mathcal{S}')^2}{\delta^3\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))}}$. *In other words, the optimal λ^* gives the same bound for Theorems 2 and 4.*

So far we have introduced theoretical results allowing to derandomize PAC-Bayesian bounds through a disintegration approach. Our bounds are general and can be instantiated to different settings. In the next section, we illustrate the instantiation and usefulness of Theorem 2 on neural networks.

The disintegration in action

Specialization to neural network classifiers. We consider neural networks (NN) parametrized by a weight vector $\mathbf{w}\in\mathbb{R}^d$ and overparametrized, *i.e.*, $d\gg m$. We aim at learning the weights of the NN leading to the lowest true risk. Practitioners usually proceed by epochs[2] and obtain one "intermediate" NN after each epoch. Then, they select the best "intermediate" NN associated with the lowest validation risk. We translate this practice into our PAC-Bayesian setting by considering one prior per epoch. Given T epochs, we hence have T priors $\mathbf{P}=\{\mathcal{P}_t\}_{t=1}^T$, where $\forall t\in\{1,\ldots,T\}, \mathcal{P}_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_d)$ is a Gaussian distribution centered at $\mathbf{v}_t$ the weight vector associated with the t-th "intermediate" NN, with a covariance matrix of $\sigma^2\mathbf{I}_d$ (where $\mathbf{I}_d$ is the $d\times d$-dimensional identity matrix). Assuming the T priors are learned from a set $\mathcal{S}_{\text{prior}}$ such that $\mathcal{S}_{\text{prior}}\bigcap\mathcal{S}=\emptyset$, then Corollaries 6 and 7 will guide us to learn a posterior $\mathcal{Q}_\mathcal{S}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$ from the best prior on $\mathcal{S}$ (we give more details on the procedure after the forthcoming corollaries).

We instantiate Theorem 2 to this setting in Corollary 6. For the sake of comparison, we instantiate in Corollary 7 the Rivasplata et al. (2020)'s bound (recalled in Equation (0.4)) in Equation (0.5), and the ones of Blanchard and Fleuret (2007) and Catoni (2007) respectively in Equations (0.6) and (0.7).

Corollary 6. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any set $\mathcal{H}$, for any set $\mathbf{P}$ of T priors on $\mathcal{H}$, for any algo. $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$,*

[1] We say that the KL-divergence is "disintegrated" since the log term is not averaged in contrast to the KL-divergence.

[2] One epoch corresponds to one pass of the entire learning set during the optimization process.

for any loss $\ell:\mathcal{H}\times\mathcal{Z}\to[0,1]$, for any $\delta>0$, we have

$$\underset{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}{\mathbb{P}}\left(\forall\mathcal{P}_t\in\mathbf{P},\ \mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le \frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2}+\ln\frac{16T\sqrt{m}}{\delta^3}\right]\right)\ge 1-\delta,$$

where $\mathrm{kl}(a\|b)=a\ln\frac{a}{b}+(1-a)\ln\frac{1-a}{1-b}$.

Corollary 7. *Under Corollary 6's assumptions, with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$, we have $\forall\mathcal{P}_t\in\mathbf{P}$*

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le\frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{2T\sqrt{m}}{\delta}\right], \tag{0.5}$$

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le\frac{1}{m}\left[\frac{m+1}{m}\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{Tm+T}{\delta}\right], \tag{0.6}$$

$$\forall c\in\mathbf{C}, R_\mathcal{D}(h)\le\frac{1-e^{-cR_\mathcal{S}(h)-\frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{T|\mathbf{C}|}{\delta}\right]}}{1-e^{-c}}, \tag{0.7}$$

where $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)$ is a Gaussian noise s.t. $\mathbf{w}+\boldsymbol{\epsilon}$ are the weights of $h\sim\mathcal{Q}_\mathcal{S}$ with $\mathcal{Q}_\mathcal{S}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$, and $\mathbf{C}$ is a set of hyperparameters fixed a priori.

Experiments we performed involve the direct minimization of Corollary 6 (or any other bounds of Corollary 7). To obtain a tight bound, the divergence between one prior $\mathcal{P}_t\in\mathbf{P}$ and $\mathcal{Q}_\mathcal{S}$ must be low, *i.e.*, $\|\mathbf{w}-\mathbf{v}_t\|2^2$ (or $\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|2^2-\|\boldsymbol{\epsilon}\|_2^2)$ has to be low. However, this can be challenging with NNs since $d\gg m$. One solution is to split the learning sample into 2 non-overlapping subsets $\mathcal{S}_{\text{prior}}$ and $\mathcal{S}$, where $\mathcal{S}_{\text{prior}}$ is used to learn the prior, while $\mathcal{S}$ is used both to learn the posterior and compute the bound. Hence, if we "pre-learn" a good enough prior $\mathcal{P}_t\in\mathbf{P}$ from $\mathcal{S}_{\text{prior}}$, then we can expect to have a low $\|\mathbf{w}-\mathbf{v}_t\|2$.

Training Method

The original training set is split in two distinct subsets: $\mathcal{S}_{\text{prior}}$ and $\mathcal{S}$. The training has two phases.
1) The prior distribution $\mathcal{P}$ is "pre-learned" with $\mathcal{S}_{\text{prior}}$ and selected by early stopping the algorithm with $\mathcal{S}$ as validation set (from an arbitrary learning algorithm A_{prior}).
2) Given $\mathcal{S}$ and $\mathcal{P}$, we learn the posterior $\mathcal{Q}_\mathcal{S}$ with the algorithm A (defined *a priori*).

At first sight, the selection of the prior weights with $\mathcal{S}$ by early stopping may appear to be "cheating". However, this procedure can be seen as: **1)** first constructing $\mathbf{P}$ from the T "intermediate" NNs learned after each epoch from $\mathcal{S}_{\text{prior}}$, **2)** optimizing the bound with the prior that leads to the best risk on $\mathcal{S}$. This gives a statistically valid result, since Corollaries 6 is valid for every $\mathcal{P}_t\in\mathbf{P}$, we can thus select the one we want, in particular the best one on $\mathcal{S}$. Usually, practitioners consider this "best" prior as the final NN. In our case, the advantage is that we refine this "best" prior on $\mathcal{S}$ to learn the posterior $\mathcal{Q}_\mathcal{S}$. Note that Pérez-Ortiz et al. (2020) have already introduced tight generalization bounds with data-dependent priors for—non-derandomized—stochastic NNs. Nevertheless, our training method to learn the prior differs greatly since **1)** we learn T NNs (*i.e.*, T priors) instead of only one, **2)** we also fix the variance of the Gaussian in the posterior $\mathcal{Q}_\mathcal{S}$. Hence, to the best of our knowledge, our training method for the prior is new.

A note about stochastic neural networks. Due to its stochastic nature, PAC-Bayesian theory has been explored to study stochastic NNs (*e.g.*, Dziugaite and Roy (2017; 2018); Zhou et al. (2019); Pérez-Ortiz et al. (2020)). In Corollary 8 below, we instantiate the bound of Equation (0.1) for stochastic NNs to empirically compare the stochastic and the deterministic NNs associated with prior(s) and posterior distributions. We recall that, in this paper, a deterministic NN is a *single* h sampled from the posterior distribution $\mathcal{Q}_\mathcal{S}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$ output by the algorithm A. This means that for each example the label prediction is performed by the same deterministic NN: The one parametrized by the weights $\mathbf{w}\in\mathbb{R}^d$. Conversely, the stochastic NN associated with a posterior distribution $\mathcal{Q}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$ predicts the label of a given example by *(i)* first sampling h according to $\mathcal{Q}$ (*i.e.*, the NN parametrized by $\mathbf{w}\in\mathbb{R}^d$), *(ii)* then returning the label predicted by h.[3] Thus, the risk of the stochastic NN is the expected risk value $\mathbb{E}_{h\sim\mathcal{Q}}R_\mathcal{D}(h)$, where the expectation is taken over *all* h sampled from $\mathcal{Q}$. We compute the empirical risk of the stochastic NN from a Monte Carlo approximation: *(i)* we sample n weight vectors, and *(ii)* we average the risk over the n associated NNs; We denote by $\mathcal{Q}^n$ the distribution of such n-sample. In this context, we propose the following randomized PAC-Bayesian bound.

Corollary 8. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any $\mathcal{H}$, for any set $\mathbf{P}$ of T priors on $\mathcal{H}$, for any loss $\ell:\mathcal{H}\times\mathcal{Z}\to\{0,1\}$, for any $\delta>0$, with proba. at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $\{h_1,\ldots,h_n\}\sim\mathcal{Q}^n$, we have simultaneously $\forall\mathcal{P}_t\in\mathbf{P},\forall\mathcal{Q}$,*

$$\mathrm{kl}\left[\underset{h\sim\mathcal{Q}}{\mathbb{E}}R_\mathcal{S}(h)\|\underset{h\sim\mathcal{Q}}{\mathbb{E}}R_\mathcal{D}(h)\right]\le\frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2}+\ln\frac{4T\sqrt{m}}{\delta}\right], \tag{0.8}$$

$$\text{and}\quad \mathrm{kl}\left[\frac{1}{n}\sum_{i=1}^n R_\mathcal{S}(h_i)\|\underset{h\sim\mathcal{Q}}{\mathbb{E}}R_\mathcal{S}(h)\right]\le\frac{1}{n}\ln\frac{4}{\delta}. \tag{0.9}$$

This result shows two key features that allow considering it as an adapted baseline for a *fair* comparison between disintegrated and classical PAC-Bayes bounds, thus between deterministic and stochastic NNs. On the one hand, it involves the same terms as Corollary 6. On the other hand, it is close to the bound of Pérez-Ortiz et al. (2020, Sec. 6.2), since *(i)* we adapt the KL-divergence to our setting (*i.e.*, $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})=\frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$), *(ii)* the bound holds for T priors thanks to a union bound argument.

[3] The stochastic classifier is called Gibbs classifier in PAC-Bayes.

		$\sigma^2=10^{-3}$				$\sigma^2=10^{-4}$				$\sigma^2=10^{-5}$				$\sigma^2=10^{-6}$			
	lr$=10^{-4}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
mnist	RIVASPLATA	.0096 ±.0004	.0480 ±.0016	.0093 ±.0003	716 ±39	.0071 ±.0002	.1024 ±.0023	.0080 ±.0001	2357 ±69	.0071 ±.0001	.2542 ±.0031	.0085 ±.0001	7589 ±122	.0078 ±.0001	.5880 ±.0029	.0116 ±.0001	24574 ±210
	BLANCHARD	.0096 ±.0004	.0483 ±.0016	.0093 ±.0003	719 ±39	.0071 ±.0002	.1028 ±.0023	.0080 ±.0001	2364 ±69	.0071 ±.0001	.2548 ±.0030	.0085 ±.0001	7608 ±119	.0079 ±.0001	.5834 ±.0030	.0116 ±.0001	24240 ±209
	CATONI	.0110 ±.0006	.0600 ±.0020	.0127 ±.0006	771 ±36	.0077 ±.0002	.1365 ±.0034	.0092 ±.0002	2424 ±71	.0081 ±.0002	.3043 ±.0030	.0108 ±.0001	7633 ±125	.0145 ±.0002	.6182 ±.0027	.0172 ±.0001	23719 ±209
	OURS	.0092 ±.0004	**.0115** ±.0004	.0079 ±.0003	0.3111	.0066 ±.0002	**.0100** ±.0002	.0064 ±.0001	3.4654	.0069 ±.0001	**.0108** ±.0001	.0072 ±.0001	2.2368	.0069 ±.0001	**.0222** ±.0001	.0097 ±.0000	117
	STOCHASTIC	—	.0338	—	0.1555	—	.0312	—	1.7327	—	.0326	—	1.1184	—	.0455	—	58
fashion	RIVASPLATA	.1123 ±.0018	.1849 ±.0032	.1051 ±.0015	716 ±39	.0979 ±.0008	.2461 ±.0027	.0904 ±.0006	2396 ±64	.0979 ±.0004	.4083 ±.0031	.0936 ±.0003	7449 ±121	.1023 ±.0003	.9938 ±.0001	.0993 ±.0002	127705 ±491
	BLANCHARD	.1122 ±.0019	.1854 ±.0032	.1051 ±.0015	720 ±39	.0980 ±.0009	.2465 ±.0029	.0905 ±.0006	2397 ±69	.0977 ±.0004	.4079 ±.0031	.0937 ±.0003	7424 ±122	.0997 ±.0003	.9694 ±.0005	.0964 ±.0002	85087 ±425
	CATONI	.1149 ±.0022	.1979 ±.0030	.1077 ±.0018	772 ±40	.1022 ±.0009	.2557 ±.0032	.0953 ±.0006	2422 ±71	.1044 ±.0007	.4733 ±.0045	.1010 ±.0004	7629 ±122	.1021 ±.0003	.8314 ±.0013	.0991 ±.0002	23654 ±219
	OURS	.1099 ±.0019	**.1132** ±.0017	.1019 ±.0016	0.4825	.0951 ±.0010	**.0967** ±.0008	.0858 ±.0008	2.1152	.0951 ±.0005	**.1014** ±.0003	.0894 ±.0003	5.0842	.0954 ±.0003	**.1087** ±.0002	.0913 ±.0002	30
	STOCHASTIC	—	.1628	—	0.2412	—	.1434	—	1.0576	—	.1485	—	2.5421	—	.1549	—	15
cifar10	RIVASPLATA	.2958 ±.0012	.4117 ±.0032	.2828 ±.0008	884 ±42	.2886 ±.0006	.5178 ±.0031	.2871 ±.0004	2722 ±72	.2880 ±.0003	.6883 ±.0028	.2815 ±.0002	8696 ±130	.3019 ±.0002	.9451 ±.0009	.2972 ±.0001	36177 ±274
	BLANCHARD	.2958 ±.0012	.4121 ±.0029	.2828 ±.0008	886 ±39	.2886 ±.0006	.5179 ±.0030	.2871 ±.0004	2719 ±70	.2881 ±.0003	.6885 ±.0028	.2815 ±.0002	8699 ±130	.3019 ±.0002	.9641 ±.0006	.2972 ±.0001	43500 ±288
	CATONI	.2950 ±.0012	.4342 ±.0020	.2817 ±.0008	969 ±43	.2886 ±.0006	.4900 ±.0027	.2872 ±.0004	2078 ±63	.2882 ±.0003	.6791 ±.0044	.2816 ±.0002	6972 ±123	.3019 ±.0002	.9779 ±.0002	.2972 ±.0001	20924 ±203
	OURS	.2990 ±.0012	**.3073** ±.0009	.2868 ±.0009	3.7321	.2894 ±.0006	**.3083** ±.0004	.2884 ±.0004	2.2824	.2883 ±.0004	**.3020** ±.0003	.2818 ±.0002	3.1390	.3019 ±.0002	**.3236** ±.0001	.2973 ±.0001	19
	STOCHASTIC	—	.3737	—	1.8661	—	.3750	—	1.1412	—	.3682	—	1.5695	—	.3886	—	9.3925
	lr$=10^{-6}$	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
mnist	RIVASPLATA	.0123 ±.0008	**.0182** ±.0010	.0145 ±.0009	0.9224 ±1.32	.0078 ±.0002	.0131 ±.0005	.0094 ±.0002	7.3079 ±3.78	.0083 ±.0002	.0187 ±.0008	.0110 ±.0001	46 ±10	.0152 ±.0002	.0384 ±.0012	.0176 ±.0001	207 ±21
	BLANCHARD	.0122 ±.0008	.0188 ±.0009	.0145 ±.0008	0.7241 ±1.36	.0078 ±.0002	.0136 ±.0005	.0094 ±.0002	7.4296 ±3.81	.0083 ±.0002	.0191 ±.0008	.0110 ±.0001	47 ±10	.0152 ±.0002	.0388 ±.0012	.0176 ±.0001	209 ±21
	CATONI	.0123 ±.0008	**.0182** ±.0010	.0146 ±.0009	0.9295 ±1.33	.0078 ±.0002	.0151 ±.0014	.0094 ±.0002	7.8751 ±3.96	.0083 ±.0002	.0205 ±.0006	.0110 ±.0001	54 ±11	.0153 ±.0002	.0404 ±.0011	.0177 ±.0001	240 ±22
	OURS	.0120 ±.0008	.0187 ±.0009	.0141 ±.0008	0.0000	.0077 ±.0002	**.0131** ±.0002	.0092 ±.0002	0.0007	.0081 ±.0002	**.0148** ±.0001	.0107 ±.0001	0.0078	.0136 ±.0002	**.0214** ±.0001	.0164 ±.0001	0.0530
	STOCHASTIC	—	.0448	—	0.0000	—	.0363	—	0.0003	—	.0389	—	0.0039	—	.0486	—	0.0265
fashion	RIVASPLATA	.1219 ±.0042	.1227 ±.0036	.1136 ±.0035	0.9212 ±1.34	.1050 ±.0011	.1082 ±.0014	.0975 ±.0008	7.3488 ±3.80	.1077 ±.0007	.1225 ±.0018	.1032 ±.0004	44 ±9	.1025 ±.0003	.1387 ±.0020	.0998 ±.0002	197 ±19
	BLANCHARD	.1215 ±.0042	.1241 ±.0036	.1133 ±.0035	0.8880 ±1.24	.1049 ±.0011	.1095 ±.0013	.0975 ±.0008	7.4644 ±3.85	.1077 ±.0007	.1236 ±.0017	.1032 ±.0004	45 ±10	.1025 ±.0003	.1394 ±.0021	.0998 ±.0002	199 ±20
	CATONI	.1216 ±.0041	**.1213** ±.0034	.1134 ±.0033	0.7573 ±1.36	.1050 ±.0011	**.1072** ±.0016	.0975 ±.0008	7.7770 ±3.77	.1077 ±.0007	.1293 ±.0036	.1032 ±.0004	54 ±11	.1025 ±.0003	.1621 ±.0010	.0999 ±.0002	239 ±21
	OURS	.1198 ±.0037	.1234 ±.0031	.1118 ±.0030	0.0002	.1035 ±.0010	.1073 ±.0008	.0964 ±.0007	0.0010	.1038 ±.0007	**.1119** ±.0004	.1008 ±.0004	0.0068	.1015 ±.0003	**.1095** ±.0002	.0985 ±.0002	0.0253
	STOCHASTIC	—	.1746	—	0.0001	—	.1560	—	0.0005	—	.1613	—	0.0034	—	.1585	—	0.0126
cifar10	RIVASPLATA	.2992 ±.0012	.3038 ±.0016	.2871 ±.0009	3.2974 ±2.61	.2895 ±.0006	.3120 ±.0024	.2884 ±.0004	20 ±6	.2883 ±.0003	.3221 ±.0028	.2818 ±.0002	82 ±13	.3019 ±.0002	.3685 ±.0030	.2973 ±.0001	268 ±23
	BLANCHARD	.2992 ±.0012	.3058 ±.0014	.2871 ±.0009	3.1688 ±2.57	.2895 ±.0006	.3136 ±.0023	.2883 ±.0004	20 ±6	.2883 ±.0004	.3228 ±.0027	.2818 ±.0002	81 ±13	.3019 ±.0002	.3690 ±.0030	.2973 ±.0001	268 ±23
	CATONI	.2992 ±.0012	**.3027** ±.0013	.2871 ±.0009	3.3855 ±2.59	.2895 ±.0006	.3113 ±.0027	.2884 ±.0004	21 ±6	.2883 ±.0003	.3308 ±.0054	.2818 ±.0002	86 ±13	.3019 ±.0002	.4210 ±.0093	.2973 ±.0001	269 ±23
	OURS	.2992 ±.0012	.3061 ±.0009	.2870 ±.0009	0.2245	.2895 ±.0006	**.3074** ±.0004	.2884 ±.0004	0.0984	.2883 ±.0004	**.3007** ±.0002	.2818 ±.0002	0.0330	.3019 ±.0002	**.3164** ±.0001	.2973 ±.0001	0.0264
	STOCHASTIC	—	.3729	—	0.1123	—	.3744	—	0.0492	—	.3673	—	0.0165	—	.3838	—	0.0132

TABLE 0.1: Comparison of OURS, RIVASPLATA, BLANCHARD and CATONI based on the disintegrated bounds and STOCHASTIC based on the randomized bounds for 400 neural networks learned with two learning rates lr $\in\{10^{-4}, 10^{-6}\}$ and different Gaussian noises ($\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$). We report the test risk ($R_{\mathcal{T}}(h)$), the value of the bound (Bnd), the training risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for OURS, the KL-divergence for STOCHASTIC, and the disintegrated KL-divergence for RIVASPLATA, BLANCHARD and CATONI). The values in **bold** are the tightest bound.

Experiments[4] on neural networks

In this section, we do not seek state-of-the-art performance, in fact, we have a twofold objective: **(i)** We show that our disintegrated bound of Corollary 6 is tighter and more stable than the of Corollary 7 based on Rivasplata et al. (2020), Blanchard and Fleuret (2007) and Catoni (2007). **(ii)** we highlight that our disintegrated bound associated with the deterministic NN is tighter than the randomized bound associated with the stochastic NN (Corollary 8).

Training method. We follow our Training Method in which we integrate the direct minimization of the bounds. We refer as OURS the training method based on our bound of Corollary 6 (STOCHASTIC denotes the randomized associated bound), as RIVASPLATA the one based on Equation (0.5), as BLANCHARD the one based on Equation (0.6), and as CATONI the one based on Equation (0.7). First of all, we replace the non-differentiable 0-1 loss by a surrogate for which the bounds still hold: The bounded cross entropy loss (Dziugaite and Roy, 2018). This latter is defined in a multiclass setting with $y\in\{1,2,\ldots\}$ by $\ell(h,(\mathbf{x},y))=-\frac{1}{Z}\ln(\Phi(h[y]))\in[0,1]$ where $h[y]$ is the y-th output of the NN, and $\forall p\in[0,1]$, $\Phi(p)=e^{-Z}+(1-2e^{-Z})p$ (we set $Z=4$). That being said, to learn the best prior $\mathcal{P}\in\mathbf{P}$ and the posterior $\mathcal{Q}_{\mathcal{S}}$, we run our Training Method with two gradient descent-based algorithms A_{prior} and A. In phase **1)** algorithm A_{prior} learns from $\mathcal{S}_{\text{prior}}$ the T priors $\mathcal{P}_1,\ldots,\mathcal{P}_T\in\mathbf{P}$ (*i.e.*, during T epochs) by minimizing the bounded cross entropy loss. In other words, at the end of the epoch t, the weights $\mathbf{w}_t$ of the classifier are used to define the prior $\mathcal{P}_t=\mathcal{N}(\mathbf{w}_t,\sigma^2\mathbf{I}_d)$. Then, the best prior $\mathcal{P}\in\mathbf{P}$ is selected by early stopping on $\mathcal{S}$. In phase **2)**, given $\mathcal{S}$ and $\mathcal{P}$, algorithm A integrates the direct optimization of the bounds with the bounded cross entropy loss.

Optimization procedure.[5] Let $\boldsymbol{\omega}$ be the mean vector of a Gaussian distribution used as NN weights that we are optimizing. In phases **1)** and **2)**, at each iteration of the optimizer (*i.e.*, Adam in our case), we sample a noise $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)$. Then, we forward the examples of the mini-batch in the NN parametrized by the weights $\boldsymbol{\omega}+\boldsymbol{\epsilon}$ and we update $\boldsymbol{\omega}$ according to the bounded cross entropy loss. Note that during phase **1)**, at the end of each epoch t, $\mathcal{P}_t=\mathcal{N}(\boldsymbol{\omega},\sigma^2\mathbf{I}_d)=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_d)$ and finally at the end of phase **2)**, we have $\mathcal{Q}_{\mathcal{S}}=\mathcal{N}(\boldsymbol{\omega},\sigma^2\mathbf{I}_d)=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$.

Datasets and experimental setting.[6] We perform our experimental study on 3 datasets. As a sanity check, we

[4] The source code of our experiments will be publicly available in case of acceptance and is provided as supplementary material. We used the PyTorch framework (Paszke et al., 2019).

[5] The details of the optimization and the evaluation of the bounds are described in Appendix.

[6] The complete description of our overparametrized NNs and the hyper-parameters are provided in Appendix.

use the MNIST dataset (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017). We also consider a more complex dataset: CIFAR-10 (Krizhevsky, 2009). For the (Fashion-)MNIST datasets we train a variant of the All Convolutional Network (Springenberg et al., 2015), and for CIFAR we train a ResNet network (He et al., 2016). We divide each original train set into 2 independent subsets $\mathcal{S}_{\text{prior}}$ and $\mathcal{S}$ (with a split of 50%/50%); The test sets denoted $\mathcal{T}$ remain the original ones. For OURS, RIVASPLATA, BLANCHARD and CATONI, *i.e.*, for each disintegrated bound, we report in Table 0.1 the test risk $R_{\mathcal{T}}(h)$ and the bound values with the associated train risk $R_{\mathcal{S}}(h)$ and the divergence value (the Rényi divergence $D_2(\mathcal{Q}_{\mathcal{S}}\|\mathcal{P}){=}\frac{1}{\sigma^2}\|\mathbf{w}{-}\mathbf{v}_t\|_2^2$ for OURS and the disintegrated KL-divergence $\ln\frac{\mathcal{Q}_{\mathcal{S}}(h)}{\mathcal{P}(h)}{=}\frac{1}{2\sigma^2}[\|\mathbf{w}{+}\boldsymbol{\epsilon}{-}\mathbf{v}_t\|_2^2{-}\|\boldsymbol{\epsilon}\|_2^2]$ for the others, note the latter can be negative). The values are averaged over 400 deterministic NNs sampled from $\mathcal{Q}_{\mathcal{S}}$ for two learning rates (10^{-4} and 10^{-6}) and different values for $\sigma^2{\in}\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$; We fixed the set of hyperparameters for CATONI as $\mathbf{C}{=}\{10^k|k{\in}\{-3,-2,\ldots,+3\}\}$ (the other optimizer parameters are set to their default values). We additionally report as STOCHASTIC (Corollary 8) the randomized bound value and KL-divergence $\mathrm{KL}(\mathcal{Q}\|\mathcal{P}){=}\frac{1}{2\sigma^2}\|\mathbf{w}{-}\mathbf{v}_t\|_2^2$ associated with the model learned by OURS, meaning that $n{=}400$ and that the test risk reported for OURS also corresponds to the risk of the stochastic NN approximated with these 400 NNs.

Analysis of the results. First of all, OURS leads to more precise bounds than the randomized STOCHASTIC. This imprecision is due to the non-avoidable sampling according to $\mathcal{Q}$ done in the randomized PAC-Bayesian bound of Corollary 8 (the higher n, the tighter is the bound). Thus, using a disintegrated PAC-Bayesian bound avoids sampling a large number of NNs to obtain a low risk. This confirms that our framework makes sense.
Moreover, OURS leads in average to the tightest bound 20 times over 24. An important point to remark is that OURS behaves differently than RIVASPLATA, BLANCHARD and CATONI. Indeed, for both learning rates, when σ^2 decreases the value of our bound remains low, while the others increase drastically due to the explosion of the disintegrated KL-divergence term. Concretely, the disintegrated KL-divergence in Corollary 7 involves the noise $\boldsymbol{\epsilon}$ through $\|\mathbf{w}{+}\boldsymbol{\epsilon}{-}\mathbf{v}_t\|_2^2{-}\|\boldsymbol{\epsilon}\|_2^2$ (our divergence simply computes $\|\mathbf{w}{-}\mathbf{v}_t\|_2^2$). Then, the sampled noise during the optimization procedure $\boldsymbol{\epsilon}$ has an influence on the disintegrated KL-divergence making it to prone to high variations and thus to a certain level of instability. This makes the objective function to optimize (*i.e.*, the bound) subjects to high variations during the optimization, implying higher final bound values. Hence, the Rényi divergence has a certain advantage compared to the disintegrated KL-divergence since it does not depend on the sampled noise $\boldsymbol{\epsilon}$.

Toward information-theoretic bounds

Before concluding, we propose another interpretation of the disintegration procedure through Theorem 9 below. Actually, the Rényi divergence between $\mathcal{P}$ and $\mathcal{Q}$ is sensitive to the choice of the learning sample $\mathcal{S}$: When the posterior $\mathcal{Q}$ learned from $\mathcal{S}$ differs greatly from the prior $\mathcal{P}$ the divergence is high. To avoid such behavior, we consider the Sibson's mutual information (Verdú, 2015) which is a measure of dependence between the random variables $\mathcal{S}\in\mathcal{Z}^m$ and $h\in\mathcal{H}$. It involves an expectation over all the learning samples of a given size m and is defined for a given $\alpha{>}1$ by

$$I_\alpha(h;\mathcal{S}) \triangleq \min_{\mathcal{P}\in\mathcal{M}(\mathcal{H})} \frac{1}{\alpha-1}\ln\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\frac{\mathcal{Q}_{\mathcal{S}}(h)}{\mathcal{P}(h)}\right]^\alpha\right].$$

The higher $I_\alpha(h;\mathcal{S})$, the higher the correlation is, meaning that the sampling of h is highly dependent on the choice of $\mathcal{S}$. This measure has two interesting properties: It generalizes the mutual information (Verdú, 2015), and it can be related to the Rényi divergence. Indeed, let $\rho(h,\mathcal{S}){=}\mathcal{Q}_{\mathcal{S}}(h){\times}\mathcal{D}^m(\mathcal{S})$, *resp.* $\pi(h,\mathcal{S}){=}\mathcal{P}(h){\times}\mathcal{D}^m(\mathcal{S})$, be the probability of sampling both $\mathcal{S}{\sim}\mathcal{D}^m$ and $h{\sim}\mathcal{Q}_{\mathcal{S}}$, *resp.* $\mathcal{S}{\sim}\mathcal{D}^m$ and $h{\sim}\mathcal{P}$. Then we can write:

$$\begin{aligned} I_\alpha(h;\mathcal{S}) &= \min_{\mathcal{P}\in\mathcal{M}(\mathcal{H})} \frac{1}{\alpha-1}\ln\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\frac{\mathcal{Q}_{\mathcal{S}}(h)\mathcal{D}^m(\mathcal{S})}{\mathcal{P}(h)\mathcal{D}^m(\mathcal{S})}\right]^\alpha\right]\\ &= \min_{\mathcal{P}\in\mathcal{M}(\mathcal{H})} D_\alpha(\rho\|\pi). \end{aligned} \tag{0.10}$$

From Verdú (2015) the optimal prior $\mathcal{P}^*$ minimizing Equation (0.10) is a *distribution-dependent* prior:

$$\mathcal{P}^*(h) = \frac{\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathcal{Q}_{\mathcal{S}'}(h)^\alpha\right]^{\frac{1}{\alpha}}}{\mathbb{E}_{h'\sim\mathcal{P}}\,\frac{1}{\mathcal{P}(h')}\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}}.$$

This leads an *Information-Theoretic generalization bound*[7]:

Theorem 9 (Disintegrated Information-Theoretic Bound)**.** *For any $\mathcal{D}$ on $\mathcal{Z}$, for any $\mathcal{H}$, for any function $\phi:\mathcal{H}{\times}\mathcal{Z}^m{\to}\mathbb{R}^+$, for any $\alpha{>}1$, for any algorithm $A{:}\,\mathcal{Z}^m{\to}\mathcal{M}(\mathcal{H})$, we have*

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim\mathcal{Q}_{\mathcal{S}}}\Bigg(&\frac{\alpha}{\alpha-1}\ln(\phi(h,\mathcal{S})) \le I_\alpha(h';\mathcal{S}')\\ &+\ln\left[\tfrac{1}{\delta^{\frac{\alpha}{\alpha-1}}}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]\Bigg) \ge 1{-}\delta.\end{aligned}$$

Note that Esposito et al. (2020, Cor.4) introduced a bound based on the Sibson's mutual information, but, as discussed in the Appendix, Theorem 9 leads to a tighter bound. From a theoretical view, Theorem 9 brings a different philosophy than the disintegrated PAC-Bayes bounds. Indeed, in Theorems 2 and 4, given $\mathcal{S}$, the Rényi divergence $D_\alpha(\mathcal{Q}_{\mathcal{S}}\|\mathcal{P})$ suggests that the learned posterior $\mathcal{Q}_{\mathcal{S}}$ should be close enough to the prior $\mathcal{P}$ to get a low bound. While in Theorem 9, the Sibson's Mutual Information $I_\alpha(h';\mathcal{S}')$ suggests that the random variable h has to be *not too much correlated* to $\mathcal{S}$ on average. However, the bound of Theorem 9 is not computable in practice due notably to the sample expectation over the unknown distribution $\mathcal{D}$ in I_α. An exciting line of future works could be to study how we can make use of Theorem 9 in practice.

[7] We provide a mutual information-based bound in Appendix.

Conclusion

We provide new and general disintegrated PAC-Bayesian bounds providing some novel perspectives in the formalization of derandomized PAC-Bayesian bounds, allowing some parametrization and offering nice opportunities for the analysis and optimization of deterministic classifiers. We notably illustrate the interest of our theoretical result on neural networks. Future work includes specializations to specific architectures and models, and the use of the perspectives carried by Theorem 9. Lastly, about ethical aspects, while our work is mainly theoretical, its practical applications require to follow the required ethical principles related in the machine learning field.

References

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes Bounds. In *NIPS*, pages 9–16, 2006.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Bounds based on the Rényi Divergence. In *AISTATS*, pages 435–444, 2016.

Gilles Blanchard and François Fleuret. Occam's Hammer. In *COLT*, pages 112–126, 2007.

Olivier Bousquet and André Elisseeff. Stability and Generalization. *JMLR*, 2:499–526, 2002.

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

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

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *NeurIPS*, pages 8440–8450, 2018.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Robust Generalization via α-Mutual Information. In *International Zurich Seminar on Information and Communication*, 2020.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian Learning of Linear Classifiers. In *ICML*, pages 353–360, 2009.

Manuel Gil, Fady Alajaji, and Tamás Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Master's thesis, University of Toronto, 2009.

John Langford and John Shawe-Taylor. Pac-bayes & margins. In *NIPS*, 2002.

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

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks. In *NeurIPS*, pages 6869–6879, 2019.

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

David A. McAllester. Some PAC-Bayesian Theorems. In *COLT*, pages 230–234, 1998.

Vaishnavh Nagarajan and J. Zico Kolter. Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. In *ICLR*, 2019a.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *NeurIPS*, pages 11611–11622, 2019b.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. In *ICLR*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, pages 8024–8035, 2019.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *CoRR*, 2020.

David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian Processes by Minimizing PAC-Bayesian Generalization Bounds. In *NeurIPS*, 2018.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes Analysis Beyond the Usual Bounds. In *NeurIPS*, 2020.

Matthias Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

John Shawe-Taylor and Robert C. Williamson. A PAC Analysis of a Bayesian Estimator. In *COLT*, pages 2–9, 1997.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. In *ICLR*, 2015.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A Strongly Quasiconvex PAC-Bayesian Bound. In *ALT*, pages 466–492, 2017.

Leslie G. Valiant. A Theory of the Learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

Tim van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

Sergio Verdú. α-mutual information. In *Information Theory and Applications Workshop*, 2015.

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

Huan Xu and Shie Mannor. Robustness and generalization. *Mach. Learn.*, 86(3):391–423, 2012.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach. In *ICLR*, 2019.

A General Framework for the Disintegration of PAC-Bayesian Bounds

Appendix

This Appendix is structured as follows. We give the proof of Theorem 2, Corollary 3, Theorem 4, Proposition 5, Corollary 6, Corollary 7, and Corollary 8. We also discuss the minimization and the evaluation of the bounds introduced in the different corollaries. Additionally, we provide more details on our experimental setup. The last section is devoted to Theorem 9.

Proof of Theorem 2

Theorem 2. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any function $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}^+$, for any $\alpha>1$, for any deterministic algorithm $A : \mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\frac{\alpha}{\alpha-1}\ln\left(\phi(h,\mathcal{S})\right) \le \frac{2\alpha-1}{\alpha-1}\ln\frac{2}{\delta} + D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) + \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\right] \ge 1-\delta,$$

where $\mathcal{Q}_\mathcal{S}$ is output by A, i.e., $\mathcal{Q}_\mathcal{S}\triangleq A(\mathcal{S})$.

Proof. For any sample $\mathcal{S}$ and deterministic algorithm A fixed which allow us to obtain the distribution $\mathcal{Q}_\mathcal{S}$, note that $\phi(h,\mathcal{S})$ is a non-negative random variable. Hence, from Markov's inequality we have

$$\mathop{\mathbb{P}}_{h\sim\mathcal{Q}_\mathcal{S}}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2} \iff \mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\mathbf{I}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2}.$$

Taking the expectation over $\mathcal{S}\sim\mathcal{D}^m$ to both sides of the inequality gives

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\mathbf{I}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2} \iff \mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2}.$$

Taking the logarithm to both sides of the equality and multiplying by $\frac{\alpha}{\alpha-1}>0$, we obtain

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\frac{\alpha}{\alpha-1}\ln\left(\phi(h,\mathcal{S})\right) \le \frac{\alpha}{\alpha-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right)\right] \ge 1-\tfrac{\delta}{2}.$$

We develop the right side of the inequality in the indicator function and make the expectation of the hypothesis over the prior distribution $\mathcal{P}$ appears. Indeed, we have

$$\begin{aligned}\forall\mathcal{P}\in\mathcal{M}(\mathcal{H}),\ \frac{\alpha}{\alpha-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right) &= \frac{\alpha}{\alpha-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\frac{\mathcal{Q}_\mathcal{S}(h')\mathcal{P}(h')}{\mathcal{P}(h')\mathcal{Q}_\mathcal{S}(h')}\phi(h',\mathcal{S})\right)\\ &= \frac{\alpha}{\alpha-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\phi(h',\mathcal{S})\right).\end{aligned}$$

Remark that $\frac{1}{r}+\frac{1}{s}=1$ with $r=\alpha$ and $s=\frac{\alpha}{\alpha-1}$. Hence, we can apply Hölder's inequality:

$$\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\phi(h',\mathcal{S}) \le \left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\left(\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\right)^\alpha\right]^{\frac{1}{\alpha}}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}.$$

Then, by taking the logarithm; adding $\ln(\frac{2}{\delta})$ and multiplying by $\frac{\alpha}{\alpha-1}>0$ to both sides of the inequality, we obtain

$$\begin{aligned}\frac{\alpha}{\alpha-1}\ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\phi(h',\mathcal{S})\right) &\le \frac{\alpha}{\alpha-1}\ln\left(\frac{2}{\delta}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\left(\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\right)^\alpha\right]^{\frac{1}{\alpha}}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\ &= \frac{1}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\left[\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\right]^\alpha\right) + \frac{\alpha}{\alpha-1}\ln\frac{2}{\delta} + \ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\\ &= D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) + \frac{\alpha}{\alpha-1}\ln\frac{2}{\delta} + \ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right).\end{aligned}$$

From this inequality, we can deduce that

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\forall\mathcal{P}\in\mathcal{M}(\mathcal{H}),\ \frac{\alpha}{\alpha-1}\ln\left(\phi(h,\mathcal{S})\right) \le D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) + \frac{\alpha}{\alpha-1}\ln\frac{2}{\delta} + \ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\right] \ge 1-\tfrac{\delta}{2}. \quad (.11)$$

Note that $\mathbb{E}_{h'\sim\mathcal{P}}\,\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}$ is a non-negative random variable, hence, we apply Markov's inequality to have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right] \ge 1-\tfrac{\delta}{2}.$$

Since the inequality does not depend on the random variable $h\sim\mathcal{Q}_\mathcal{S}$, we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right] &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathbf{I}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\mathbf{I}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]\\
&= \mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right].
\end{aligned}$$

Taking the logarithm to both sides of the inequality and adding $\frac{\alpha}{\alpha-1}\ln\frac{2}{\delta}$ give us

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right] \ge 1-\tfrac{\delta}{2} \iff\\
&\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\frac{\alpha}{\alpha-1}\ln\frac{2}{\delta} + \ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right) \le \frac{2\alpha-1}{\alpha-1}\ln\frac{2}{\delta} + \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\right] \ge 1-\tfrac{\delta}{2}.
\end{aligned} \tag{.12}$$

Combining Equation (.11) and Equation (.12) with a union bound gives us the desired result. ☐

Proof of Corollary 3

Corollary 3 *Under the assumptions of Theorem 2, when $\alpha\to 1^+$, with probability at least $1-\delta$ we have*

$$\ln\phi(h,\mathcal{S}) \le \ln\frac{2}{\delta} + \ln\left[\mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z}, h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right].$$

When $\alpha\to+\infty$, with probability at least $1-\delta$ we have

$$\ln\phi(h,\mathcal{S}) \le \ln\mathop{\mathrm{esssup}}_{h'\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')} + \ln\Big[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\Big],$$

where esssup *is the essential supremum.*

Proof. Starting from Theorem 2 and rearranging, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\ln(\phi(h,\mathcal{S})) \le \frac{2\alpha-1}{\alpha}\ln\frac{2}{\delta} + \frac{\alpha-1}{\alpha}D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) + \ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\right] \ge 1-\delta.$$

Then, we will prove the case when $\alpha\to 1$ and $\alpha\to+\infty$ separately.

When $\alpha\to 1$. First, we have $\lim_{\alpha\to 1^+}\frac{2\alpha-1}{\alpha}\ln\frac{2}{\delta} = \ln\frac{2}{\delta}$ and $\lim_{\alpha\to 1^+}\frac{\alpha-1}{\alpha}D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) = 0$.
Furthermore, note that

$$\|\phi\|_{\frac{\alpha}{\alpha-1}} = \left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}|\phi(h',\mathcal{S}')|^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} = \left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$$

is the $L^{\frac{\alpha}{\alpha-1}}$-norm of the function $\phi:\mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}^+$, where $\lim_{\alpha\to 1}\|\phi\|_{\frac{\alpha}{\alpha-1}} = \lim_{\alpha'\to+\infty}\|\phi\|_{\alpha'}$ (since we have $\lim_{\alpha\to 1^+}\frac{\alpha}{\alpha-1} = (\lim_{\alpha\to 1}\alpha)(\lim_{\alpha\to 1}\frac{1}{\alpha-1}) = +\infty$). Then, it is well known that

$$\|\phi\|_\infty = \lim_{\alpha'\to+\infty}\|\phi\|_{\alpha'} = \mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z}, h'\in\mathcal{H}}\phi(h',\mathcal{S}').$$

Hence, we have

$$\begin{aligned}
\lim_{\alpha\to 1}\ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right) &= \ln\left(\lim_{\alpha\to 1}\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\
&= \ln\left(\lim_{\alpha\to 1}\|\phi\|_{\frac{\alpha}{\alpha-1}}\right) = \ln\left(\lim_{\alpha'\to+\infty}\|\phi\|_{\alpha'}\right)\\
&= \ln(\|\phi\|_\infty) = \ln\left(\mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z}, h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right).
\end{aligned}$$

Finally, we can deduce that

$$\lim_{\alpha\to 1}\left[\frac{2\alpha-1}{\alpha}\ln\frac{2}{\delta}+\frac{\alpha-1}{\alpha}D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\right]=\ln\frac{2}{\delta}+\ln\left[\mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right].$$

When $\alpha\to+\infty$**.** First, we have $\lim_{\alpha\to+\infty}\frac{2\alpha-1}{\alpha}\ln\frac{2}{\delta}=\ln\frac{2}{\delta}\left[2-\lim_{\alpha\to+\infty}\frac{1}{\alpha}\right]=2$ and $\lim_{\alpha\to+\infty}\|\phi\|_{\frac{\alpha}{\alpha-1}}=\lim_{\alpha'\to 1}\|\phi\|_{\alpha'}=\|\phi\|_1$ (since $\lim_{\alpha\to+\infty}\frac{\alpha}{\alpha-1}=\lim_{\alpha\to+\infty}\frac{1}{1-\frac{1}{\alpha}}=1$). Hence, we have

$$\begin{aligned}\lim_{\alpha\to+\infty}\ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)&=\ln\left(\lim_{\alpha\to+\infty}\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\&=\ln\left(\lim_{\alpha\to+\infty}\|\phi\|_{\frac{\alpha}{\alpha-1}}\right)=\ln\left(\lim_{\alpha'\to 1}\|\phi\|_{\alpha'}\right)\\&=\ln\left(\|\phi\|_1\right)=\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\right).\end{aligned}$$

Moreover, by rearranging the terms in $\frac{\alpha-1}{\alpha}D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$, we have

$$\frac{\alpha-1}{\alpha}D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})=\frac{1}{\alpha}\ln\left(\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right]^\alpha\right)=\ln\left(\left[\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right]^\alpha\right]^{\frac{1}{\alpha}}\right)=\ln\left(\left[\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\Delta(h)^\alpha\right]^{\frac{1}{\alpha}}\right)=\ln(\|\Delta\|_\alpha),$$

where $\|\Delta\|_\alpha$ is the L^α-norm of the function Δ defined as $\Delta(h)=\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}$. Hence, we have

$$\begin{aligned}\lim_{\alpha\to+\infty}\frac{\alpha-1}{\alpha}D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})&=\lim_{\alpha\to+\infty}\ln(\|\Delta\|_\alpha)=\ln\left(\lim_{\alpha\to+\infty}\|\Delta\|_\alpha\right)\\&=\ln\left(\|\Delta\|_\infty\right)=\ln\left(\mathop{\mathrm{esssup}}_{h\in\mathcal{H}}\Delta(h)\right)=\ln\left(\mathop{\mathrm{esssup}}_{h\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right).\end{aligned}$$

Finally, we can deduce that

$$\begin{aligned}&\lim_{\alpha\to+\infty}\left[\frac{2\alpha-1}{\alpha}\ln\frac{2}{\delta}+\frac{\alpha-1}{\alpha}D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\right]\\=\;&\ln\mathop{\mathrm{esssup}}_{h'\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}+\ln\left[\frac{4}{\delta^2}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\right].\end{aligned}$$

□

Proof of Theorem 4

For the sake of completeness, we first prove an upper bound on $\sqrt{ab}$ (see *e.g.*, Thiemann et al., 2017).

Lemma 10. *For any $a>0,b>0$, we have*

$$\sqrt{\tfrac{a}{b}}=\mathop{\mathrm{arginf}}_{\lambda>0}\left(\frac{a}{\lambda}+\lambda b\right),\quad 2\sqrt{ab}=\inf_{\lambda>0}\left(\frac{a}{\lambda}+\lambda b\right)\quad\textit{and}\quad\forall\lambda>0,\sqrt{ab}\le\frac{1}{2}\left(\frac{a}{\lambda}+\lambda b\right).$$

Proof. Let $f(\lambda)=\left(\frac{a}{\lambda}+\lambda b\right)$. The first and second derivatives of f *w.r.t.* λ is

$$\frac{\partial f}{\partial\lambda}(\lambda)=\left(b-\frac{a}{\lambda^2}\right)\quad\text{and}\quad\frac{\partial f}{\partial\lambda^2}(\lambda)=\frac{2a}{\lambda^3}.$$

Hence, for $\lambda>0$, we have $\frac{\partial f}{\partial\lambda^2}(\lambda)>0$; f is strictly convex and admit a unique minimum. Solving $\frac{\partial f}{\partial\lambda}(\lambda)=0$ we have $\lambda^*=\sqrt{\frac{a}{b}}$. Additionally, $f(\lambda^*)=2\sqrt{ab}$ which proves the claim. □

We can now prove Theorem 4 with Lemma 10.

Theorem 4. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any function $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}^+$, for any deterministic algorithm $A : \mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\forall\lambda>0, \ln(\phi(h,\mathcal{S})) \le \ln\!\left(\frac{1}{2}\left[\lambda e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})} + \frac{8}{\lambda\delta^3}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\right)\right] \ge 1-\delta,$$

where $\mathcal{Q}_\mathcal{S} \triangleq A(\mathcal{S})$.

Proof. The proof is similar to the one of Theorem 2. Given a sample $\mathcal{S}$ and a deterministic algorithm A (which allow us to obtain the distribution $\mathcal{Q}_\mathcal{S}$), the value $\phi(h,\mathcal{S})$ is a non-negative random variable. Hence, from Markov's inequality we have

$$\mathop{\mathbb{P}}_{h\sim\mathcal{Q}_\mathcal{S}}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2} \iff \mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\mathbf{I}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2}.$$

Taking the expectation over $\mathcal{S}\sim\mathcal{D}^m$ to both sides of the inequality gives

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\mathbf{I}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2} \iff \mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\phi(h,\mathcal{S}) \le \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right] \ge 1-\tfrac{\delta}{2}.$$

Using the fact that Lemma 10 with $a = \frac{4}{\delta^2}\phi(h',\mathcal{S})^2$ and $b = \frac{\mathcal{Q}_\mathcal{S}(h')^2}{\mathcal{P}(h')^2}$, we have

$$\forall\mathcal{P}\in\mathcal{M}(\mathcal{H}), \forall\lambda>0, \frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S}) = \mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\sqrt{\frac{\mathcal{Q}_\mathcal{S}(h')^2}{\mathcal{P}(h')^2}\frac{4}{\delta^2}\phi(h',\mathcal{S})^2} \le \frac{1}{2}\left[\lambda\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\left(\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\right)^2 + \frac{4}{\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2\right].$$

Then, taking the logarithm to both sides of the inequality, we obtain

$$\begin{aligned}\forall\lambda>0, \ln\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\phi(h',\mathcal{S})\right) &\le \ln\left(\frac{1}{2}\left[\lambda\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\left(\frac{\mathcal{Q}_\mathcal{S}(h')}{\mathcal{P}(h')}\right)^2 + \frac{4}{\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2\right]\right)\\ &= \ln\left(\frac{1}{2}\left[\lambda\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})) + \frac{4}{\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2\right]\right).\end{aligned}$$

Hence, we can deduce that

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\Bigg[&\forall\mathcal{P}\in\mathcal{M}(\mathcal{H}), \forall\lambda>0, \ln(\phi(h,\mathcal{S}))\\ &\le \ln\left(\frac{1}{2}\left[\lambda e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})} + \frac{4}{\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2\right]\right)\Bigg] \ge 1-\tfrac{\delta}{2}.\end{aligned} \tag{.13}$$

Note that $\mathbb{E}_{h'\sim\mathcal{P}}\,\phi(h',\mathcal{S})^2$ is a non-negative random variable, hence, we apply Markov's inequality:

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2 \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right] \ge 1-\tfrac{\delta}{2}.$$

Since the inequality does not depend on the random variable $h\sim\mathcal{Q}_\mathcal{S}$, we have

$$\begin{aligned}&\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2 \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\\ =&\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathbf{I}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2 \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\\ =&\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\mathbf{I}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2 \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\\ =&\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2 \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right].\end{aligned}$$

Additionally, note that multiplying by $\frac{4}{2\lambda\delta^2} > 0$, adding $\frac{\lambda}{2}\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))$, and taking the logarithm to both sides of the inequality results in the same indicator function. Indeed,

$$
\begin{aligned}
&\mathbf{I}\left[\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2 \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\\
&=\mathbf{I}\left[\forall\lambda>0,\, \tfrac{4}{2\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2 \le \tfrac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\\
&=\mathbf{I}\Bigg[\forall\lambda>0, \ln\!\left(\tfrac{\lambda}{2}\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))+\tfrac{4}{2\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2\right)\\
&\qquad\qquad\le \ln\!\left(\tfrac{\lambda}{2}\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))+\tfrac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right)\Bigg].
\end{aligned}
$$

Hence, we can deduce that

$$
\begin{aligned}
\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}&\Bigg[\forall\lambda>0,\ \ln\left(\frac{1}{2}\left[\lambda\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))+\frac{4}{\lambda\delta^2}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^2\right]\right)\\
&\le \ln\left(\frac{1}{2}\left[\lambda\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))+\frac{8}{\lambda\delta^3}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right]\right)\Bigg] \ge 1-\tfrac{\delta}{2}.
\end{aligned}
\tag{.14}
$$

Combining Equation (.13) and Equation (.14) with a union bound gives us the desired result. □

Proof of Proposition 5

Proposition 5. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any function $\phi:\mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}^+$ and for any deterministic algorithm $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, let*

$$
\lambda^* = \mathop{\mathrm{argmin}}_{\lambda>0}\ \ln\left(\frac{\lambda}{2}\exp[D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})]+\frac{8}{2\lambda\delta^3}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2\right),
$$

$$
\text{then}\quad \overbrace{2\ln\left[\frac{\lambda^*}{2}\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))+\mathop{\mathbb{E}}_{\substack{\mathcal{S}'\sim\mathcal{D}^m\\ h'\sim\mathcal{P}}}\frac{8\phi(h',\mathcal{S}')^2}{2\lambda^*\delta^3}\right]}^{\text{Theorem 4}} = \overbrace{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\ln\left[\frac{8}{\delta^3}\mathop{\mathbb{E}}_{\substack{\mathcal{S}'\sim\mathcal{D}^m\\ h'\sim\mathcal{P}}}\phi(h',\mathcal{S}')^2\right]}^{\text{Theorem 2 with } \alpha=2.},
$$

$$
\text{where}\quad \lambda^* = \sqrt{\frac{8\,\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2}{\delta^3\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))}}.
$$

Proof. We first prove the closed form of λ^*, then, we compute the difference between the two bounds. For the sake of clarity, let $d{=}\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))$, and $p=\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2$ and $f(\lambda)=\ln\!\big(\frac{\lambda}{2}d+\frac{8p}{2\lambda\delta^3}\big)$. We can find the first and second derivatives of f *w.r.t.* λ, indeed, we have

$$
\frac{\partial f}{\partial\lambda}(\lambda)=\frac{\delta^3\lambda^2 d-8p}{\delta^3\lambda^3 d+8\lambda p}\quad\text{and}\quad \frac{\partial f}{\partial\lambda^2}(\lambda)=\frac{-[\delta^3\lambda^2 d]^2+32p[\delta^3\lambda^2 d]+64p^2}{(\delta^3\lambda^3 d+8\lambda p)^2}.
$$

Furthermore, when $\frac{\partial f}{\partial\lambda}(\lambda){=}0$, we have $\lambda{=}\sqrt{\frac{8p}{\delta^3 d}}$, and we can prove that the associated minimum is unique (*i.e.*, $\lambda{=}\lambda^*$) by studying the function. Actually, we have $\frac{\partial f}{\partial\lambda^2}>0 \iff 64p^2{+}32px{-}x^2>0$ where $x=\delta^3\lambda^2 d$. Solving this polynomial of order 2 ($ax^2+bx+c>0$ with $a=-1, b=32p, c=64p^2$), we have that

$$
\tfrac{\partial f}{\partial\lambda^2}(\lambda)>0 \iff 0<\delta^3\lambda^2 d<8p(2+\sqrt{5}) \iff 0<\lambda<\sqrt{\tfrac{8p}{\delta^3 d}}\sqrt{2+\sqrt{5}}.
$$

Hence, for $\lambda\in\left]0,\sqrt{\frac{8p}{\delta^3 d}}\sqrt{2+\sqrt{5}}\right[$, f is strictly convex and λ^* is the unique minimizer on that interval. Furthermore, from the first derivative, we know that

$$
\frac{\partial f}{\partial\lambda}(\lambda)>0 \iff \lambda>\sqrt{\frac{8p}{d\delta^3}}.
$$

Hence, f is strictly increasing on $\left]\sqrt{\frac{8p}{d\delta^3}},+\infty\right[$, hence, $\lambda^*{=}\sqrt{\frac{8p}{d\delta^3}}$ is the unique minimizer on $]0,+\infty[$.

Lastly, substituting λ by λ^* in the right term of the inequality in Theorem 4 gives the equality. □

Proof of Corollary 6

We introduce Theorem 2' which takes into account a set of priors $\mathbf{P}$ while Theorem 2 handles a unique prior $\mathcal{P}$.

Theorem 2'. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any priors set $\mathbf{P}=\{\mathcal{P}_t\}_{t=1}^T$ of T prior $\mathcal{P}$ on $\mathcal{H}$, for any function $\phi:\mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}^+$, for any $\alpha>1$, for any deterministic algorithm $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_\mathcal{S}}\left[\begin{array}{l}\text{For all prior } \mathcal{P}_t\in\mathbf{P},\\ \frac{\alpha}{\alpha-1}\ln(\phi(h,\mathcal{S}))\le D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\frac{\alpha}{\alpha-1}\ln\frac{2}{\delta}+\ln\frac{2T}{\delta}+\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\end{array}\right]\ge 1-\delta,$$

Proof. The proof is essentially the same as Theorem 2. Indeed, we first derive the same equation as Equation (.11), we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_\mathcal{S}}\left[\forall\mathcal{P}\in\mathcal{M}(\mathcal{H}),\ \frac{\alpha}{\alpha-1}\ln(\phi(h,\mathcal{S}))\le D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\frac{\alpha}{\alpha-1}\ln\frac{2}{\delta}+\ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\right]\ge 1-\tfrac{\delta}{2}.$$

Then, we apply Markov's inequality (as in Theorem 2) T times with the T priors $\mathcal{P}_t$ belonging to $\mathbf{P}$, however, we set the confidence to $\frac{\delta}{2T}$ instead of $\frac{\delta}{2}$, we have

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_\mathcal{S}}\left[\ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}_t}\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\le\ln\frac{2T}{\delta}+\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}_t}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\right]\ge 1-\tfrac{\delta}{2T}.$$

Finally, combining the $T+1$ bounds with a union bound give us the desired result. □

We now prove Corollary 6 from Theorem 2'.

Corollary 6. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any priors set $\mathbf{P}=\{\mathcal{P}_t\}_{t=1}^T$ of T prior $\mathcal{P}$ on $\mathcal{H}$, for any algorithm $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, for any loss $\ell:\mathcal{H}\times\mathcal{Z}\to[0,1]$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_\mathcal{S}}\left[\forall\mathcal{P}_t\in\mathbf{P},\ \mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le\frac{1}{m}\left(\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2}+\ln\frac{16T\sqrt{m}}{\delta^3}\right)\right]\ge 1-\delta,$$

Proof. We instantiate Theorem 2' with $\phi(h,\mathcal{S})=\exp\left[\frac{\alpha-1}{\alpha}m\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\right]$ and $\alpha=2$: we have with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$, for all prior $\mathcal{P}_t\in\mathbf{P}$

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le\frac{1}{m}\left[D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}_t)+\ln\left(\frac{8T}{\delta^3}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}_t}e^{m\mathrm{kl}(R_{\mathcal{S}'}(h')\|R_\mathcal{D}(h'))}\right)\right].$$

From Maurer (2004) we upper-bound $\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}_t}e^{m\mathrm{kl}(R_{\mathcal{S}'}(h')\|R_\mathcal{D}(h'))}$ by $2\sqrt{m}$ for each prior $\mathcal{P}_t$. Hence, we have, for all prior $\mathcal{P}_t\in\mathbf{P}$

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le\frac{1}{m}\left[D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}_t)+\ln\left(\frac{16T\sqrt{m}}{\delta^3}\right)\right].$$

Additionally, the Rényi divergence $D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}_t)$ between two multivariate Gaussians $\mathcal{Q}_\mathcal{S}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$ and $\mathcal{P}_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_d)$ is well known: its closed-form solution is $D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}_t)=\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2}$ (see, for example, (Gil et al., 2013)). □

Proof of Corollary 7

We first prove the following Lemma in order to prove Corollary 7.

Lemma 11. *If $\mathcal{Q}_\mathcal{S}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$ and $\mathcal{P}=\mathcal{N}(\mathbf{v},\sigma^2\mathbf{I}_d)$, we have*

$$\ln\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}=\frac{1}{2\sigma^2}\left[\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2-\|\boldsymbol{\epsilon}\|_2^2\right],$$

where $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)$ is a Gaussian noise such that $\mathbf{w}+\boldsymbol{\epsilon}$ are the weights of $h\sim\mathcal{Q}_\mathcal{S}$ with $\mathcal{Q}_\mathcal{S}=\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$.

Proof. The probability density functions of $\mathcal{Q}_\mathcal{S}$ and $\mathcal{P}$ for $h\sim\mathcal{Q}_\mathcal{S}$ (with the weights $\mathbf{w}+\boldsymbol{\epsilon}$) can be rewritten as

$$\mathcal{Q}_\mathcal{S}(h)=\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^d\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|_2^2\right)\ \text{and}\ \mathcal{P}(h)=\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^d\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2\right).$$

We can derive a closed-form expression of $\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right]$. Indeed, we have

$$\begin{aligned}\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right]&=\ln\left[\mathcal{Q}_\mathcal{S}(h)\right]-\ln\left[\mathcal{P}(h)\right]\\&=\ln\left(\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^d\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|_2^2\right)\right)-\ln\left(\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^d\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2\right)\right)\\&=-\frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|_2^2+\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2=\frac{1}{2\sigma^2}\left[\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}\|_2^2-\|\boldsymbol{\epsilon}\|_2^2\right].\end{aligned}$$

□

We can now prove Corollary 7.

Corollary 7. *Under Corollary 6's assumptions, with probability at least* $1{-}\delta$ *over* $\mathcal{S}{\sim}\mathcal{D}^m$ *and* $h{\sim}\mathcal{Q}_\mathcal{S}$*,* we have $\forall \mathcal{P}_t{\in}\mathbf{P}$

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h)) \le \frac{1}{m}\left[\frac{\|\mathbf{w}{+}\boldsymbol{\epsilon}{-}\mathbf{v}_t\|_2^2{-}\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{2T\sqrt{m}}{\delta}\right],$$

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h)) \le \frac{1}{m}\left[\frac{m{+}1}{m}\frac{\|\mathbf{w}{+}\boldsymbol{\epsilon}{-}\mathbf{v}_t\|_2^2{-}\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{Tm{+}T}{\delta}\right],$$

$$\forall c{\in}\mathbf{C}, R_\mathcal{D}(h) \le \frac{1{-}e^{-cR_\mathcal{S}(h)-\frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{T|\mathbf{C}|}{\delta}\right]}}{1{-}e^{-c}},$$

where $\boldsymbol{\epsilon}{\sim}\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)$ *is a Gaussian noise s.t.* $\mathbf{w}{+}\boldsymbol{\epsilon}$ *are the weights of* $h{\sim}\mathcal{Q}_\mathcal{S}$ *with* $\mathcal{Q}_\mathcal{S}{=}\mathcal{N}(\mathbf{w},\sigma^2\mathbf{I}_d)$*, and* $\mathbf{C}$ *is a set of hyperparameters fixed a priori.*

Proof. We will prove separately the three bounds. The proofs share the same proof technique.

Equation (0.5). We instantiate Theorem 1*(i)* of Rivasplata et al. (2020) with $\phi(h,\mathcal{S}) = \exp[m\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))]$, however, we apply the theorem T times for each prior $\mathcal{P}_t \in \mathbf{P}$ (with a confidence $\frac{\delta}{T}$ instead of δ). Hence, for each prior $\mathcal{P}_t \in \mathbf{P}$, we have with probability at least $1 - \frac{\delta}{T}$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{Q}_\mathcal{S}$

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h)) \le \frac{1}{m}\left[\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}_t(h)}\right]+\ln\left(\frac{T}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}e^{m\mathrm{kl}(R_{\mathcal{S}'}(h')\|R_\mathcal{D}(h'))}\right)\right].$$

From Maurer (2004), we upper-bound $\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}_t}e^{m\mathrm{kl}(R_{\mathcal{S}'}(h')\|R_\mathcal{D}(h'))}$ by $2\sqrt{m}$ and using Lemma 11 we rewrite the disintegrated KL divergence. Finally, a union bound argument on the T bounds gives us the claim.

Equation (0.6). We apply T times Proposition 3.1 of Blanchard and Fleuret (2007) with a confidence $\frac{\delta}{T}$ instead of δ and with their parameter k defined as $k = m$. Hence, for each prior $\mathcal{P}_t \in \mathbf{P}$, we have with probability at least $1 - \frac{\delta}{T}$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{Q}_\mathcal{S}$

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h)) \le \frac{1}{m}\left[\left(1{+}\frac{1}{m}\right)\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}_t(h)}\right]+\ln\left(\frac{T(m{+}1)}{\delta}\right)\right].$$

From Lemma 11 and a union bound argument on the T bounds, we obtain the claim.

Equation (0.7). We apply $T|\mathbf{C}|$ times Theorem 1.2.7 of Catoni (2007) with a confidence $\frac{\delta}{T|\mathbf{C}|}$ instead of δ. For each prior $\mathcal{P}_t \in \mathbf{P}$ and hyperparameter $c \in \mathbf{C}$, we have with probability at least $1 - \frac{\delta}{T|\mathbf{C}|}$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{Q}_\mathcal{S}$

$$R_\mathcal{D}(h) \le \frac{1}{1{-}e^{-c}}\left[1{-}\exp\left(-cR_\mathcal{S}(h){-}\frac{1}{m}\left[\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}_t(h)}\right]+\ln\frac{T|\mathbf{C}|}{\delta}\right]\right)\right].$$

From Lemma 11 and a union bound argument on the T bounds, we obtain the claim. □

Proof of Corollary 8

Corollary 8. *For any distribution* $\mathcal{D}$ *on* $\mathcal{Z}$*, for any hypothesis set* $\mathcal{H}$*, for any priors' set* $\mathbf{P} = \{\mathcal{P}_t\}_{t=1}^T$ *of* T *priors* $\mathcal{P}$ *on* $\mathcal{H}$*, for any loss* $\ell{:}\mathcal{H}{\times}\mathcal{Z}{\to}\{0,1\}$ *we have, with probability at least* $1{-}\delta$ *over the random choice of* $\mathcal{S}{\sim}\mathcal{D}^m$ *and* $\{h_1,\dots,h_n\}{\sim}\mathcal{Q}^n$*, we have simultaneously*

$$\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n R_\mathcal{S}(h_i)\|\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}R_\mathcal{S}(h)\right) \le \frac{1}{n}\ln\frac{4}{\delta}, \qquad \text{(Equation (0.9))}$$

and, for all prior $\mathcal{P}_t{\in}\mathbf{P}$, for all posterior $\mathcal{Q}$,

$$\mathrm{kl}\left(\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}R_\mathcal{S}(h)\|\mathop{\mathbb{E}}_{h\sim\mathcal{Q}}R_\mathcal{D}(h)\right) \le \frac{1}{m}\left[\frac{\|\mathbf{w}{-}\mathbf{v}_t\|_2^2}{2\sigma^2}+\ln\frac{4T\sqrt{m}}{\delta}\right]. \qquad \text{(Equation (0.8))}$$

Proof. We instantiate Equation (0.3) (and apply Jensen's inequality on the right side of the inequation) for each prior $\mathcal{P}_t$ with $\mathcal{Q}{=}\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_d)$ and $\mathcal{P}_t{=}\mathcal{N}(\mathbf{v}_t, \sigma^2\mathbf{I}_d)$ with a confidence $\frac{\delta}{2T}$ instead of δ. Indeed, for each prior $\mathcal{P}_t$, with probability at least $1{-}\frac{\delta}{2T}$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$, we have for all posterior $\mathcal{Q}$ on $\mathcal{H}$,

$$\mathrm{kl}\left(\mathop{\mathbb{E}}_{h\sim\mathcal{Q}} R_{\mathcal{S}}(h) \| \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} R_{\mathcal{D}}(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\mathcal{Q}\|\mathcal{P}_t) + \ln\frac{4T\sqrt{m}}{\delta}\right].$$

Note that the closed form solution of the KL divergence between the Gaussian distributions $\mathcal{Q}$ and $\mathcal{P}_t$ is well known, we have $\mathrm{KL}(\mathcal{Q}\|\mathcal{P}_t){=}\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2}$. Then, by applying a union bound argument over the T bounds obtained with the T priors $\mathcal{P}_t$, we have with probability at least $1{-}\frac{\delta}{2}$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$, for all prior $\mathcal{P}_t \in \mathbf{P}$, for all posterior $\mathcal{Q}$

$$\mathrm{kl}\left(\mathop{\mathbb{E}}_{h\sim\mathcal{Q}} R_{\mathcal{S}}(h) \| \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} R_{\mathcal{D}}(h)\right) \leq \frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2} + \ln\frac{4T\sqrt{m}}{\delta}\right]. \quad \text{(Equation (0.8))}$$

Additionally, we obtained Equation (0.9) by a direct application the Theorem 2.2 of Dziugaite and Roy (2017) (with confidence $\frac{\delta}{2}$ instead of δ). Finally, from a union bound of the two bounds in Equations (0.9) and (0.8) gives the claimed result. □

Evaluation and Minimization of the Bounds in Corollaries 6, 7 and 8

We **optimize** and **evaluate** the bounds of the corollaries (except Eq. (0.7)) thanks to the inverse binary kl divergence defined as

$$\mathrm{kl}^{-1}(q|\psi) = \max\left\{p \in (0,1) \,\middle|\, \mathrm{kl}(q\|p) \leq \psi\right\},$$

where q is typically the empirical risk and ψ the PAC-Bayesian bound. Here, the function $\mathrm{kl}^{-1}(q|\psi)$ outputs the worst true risk p where the inequality $\mathrm{kl}(q\|p) \leq \psi$ holds. We can actually instantiate p, q and ψ for the different corollaries. Indeed, we have

$$R_{\mathcal{D}}(h) \leq \underbrace{\mathrm{kl}^{-1}\left(R_{\mathcal{S}}(h) \,\middle|\, \frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{\sigma^2} + \ln\frac{2T\sqrt{m}}{\delta^3}\right]\right)}_{\text{Corollary 6}},$$

$$R_{\mathcal{D}}(h) \leq \underbrace{\mathrm{kl}^{-1}\left(R_{\mathcal{S}}(h) \,\middle|\, \frac{1}{m}\left[\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln\frac{2T\sqrt{m}}{\delta}\right]\right)}_{\text{Equation (0.5)}},$$

$$R_{\mathcal{D}}(h) \leq \underbrace{\mathrm{kl}^{-1}\left(R_{\mathcal{S}}(h) \,\middle|\, \frac{1}{m}\left[\left(1+\frac{1}{m}\right)\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln\frac{T(m+1)}{\delta}\right]\right)}_{\text{Equation (0.6)}},$$

$$\text{and} \quad \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} R_{\mathcal{D}}(h) \leq \underbrace{\mathrm{kl}^{-1}\left(\mathrm{kl}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} R_{\mathcal{S}}(h_i) \,\middle|\, \frac{1}{n}\ln\frac{4}{\delta}\right) \,\middle|\, \frac{1}{m}\left[\frac{\|\mathbf{w}-\mathbf{v}_t\|_2^2}{2\sigma^2} + \ln\frac{4T\sqrt{m}}{\delta}\right]\right)}_{\text{Corollary 8}}.$$

Hence, kl^{-1} has to be evaluated in order to obtain the value of the upper-bound on $R_{\mathcal{D}}(h)$ or $\mathbb{E}_{h\sim\mathcal{Q}} R_{\mathcal{D}}(h)$: the evaluation of $\mathrm{kl}^{-1}(q|\psi)$ is performed by the bisection method. From these new formulation of the bounds, we can remark that the objective is to minimize the function $\mathrm{kl}^{-1}(q|\psi)$ in order to minimize the true risk p. To do so, Reeb et al. (2018) introduced an analytical expression of the derivative of kl^{-1} with respect to the empirical risk q and the PAC-Bayesian bound ψ. The two partial derivatives are defined in the following way:

$$\frac{\partial \mathrm{kl}^{-1}(q|\psi)}{\partial q} = \frac{\ln\frac{1-q}{1-\mathrm{kl}^{-1}(q|\psi)} - \ln\frac{q}{\mathrm{kl}^{-1}(q|\psi)}}{\frac{1-q}{1-\mathrm{kl}^{-1}(q|\psi)} - \frac{q}{\mathrm{kl}^{-1}(q|\psi)}}, \text{ and } \frac{\partial \mathrm{kl}^{-1}(q|\psi)}{\partial \psi} = \frac{1}{\frac{1-q}{1-\mathrm{kl}^{-1}(q|\psi)} - \frac{q}{\mathrm{kl}^{-1}(q|\psi)}}.$$

Note that these partial derivatives need the evaluation of $\mathrm{kl}^{-1}(q|\psi)$ for a given empirical risk q and a PAC-Bayesian bound ψ. Then, by computing the derivatives of q and ψ with respect to the parameters and by using the chain rule of differentiation, a library like PyTorch (see Paszke et al. (2019)) can compute automatically the derivatives of kl^{-1} with respect to the parameters.

At each iteration in phase **2)**, after sampling the noise $\boldsymbol{\epsilon}$, the algorithm update the weights $\boldsymbol{\omega}$ (*i.e.*, the hypothesis h) by optimizing

$$\underbrace{\mathrm{kl}^{-1}\!\left(R_{\mathcal{S}}(h)\,\middle|\,\frac{1}{m}\!\left[\frac{\|\boldsymbol{\omega}-\mathbf{v}_t\|_2^2}{\sigma^2}+\ln\frac{2T\sqrt{m}}{\delta^3}\right]\right)}_{\text{Objective function for Corollary 6}},$$

$$\underbrace{\mathrm{kl}^{-1}\!\left(R_{\mathcal{S}}(h)\,\middle|\,\frac{1}{m}\!\left[\frac{\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\frac{2T\sqrt{m}}{\delta}\right]\right)}_{\text{Objective function for Equation (0.5)}},$$

$$\underbrace{\mathrm{kl}^{-1}\!\left(R_{\mathcal{S}}(h)\,\middle|\,\frac{1}{m}\!\left[\left(1+\frac{1}{m}\right)\frac{\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}+\ln\tfrac{T(m+1)}{\delta}\right]\right)}_{\text{Objective function for Equation (0.6)}}.$$

with the bounded cross entropy loss $\ell(h,(\mathbf{x},y)){=}-\frac{1}{Z}\ln(\Phi(h[y]))$ (see Dziugaite and Roy (2018)).

To optimize Equation (0.7), we *(a)* initialize $c\in\mathbf{C}$ with the one that performs best on the first mini-batch and *(b)* optimize by gradient descent the parameter. Note that the loss is the bounded cross entropy loss $\ell(h,(\mathbf{x},y)){=}-\frac{1}{Z}\ln(\Phi(h[y]))$ of Dziugaite and Roy (2018) during the optimization. To evaluate Equation (0.7), we take $c\in\mathbf{C}$ that leads to the tightest bound.

Experiments

We give more details on the architectures and the hyperparameters that we consider in the experiments.

Training the MNIST-like datasets. We train a variant of the *All Convolutional Network* of Springenberg et al. (2015). The model is a 3-hidden layers convolutional network with 96 channels. We use 5×5 convolutions with a padding of size 1, and a stride of size 1 everywhere except on the second convolution where we use a stride of size 2. We adopt the Leaky ReLU activation functions after each convolution. Lastly, we use a global average pooling of size 8×8 in order to obtain the desired output size. Furthermore, the weights are initialized with Xavier Normal initializer while the biases are left initialized with the default initializer. We learn the prior weights by using Adam optimizer for 10 epochs with a learning rate of .001 and a batch size of 32. Moreover, the posterior weights are learned for one epoch with the same batch size and optimizer (except that the learning rate is different).

Training the CIFAR datasets. We train ResNet-20, *i.e.*, a ResNet network from He et al. (2016) with 20 layers. The weights are initialized with Kaiming Normal initializer and the initialization of the bias is the default one. The prior weights are learned for 100 epochs and the posterior weights for 10 epochs with a batch size of 32 by using Adam optimizer as well. Additionally, the learning rate to learn the prior for CIFAR-10 is .01.

About Theorem 9

This section is devoted to *(i)* the proof of a bound that is easier to interpret than Theorem 9, *(ii)* the proof of Theorem 9 and *(iii)* a discussion about Theorem 9.

A bound easier to interpret

Since the mutual information is well known, a bound based on this quantity will be more interpretable than the one with the Sibson's. Hence, we propose a mutual-information-based bound in Theorem 13. However, in order to prove this theorem, we need to prove Lemma 12.

Lemma 12. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any function $\phi:\mathcal{H}\times\mathcal{Z}^m\to[1,+\infty[$, for any deterministic algorithm $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim\mathcal{Q}_{\mathcal{S}}}\left[\forall\mathcal{P}{\in}\mathcal{M}(\mathcal{H}),\ \ln\phi(h{,}\mathcal{S})\leq\frac{1}{\delta}\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_{\mathcal{S}}\|\mathcal{P})+\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\phi(h,\mathcal{S})\right)\right]\right]\geq 1-\delta.$$

Proof. By developing $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathbb{E}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\phi(h,\mathcal{S})$, we have for all prior $\mathcal{P}\in\mathcal{M}(\mathcal{H})$

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\phi(h,\mathcal{S}) &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)\mathcal{P}(h)}{\mathcal{P}(h)\mathcal{Q}_\mathcal{S}(h)}\phi(h,\mathcal{S})\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right]+\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{P}(h)}{\mathcal{Q}_\mathcal{S}(h)}\phi(h,\mathcal{S})\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{P}(h)}{\mathcal{Q}_\mathcal{S}(h)}\phi(h,\mathcal{S})\right].\end{aligned}$$

From Jensen's inequality, we have for all prior $\mathcal{P}\in\mathcal{M}(\mathcal{H})$

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{P}(h)}{\mathcal{Q}_\mathcal{S}(h)}\phi(h,\mathcal{S})\right] &\le \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\ln\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\frac{\mathcal{P}(h)}{\mathcal{Q}_\mathcal{S}(h)}\phi(h,\mathcal{S})\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})+\ln\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\phi(h,\mathcal{S})\right]. \qquad (.15)\end{aligned}$$

Since we assume in this case that $\phi(h,\mathcal{S})\ge 1$ for all $h\in\mathcal{H}$ and $\mathcal{S}\in\mathcal{Z}^m$, we have $\ln\phi(h,\mathcal{S})\ge 0$; we can apply Markov's inequality to obtain

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\ln\phi(h,\mathcal{S})\le\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{Q}_\mathcal{S}}\ln\phi(h,\mathcal{S})\right]\ge 1-\delta. \qquad (.16)$$

Then, from Equations (.15) and (.16), we can deduce that stated result. □

We are now ready to prove Theorem 13.

Theorem 13. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any function $\phi:\mathcal{H}\times\mathcal{Z}^m\to[1,+\infty[$, for any deterministic algorithm $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{S}\sim D^m, h\sim\mathcal{Q}_\mathcal{S}}\left[\ln\phi(h,\mathcal{S})\le\frac{1}{\delta}\left[I(h;\mathcal{S})+\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}^*}\phi(h,\mathcal{S})\right)\right]\right]\ge 1-\delta,$$

where $\mathcal{P}^*$ is defined such that $\mathcal{P}^*(h)=\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)$ and $I(h;\mathcal{S})=\min_{\mathcal{P}\in\mathcal{M}(\mathcal{H})}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$.

Proof. Note that the mutual information is defined by $I(h;\mathcal{S})=\min_{\mathcal{P}\in\mathcal{M}(\mathcal{H})}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$. Hence, to prove Theorem 13, we have to instantiate Lemma 12 with the optimal prior, *i.e.*, the prior $\mathcal{P}$ which minimizes $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$. The optimal prior is well-known: for the sake of completeness, we derive it. First, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\\ &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)]}{\mathcal{P}(h)[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)]}\right]\\ &= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)}\right]+\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)}{\mathcal{P}(h)}\right].\end{aligned}$$

Hence,

$$\begin{aligned}\mathop{\mathrm{argmin}}_{\mathcal{P}\in\mathcal{M}(\mathcal{H})}\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) &= \mathop{\mathrm{argmin}}_{\mathcal{P}\in\mathcal{M}(\mathcal{H})}\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)}\right]+\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)}{\mathcal{P}(h)}\right]\right]\\ &= \mathop{\mathrm{argmin}}_{\mathcal{P}\in\mathcal{M}(\mathcal{H})}\left[\mathop{\mathbb{E}}_{h\sim\mathcal{Q}_\mathcal{S}}\ln\left[\frac{\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)}{\mathcal{P}(h)}\right]\right]=\mathcal{P}^*,\end{aligned}$$

where $\mathcal{P}^*(h)=\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h)$. Note that $\mathcal{P}^*$ is defined from the data distribution $\mathcal{D}$, hence, $\mathcal{P}^*$ is a valid prior when instantiating Lemma 12 with $\mathcal{P}^*$. Then, we have with *prob.* at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$

$$\begin{aligned}\ln\phi(h,\mathcal{S}) &\le \frac{1}{\delta}\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P}^*)+\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\phi(h,\mathcal{S})\right)\right]\\ &= \frac{1}{\delta}\left[I(h;\mathcal{S})+\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\phi(h,\mathcal{S})\right)\right].\end{aligned}$$

□

As you can remark, this bound is looser than Theorem 9 which is based on the Sibson's mutual information. For example, when we instantiate this bound with $\phi(h,\mathcal{S})=\exp[m\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))]$, the bound will be multiplied by $\frac{1}{\delta m}$, while the bound of Theorem 9 is only multiplied by $\frac{1}{m}$ (but we add the term $\frac{1}{m}\ln\frac{1}{\delta}$ to the bound which is small even for small m).

Proof of Theorem 9

We first introduce Lemma 14 in order to prove Theorem 9.

Lemma 14. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any prior distribution $\mathcal{P}$ on $\mathcal{H}$, for any function $\phi : \mathcal{H} \times \mathcal{Z}^m$, for any $\alpha > 1$, for any deterministic algorithm $A : \mathcal{Z}^m \to \mathcal{M}(\mathcal{H})$, we have*

$$\underset{\mathcal{S}\sim D^m, h\sim\mathcal{Q}_\mathcal{S}}{\mathbb{P}}\left[\begin{array}{l}\forall \mathcal{P}\in\mathcal{M}(\mathcal{H}),\\ \dfrac{\alpha}{\alpha-1}\ln(\phi(h,\mathcal{S})) \leq D_\alpha(\rho\|\pi) + \ln\left(\frac{1}{\delta^{\frac{\alpha}{\alpha-1}}} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\end{array}\right] \geq 1-\delta.$$

where $\rho(h,\mathcal{S})=\mathcal{Q}_\mathcal{S}(h)\times\mathcal{D}^m(\mathcal{S})$; $\pi(h,\mathcal{S})=\mathcal{P}(h)\times\mathcal{D}^m(\mathcal{S})$ and $\mathcal{Q}_\mathcal{S}\triangleq A(\mathcal{S})$.

Proof. For any deterministic algorithm A fixed which allow us to obtain the distribution $\mathcal{Q}_\mathcal{S}$ from a sample $\mathcal{S}$, note that $\phi(h,\mathcal{S})$ is a non-negative random variable. From Markov's inequality, we have

$$\underset{\mathcal{S}\sim D^m, h\sim\mathcal{Q}_\mathcal{S}}{\mathbb{P}}\left[\phi(h,\mathcal{S}) \leq \frac{1}{\delta} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{Q}_\mathcal{S}}{\mathbb{E}} \phi(h',\mathcal{S}')\right] \geq 1-\delta.$$

Taking the logarithm to both sides of the equality and multiplying by $\frac{\alpha}{\alpha-1} > 0$, we have

$$\underset{\mathcal{S}\sim \mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}{\mathbb{P}}\left[\frac{\alpha}{\alpha-1}\ln\left(\phi(h,\mathcal{S})\right) \leq \frac{\alpha}{\alpha-1}\ln\left(\frac{1}{\delta} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{Q}_{\mathcal{S}'}}{\mathbb{E}} \phi(h',\mathcal{S}')\right)\right] \geq 1-\delta.$$

We develop the right side of the inequality in the indicator function and make the expectation of the hypothesis over the distribution $\mathcal{P}$ appears: for all priors $\mathcal{P}\in\mathcal{M}(\mathcal{H})$,

$$\begin{aligned}\frac{\alpha}{\alpha-1}\ln\left(\frac{1}{\delta} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{Q}_{\mathcal{S}'}}{\mathbb{E}} \phi(h',\mathcal{S}')\right) &= \frac{\alpha}{\alpha-1}\ln\left(\frac{1}{\delta} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{Q}_{\mathcal{S}'}}{\mathbb{E}} \frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')}\frac{\mathcal{P}(h')}{\mathcal{Q}_{\mathcal{S}'}(h')}\phi(h',\mathcal{S}')\right)\\ &= \frac{\alpha}{\alpha-1}\ln\left(\frac{1}{\delta} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')}\phi(h',\mathcal{S}')\right).\end{aligned}$$

Then, remark that $\frac{1}{r}+\frac{1}{s}=1$ where $r{=}\alpha$ and $s{=}\frac{\alpha}{\alpha-1}$. Hence, Hölder's inequality gives

$$\underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{Q}_{\mathcal{S}'}}{\mathbb{E}} \phi(h',\mathcal{S}') \leq \left[\underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \left(\frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')}\right)^\alpha\right]^{\frac{1}{\alpha}} \left[\underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}.$$

Taking the log, adding $\ln(\frac{1}{\delta})$, and multiplying by $\frac{\alpha}{\alpha-1} > 0$ to both sides of the inequality, we have

$$\begin{aligned}&\frac{\alpha}{\alpha-1}\ln\left(\frac{1}{\delta} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{Q}_{\mathcal{S}'}}{\mathbb{E}} \phi(h',\mathcal{S}')\right)\\ \leq&\frac{\alpha}{\alpha-1}\ln\left(\frac{1}{\delta}\left[\underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \left(\frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')}\right)^\alpha\right]^{\frac{1}{\alpha}} \left[\underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\ =&\frac{1}{\alpha-1}\ln\left(\underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \left(\frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')}\right)^\alpha\right) + \ln\left(\frac{1}{\delta^{\frac{\alpha}{\alpha-1}}} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right).\end{aligned}$$

Hence, we can deduce that

$$\begin{aligned}\underset{\mathcal{S}\sim D^m, h\sim\mathcal{Q}_\mathcal{S}}{\mathbb{P}}\Bigg[\forall\mathcal{P}\in\mathcal{M}(\mathcal{H}), \frac{\alpha}{\alpha-1}\ln(\phi(h,\mathcal{S})) \leq \frac{1}{\alpha-1}\ln&\left(\underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \left[\frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')}\right]^\alpha\right)\\ &+\ln\left(\frac{1}{\delta^{\frac{\alpha}{\alpha-1}}} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\Bigg] \geq 1-\delta,\end{aligned}$$

where by definition we have $D_\alpha(\rho\|\pi) = \frac{1}{\alpha-1}\ln\left(\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\left[\frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')}\right]^\alpha\right)$. □

From Lemma 14, we prove Theorem 9.

Theorem 9. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any function $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}^+$, for any $\alpha{>}1$, for any deterministic algorithm $A : \mathcal{Z}^m \to \mathcal{M}(\mathcal{H})$, we have*

$$\underset{\substack{\mathcal{S}\sim D^m\\ h\sim\mathcal{Q}_\mathcal{S}}}{\mathbb{P}}\left[\frac{\alpha}{\alpha-1}\ln(\phi(h,\mathcal{S})) \leq I_\alpha(h';\mathcal{S}') + \ln\left(\frac{1}{\delta^{\frac{\alpha}{\alpha-1}}} \underset{\mathcal{S}'\sim\mathcal{D}^m}{\mathbb{E}} \underset{h'\sim\mathcal{P}}{\mathbb{E}} \phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\right] \geq 1-\delta,$$

where $\mathcal{Q}_\mathcal{S} \triangleq A(\mathcal{S})$.

Proof. Note that Sibson's mutual information is defined as $I_\alpha(h;\mathcal{S}) = \min_{\mathcal{P}\in\mathcal{M}(\mathcal{H})} D_\alpha(\rho\|\pi)$. Hence, in order to prove Theorem 9, we have to instantiate Lemma 14 with the optimal prior, *i.e.*, the prior $\mathcal{P}$ which minimizes $D_\alpha(\rho\|\pi)$. Actually, this optimal prior has a closed-form solution (Verdú, 2015). For the sake of completeness, we derive it. First, we have

$$
\begin{aligned}
&D_\alpha(\rho\|\pi)\\
=&\frac{1}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right]^\alpha\right)\\
=&\frac{1}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathcal{Q}_\mathcal{S}(h)^\alpha\right]\mathcal{P}(h)^{-\alpha}\right)\\
=&\frac{1}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathcal{Q}_\mathcal{S}(h)^\alpha\right]\mathcal{P}(h)^{-\alpha}\left[\frac{\mathbb{E}_{h'\sim\mathcal{P}}\frac{1}{\mathcal{P}(h')}\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}}{\mathbb{E}_{h'\sim\mathcal{P}}\frac{1}{\mathcal{P}(h')}\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}}\right]^\alpha\right)\\
=&\frac{\alpha}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{1}{\mathcal{P}(h')}\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}\right)+\frac{1}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\frac{1}{\mathcal{P}(h)^\alpha}\left[\frac{\left[\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathcal{Q}_\mathcal{S}(h)^\alpha\right]^{\frac{1}{\alpha}}}{\mathbb{E}_{h'\sim\mathcal{P}}\frac{1}{\mathcal{P}(h')}\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}}\right]^\alpha\right)\\
=&\frac{\alpha}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{1}{\mathcal{P}(h')}\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}\right)+D_\alpha(\mathcal{P}^*\|\mathcal{P}),
\end{aligned}
$$

where $\mathcal{P}^*(h)=\left[\frac{\left[\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathcal{Q}_\mathcal{S}(h)^\alpha\right]^{\frac{1}{\alpha}}}{\mathbb{E}_{h'\sim\mathcal{P}}\frac{1}{\mathcal{P}(h')}\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}}\right]$.
From these equalities and using the fact that $D_\alpha(\mathcal{P}^*\|\mathcal{P})$ is minimal (*i.e.*, equal to zero) when $\mathcal{P}^*=\mathcal{P}$, we can deduce that

$$
\begin{aligned}
\mathop{\mathrm{argmin}}_{\mathcal{P}\in\mathcal{M}(\mathcal{H})} D_\alpha(\rho\|\pi)&=\mathop{\mathrm{argmin}}_{\mathcal{P}\in\mathcal{M}(\mathcal{H})}\left[\frac{\alpha}{\alpha-1}\ln\left(\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\frac{1}{\mathcal{P}(h')}\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathcal{Q}_{\mathcal{S}'}(h')^\alpha\right]^{\frac{1}{\alpha}}\right)+D_\alpha(\mathcal{P}^*\|\mathcal{P})\right]\\
&=\mathop{\mathrm{argmin}}_{\mathcal{P}\in\mathcal{M}(\mathcal{H})} D_\alpha(\mathcal{P}^*\|\mathcal{P})=\mathcal{P}^*.
\end{aligned}
$$

Note that $\mathcal{P}^*$ is defined from the data distribution $\mathcal{D}$, hence, $\mathcal{P}^*$ is a valid prior when instantiating Lemma 14 with $\mathcal{P}^*$. Then, we have with *prob.* at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$

$$
\begin{aligned}
\frac{\alpha}{\alpha-1}\ln(\phi(h,\mathcal{S}))&\le D_\alpha(\rho\|\pi^*)+\ln\left(\frac{1}{\delta^{\frac{\alpha}{\alpha-1}}}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\\
&=I_\alpha(h';\mathcal{S}')+\ln\left(\frac{1}{\delta^{\frac{\alpha}{\alpha-1}}}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right).
\end{aligned}
$$

where $\pi^*(h,\mathcal{S})=\mathcal{P}^*(h)\times\mathcal{D}^m(\mathcal{S})$. □

About Theorem 9

For the sake of comparison, we introduce the following corollary of Theorem 9.

Corollary 15. *Under the assumptions of Theorem 9, when $\alpha\to1^+$, with probability at least $1-\delta$ we have*

$$\ln\phi(h,\mathcal{S})\le\ln\frac{1}{\delta}+\ln\left[\mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right].$$

When $\alpha\to+\infty$, with probability at least $1-\delta$ we have

$$\ln\phi(h,\mathcal{S})\le\ln\left(\mathop{\mathrm{esssup}}_{\mathcal{S}\in\mathcal{S},h\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}^*(h)}\right)+\ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\right],$$

where esssup *is the essential supremum.*

Proof. The proof is similar to Corollary 3. Starting from Theorem 9 and rearranging, we have

$$\mathop{\mathbb{P}}_{\substack{\mathcal{S}\sim D^m\\ h\sim\mathcal{Q}_\mathcal{S}}}\left[\ln(\phi(h,\mathcal{S}))\le\frac{\alpha-1}{\alpha}I_\alpha(h';\mathcal{S}')+\ln\frac{1}{\delta}+\ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\right]\ge1-\delta,$$

Then, we will prove separately the case when $\alpha\to1$ and $\alpha\to+\infty$.

When $\alpha \to 1$**.** First, we have $\lim_{\alpha\to1^+} \frac{\alpha-1}{\alpha} I_\alpha(h';\mathcal{S}') = 0$. Furthermore, note that

$$\|\phi\|_{\frac{\alpha}{\alpha-1}} = \left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*} |\phi(h',\mathcal{S}')|^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} = \left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*} \phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$$

is the $L^{\frac{\alpha}{\alpha-1}}$-norm of the function $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}^+$, where $\lim_{\alpha\to1}\|\phi\|_{\frac{\alpha}{\alpha-1}} = \lim_{\alpha'\to+\infty}\|\phi\|_{\alpha'}$ (since we have $\lim_{\alpha\to1^+}\frac{\alpha}{\alpha-1} = (\lim_{\alpha\to1}\alpha)(\lim_{\alpha\to1}\frac{1}{\alpha-1}) = +\infty$). Then, it is well known that

$$\|\phi\|_\infty = \lim_{\alpha'\to+\infty}\|\phi\|_{\alpha'} = \mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}} \phi(h',\mathcal{S}').$$

Hence, we have

$$\begin{aligned}
\lim_{\alpha\to1}\ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right) &= \ln\left(\lim_{\alpha\to1}\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\
&= \ln\left(\lim_{\alpha\to1}\|\phi\|_{\frac{\alpha}{\alpha-1}}\right) = \ln\left(\lim_{\alpha'\to+\infty}\|\phi\|_{\alpha'}\right)\\
&= \ln\left(\|\phi\|_\infty\right) = \ln\left(\mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right).
\end{aligned}$$

Finally, we can deduce that

$$\lim_{\alpha\to1}\left[\frac{\alpha-1}{\alpha}I_\alpha(h';\mathcal{S}') +\ \ln\frac{1}{\delta} + \ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\right] = \ln\frac{1}{\delta} + \ln\left[\mathop{\mathrm{esssup}}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right].$$

When $\alpha \to +\infty$**.** First, we have $\lim_{\alpha\to+\infty}\|\phi\|_{\frac{\alpha}{\alpha-1}} = \lim_{\alpha'\to1}\|\phi\|_{\alpha'} = \|\phi\|_1$ Hence, we have

$$\begin{aligned}
\lim_{\alpha\to+\infty}\ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right) &= \ln\left(\lim_{\alpha\to+\infty}\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\
&= \ln\left(\lim_{\alpha\to+\infty}\|\phi\|_{\frac{\alpha}{\alpha-1}}\right) = \ln\left(\lim_{\alpha'\to1}\|\phi\|_{\alpha'}\right)\\
&= \ln\left(\|\phi\|_1\right) = \ln\left(\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\right).
\end{aligned}$$

Moreover, by rearranging the terms in $\frac{\alpha-1}{\alpha}I_\alpha(h';\mathcal{S}')$, we have

$$\frac{\alpha-1}{\alpha}I_\alpha(h';\mathcal{S}') = \frac{1}{\alpha}\ln\left(\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}^*}\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}^*(h)}\right]^\alpha\right) = \ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}^*(h)}\right]^\alpha\right]^{\frac{1}{\alpha}}\right) = \ln\left(\left[\mathop{\mathbb{E}}_{h\sim\mathcal{P}}\Delta(h)^\alpha\right]^{\frac{1}{\alpha}}\right) = \ln(\|\Delta\|_\alpha),$$

where $\|\Delta\|_\alpha$ is the L^α-norm of the function Δ defined as $\Delta(h) = \frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}^*(h)}$. Hence, we have

$$\begin{aligned}
\lim_{\alpha\to+\infty}\frac{\alpha-1}{\alpha}I_\alpha(h';\mathcal{S}') &= \lim_{\alpha\to+\infty}\ln(\|\Delta\|_\alpha) = \ln\left(\lim_{\alpha\to+\infty}\|\Delta\|_\alpha\right)\\
&= \ln\left(\|\Delta\|_\infty\right) = \ln\left(\mathop{\mathrm{esssup}}_{\mathcal{S}\in\mathcal{S},h\in\mathcal{H}}\Delta(h)\right) = \ln\left(\mathop{\mathrm{esssup}}_{\mathcal{S}\in\mathcal{S},h\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}^*(h)}\right).
\end{aligned}$$

Finally, we can deduce that

$$\begin{aligned}
&\lim_{\alpha\to1}\left[\frac{\alpha-1}{\alpha}I_\alpha(h';\mathcal{S}') +\ \ln\frac{1}{\delta} + \ln\left(\left[\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}^*}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\right]\\
=\ &\ln\left(\mathop{\mathrm{esssup}}_{\mathcal{S}\in\mathcal{S},h\in\mathcal{H}}\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}^*(h)}\right) + \ln\left[\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\right].
\end{aligned}$$

□

As for Theorem 2, this corollary illustrate a trade-off introduced by α between the Sibson's mutual information $I_\alpha(h';\mathcal{S}')$ and the term $\ln\big(\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\mathcal{P}}\,\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\big)$.

Furthermore, Esposito et al. (2020, Cor.4) introduced a bound involving the Sibson's mutual information. Their bound holds with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$:

$$2(R_\mathcal{S}(h)-R_\mathcal{D}(h))^2 \le \frac{1}{m}\left[I_\alpha(h';\mathcal{S}') + \ln\frac{2}{\delta^{\frac{\alpha}{\alpha-1}}}\right]. \tag{.17}$$

Hence, we compare Equation .17 with the equations of the following corollary.

Corollary 16. *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any deterministic algorithm $A:\mathcal{Z}^m\to\mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$, we have*

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le\frac{1}{m}\left[I_\alpha(h';\mathcal{S}')+\ln\frac{2\sqrt{m}}{\delta^{\frac{\alpha}{\alpha-1}}}\right] \tag{.18}$$

$$\textit{and}\quad 2(R_\mathcal{S}(h)-R_\mathcal{D}(h))^2\le\frac{1}{m}\left[I_\alpha(h';\mathcal{S}')+\ln\frac{2\sqrt{m}}{\delta^{\frac{\alpha}{\alpha-1}}}\right]. \tag{.19}$$

Proof. First of all, we instantiate Theorem 9 with $\phi(h,\mathcal{S}) = \exp\big[\frac{\alpha-1}{\alpha}m\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\big]$, we have (by rearranging the terms)

$$\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\le\frac{1}{m}\left[I_\alpha(h';\mathcal{S}')+\ln\left(\frac{1}{\delta^{\frac{\alpha}{\alpha-1}}}\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h'\sim\mathcal{P}}e^{m\mathrm{kl}(R_{\mathcal{S}'}(h')\|R_\mathcal{D}(h'))}\right)\right].$$

Then, from Maurer (2004), we upper-bound $\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\,\mathbb{E}_{h'\sim\mathcal{P}}\,e^{m\mathrm{kl}(R_{\mathcal{S}'}(h')\|R_\mathcal{D}(h'))}$ by $2\sqrt{m}$ to obtain Equation (.18). Finally, to obtain Equation (.19), we apply Pinsker's inequality, *i.e.*, $2(R_\mathcal{S}(h)-R_\mathcal{D}(h))^2\le\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))$ on Equation (.18). ☐

Equation (.19) is slightly looser than Equation (.17) since it involves an extra term of $\frac{1}{m}\ln\sqrt{m}$. However, Equation (.18) is tighter than Equation (.17) when $\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))-2(R_\mathcal{S}(h)-R_\mathcal{D}(h))^2 \ge \frac{1}{m}\ln\sqrt{m}$ (which becomes more frequent as m grows).