



HAL
open science

A General Framework for the Practical Disintegration of PAC-Bayesian Bounds

Paul Viillard, Pascal Germain, Amaury Habrard, Emilie Morvant

► **To cite this version:**

Paul Viillard, Pascal Germain, Amaury Habrard, Emilie Morvant. A General Framework for the Practical Disintegration of PAC-Bayesian Bounds. Machine Learning, In press. hal-03143025v3

HAL Id: hal-03143025

<https://hal.science/hal-03143025v3>

Submitted on 18 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A General Framework for the Practical Disintegration of PAC-Bayesian Bounds

Paul Viillard^{1*†}, Pascal Germain², Amaury Habrard^{3,4}
and Emilie Morvant³

¹Inria, CNRS, Ecole Normale Supérieure, PSL Research University, Paris, France.

²Département d'informatique et de génie logiciel, Université Laval, Québec, Canada.

³Université Jean Monnet Saint-Étienne, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ÉTIENNE, FRANCE.

⁴Institut Universitaire de France (IUF).

*Corresponding author(s). E-mail(s): paul.viillard@inria.fr;

Contributing authors: pascal.germain@ift.ulaval.ca;

amaury.habrard@univ-st-etienne.fr;

emilie.morvant@univ-st-etienne.fr;

†This work was done when P. Viillard was affiliated to Laboratoire Hubert Curien

Abstract

PAC-Bayesian bounds are known to be tight and informative when studying the generalization ability of randomized classifiers. However, they require a loose and costly derandomization step when applied to some families of deterministic models such as neural networks. 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 this behavior on neural networks, and we show a significant practical improvement over the state-of-the-art framework.

Keywords: Disintegration, PAC-Bayesian, Generalization Bound, Neural Networks

1 Introduction

In statistical learning theory, PAC-Bayesian theory¹ (McAllester, 1998; Shawe-Taylor & Williamson, 1997) provides a powerful framework for analyzing the generalization ability of machine learning models such as linear classifiers (Germain, Lacasse, Lavolette, & Marchand, 2009), SVM (Ambroladze, Parrado-Hernández, & Shawe-Taylor, 2006), or neural networks (Dziugaite & Roy, 2017; Pérez-Ortiz, Rivasplata, Shawe-Taylor, & Szepesvári, 2021). In the PAC-Bayesian theory, the machine learning models are considered *randomized* (or *stochastic*), *i.e.*, a model is sampled from a *posterior* probability distribution for each prediction. The analysis of such a randomized classifier usually takes the form of bounds on the average risk with respect to a learned *posterior* distribution given a learning sample and a chosen *prior* distribution defined over a set of hypotheses. Note that the prior distribution can encode an *a priori* belief on the set of hypotheses, or if we have no belief, it can be set to a non-informative distribution, such as the uniform distribution. While such bounds are very effective for analyzing randomized/stochastic classifiers, the vast majority of machine learning methods nevertheless need guarantees on deterministic models. In this case, a *derandomization step* of the bound is required to get a bound on the risk of the deterministic model. In general, the *derandomization step* consists in obtaining a bound on the risk of a deterministic model from a bound that is originally for randomized/stochastic models. Different forms of derandomization have been introduced in the literature for specific settings. Among them, Langford and Shawe-Taylor (2002) proposed 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, Germain, Guedj, and Lavolette (2019) derived a PAC-Bayesian bound for a very specific deterministic network architecture using sign functions as activations; this approach has been further extended by Biggs and Guedj (2021, 2022). Another line of works derandomizes neural networks (Nagarajan & Kolter, 2019a; Neyshabur, Bhojanapalli, & Srebro, 2018). While technically different, it starts from PAC-Bayesian guarantees on the randomized classifier and uses an “output perturbation” bound to convert a guarantee from a random classifier to the mean classifier. These works highlight the need for 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 *disintegrated PAC-Bayesian bounds* upper-bound the *risk of a sampled* (unique) classifier from the posterior distribution. Despite

¹The reader can refer to Guedj (2019) or Alquier (2021) for recent surveys on PAC-Bayes.

their interest in derandomizing PAC-Bayesian bounds, these kinds of bounds have only received little study in the literature; especially, we can cite the recent work of Rivasplata, Kuzborskij, Szepesvári, and Shawe-Taylor (2020, Th.1(i)) who have derived a general disintegrated PAC-Bayesian theorem. It is important to note that these bounds have never been used in practice. Driven by machine learning practical purposes, our objective is thus twofold. We derive new tight and usable *disintegrated* PAC-Bayesian bounds (i) that directly derandomize any classifiers without any other additional step and with *almost* no impact on the guarantee, and (ii) that can be easily optimized to learn classifiers with strong guarantees. To achieve this objective, our contribution consists in providing a new general disintegration framework based on the Rényi divergence (in Theorem 2), allowing us to meet the practical goal of efficient learning. From the theoretical standpoint, due to the Rényi divergence term, our bound is expected to be looser than the one of Rivasplata et al. (2020, Th.1(i)) in which the divergence term is “disintegrated” but depends on the sampled hypothesis only. However, as we show in our experimental evaluation on neural networks, their “disintegrated” term is, in practice, subject to high variance, making their bound harder to optimize. This variance arises because the sampled hypothesis does not influence our Rényi divergence term. Our bound has then the main advantage of leading to a more stable learning algorithm with better empirical results. In addition, we derive a new theoretical result in the form of an information-theoretic bound, giving new insights into disintegration procedures.

The rest of the paper is organized as follows. Section 2 introduces the notations we follow and recalls some basics on generalization bounds. In Section 3, we derive our main contribution relying on *disintegrated* PAC-Bayesian bounds. Then, we illustrate the practical usefulness of this disintegration on deterministic neural networks in Section 5. Before concluding in Section 7, we discuss in Section 6 another point of view of the disintegrated through an information-theoretic bound. For readability, we deferred the proofs of our theoretical results to the Appendix.

2 Setting and basics

2.1 General notations

We denote by $\mathcal{M}(\mathcal{A})$ the set of probability densities on the measurable space $(\mathcal{A}, \Sigma_{\mathcal{A}})$ with respect to a reference measure² where $\Sigma_{\mathcal{A}}$ is the σ -algebra on the set \mathcal{A} . In this paper, we consider supervised classification tasks with \mathcal{X} the *input space*, \mathcal{Y} the *label set*, and $\mathcal{D} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ 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 an m -sample being $\mathcal{D}^m \in \mathcal{M}(\mathcal{Z}^m)$. We consider a *hypothesis set* \mathcal{H} of

²The measure considered for $(\mathcal{A}, \Sigma_{\mathcal{A}})$ is usually the Lebesgue or the counting measure.

functions $h: \mathcal{X} \rightarrow \mathcal{Y}$. 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* $\ell: \mathcal{H} \times \mathcal{Z} \rightarrow [0, 1]$ evaluating to which extent the prediction is accurate. Given a loss function ℓ , the *true risk* $R_{\mathcal{D}}(h)$ of a hypothesis h on the distribution \mathcal{D} and its empirical counterpart, the *empirical risk*, $R_{\mathcal{S}}(h)$ estimated on \mathcal{S} are defined as

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

Then, the learner wants to find the hypothesis h from \mathcal{H} that minimizes $R_{\mathcal{D}}(h)$. However, we cannot compute $R_{\mathcal{D}}(h)$ since \mathcal{D} is unknown. In practice, one could work under the Empirical Risk Minimization principle (ERM) 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 $R_{\mathcal{D}}(h)$. Statistical machine learning theory (see, *e.g.*, Vapnik, 2000) studies the conditions of consistency and convergence of ERM towards the true risk. This kind of result is called *generalization bound*, often referred to as PAC (Probably Approximately Correct) bound (Valiant, 1984), and takes the form:

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left[\left| R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h) \right| \leq \varepsilon \left(\frac{1}{\delta}, \frac{1}{m} \right) \right] \geq 1 - \delta.$$

Put into words, with high probability (at least $1 - \delta$) on the random choice of the learning sample \mathcal{S} , good generalization guarantees are obtained when the deviation between the true risk $R_{\mathcal{D}}(h)$ and its empirical estimate $R_{\mathcal{S}}(h)$ is low, *i.e.*, $\varepsilon(\frac{1}{\delta}, \frac{1}{m})$ should be as small as possible. The function ε depends mainly on two quantities: (i) the number of examples m for statistical precision, and (ii) the confidence parameter δ . We now recall three classical bounds while focusing on the PAC-Bayesian theory at the heart of our contribution. By abuse of notation, in the following, we use the function ε for the different presented frameworks: we consider an additional argument of ε to pinpoint the differences between the frameworks.

2.2 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} (such as the VC-dimension or the Rademacher complexity) and hold for all the hypotheses of \mathcal{H} . This type of bound takes the form:

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left[\sup_{h \in \mathcal{H}} |R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)| \leq \varepsilon \left(\frac{1}{\delta}, \frac{1}{m}, \mathcal{H} \right) \right] \geq 1 - \delta.$$

Due to $\sup_{h \in \mathcal{H}}$, this bound can be seen as a *worst-case* analysis. Indeed, it means that the bound $|R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)| \leq \varepsilon(\frac{1}{\delta}, \frac{1}{m}, \mathcal{H})$ holds with a high probability for all $h \in \mathcal{H}$, including the best but also the worst. This *worst-case* analysis makes it hard to obtain a non-vacuous bound *i.e.*, with $\varepsilon(\frac{1}{\delta}, \frac{1}{m}, \mathcal{H}) < 1$. Note that the ability of such bounds to explain the generalization of deep learning has been recently challenged (Nagarajan & Kolter, 2019b).

2.3 Algorithmic-dependent bounds

A potential drawback of the Uniform Convergence bounds is that they are independent of the learning algorithm, *i.e.*, they do not take into account the way the hypothesis space is explored. To tackle this issue, algorithmic-dependent bounds have been proposed to take advantage of some particularities of the learning algorithm, such as its uniform stability (Bousquet & Elisseeff, 2002) or robustness (Xu & Mannor, 2012). In this case, the bounds obtained hold for a single hypothesis $h_{L(\mathcal{S})}$, the one learned with the algorithm L from the learning sample \mathcal{S} . The form of such bounds is:

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left[|R_{\mathcal{D}}(h_{L(\mathcal{S})}) - R_{\mathcal{S}}(h_{L(\mathcal{S})})| \leq \varepsilon(\frac{1}{\delta}, \frac{1}{m}, L) \right] \geq 1 - \delta.$$

For example, this approach has been used by Hardt, Recht, and Singer (2016) to derive generalization bounds for hypotheses learned by stochastic gradient descent.

2.4 PAC-Bayesian bound

This paper leverages PAC-Bayesian bounds that stand in the PAC framework but borrows inspiration from the Bayesian probabilistic view that deals with randomness and uncertainty in machine learning (McAllester, 1998). In the PAC-Bayesian setting, we consider a *prior* distribution $\mathcal{P} \in \mathcal{M}^*(\mathcal{H}) \subseteq \mathcal{M}(\mathcal{H})$ on \mathcal{H} , with $\mathcal{M}^*(\mathcal{H})$ the set of strictly positive probability densities. This distribution encodes an *a priori* belief on \mathcal{H} before observing the learning sample \mathcal{S} . Then, given \mathcal{S} and the prior \mathcal{P} , we learn a *posterior* distribution $\mathcal{Q} \in \mathcal{M}(\mathcal{H})$. In this case, the bounds take the form:

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left[\forall \mathcal{Q} \in \mathcal{M}(\mathcal{H}), \quad \mathbb{E}_{h \sim \mathcal{Q}} |R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)| \leq \varepsilon(\frac{1}{\delta}, \frac{1}{m}, \mathcal{Q}) \right] \geq 1 - \delta.$$

A key notion is that the function $\varepsilon(\cdot)$ upper-bounds a \mathcal{Q} -*weighted expectation* over the risks of all classifiers in \mathcal{H} . Hence, it upper-bounds the risk of a *randomized classifier*.³ Such a randomized classifier can be described as follows: to predict the label of an input $\mathbf{x} \in \mathcal{X}$, (i) a hypothesis $h \in \mathcal{H}$ is sampled from \mathcal{Q} and (ii) the classifier predicts the label given by $h(\mathbf{x})$.

³The risk of the randomized classifier $\mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h)$ is sometimes referred to as the Gibbs risk in the PAC-Bayes literature.

We recall below the classical PAC-Bayesian bounds in a general form as proposed by [Bégin, Germain, Laviolette, and Roy \(2016\)](#); [Germain et al. \(2009\)](#). The idea is to express the bound in terms of a generic function $\phi: \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$ that is meant to capture the deviation between the true and the empirical risks, instead of deriving a theorem by settling on a specific measure of deviation such as $|R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)|$. Note that, [Theorem 1](#) is expressed in a slightly different form than the original ones; we prove [Theorem 1](#) in [Appendix A](#) for the sake of completeness.

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

$$\underbrace{s_{\mathcal{S} \sim \mathcal{D}^m}^{\mathbb{P}} \left(\forall Q \in \mathcal{M}(\mathcal{H}), \mathbb{E}_{h \sim Q} \ln(\phi(h, \mathcal{S})) \leq \text{KL}(Q \| \mathcal{P}) + \ln \left[\frac{1}{\delta} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S}) \right] \right)}_{(\text{Germain et al., 2009})} \geq 1 - \delta, \quad (1)$$

and

$$\underbrace{s_{\mathcal{S} \sim \mathcal{D}^m}^{\mathbb{P}} \left(\forall Q \in \mathcal{M}(\mathcal{H}), \frac{\alpha}{\alpha-1} \ln \left[\mathbb{E}_{h \sim Q} \phi(h, \mathcal{S}) \right] \leq D_{\alpha}(Q \| \mathcal{P}) + \ln \left[\frac{1}{\delta} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S})^{\frac{\alpha}{\alpha-1}} \right] \right)}_{(\text{Bégin et al., 2016})} \geq 1 - \delta, \quad (2)$$

with $\text{KL}(Q \| \mathcal{P}) \triangleq \mathbb{E}_{h \sim Q} \ln \frac{Q(h)}{\mathcal{P}(h)}$ the Kullback-Leibler (KL-)divergence between Q and \mathcal{P} , and $D_{\alpha}(Q \| \mathcal{P}) \triangleq \frac{1}{\alpha-1} \ln \left[\mathbb{E}_{h \sim \mathcal{P}} \left[\frac{Q(h)}{\mathcal{P}(h)} \right]^{\alpha} \right]$ the Rényi divergence between Q and \mathcal{P} ($\alpha > 1$).

Note that [Equation \(2\)](#) is more general than [Equation \(1\)](#). Indeed, the former is obtained from the latter by the three following steps: (i) substituting $\phi(h, \mathcal{S})$ by $\phi(h, \mathcal{S})^{\frac{\alpha-1}{\alpha}}$ in [Equation \(2\)](#), (ii) applying Jensen's inequality in order to move the expectation over Q in front of the logarithm, and (iii) taking the limit when α tends to 1. Note also the original bound statements of [Bégin et al. \(2016\)](#); [Germain et al. \(2009\)](#) are recovered by choosing a convex function $\Delta: [0, 1]^2 \rightarrow \mathbb{R}$ that captures a deviation between the true risk $R_{\mathcal{D}}(h)$ and the empirical risk $R_{\mathcal{S}}(h)$. Then, two steps are required: (i) setting $\phi(h, \mathcal{S}) = \exp(m\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h)))$ in [Equation \(1\)](#), or $\phi(h, \mathcal{S}) = \Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h))$ in [Equation \(2\)](#), and then (ii) applying Jensen's inequality on the left-hand side of the inequation. In fact, our proofs follow the exact same steps than those of [Germain et al. \(2009, Th.2.1\)](#) and [Bégin et al. \(2016, Th.9\)](#), but instead of starting from $\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h))$, we consider the slightly more general expression $\phi(h, \mathcal{S})$ from the beginning.⁴

⁴We refer the reader to the proof sketches given by [Figure 1](#) of [Bégin et al. \(2016\)](#) for more insights.

The advantage of Theorem 1 is that it can be used as a starting point for deriving different forms of bounds. For instance, for a loss function $\ell: \mathcal{H} \times \mathcal{Z} \rightarrow [0, 1]$ with $\phi(h, \mathcal{S}) = \exp(m \Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h)))$ and $\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h)) = 2[R_{\mathcal{S}}(h) - R_{\mathcal{D}}(h)]^2$ we retrieve from Equation (1) the bound proposed by McAllester (1998):

$$\begin{aligned} & \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left(\forall \mathcal{Q}, \left| \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h) - \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h) \right| \leq \sqrt{\frac{\text{KL}(\mathcal{Q} \parallel \mathcal{P}) + \ln \frac{2\sqrt{m}}{\delta}}{2m}} \right) \geq 1 - \delta \\ \implies & \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left(\forall \mathcal{Q}, \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h) \leq \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h) + \sqrt{\frac{\text{KL}(\mathcal{Q} \parallel \mathcal{P}) + \ln \frac{2\sqrt{m}}{\delta}}{2m}} \right) \geq 1 - \delta. \end{aligned}$$

This bound illustrates the trade-off between the average empirical risk and $\varepsilon(\frac{1}{\delta}, \frac{1}{m}, \mathcal{Q}) = \sqrt{\frac{1}{2m}(\text{KL}(\mathcal{Q} \parallel \mathcal{P}) + \ln \frac{2\sqrt{m}}{\delta})}$. More precisely, the higher m is, the lower $\varepsilon(\frac{1}{\delta}, \frac{1}{m}, \mathcal{Q})$ is therefore the smaller the difference between the true risk $\mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h)$ and the empirical risk $\mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h)$.

Another example leading to a slightly tighter but less interpretable bound is the Maurer (2004); Seeger (2002)'s bound that we retrieve with $\phi(h, \mathcal{S}) = \exp(m \Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h)))$ and $\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h)) = \text{kl}[R_{\mathcal{S}}(h) \parallel R_{\mathcal{D}}(h)]$:

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left(\forall \mathcal{Q}, \mathbb{E}_{h \sim \mathcal{Q}} \text{kl}(R_{\mathcal{S}}(h) \parallel R_{\mathcal{D}}(h)) \leq \frac{\text{KL}(\mathcal{Q} \parallel \mathcal{P}) + \ln \frac{2\sqrt{m}}{\delta}}{m} \right) \geq 1 - \delta, \quad (3)$$

where

$$\text{kl}(q \parallel p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \quad (4)$$

is the KL divergence between two Bernoulli distributions of parameters q and p . Such PAC-Bayesian bounds are known to be tight (*e.g.*, Pérez-Ortiz et al. (2021); Zantedeschi et al. (2021)), but they hold 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 PAC-Bayesian bounds to obtain a guarantee for a deterministic classifier instead of a randomized one (by removing the expectation on \mathcal{H}). In some cases, this derandomization results from the structure of the hypotheses, such as for randomized linear classifiers 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 the class of hypotheses, such as for neural networks (*e.g.*, Neyshabur et al. (2018), Nagarajan and Kolter (2019b, Ap. J), Biggs and Guedj (2022)). The next section states our main contribution, which is a general derandomization framework (based on the Rényi divergence) for disintegrating PAC-Bayesian bounds into a bound for a single hypothesis from \mathcal{H} .

3 Disintegrated PAC-Bayesian theorems

3.1 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:

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} \left(|R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)| \leq \varepsilon \left(\frac{1}{\delta}, \frac{1}{m}, \mathcal{Q}_{\mathcal{S}} \right) \right) \geq 1 - \delta, \quad (5)$$

where $\mathcal{Q}_{\mathcal{S}} \triangleq A(\mathcal{S}, \mathcal{P})$ with $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$ a *deterministic* algorithm chosen *a priori* which (i) takes a learning sample $\mathcal{S} \in \mathcal{Z}^m$ and a prior distribution \mathcal{P} as inputs, and (ii) outputs a *data-dependent* distribution $\mathcal{Q}_{\mathcal{S}} \triangleq A(\mathcal{S}, \mathcal{P})$ from the set $\mathcal{M}(\mathcal{H})$ of all possible probability densities on \mathcal{H} . Concretely, this kind of generalization bound allows one to derandomize the usual PAC-Bayes bounds as follows. Instead of considering a bound holding for all the posterior distributions on \mathcal{H} as usually done in PAC-Bayes (the “ $\forall \mathcal{Q}$ ” in Theorem 1), we consider only the posterior distribution $\mathcal{Q}_{\mathcal{S}}$ obtained through a deterministic algorithm A taking the learning sample \mathcal{S} and the prior \mathcal{P} as inputs. Then, the above bound holds for a unique hypothesis $h \sim \mathcal{Q}_{\mathcal{S}}$ instead of the 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 introduce in Theorems 2 and 4 below two new general disintegrated PAC-Bayesian bounds. A key asset of our results is that the bounds are instantiable to specific settings as for the “classical” PAC-Bayesian bounds (e.g., with *i.i.d.*/non-*i.i.d.* data, unbounded losses, etc.): to instantiate the bound, one has to instantiate the function ϕ . Note that, except our bound and the one of [Rivasplata et al. \(2020, Th.1\(i\)\)](#), the disintegrated bounds of the literature introduced by [Blanchard and Fleuret \(2007\)](#) and [Catoni \(2007, Th.1.2.7\)](#) do not depend on such a general function ϕ . With an appropriate instantiation, we obtain an easily optimizable bound, leading to a self-bounding⁵ algorithm ([Freund, 1998](#)) with theoretical guarantees. As an illustration of the usefulness of our results, we provide, in Section 4, such an instantiation for neural networks.

3.2 Disintegrated PAC-Bayesian bounds with the Rényi divergence

3.2.1 Our main contribution: a general disintegrated bound

In the same spirit as Equation (2) our main result stated in Theorem 2 is a general bound involving the Rényi divergence $D_{\alpha}(\mathcal{Q}_{\mathcal{S}} \parallel \mathcal{P})$ of order $\alpha > 1$.

⁵A self-bounding algorithm minimizes a generalization bound to obtain a model with a generalization guarantee.

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

$$\begin{aligned} \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} \left(\frac{\alpha}{\alpha-1} \ln(\phi(h, \mathcal{S})) \right) \\ \leq \frac{2\alpha-1}{\alpha-1} \ln \frac{2}{\delta} + D_{\alpha}(\mathcal{Q}_{\mathcal{S}} \parallel \mathcal{P}) + \ln \left[\mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', \mathcal{S}')^{\frac{\alpha}{\alpha-1}} \right) \right] \geq 1 - \delta, \end{aligned}$$

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

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

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

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

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

Lastly, we combine the two bounds with a union-bound argument. \square

As for the general classical PAC-Bayesian bounds (Theorem 1), the above theorem can be seen as the starting point of the derivation of generalization bounds depending on the choice of the function ϕ , as done in Corollary 6 in Section 4.1; this property makes it the main result of our paper.

In its proof, Hölder's inequality is used differently than in the classic PAC-Bayes bound's proofs. Indeed, in Bégin et al. (2016, Th. 8), the change of measure based on Hölder's inequality is key for deriving 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} and the prior \mathcal{P} . In fact, we use Hölder's inequality to introduce a prior \mathcal{P} independent from \mathcal{S} : a crucial point for our bound instantiated in Corollary 6.

Compared to Equation (2), our bound involves the term $\frac{2\alpha-1}{\alpha-1} \ln \frac{2}{\delta}$ instead of $\ln \frac{1}{\delta}$, that is an additional constant value of $\frac{2\alpha-1}{\alpha-1} \ln \frac{2}{\delta} - \ln \frac{1}{\delta} = \ln 2 + \frac{\alpha}{\alpha-1} \ln \frac{2}{\delta}$. When $\alpha = 2$, this constant equals $\ln \frac{8}{\delta^2}$, which turns out to be a reasonable cost to “derandomize” a bound into a disintegrated one, as typical choices for $\phi(h, \mathcal{S})$ will make the constant imprint on the bound value decay with m . This is similar to the bounds of Theorem 2 that tighten as m increases, provided that $\phi(h, \mathcal{S})$ is chosen wisely. For instance, by setting $\phi(h, \mathcal{S}) =$

$\exp(\frac{\alpha-1}{\alpha} m \text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)))$ with $\text{kl}(\cdot \| \cdot)$ defined by Equation (4), the bound depends on m and converges as m increases (see Section 4). Moreover, the tightness of the bound depends also on the deviation between $\mathcal{Q}_{\mathcal{S}}$ and \mathcal{P} , which makes the bound tighter when $\mathcal{Q}_{\mathcal{S}} = \mathcal{P}$.

We instantiate below Theorem 2 for $\alpha \rightarrow 1^+$ and $\alpha \rightarrow +\infty$ showing that the bound converges when $\alpha \rightarrow 1^+$ and $\alpha \rightarrow +\infty$.

Corollary 3 *Under the assumptions of Theorem 2, when $\alpha \rightarrow 1^+$, we have*

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} \left(\ln \phi(h, \mathcal{S}) \leq \ln \frac{2}{\delta} + \ln \left[\text{esssup}_{\mathcal{S}' \in \mathcal{Z}, h' \in \mathcal{H}} \phi(h', \mathcal{S}') \right] \right) \geq 1 - \delta,$$

when $\alpha \rightarrow +\infty$, we have

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} \left(\ln \phi(h, \mathcal{S}) \leq \ln \text{esssup}_{h' \in \mathcal{H}} \frac{\mathcal{Q}_{\mathcal{S}}(h')}{\mathcal{P}(h')} + \ln \left[\frac{4}{\delta^2} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \phi(h', \mathcal{S}') \right] \right) \geq 1 - \delta,$$

where esssup is the essential supremum defined as the supremum on a set with non-zero probability measures, i.e.,

$$\begin{aligned} \text{esssup}_{\mathcal{S}' \in \mathcal{Z}, h' \in \mathcal{H}} \phi(h', \mathcal{S}') &= \inf \left\{ \tau \in \mathbb{R}, \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} [\phi(h, \mathcal{S}) > \tau] = 0 \right\}, \\ \text{and } \text{esssup}_{h' \in \mathcal{H}} \frac{\mathcal{Q}_{\mathcal{S}}(h')}{\mathcal{P}(h')} &= \inf \left\{ \tau \in \mathbb{R}, \mathbb{P}_{h \sim \mathcal{Q}_{\mathcal{S}}} \left[\frac{\mathcal{Q}_{\mathcal{S}}(h)}{\mathcal{P}(h)} > \tau \right] = 0 \right\}. \end{aligned}$$

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

3.2.2 Comparison with the bound of Rivasplata et al. (2020)

For the sake of comparison, we recall in Equation (6) the bound proposed by Rivasplata et al. (2020, Th.1(i)), that is more general than the bounds of Blanchard and Fleuret (2007) and Catoni (2007, Th.1.2.7):

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} \left(\ln(\phi(h, \mathcal{S})) \leq \ln \frac{\mathcal{Q}_{\mathcal{S}}(h)}{\mathcal{P}(h)} + \ln \left(\frac{1}{\delta} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \phi(h', \mathcal{S}') \right) \right) \geq 1 - \delta. \quad (6)$$

The term $\ln \frac{\mathcal{Q}_{\mathcal{S}}(h)}{\mathcal{P}(h)}$ (also involved in Blanchard and Fleuret (2007); Catoni (2007)) can be seen as a “disintegrated⁶ KL divergence” depending only on

⁶We say that the KL divergence is “disintegrated” since the log term is not averaged in contrast to the KL divergence.

the sampled $h \sim \mathcal{Q}_S$. In contrast, our bound involves the Rényi divergence $D_\alpha(\mathcal{Q}_S \parallel \mathcal{P})$ between the prior \mathcal{P} and the posterior \mathcal{Q}_S , meaning our bound involves only one term that depends on the sampled hypothesis (the risk): the divergence value is the same whatever the hypothesis. Our bound is expected to be looser because of the Rényi divergence (see [van Erven & Harremoës, 2014](#)) and the dependence in δ (which is worse than Equation (6)). Nevertheless, our divergence term is the main advantage of our bound. Indeed, as confirmed by our experiments (Section 5), our bound with $D_\alpha(\mathcal{Q}_S \parallel \mathcal{P})$ makes the learning procedure (in our self-bounding algorithm) more stable and efficient compared to the optimization of Equation (6) with $\ln \frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)}$ that is subject to high variance.

3.2.3 A parameterizable general disintegrated bound

In the PAC-Bayesian literature, parametrized bounds have been introduced (*e.g.*, [Catoni \(2007\)](#); [Thiemann, Igel, Wintenberger, and Seldin \(2017\)](#)) to control the trade-off between the empirical risk and the divergence along with the additional term. For the sake of completeness, we now provide a parametrized version of our bound, enlarging its practical scope. We follow a similar approach to introduce a version of a disintegrated Rényi divergence-based bound that has the advantage of being parameterizable.

Theorem 4 (Parameterizable Disintegrated PAC-Bayes Bound) *For any distribution \mathcal{D} on \mathcal{Z} , for any hypothesis set \mathcal{H} , for any prior distribution $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$, for any measurable function $\phi: \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$, for any $\delta \in (0, 1]$, for any algorithm $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have*

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

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

Note that $e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})}$ is closely related to the χ^2 -distance. Indeed we have: $\chi^2(\mathcal{Q}_S \parallel \mathcal{P}) \triangleq \mathbb{E}_{h \sim \mathcal{P}} \left[\frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)} \right]^2 - 1 = e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})} - 1$. An asset of Theorem 4 is the parameter λ controlling the trade-off between the exponentiated Rényi divergence $e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})}$ 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 $\lambda > 0$, thus, from a practical view, we can learn/tune the parameter λ to minimize the bound and control the possible numerical instability due to $e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})}$. Indeed, if $D_2(\mathcal{Q}_S \parallel \mathcal{P})$ is large, the numerical computation can lead to an infinite value due to finite precision arithmetic. It is important to notice that, like other parametrized bounds (*e.g.*, [Thiemann et al., 2017](#)), there exists a closed-form solution of the optimal parameter λ (for a fixed \mathcal{P} and \mathcal{Q}_S); the solution is derived in Proposition 5 and shows that the optimal bound of Theorem 4 corresponds to 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 $\delta \in (0, 1]$, for any measurable function $\phi: \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$, for any algorithm $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, let*

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

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

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

Put into words: the optimal λ^* gives the same bound for Theorem 2 and Theorem 4.

4 The disintegration in action

So far, we have introduced theoretical results to derandomize PAC-Bayesian bounds through a disintegration approach. Indeed, the disintegration allows us to obtain a bound for a unique model sampled from the distribution \mathcal{Q}_S instead of having a bound on the averaged risk of the models. We propose in this section to illustrate the instantiation and the usefulness of Theorem 2 on neural networks compared to the classical PAC-Bayesian bounds.

4.1 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 to learn the weights of the NN leading to the lowest true risk. Practitioners usually proceed by epochs⁷ and obtain one “intermediate” NN after each epoch. Then, they select the “intermediate” NN associated with the lowest validation risk. We propose translating 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 $\forall t \in \{1, \dots, T\}$, $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$ is a Gaussian distribution centered at \mathbf{v}_t (the weights 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 $\mathcal{S}_{\text{prior}} \cap \mathcal{S} = \emptyset$, then Corollaries 6 and 7 will guide us to learn a posterior $\mathcal{Q}_S = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$ from a prior $\mathcal{P} \in \mathbf{P}$ minimizing the empirical risk on \mathcal{S} (we give more details on the procedure after the forthcoming corollaries). Note that considering Gaussian distributions has

⁷One epoch corresponds to one pass of the entire learning set during the optimization process.

the advantage of simplifying the expression of the KL divergence, and thus is commonly used in the PAC-Bayesian literature for neural networks (e.g., Dziugaite & Roy, 2017; Letarte et al., 2019; Zhou, Veitch, Austern, Adams, & Orbanz, 2019).⁸

Corollary 6 below instantiates Theorem 2 to this neural networks setting. Then, for the sake of comparison, Corollary 7 instantiates other disintegrated bounds from the literature; more precisely, Equation (7) corresponds to Rivasplata et al. (2020)’s bound of Equation (6), Equation (8) to Blanchard and Fleuret (2007)’s one, and Equation (9) to Catoni (2007)’s one.

Corollary 6 *For any distribution \mathcal{D} on \mathcal{Z} , for any hypothesis set \mathcal{H} , for any set $\mathbf{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ of T priors on \mathcal{H} where $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, 1]$, for any $\delta \in (0, 1]$, we have*

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

where $\text{kl}(a \| b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$, $\mathcal{Q}_{\mathcal{S}} = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$, and the hypothesis $h \sim \mathcal{Q}_{\mathcal{S}}$ is parametrized by $\mathbf{w} + \epsilon$.

Corollary 7 *For any distribution \mathcal{D} on \mathcal{Z} , for any set \mathcal{H} , for any set $\mathbf{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ of T priors on \mathcal{H} where $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \{0, 1\}$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the learning sample $\mathcal{S} \sim \mathcal{D}^m$ and the hypothesis $h \sim \mathcal{Q}_{\mathcal{S}}$ parametrized by $\mathbf{w} + \epsilon$, we have $\forall \mathcal{P}_t \in \mathbf{P}$*

$$\text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[\frac{\|\mathbf{w} + \epsilon - \mathbf{v}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \ln \frac{2T\sqrt{m}}{\delta} \right], \quad (7)$$

$$\forall b \in \mathbf{B}, \quad \text{kl}_+(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[\frac{b+1}{b} \left[\frac{\|\mathbf{w} + \epsilon - \mathbf{v}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} \right]_+ + \ln \frac{(b+1)T|\mathbf{B}|}{\delta} \right], \quad (8)$$

$$\forall c \in \mathbf{C}, \quad R_{\mathcal{D}}(h) \leq \frac{1 - \exp \left(-cR_{\mathcal{S}}(h) - \frac{1}{m} \left[\frac{\|\mathbf{w} + \epsilon - \mathbf{v}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \ln \frac{T|\mathbf{C}|}{\delta} \right] \right)}{1 - e^{-c}}, \quad (9)$$

with $[x]_+ = \max(x, 0)$, and $\text{kl}_+(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) = \text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h))$ if $R_{\mathcal{S}}(h) < R_{\mathcal{D}}(h)$ and 0 otherwise. Moreover, $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ is a Gaussian noise such that $\mathbf{w} + \epsilon$ 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}, \mathbf{B} are two sets of hyperparameters fixed a priori.

As the parameter λ of the Theorem 4, $c \in \mathbf{C}$ is a hyperparameter that controls a trade-off between the empirical risk $R_{\mathcal{S}}(h)$ and the term

⁸Gaussian distributions have been first studied in PAC-Bayes in the context of linear classifiers (e.g., Ambroladze et al., 2006; Germain, Habrard, Laviolette, & Morvant, 2020; Germain et al., 2009), but in this context, the symmetry of the Gaussian distribution also ease the derandomization.

$\frac{1}{m} \left[\frac{\|\mathbf{w} + \boldsymbol{\epsilon} - \mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln \frac{T|\mathbf{C}|}{\delta} \right]$. Besides, the parameter $b \in \mathbf{B}$ controls the tightness of the bound. In general, these parameters can be tuned to minimize the bound of Equation (8) and Equation (9); however, there is no closed-form solution for the expression of the minimum of this equation. In consequence, our experimental protocol requires minimizing the bounds by gradient descent for each $b \in \mathbf{B}$, respectively $c \in \mathbf{C}$, in order to learn the distribution \mathcal{Q}_S leading to the lowest bound value. To obtain a tight bound, the divergence between one prior $\mathcal{P}_t \in \mathbf{P}$ and \mathcal{Q}_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 small. 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 $\|\mathbf{w} - \mathbf{v}_t\|_2$.

Algorithm 1 Training Method

The original training set is split into two distinct subsets: $\mathcal{S}_{\text{prior}}$ and \mathcal{S} (respectively of size m_{prior} and m , that can be different).

The training has two phases.

- 1) The prior distribution \mathcal{P} is “pre-learned” with $\mathcal{S}_{\text{prior}}$ and selected by early stopping, with \mathcal{S} as validation set, using the algorithm A_{prior} (an arbitrary learning algorithm).
 - 2) Given \mathcal{S} and \mathcal{P} , we learn the posterior \mathcal{Q}_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 $\mathcal{S}_{\text{prior}}$, then **2)** optimizing the bound with the prior that leads to the best risk on \mathcal{S} . This gives a statistically valid result: since Corollary 6 is valid for every $\mathcal{P}_t \in \mathbf{P}$, we can select the one we want, in particular the one minimizing $R_S(h)$ for a sampled $h \sim \mathcal{P}_t$. This heuristic makes sense: it allows us to detect if a prior is concentrated around hypotheses that potentially overfit the learning sample $\mathcal{S}_{\text{prior}}$. Usually, practitioners consider this “best” prior as the final NN. In our case, the advantage is that we refine this “best” prior with \mathcal{S} to learn the posterior \mathcal{Q}_S . Note that Pérez-Ortiz et al. (2021) have already introduced tight generalization bounds with data-dependent priors for—non-derandomized—stochastic NNs.⁹ Nevertheless, the weights of the stochastic NNs are, by definition, sampled from the posterior distribution \mathcal{Q} for each prediction. In that sense, it is important to mention that stochastic NNs differ from *derandomized* NNs where only one model is sampled from \mathcal{Q}_S . Moreover, our training method to learn the prior differs greatly since **1)** we learn T NNs

⁹Stochastic NNs were introduced in the PAC-Bayesian literature by Langford and Caruana (2001).

(i.e., T priors) instead of only one, **2**) we fix the variance of the Gaussian in the posterior $\mathcal{Q}_{\mathcal{S}}$. Note that, as illustrated in Section 5, fixing the variance is not restrictive: the advantage is that it simplifies the expression of the KL divergence while keeping the bounds tight. To the best of our knowledge, our training method for the prior is new.

4.2 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); Langford and Caruana (2001); Pérez-Ortiz et al. (2021); Zhou et al. (2019)). In Corollary 8 below, we instantiate the bound of Equation (1) for stochastic NNs to empirically compare the stochastic and the deterministic NNs associated to the same prior 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 $\mathbf{w} + \epsilon \in \mathbb{R}^d$. 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} , (ii) then returning the label predicted by h . 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 obtain the following PAC-Bayesian bound.

Corollary 8 *For any distribution \mathcal{D} on \mathcal{Z} , for any \mathcal{H} , for any set $\mathbf{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ of T priors on \mathcal{H} where $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$, for any loss $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \{0, 1\}$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over $\mathcal{S} \sim \mathcal{D}^m$ and $\{h_1, \dots, h_n\} \sim \mathcal{Q}^n$, we have simultaneously $\forall \mathcal{P}_t \in \mathbf{P}$,*

$$\text{kl} \left(\mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h) \parallel \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h) \right) \leq \frac{1}{m} \left[\frac{\|\mathbf{w} - \mathbf{v}_t\|_2^2}{2\sigma^2} + \ln \frac{4T\sqrt{m}}{\delta} \right], \quad (10)$$

$$\text{and } \text{kl} \left(\frac{1}{n} \sum_{i=1}^n R_{\mathcal{S}}(h_i) \parallel \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h) \right) \leq \frac{1}{n} \ln \frac{4}{\delta}, \quad (11)$$

where $\mathcal{Q} = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$ and the hypothesis h sampled from \mathcal{Q} is parametrized by $\mathbf{w} + \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$.

This result shows two key features that allow considering it as an adapted baseline for a *fair* comparison between disintegrated and classical PAC-Bayesian 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. (2021, Sec. 6.2), since (i) we adapt the KL divergence to our setting (i.e., $\text{KL}(\mathcal{Q} \parallel \mathcal{P}) = \frac{1}{2\sigma^2} \|\mathbf{w} - \mathbf{v}_t\|_2^2$), (ii) the bound holds for T priors thanks to a union-bound argument.

5 Experiments with neural networks¹⁰

In this section, we do not seek state-of-the-art performance; in fact, we have a threefold objective: **(a)** we check if 50%/50% is a good choice for splitting the original train set into $(\mathcal{S}_{\text{prior}}, \mathcal{S})$ (which is the most common split in the PAC-Bayesian literature (Germain et al., 2009; Pérez-Ortiz et al., 2021)); **(b)** we highlight that our disintegrated bound associated with the deterministic NN is tighter than the randomized bound associated with the stochastic NN (Corollary 8); **(c)** we show that our disintegrated bound (Corollary 6) is tighter and more stable than the ones based on Rivasplata et al. (2020), Blanchard and Fleuret (2007) and Catoni (2007) (Corollary 7).

5.1 Training method

We follow our Training Method (Section 4.1) in which we integrate the direct minimization of all the bounds. We refer as *OURS* the training method based on the minimization of our bound in Corollary 6, as *RIVASPLATA* the one based on Equation (7), as *BLANCHARD* the one based on Equation (8), and as *CATONI* the one based on Equation (9). *STOCHASTIC* denotes the PAC-Bayesian bound with the prior and posterior distributions obtained from *OURS*. To optimize the bound with gradient descent, we replace the non-differentiable 0-1 loss with a surrogate: the bounded cross-entropy loss (Dziugaite & Roy, 2018). We made this replacement since cross-entropy minimization works well in practice for neural networks (Goodfellow, Bengio, & Courville, 2016) and because this loss is bounded between 0 and 1, which is required for the $\text{kl}()$ function. The cross-entropy is defined in a multiclass setting with $y \in \{1, 2, \dots\}$ by $\ell(h, (\mathbf{x}, y)) = -\frac{1}{Z} \ln(\Phi(h(\mathbf{x})[y])) \in [0, 1]$ where $h(\mathbf{x})[y]$ is the y -th output of the NN, and $\forall p \in [0, 1], \Phi(p) = e^{-Z} + (1 - 2e^{-Z})p$ (we set $Z=4$, the default parameter of Dziugaite and Roy (2018)). That being said, to learn a good enough prior $\mathcal{P} \in \mathbf{P}$ and the posterior $\mathcal{Q}_{\mathcal{S}}$, we run our Training Method with two stochastic gradient descent-based algorithms A_{prior} and A . Note that the randomness in the stochastic gradient descent algorithm is fixed to have deterministic algorithms. In phase **1**) algorithm A_{prior} learns from $\mathcal{S}_{\text{prior}}$ the T priors $\mathcal{P}_1, \dots, \mathcal{P}_T \in \mathbf{P}$ (*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.

¹⁰The source code of our experiments is available at <https://github.com/paulviallard/MLJ-Disintegrated-PB>. We used the PyTorch framework (Paszke et al., 2019).

5.2 Optimization procedure in algorithms A and A_{prior} ¹¹

Let ω be the mean vector of a Gaussian distribution used as NN weights that we are optimizing. In algorithms A and A_{prior} , we use the Adam optimizer (Kingma & Ba, 2015), and we sample a noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ at each iteration of the optimizer. Then, we forward the examples of the mini-batch to the NN parametrized by the weights $\omega + \epsilon$, and we update ω according to the bounded cross-entropy loss. Note that during phase **1**), at the end of each epoch t , $\mathcal{P}_t = \mathcal{N}(\omega, \sigma^2 \mathbf{I}_d) = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$ and finally at the end of phase **2**) we have $\mathcal{Q}_S = \mathcal{N}(\omega, \sigma^2 \mathbf{I}_d) = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$.

5.3 Experimental setting

5.3.1 Datasets

We perform our experimental study on three datasets: MNIST (LeCun, Cortes, & Burges, 1998), Fashion-MNIST (Xiao, Rasul, & Vollgraf, 2017), and CIFAR-10 (Krizhevsky, 2009). We divide each original train set into two independent subsets $\mathcal{S}_{\text{prior}}$ of size m_{prior} and \mathcal{S} of size m with varying split ratios defined as $\frac{m_{\text{prior}}}{m+m_{\text{prior}}} \in \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9\}$. The test sets denoted by \mathcal{T} remain the original ones.

5.3.2 Models

For the (Fashion-)MNIST datasets, we train a variant of the All Convolutional Network (Springenberg, Dosovitskiy, Brox, & Riedmiller, 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 to obtain the desired output size. Furthermore, the weights are initialized with Xavier Normal initializer (Glorot & Bengio, 2010) while each bias of size l is initialized uniformly between $-1/\sqrt{l}$ and $1/\sqrt{l}$.

For the CIFAR-10 dataset, we train a ResNet-20 network, *i.e.*, a ResNet network from He, Zhang, Ren, and Sun (2016) with 20 layers. The weights are initialized with Kaiming Normal initializer (He, Zhang, Ren, & Sun, 2015) and each bias of size l is initialized uniformly between $-1/\sqrt{l}$ and $1/\sqrt{l}$.

5.3.3 Optimization

For the (Fashion-)MNIST datasets, we learn the parameters of our prior distributions $\mathcal{P}_1, \dots, \mathcal{P}_T$ by using Adam optimizer for $T = 10$ epochs with a learning rate of 10^{-3} and a batch size of 32 (the other parameters of Adam are left by default). Moreover, the parameters of the posterior distribution \mathcal{Q}_S are learned for one epoch with the same batch size and optimizer (except that the learning rate is either 10^{-4} or 10^{-6}). For the CIFAR-10 dataset, the parameters of the priors $\mathcal{P}_1, \dots, \mathcal{P}_T$ are learned for $T = 100$ epochs, and the posterior

¹¹The details of the optimization and the evaluation of the bounds are described in Appendix I.

distribution \mathcal{Q}_S 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 10^{-2} .

5.3.4 Bounds

For BLANCHARD’s bounds, the set of hyperparameters is defined as $\mathbf{B}=\{b\in\mathbb{N} \mid b=\sqrt{x}, (x+1)\leq 2\sqrt{m}\}$, *i.e.*, such that BLANCHARD’s bounds can be tighter than RIVASPLATA’s ones. We fixed the set of hyperparameters for CATONI as $\mathbf{C}=\{10^k \mid k\in\{-3, -2, \dots, +3\}\}$. We try different values for $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ associated with the disintegrated KL divergence $\ln \frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)} = \frac{1}{2\sigma^2}(\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2)$, the “normal” Rényi divergence $D_2(\mathcal{Q}\|\mathcal{P})=\frac{1}{\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ and the KL divergence $\text{KL}(\mathcal{Q}\|\mathcal{P})=\frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$. For all the figures, the values are averaged over 400 deterministic NNs sampled from \mathcal{Q}_S (the standard deviation is small and presented in the Appendix K). We additionally report as STOCHASTIC (Corollary 8) the randomized bound value and KL divergence $\text{KL}(\mathcal{Q}\|\mathcal{P})=\frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ 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.

5.4 Results

5.4.1 Analysis of the influence of the split ratio between $\mathcal{S}_{\text{prior}}$ and \mathcal{S}

Figures 1 and 2 study the evolution of the bound values after optimizing the bounds with our Training Method for different parameters. Specifically, the split ratio of the original train set varies from 0.1 to 0.9 (0.1 means that $m_{\text{prior}} = 0.1(m + m_{\text{prior}})$), for four variances values σ^2 and the two learning rates (10^{-6} and 10^{-4}). For the sake of readability, we present detailed results when the split ratio is 0 in Table K1. We first remark that the behavior is different for the two learning rates. On the one hand, for $\text{lr}=10^{-6}$, the mean bound values are close to each other, which is not surprising since the disintegrated KL divergences and the Rényi divergences are close to zero (see Tables K2 to K10). Moreover, for MNIST and Fashion-MNIST, there is a trade-off between learning a good prior with $\mathcal{S}_{\text{prior}}$ and minimizing a generalization bound with \mathcal{S} . In this case, the split ratio 0.5 appears to be a good choice to obtain a tight (disintegrated) PAC-Bayesian bound. This ratio is widely used in the PAC-Bayesian literature (see, *e.g.*, in the context of linear classifiers (Germain et al., 2009), majority votes (Zantedeschi et al., 2021), and neural networks (Letarte et al., 2019; Pérez-Ortiz et al., 2021)). On the other hand, when $\text{lr}=10^{-4}$, the mean bound values tend to increase when the split ratio increases as well for the bounds introduced in the literature (*i.e.*, for BLANCHARD, CATONI, and RIVASPLATA), while the mean bound values of our bound remain low. Indeed, m decreases as long as the split ratio increases, which has the effect of increasing the bound value drastically when the disintegrated KL

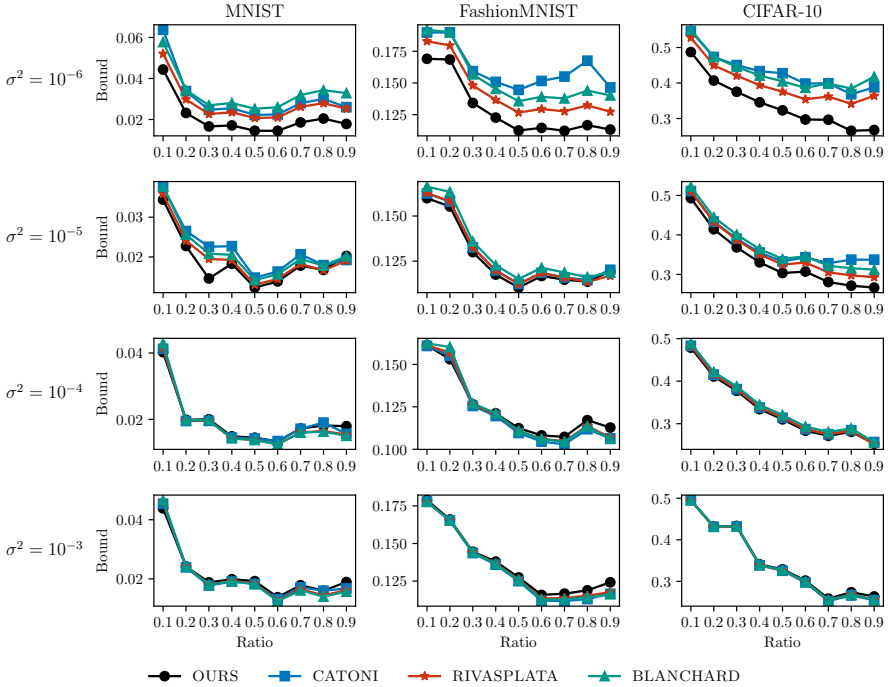


Fig. 1: Evolution of the bound values in terms of the split ratio. The x-axis represents the different split ratios, and the y-axis represents the bound values obtained after their optimization using our Training Method. Each row corresponds to a given variance σ^2 , and each column corresponds to a dataset (MNIST, Fashion-MNIST, or CIFAR-10). In this figure, we consider a learning rate of 10^{-6} .

divergence is high. We further explain why the disintegrated KL divergence can become high for the disintegrated bounds of the literature. To do so, we will now restrict our study to a split ratio of 0.5 in order to (i) compare the tightness of the bounds, (ii) understand why the disintegrated bounds of the literature diverge.

5.4.2 Comparison between disintegrated and “classic” bounds

We first compare the “classic” PAC-Bayesian bound (Corollary 8) and our disintegrated PAC-Bayesian bound (Corollary 6). To do so, we fix the variance $\sigma^2=10^{-3}$ (along with the split ratio equals 0.5). We report in Figure 3, the mean bound values associated with ours (i.e., the Training Method that minimizes our bound) and STOCHASTIC (we recall that STOCHASTIC is the PAC-Bayesian bound of Corollary 8 on the model learned by ours). Actually, ours leads to more precise bounds than the randomized STOCHASTIC even if the two empirical risks are the same and the KL divergence is smaller than the Rényi

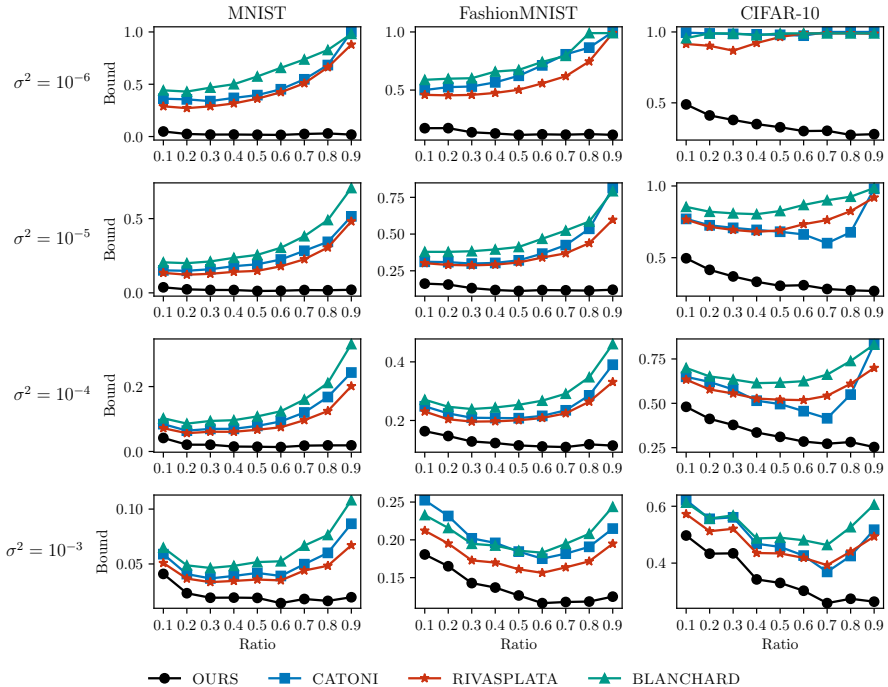


Fig. 2: Evolution of the bound values in terms of the split ratio. The x-axis represents the different split ratios, and the y-axis represents the bound values obtained after their optimization using our Training Method. Each row corresponds to a given variance σ^2 , and each column corresponds to a dataset (MNIST, Fashion-MNIST, or CIFAR-10). In this figure, we consider a learning rate of 10^{-4} .

one. 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 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 for practical purposes and has a great advantage in terms of time complexity when computing the bounds.

5.4.3 Analysis of the tightness of the disintegrated bounds

We now compare the tightness of the different disintegrated PAC-Bayesian bounds (*i.e.*, our bound and the ones in the literature). We study, as before, the case where the split ratio is 0.5 and the variance $\sigma^2 = 10^{-3}$. We report in Figure 4 for OURS, RIVASPLATA, BLANCHARD and CATONI, the mean bounds values; the mean test risk $R_{\mathcal{T}}(h)$ before (*i.e.*, with the prior \mathcal{P}) and after applying Step 2) (*i.e.*, with the posterior \mathcal{Q}_S). Moreover, we report above the bars the mean train risks $R_S(h)$ and the mean/standard deviation divergence values

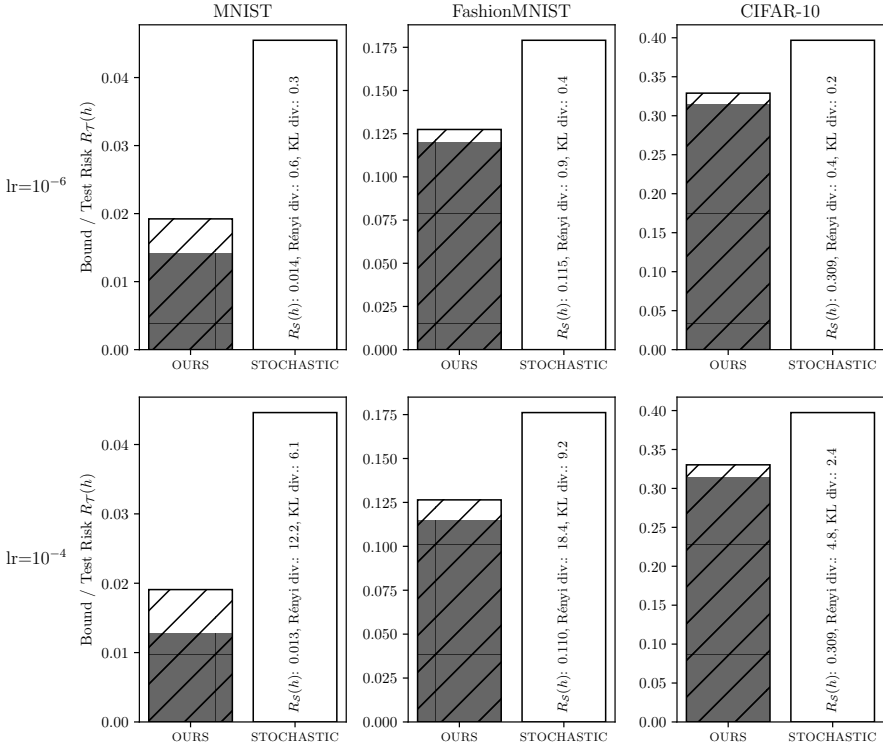


Fig. 3: The values of the PAC-Bayes bound (Corollary 8) and the values of the disintegrated bound (Corollary 6) for learning rates of 10^{-4} and 10^{-6} , and a split ratio is 0.5. The y-axis shows the values of the bounds (the hatched bar for ours (Corollary 6) and the white bar for STOCHASTIC (Corollary 8)) and the test risks $R_{\mathcal{T}}(h)$ (gray shaded bar). We also report the values of the empirical risk $R_{\mathcal{S}}(h)$, the Rényi divergence (associated with ours’ bound), and the KL divergence (associated with STOCHASTIC’s bound).

obtained after Step 2), *i.e.*, the Rényi divergence $D_2(Q_{\mathcal{S}}\|\mathcal{P}) = \frac{1}{\sigma^2} \|\mathbf{w} - \mathbf{v}_t\|_2^2$ for ours and the disintegrated KL divergence $\ln \frac{Q_{\mathcal{S}}(h)}{\mathcal{P}(h)} = \frac{1}{2\sigma^2} [\|\mathbf{w} + \boldsymbol{\epsilon} - \mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2]$ for the others. First of all, we can remark that we observe two different behaviors for $\text{lr}=10^{-4}$ and $\text{lr}=10^{-6}$. For $\text{lr}=10^{-6}$, the bound values are close to each other, as well as the empirical risks and the divergences (which are close to 0). In Figure 4, we observe that the bound values and the test risks are close to the one associated with the prior distribution because the divergence is close to 0. This is probably due to the fact that the learning rate is too small, implying that the bounds are not optimized. With a higher learning rate of $\text{lr}=10^{-4}$, we observe that our bound remains tight while the disintegrated bounds of the literature are looser. Hopefully, our bound is improved after performing Step 2) of our Training Method. However, for the bounds of the literature, the

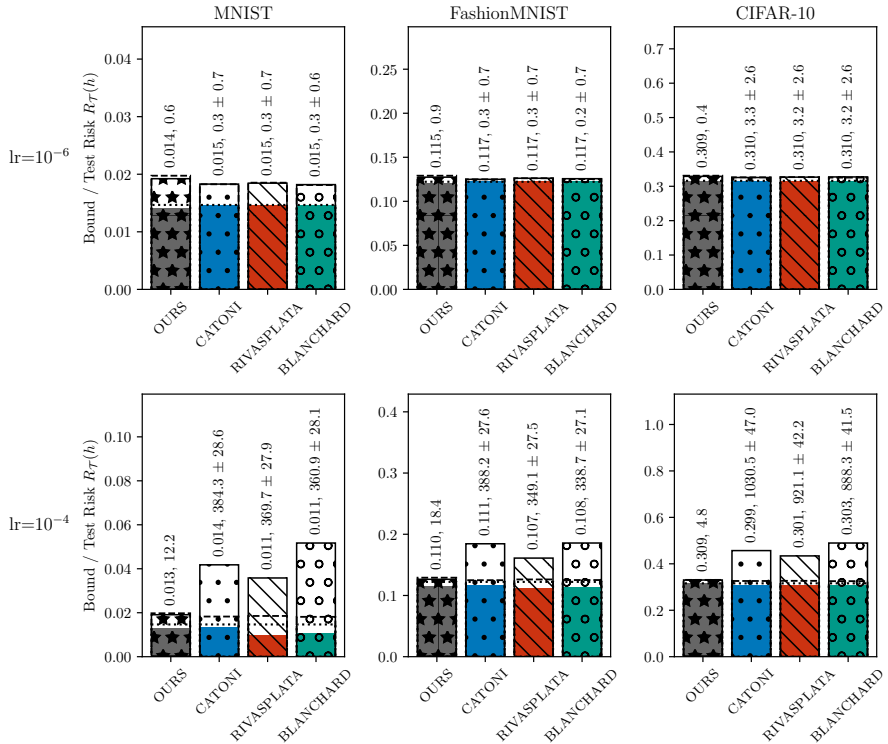


Fig. 4: The value of the bounds (hatched bars) and the test risks (colored bars) for Corollary 6 (“ours”) and Corollary 7 (“CATONI”, “RIVASPLATA” and “BLANCHARD”) in two different settings, *i.e.*, with a learning rate of 10^{-6} and 10^{-4} and with split ratio of 0.5. We also plot the value of the bounds (the dashed lines) and the test risks (the dotted lines) before executing Step 2) of our Training Method. The y-axis shows the values of the bounds and the test risks $R_T(h)$. The empirical risk $R_S(h)$ is presented above each bar. Moreover, the second value represents the mean value of the divergence (the standard deviations are also given for the disintegrated bounds of the literature).

value of the disintegrated KL divergence is large, making the bounds looser after executing Step 2). We now investigate the reasons for the divergence of the bounds by looking at the influence of the variance σ^2 .

5.4.4 Analysis of the influence of the variance

Given a split ratio of 0.5 and $\text{lr} \in \{10^{-6}, 10^{-4}\}$, we report in Figure 5 the evolution of the bound values associated with Ours, RIVASPLATA, BLANCHARD, and CATONI when the variance varies from 10^{-6} to 10^{-3} . First of all, the important point is that Ours behaves differently than RIVASPLATA, BLANCHARD, and CATONI. Indeed, for both learning rates, when σ^2 decreases, the value of our

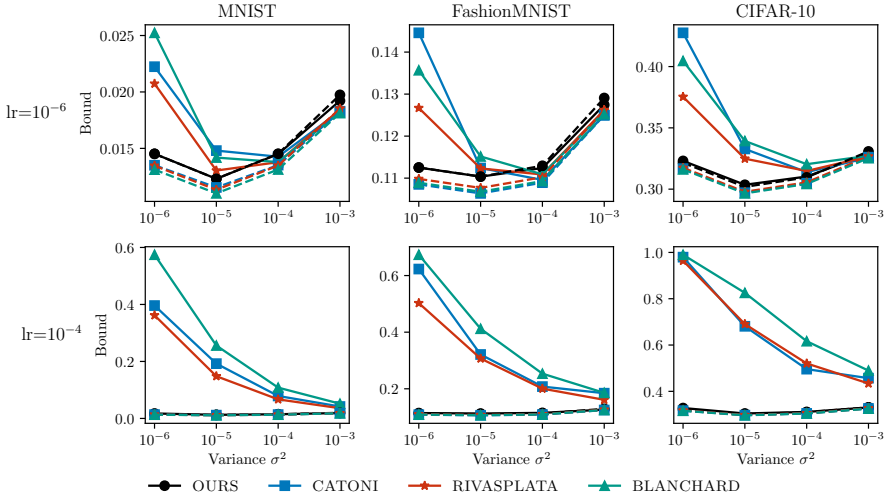


Fig. 5: We plot the evolution of the mean bound values (the plain lines) in terms of the variance σ^2 after optimizing the bounds with our Training Method. Moreover, we plot the mean bound values (the dashed lines) obtained before executing the Step 2) of our Training Method. The variance is represented on the x-axis, while the bound values are represented on the y-axis. Furthermore, each row corresponds to a given learning rate (10^{-6} or 10^{-4}), and each column corresponds to a dataset (either MNIST, FashionMNIST, or CIFAR-10). The split ratio considered is 0.5.

bound remains low, while the others increase drastically due to the explosion of the disintegrated KL divergence term (see Table K6 in Appendix K for more details). Concretely, the disintegrated KL divergence in Corollary 7 involves the noise ϵ through $\frac{1}{2\sigma^2} \|\mathbf{w} + \epsilon - \mathbf{v}_i\|_2^2 - \|\epsilon\|_2^2$ compared to our divergence which is $\frac{1}{\sigma^2} \|\mathbf{w} - \mathbf{v}_i\|_2^2$ (without noise). Then, the sampled noise during the optimization procedure ϵ influences the disintegrated KL divergence, making it prone to high variations during training (depending thus σ^2). To illustrate the difference during the optimization, we focus on the objective function (detailed in Appendix I) of Corollary 6 and Corollary 7 (Equation (7)). Roughly speaking, the divergence in Corollary 6 does not depend on the sampled hypothesis h (with weights $\omega + \epsilon$), while the divergence of Equation (7) does. In consequence, the derivatives are less dependent on h for Corollary 6 than for Equation (7). To be convinced of this, we propose to study the gradient with respect to the current mean vector ω . On the one hand, the gradient $\frac{\partial R_S(h)}{\partial \omega}$ of the risk *w.r.t.* ω is the same for both bounds; hence, the phenomenon cannot come from this derivative. On the other hand, the gradients of the divergence in Equation (7)

and Corollary 6 are respectively

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\omega}} \left[\frac{1}{m} \left(\frac{\|\boldsymbol{\omega} + \boldsymbol{\epsilon} - \mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} \right) \right] &= \frac{\partial}{\partial \boldsymbol{\omega}} \left[\frac{1}{m2\sigma^2} \|\boldsymbol{\omega} + \boldsymbol{\epsilon} - \mathbf{v}_t\|_2^2 \right] \\ &= \frac{1}{m\sigma^2} (\boldsymbol{\omega} + \boldsymbol{\epsilon} - \mathbf{v}_t) = \diamond, \\ \text{and} \quad \frac{\partial}{\partial \boldsymbol{\omega}} \left[\frac{1}{m} \left(\frac{\|\boldsymbol{\omega} - \mathbf{v}_t\|_2^2}{\sigma^2} \right) \right] &= \frac{\partial}{\partial \boldsymbol{\omega}} \left[\frac{1}{m\sigma^2} \|\boldsymbol{\omega} - \mathbf{v}_t\|_2^2 \right] \\ &= \frac{2}{m\sigma^2} (\boldsymbol{\omega} - \mathbf{v}_t) = \heartsuit. \end{aligned}$$

From the two derivatives, we deduce that $\diamond = \frac{1}{2}\heartsuit + \frac{1}{m\sigma^2}\boldsymbol{\epsilon}$. Hence, each gradient step involves a noise in the gradient of the disintegrated KL divergence $\frac{1}{m\sigma^2}\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \frac{1}{m}\mathbf{I}_d)$, which is high for a small m . This randomness causes the disintegrated KL divergence $\frac{1}{2\sigma^2} \|\boldsymbol{\omega} + \boldsymbol{\epsilon} - \mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2$ to be larger when σ^2 decreases since (i) the divergence is divided by $2m\sigma^2$ and (ii) the deviation between $\boldsymbol{\omega}$ and \mathbf{v}_t increases. In conclusion, this makes the objective function (i.e., the bound) subject to high variations during the optimization, implying higher final bound values. Thus, the Rényi divergence has a valuable asset over the disintegrated KL divergence since it does not depend on the sampled noise $\boldsymbol{\epsilon}$.

5.4.5 Take-home message from the experiments

To summarize, our experiments show that our disintegrated bound is, in practice, tighter than the ones in the literature. This tightness allows us to precisely bound the true risk $R_{\mathcal{D}}(h)$ (or the test risk $R_{\mathcal{T}}(h)$); thus, the model selection from the disintegrated bound is effective. Moreover, we show that our bound is more easily optimizable than the others. This is mainly due to the disintegrated KL divergence, which depends on the sampled hypothesis h with weights $\boldsymbol{\omega} + \boldsymbol{\epsilon}$. Indeed, the gradients of the disintegrated KL divergence with respect to $\boldsymbol{\omega}$ include the noise $\boldsymbol{\epsilon}$, making the gradient inaccurate (especially with “high” learning rate and small variance σ^2).

6 Toward information-theoretic bounds

Before concluding, we discuss 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 a behavior, we consider 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

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

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 $\rho(h, \mathcal{S}) = Q_{\mathcal{S}}(h) \mathcal{D}^m(\mathcal{S})$, resp. $\pi(h, \mathcal{S}) = \mathcal{P}(h) \mathcal{D}^m(\mathcal{S})$, be the probability of sampling both $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim Q_{\mathcal{S}}$, resp. $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{P}$. Then we can write:

$$\begin{aligned} I_\alpha(h; \mathcal{S}) &= \min_{\mathcal{P} \in \mathcal{M}^*(\mathcal{H})} \frac{1}{\alpha-1} \ln \left[\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \left[\frac{Q_{\mathcal{S}}(h) \mathcal{D}^m(\mathcal{S})}{\mathcal{P}(h) \mathcal{D}^m(\mathcal{S})} \right]^\alpha \right] \\ &= \min_{\mathcal{P} \in \mathcal{M}^*(\mathcal{H})} D_\alpha(\rho \| \pi). \end{aligned} \tag{12}$$

From Verdú (2015) the optimal prior \mathcal{P}^* minimizing Equation (12) is a *distribution-dependent* prior:

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

This leads to an *Information-Theoretic generalization bound*¹².

Theorem 9 (Disintegrated Information-Theoretic Bound) *For any distribution \mathcal{D} on \mathcal{Z} , for any hypothesis set \mathcal{H} , for any measurable function $\phi : \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$, for any $\alpha > 1$, for any $\delta \in (0, 1]$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have*

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

Note that Esposito, Gastpar, and Issa (2020, Cor.4) introduced a bound based on the Sibson’s mutual information, but, as discussed in Appendix J, 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 $D_\alpha(Q_{\mathcal{S}} \| \mathcal{P})$ suggests that the learned posterior $Q_{\mathcal{S}}$ 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} . 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.

¹²We provide a mutual information-based bound in Appendix J.

7 Conclusion and future works

We provide a new and general disintegrated PAC-Bayesian bound (Theorem 2) in the family of Equation (5), *i.e.*, when the derandomization step consists in (i) learning a posterior distribution $\mathcal{Q}_{\mathcal{S}}$ on the classifiers set (given an algorithm, a learning sample \mathcal{S} and a prior distribution \mathcal{P}) and (ii) sampling a hypothesis h from this posterior $\mathcal{Q}_{\mathcal{S}}$. While our bound can be looser than the ones of Blanchard and Fleuret (2007); Catoni (2007); Rivasplata et al. (2020), it provides nice opportunities for learning deterministic classifiers. Indeed, our bound can be used not only to study the theoretical guarantees of deterministic classifiers but also to derive self-bounding algorithms (based on the bound optimization) that are more stable and efficient than the ones we obtain from the bounds of the literature. Concretely, the bounds of Blanchard and Fleuret (2007); Catoni (2007); Rivasplata et al. (2020) depend on two terms related to the classifier drawn: the risk and the “disintegrated KL divergence”, while in our bound the (Rényi) divergence term depends on the hypothesis set, implying that the divergence remains the same whatever which classifier is drawn. In this sense, our bound is more stable as the learning algorithm seeking to minimize the bound allows, in practice, to choose a better hypothesis than with the bounds of Blanchard and Fleuret (2007); Catoni (2007); Rivasplata et al. (2020). We have illustrated the interest of our bound on neural networks, but our result could be instantiated to other well-known settings such as linear classifiers (Germain et al., 2009) or the majority vote classifier (Zantedeschi et al., 2021).

Our general framework opens the way to the study of other machine learning settings by exploiting the proven *randomized* PAC-Bayesian theorems, for example, for Domain Adaptation (Germain et al., 2020), Adversarial Robustness (Viallard, Vidot, Habrard, & Morvant, 2021) or Transductive Learning (Bégin, Germain, Laviolette, & Roy, 2014).

Despite being an important step towards the practical use of PAC-Bayes guarantees, our disintegrated bounds arguably have a drawback: we sample a hypothesis from a distribution instead of obtaining a bound for all the possible hypotheses, like for uniform convergence bounds. While uniform convergence bounds can be vacuous (Nagarajan & Kolter, 2019b), they hold (with high probability on the choice of the learning sample) for all hypotheses including the one with the best guarantee (*i.e.*, the one minimizing the bound). In the case of disintegrated bounds, we learn a distribution on the hypothesis set, and then we sample a hypothesis according to this distribution. Hence, there is a small probability (*i.e.*, less than δ) of sampling a *bad* hypothesis. An interesting research direction is comparing disintegrated and uniform convergence bounds to understand in which cases using disintegrated bounds can be better than using uniform convergence bounds. Knowing that there are connections between (agnostic) PAC-learnability and uniform convergence (see, *e.g.*, Shalev-Shwartz and Ben-David (2014)), we believe that defining a new notion of PAC-learnability, which better fits with the disintegrated framework, could help to provide such a comparison.

Acknowledgments. This work was partially funded by the French ANR Project APRIORI ANR-18-CE23-0015. Pascal Germain is supported by the Canada CIFAR AI Chair Program, and the NSERC Discovery grant RGPIN-2020-07223. We would like to thank the reviewers for their valuable comments and their suggestions to improve the paper.

Declarations

Funding. This work was partially funded by the French ANR Project APRIORI ANR-18-CE23-0015. Pascal Germain is supported by the Canada CIFAR AI Chair Program, and the NSERC Discovery grant RGPIN-2020-07223.

Conflict of interest/Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. Not applicable.

Code availability. The code is available on Github at <https://github.com/paulviillard/MLJ-Disintegrated-PB>.

Authors' contributions. Conceptualization: *Dr. P. Viillard, Dr. E. Morvant, Dr. P. Germain, Pr. A. Habrard*; Formal analysis and investigation: *Dr. P. Viillard*; Software: *Dr. P. Viillard*; Writing - original draft preparation: *Dr. P. Viillard, Dr. E. Morvant*; Writing - review and editing: *Dr. P. Germain, Pr. A. Habrard*; Funding acquisition: *Dr. E. Morvant, Dr. P. Germain, Pr. A. Habrard*; Supervision: *Dr. E. Morvant, Dr. P. Germain, Pr. A. Habrard*.

References

- Alquier, P. (2021). User-friendly introduction to PAC-Bayes bounds. *CoRR*, *abs/2110.11216*.
- Ambroladze, A., Parrado-Hernández, E., Shawe-Taylor, J. (2006). Tighter PAC-Bayes bounds. *Advances in neural information processing systems (NIPS)* (pp. 9–16). MIT Press.
- Bégin, L., Germain, P., Laviolette, F., Roy, J. (2014). PAC-Bayesian theory for transductive learning. *International conference on artificial intelligence and statistics (AISTATS)* (Vol. 33, pp. 105–113). JMLR.org.

- Bégin, L., Germain, P., Laviolette, F., Roy, J. (2016). PAC-Bayesian bounds based on the Rényi divergence. *International conference on artificial intelligence and statistics (AISTATS)* (Vol. 51, pp. 435–444). JMLR.org.
- Biggs, F., & Guedj, B. (2021). Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10), 1280.
- Biggs, F., & Guedj, B. (2022). On margins and derandomisation in PAC-Bayes. *International conference on artificial intelligence and statistics (AISTATS)* (Vol. 151, pp. 3709–3731). PMLR.
- Blanchard, G., & Fleuret, F. (2007). Occam’s hammer. *Annual conference on learning theory (COLT)* (Vol. 4539, pp. 112–126). Springer.
- Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2, 499–526.
- Catoni, O. (2007). PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *CoRR*, abs/0712.0248.
- Dziugaite, G.K., & Roy, D. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Conference on uncertainty in artificial intelligence (UAI)*. AUAI Press.
- Dziugaite, G.K., & Roy, D. (2018). Data-dependent PAC-Bayes priors via differential privacy. *Advances in neural information processing systems (NeurIPS)* (pp. 8440–8450).
- Esposito, A.R., Gastpar, M., Issa, I. (2020). Robust generalization via α -mutual information. *CoRR*, abs/2001.06399.
- Freund, Y. (1998). Self bounding learning algorithms. *Annual conference on computational learning theory (COLT)* (pp. 247–258). ACM.
- Germain, P., Habrard, A., Laviolette, F., Morvant, E. (2020). PAC-Bayes and domain adaptation. *Neurocomputing*, 379, 379–397.
- Germain, P., Lacasse, A., Laviolette, F., Marchand, M. (2009). PAC-Bayesian learning of linear classifiers. *Annual international conference on machine learning (ICML)* (Vol. 382, pp. 353–360). ACM.

- Gil, M., Alajaji, F., Linder, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249, 124–131.
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *International conference on artificial intelligence and statistics (AISTATS)* (Vol. 9, pp. 249–256). JMLR.org.
- Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep learning*. MIT Press.
- Guedj, B. (2019). A primer on PAC-Bayesian learning. *CoRR*, abs/1901.05353.
- Hardt, M., Recht, B., Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. *International conference on machine learning (ICML)* (Vol. 48, pp. 1225–1234). JMLR.org.
- He, K., Zhang, X., Ren, S., Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *IEEE international conference on computer vision (ICCV)* (pp. 1026–1034). IEEE Computer Society.
- He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). IEEE Computer Society.
- Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. *International conference on learning representations (ICLR)*.
- Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images* (Unpublished master’s thesis). University of Toronto.
- Langford, J., & Caruana, R. (2001). (Not) bounding the true error. *Advances in neural information processing systems (NIPS)* (pp. 809–816). MIT Press.
- Langford, J., & Shawe-Taylor, J. (2002). PAC-Bayes & margins. *Advances in neural information processing systems (NIPS)* (pp. 423–430). MIT Press.
- LeCun, Y., Cortes, C., Burges, C. (1998). *The MNIST dataset of handwritten digits*. Retrieved from <http://yann.lecun.com/exdb/mnist/>
- Letarte, G., Germain, P., Guedj, B., Laviolette, F. (2019). Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks.

Advances in neural information processing systems (NeurIPS) (pp. 6869–6879).

- Lever, G., Laviolette, F., Shawe-Taylor, J. (2013). Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473, 4–28.
- Maurer, A. (2004). A note on the PAC Bayesian theorem. *CoRR*, cs.LG/0411099.
- McAllester, D. (1998). Some PAC-Bayesian theorems. *Annual conference on computational learning theory (COLT)* (pp. 230–234). ACM.
- Nagarajan, V., & Kolter, Z. (2019a). Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience. *International conference on learning representations (ICLR)*. OpenReview.net.
- Nagarajan, V., & Kolter, Z. (2019b). Uniform convergence may be unable to explain generalization in deep learning. *Advances in neural information processing systems (NeurIPS)* (pp. 11611–11622).
- Neyshabur, B., Bhojanapalli, S., Srebro, N. (2018). A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *International conference on learning representations (ICLR)*. OpenReview.net.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems (NeurIPS)* (pp. 8024–8035).
- Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., Szepesvári, C. (2021). Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 227:1–227:40.
- Reeb, D., Doerr, A., Gerwinn, S., Rakitsch, B. (2018). Learning gaussian processes by minimizing PAC-Bayesian generalization bounds. *Advances in neural information processing systems (NeurIPS)* (pp. 3341–3351).
- Rivasplata, O., Kuzborskij, I., Szepesvári, C., Shawe-Taylor, J. (2020). PAC-Bayes analysis beyond the usual bounds. *Advances in neural information processing systems (NeurIPS)*.

- Seeger, M. (2002). PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3, 233–269.
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning - from theory to algorithms*. Cambridge University Press.
- Shawe-Taylor, J., & Williamson, R. (1997). A PAC analysis of a bayesian estimator. *Annual conference on computational learning theory (COLT)* (pp. 2–9). ACM.
- Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. *International conference on learning representations (ICLR)*.
- Thiemann, N., Igel, C., Wintenberger, O., Seldin, Y. (2017). A strongly quasiconvex PAC-Bayesian bound. *International conference on algorithmic learning theory (ALT)* (Vol. 76, pp. 466–492). PMLR.
- Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134–1142.
- van Erven, T., & Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7), 3797–3820.
- Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.
- Verdú, S. (2015). α -mutual information. *Information theory and applications workshop (ITA)* (pp. 1–6). IEEE.
- Viallard, P., Vidot, G., Habrard, A., Morvant, E. (2021). A PAC-Bayes analysis of adversarial robustness. *Advances in neural information processing systems (NeurIPS)* (pp. 14421–14433).
- Xiao, H., Rasul, K., Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.
- Xu, H., & Mannor, S. (2012). Robustness and generalization. *Machine Learning*, 86(3), 391–423.

- Zantedeschi, V., Viillard, P., Morvant, E., Emonet, R., Habrard, A., Germain, P., Guedj, B. (2021). Learning stochastic majority votes by minimizing a PAC-Bayes generalization bound. *Advances in neural information processing systems (NeurIPS)* (pp. 455–467).
- Zhou, W., Veitch, V., Austern, M., Adams, R., Orbanz, P. (2019). Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. *International conference on learning representations (ICLR)*. OpenReview.net.

This Appendix is structured as follows. We give the proof of Theorem 1, Theorem 2, Corollary 3, Theorem 4, Proposition 5, Corollary 6, Corollary 7, and Corollary 8 in Appendix A, Appendix B, Appendix C, Appendix D, Appendix D, Appendix F, Appendix G, and Appendix H respectively. We also discuss the minimization and the evaluation of the bounds introduced in the different corollaries in Appendix I. Additionally, Appendix J is devoted to Theorem 9. Appendix K provides an exhaustive list of numerical results.

Appendix A Proof of Theorem 1

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

$$\underbrace{\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left(\forall \mathcal{Q} \in \mathcal{M}(\mathcal{H}), \mathbb{E}_{h \sim \mathcal{Q}} \ln(\phi(h, \mathcal{S})) \leq \text{KL}(\mathcal{Q} \parallel \mathcal{P}) + \ln \left[\frac{1}{\delta} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S}) \right] \right)}_{(\text{Germain et al., 2009})} \geq 1 - \delta, \quad (1)$$

and

$$\underbrace{\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left(\forall \mathcal{Q} \in \mathcal{M}(\mathcal{H}), \frac{\alpha}{\alpha-1} \ln \left[\mathbb{E}_{h \sim \mathcal{Q}} \phi(h, \mathcal{S}) \right] \leq D_\alpha(\mathcal{Q} \parallel \mathcal{P}) + \ln \left[\frac{1}{\delta} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S})^{\frac{\alpha}{\alpha-1}} \right] \right)}_{(\text{Bégin et al., 2016})} \geq 1 - \delta, \quad (2)$$

with $\text{KL}(\mathcal{Q} \parallel \mathcal{P}) \triangleq \mathbb{E}_{h \sim \mathcal{Q}} \ln \frac{\mathcal{Q}(h)}{\mathcal{P}(h)}$ the Kullback-Leibler (KL-)divergence between \mathcal{Q} and \mathcal{P} , and $D_\alpha(\mathcal{Q} \parallel \mathcal{P}) \triangleq \frac{1}{\alpha-1} \ln \left[\mathbb{E}_{h \sim \mathcal{P}} \left[\frac{\mathcal{Q}(h)}{\mathcal{P}(h)} \right]^\alpha \right]$ the Rényi divergence between \mathcal{Q} and \mathcal{P} ($\alpha > 1$).

Proof By the Donsker-Varadhan’s variational formula (see e.g., Bégin et al., 2016, Lemma 3), we have

$$\forall \mathcal{Q} \in \mathcal{M}(\mathcal{H}), \quad \mathbb{E}_{h \sim \mathcal{Q}} \ln(\phi(h, \mathcal{S})) \leq \text{KL}(\mathcal{Q} \parallel \mathcal{P}) + \ln \left[\mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S}) \right]. \quad (\text{A1})$$

By Markov’s inequality and taking the logarithm to both sides, we have

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left[\ln \left[\mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S}) \right] \leq \ln \left[\frac{1}{\delta} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S}) \right] \right] \geq 1 - \delta. \quad (\text{A2})$$

By merging Equations (A1) and (A2), we obtain Equation (1).

The proof of Equation (2) is similar to the one of Equation (1). Indeed, from the Rényi change of measure (see e.g., Bégin et al., 2016, Theorem 8), we have

$$\forall \mathcal{Q} \in \mathcal{M}(\mathcal{H}), \quad \frac{\alpha}{\alpha-1} \ln \left[\mathbb{E}_{h \sim \mathcal{Q}} \phi(h, \mathcal{S}) \right] \leq D_\alpha(\mathcal{Q} \parallel \mathcal{P}) + \ln \left[\mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S})^{\frac{\alpha}{\alpha-1}} \right]. \quad (\text{A3})$$

By Markov's inequality and taking the logarithm to both sides, we have

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m} \left[\ln \left[\mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S})^{\frac{\alpha}{\alpha-1}} \right] \leq \ln \left[\frac{1}{\delta} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S})^{\frac{\alpha}{\alpha-1}} \right] \right] \geq 1 - \delta. \quad (\text{A4})$$

By merging Equations (A3) and (A4), Equation (2) is obtained. \square

Appendix B Proof of Theorem 2

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

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

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

Proof For any sample $\mathcal{S} \in \mathcal{Z}^m$, prior $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$ and deterministic algorithm A fixed a priori, let $\mathcal{Q}_S = A(\mathcal{S}, \mathcal{P})$ the distribution obtained from the algorithm A . Note that $\phi(h, \mathcal{S})$ is a strictly positive random variable. Hence, from Markov's inequality, we have

$$\begin{aligned} \mathbb{P}_{h \sim \mathcal{Q}_S} \left[\phi(h, \mathcal{S}) \leq \frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{Q}_S} \phi(h', \mathcal{S}) \right] &\geq 1 - \frac{\delta}{2} \\ \iff \mathbb{E}_{h \sim \mathcal{Q}_S} \mathbf{I} \left[\phi(h, \mathcal{S}) \leq \frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{Q}_S} \phi(h', \mathcal{S}) \right] &\geq 1 - \frac{\delta}{2}. \end{aligned}$$

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

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

Since both sides of the inequality are strictly positive, we can take the logarithm and multiply by $\frac{\alpha}{\alpha-1} > 0$ to obtain

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

We develop the right-hand side of the inequality and take the expectation of the hypothesis over the prior distribution \mathcal{P} . We have for all prior $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$,

$$\frac{\alpha}{\alpha-1} \ln \left(\frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{Q}_S} \phi(h', \mathcal{S}) \right) = \frac{\alpha}{\alpha-1} \ln \left(\frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{P}} \frac{\mathcal{Q}_S(h')}{\mathcal{P}(h')} \phi(h', \mathcal{S}) \right),$$

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:

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

Then, since both sides of the inequality are strictly positive, we take the logarithm, add $\ln(\frac{2}{\delta})$ and multiply by $\frac{\alpha}{\alpha-1} > 0$ to both sides of the inequality, to obtain

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

From this inequality, we can deduce that

$$\begin{aligned} \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\forall \mathcal{P} \in \mathcal{M}^*(\mathcal{H}), \frac{\alpha}{\alpha-1} \ln(\phi(h, \mathcal{S})) \leq D_\alpha(\mathcal{Q}_S \| \mathcal{P}) \right. \\ \left. + \frac{\alpha}{\alpha-1} \ln \frac{2}{\delta} + \ln \left(\mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', \mathcal{S})^{\frac{\alpha-1}{\alpha}} \right) \right) \right] \geq 1 - \frac{\delta}{2}. \end{aligned} \tag{B5}$$

Given a prior $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$, note that $\mathbb{E}_{h' \sim \mathcal{P}} \phi(h', \mathcal{S})^{\frac{\alpha-1}{\alpha}}$ is a strictly positive random variable. Hence, we apply Markov's inequality to have

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

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

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

Since both sides of the inequality are strictly positive, we take the logarithm to both sides of the inequality, and we add $\frac{\alpha}{\alpha-1} \ln \frac{2}{\delta}$ to have

$$\begin{aligned} & \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', \mathcal{S})^{\frac{\alpha-1}{\alpha}} \right) \leq \frac{2}{\delta} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', \mathcal{S}')^{\frac{\alpha-1}{\alpha}} \right) \right] \geq 1 - \frac{\delta}{2} \iff \\ & \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\frac{\alpha}{\alpha-1} \ln \frac{2}{\delta} + \ln \left(\mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', \mathcal{S})^{\frac{\alpha-1}{\alpha}} \right) \right) \leq \frac{2\alpha-1}{\alpha-1} \ln \frac{2}{\delta} \right. \\ & \quad \left. + \ln \left(\mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', \mathcal{S}')^{\frac{\alpha-1}{\alpha}} \right) \right) \right] \geq 1 - \frac{\delta}{2}. \end{aligned} \tag{B6}$$

Combining Equations (B5) and (B6) with a union bound gives us the desired result. \square

Appendix C Proof of Corollary 3

Corollary 3 *Under the assumptions of Theorem 2, when $\alpha \rightarrow 1^+$, we have*

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left(\ln \phi(h, \mathcal{S}) \leq \ln \frac{2}{\delta} + \ln \left[\operatorname{esssup}_{S' \in \mathcal{Z}, h' \in \mathcal{H}} \phi(h', S') \right] \right) \geq 1 - \delta,$$

when $\alpha \rightarrow +\infty$, we have

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left(\ln \phi(h, \mathcal{S}) \leq \ln \operatorname{esssup}_{h' \in \mathcal{H}} \frac{\mathcal{Q}_S(h')}{\mathcal{P}(h')} + \ln \left[\frac{4}{\delta^2} \mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \phi(h', S') \right] \right) \geq 1 - \delta,$$

where esssup is the essential supremum defined as the supremum on a set with non-zero probability measures, i.e.,

$$\operatorname{esssup}_{S' \in \mathcal{Z}, h' \in \mathcal{H}} \phi(h', S') = \inf \left\{ \tau \in \mathbb{R}, \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} [\phi(h, \mathcal{S}) > \tau] = 0 \right\},$$

and

$$\operatorname{esssup}_{h' \in \mathcal{H}} \frac{\mathcal{Q}_S(h')}{\mathcal{P}(h')} = \inf \left\{ \tau \in \mathbb{R}, \mathbb{P}_{h \sim \mathcal{Q}_S} \left[\frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)} > \tau \right] = 0 \right\}.$$

Proof Starting from Theorem 2 and rearranging, we have

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

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

When $\alpha \rightarrow 1$.

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

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

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

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

Hence, we have

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

$$= \ln(\|\phi\|_\infty) = \ln\left(\operatorname{esssup}_{S' \in \mathcal{Z}, h' \in \mathcal{H}} \phi(h', S')\right).$$

Finally, we can deduce that

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

When $\alpha \rightarrow +\infty$.

First, we have $\lim_{\alpha \rightarrow +\infty} \frac{2\alpha-1}{\alpha} \ln \frac{2}{\delta} = \ln \frac{2}{\delta} [2 - \lim_{\alpha \rightarrow +\infty} \frac{1}{\alpha}] = 2 \ln \frac{2}{\delta} = \ln \frac{4}{\delta^2}$ and $\lim_{\alpha \rightarrow +\infty} \|\phi\|_{\frac{\alpha}{\alpha-1}} = \lim_{\alpha' \rightarrow 1} \|\phi\|_{\alpha'} = \|\phi\|_1$ (since $\lim_{\alpha \rightarrow +\infty} \frac{\alpha}{\alpha-1} = \lim_{\alpha \rightarrow +\infty} \frac{1}{1-\frac{1}{\alpha}} = 1$). Hence, we have

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

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

$$\begin{aligned} \frac{\alpha-1}{\alpha} D_\alpha(\mathcal{Q}_S \| \mathcal{P}) &= \frac{1}{\alpha} \ln \left(\mathbb{E}_{h \sim \mathcal{P}} \left(\left[\frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)} \right]^\alpha \right) \right) = \ln \left(\left[\mathbb{E}_{h \sim \mathcal{P}} \left(\left[\frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)} \right]^\alpha \right) \right]^{\frac{1}{\alpha}} \right) \\ &= \ln \left(\left[\mathbb{E}_{h \sim \mathcal{P}} (\gamma(h)^\alpha) \right]^{\frac{1}{\alpha}} \right) = \ln(\|\gamma\|_\alpha), \end{aligned}$$

where $\|\gamma\|_\alpha$ is the L^α -norm of the function γ defined as $\gamma(h) = \frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)}$. We have

$$\begin{aligned} \lim_{\alpha \rightarrow +\infty} \frac{\alpha-1}{\alpha} D_\alpha(\mathcal{Q}_S \| \mathcal{P}) &= \lim_{\alpha \rightarrow +\infty} \ln(\|\gamma\|_\alpha) = \ln \left(\lim_{\alpha \rightarrow +\infty} \|\gamma\|_\alpha \right) \\ &= \ln(\|\gamma\|_\infty) = \ln \left(\operatorname{esssup}_{h \in \mathcal{H}} \gamma(h) \right) = \ln \left(\operatorname{esssup}_{h \in \mathcal{H}} \frac{\mathcal{Q}_S(h)}{\mathcal{P}(h)} \right). \end{aligned}$$

Finally, we can deduce that

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

□

Appendix D 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*

$$\sqrt{\frac{a}{b}} = \operatorname{argmin}_{\lambda > 0} \left(\frac{a}{\lambda} + \lambda b \right), \text{ and } 2\sqrt{ab} = \min_{\lambda > 0} \left(\frac{a}{\lambda} + \lambda b \right),$$

$$\text{and } \forall \lambda > 0, \sqrt{ab} \leq \frac{1}{2} \left(\frac{a}{\lambda} + \lambda b \right).$$

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

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

Moreover, from the derivative we can deduce that we have $\frac{\partial f}{\partial \lambda}(\lambda) < 0 \iff \lambda \in (0, \sqrt{\frac{a}{b}})$, and $\frac{\partial f}{\partial \lambda}(\lambda) > 0 \iff \lambda > \sqrt{\frac{a}{b}}$ and $\frac{\partial f}{\partial \lambda}(\lambda) = 0 \iff \lambda = \sqrt{\frac{a}{b}}$. It implies that the function is strictly decreasing on $\lambda \in (0, \sqrt{\frac{a}{b}})$, strictly increasing for $\lambda > \sqrt{\frac{a}{b}}$ and admit a unique minimum at $\lambda^* = \sqrt{\frac{a}{b}}$. Additionally, $f(\lambda^*) = 2\sqrt{ab}$ which proves the claim. \square

We can now prove Theorem 4 with Lemma 10.

Theorem 4 (Parametrizable Disintegrated PAC-Bayes Bound) *For any distribution \mathcal{D} on \mathcal{Z} , for any hypothesis set \mathcal{H} , for any prior distribution $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$, for any measurable function $\phi: \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$, for any $\delta \in (0, 1]$, for any algorithm $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have*

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

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

Proof The proof is similar to the one of Theorem 2. Since $\phi(h, \mathcal{S})$ is a strictly positive random variable, from Markov's inequality, we have

$$\mathbb{P}_{h \sim \mathcal{Q}_{\mathcal{S}}} \left[\phi(h, \mathcal{S}) \leq \frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{Q}_{\mathcal{S}}} \phi(h', \mathcal{S}) \right] \geq 1 - \frac{\delta}{2}$$

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

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

$$\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{Q}_{\mathcal{S}}} \mathbf{I} \left[\phi(h, \mathcal{S}) \leq \frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{Q}_{\mathcal{S}}} \phi(h', \mathcal{S}) \right] \geq 1 - \frac{\delta}{2}$$

$$\iff \mathbb{P}_{S \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\phi(h, S) \leq \frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{Q}_S} \phi(h', S) \right] \geq 1 - \frac{\delta}{2}.$$

Using Lemma 10 with $a = \frac{4}{\delta^2} \phi(h', S)^2$ and $b = \frac{\mathcal{Q}_S(h')^2}{\mathcal{P}(h')^2}$, we have for all prior $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$

$$\begin{aligned} \forall \lambda > 0, \quad \frac{2}{\delta} \mathbb{E}_{h' \sim \mathcal{Q}_S} \phi(h', S) &= \mathbb{E}_{h' \sim \mathcal{P}} \sqrt{\frac{\mathcal{Q}_S(h')^2}{\mathcal{P}(h')^2} \frac{4}{\delta^2} \phi(h', S)^2} \\ &\leq \frac{1}{2} \left[\lambda \mathbb{E}_{h' \sim \mathcal{P}} \left(\frac{\mathcal{Q}_S(h')}{\mathcal{P}(h')} \right)^2 + \frac{4}{\lambda \delta^2} \mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', S)^2 \right) \right]. \end{aligned}$$

Then, since both sides of the inequality are strictly positive, we take the logarithm to obtain

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

Hence, we can deduce that

$$\begin{aligned} \mathbb{P}_{S \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\forall \mathcal{P} \in \mathcal{M}^*(\mathcal{H}), \forall \lambda > 0, \ln(\phi(h, S)) \right. \\ \left. \leq \ln \left(\frac{1}{2} \left[\lambda e^{D_2(\mathcal{Q}_S \| \mathcal{P})} + \frac{4}{\lambda \delta^2} \mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', S)^2 \right) \right] \right) \right] \geq 1 - \frac{\delta}{2}. \end{aligned} \quad (D7)$$

Given a prior $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$, note that $\mathbb{E}_{h' \sim \mathcal{P}} \phi(h', S)^2$ is a strictly positive random variable. Hence, we apply Markov's inequality:

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

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

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

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

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

Hence, we can deduce that

$$\begin{aligned} & \mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\forall \lambda > 0, \ln \left(\frac{1}{2} \left[\lambda \exp(D_2(\mathcal{Q}_S \parallel \mathcal{P})) + \frac{4}{\lambda \delta^2} \mathbb{E}_{h' \sim \mathcal{P}} (\phi(h', \mathcal{S})^2) \right] \right) \right] \\ & \leq \ln \left(\frac{1}{2} \left[\lambda \exp(D_2(\mathcal{Q}_S \parallel \mathcal{P})) + \frac{8}{\lambda \delta^3} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} (\phi(h', \mathcal{S}')^2) \right] \right) \geq 1 - \frac{\delta}{2}. \quad (\text{D8}) \end{aligned}$$

Combining Equations (D7) and (D8) with a union bound gives us the desired result. \square

Appendix E 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 $\delta \in (0, 1]$, for any measurable function $\phi: \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$, for any algorithm $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, let*

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

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

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

Put into words: the optimal λ^* gives the same bound for Theorem 2 and Theorem 4.

Proof We consider the right-hand side of the inequality of Theorem 4 (which is strictly positive): we have

$$\ln \left[\frac{\lambda}{2} e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})} + \frac{8}{2\lambda\delta^3} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} [\phi(h', \mathcal{S}')^2] \right]. \quad (\text{E9})$$

Since \ln is a strictly increasing function, we have

$$\begin{aligned} & \min_{\lambda > 0} \left\{ \ln \left[\frac{\lambda}{2} e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})} + \frac{8}{2\lambda\delta^3} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} [\phi(h', \mathcal{S}')^2] \right] \right\} \\ & = \ln \left[\min_{\lambda > 0} \left\{ \frac{\lambda}{2} e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})} + \frac{8}{2\lambda\delta^3} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} [\phi(h', \mathcal{S}')^2] \right\} \right]. \end{aligned}$$

Then, we apply Lemma 10 by taking $a = \frac{8}{2\delta^3} \mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} [\phi(h', \mathcal{S}')^2]$ and $b = \frac{1}{2} e^{D_2(\mathcal{Q}_S \parallel \mathcal{P})}$ to obtain $\lambda^* = \sqrt{\frac{a}{b}} = \sqrt{\frac{\mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} [8\phi(h', \mathcal{S}')^2]}{\delta^3 \exp(D_2(\mathcal{Q}_S \parallel \mathcal{P}))}}$. Finally, by substituting λ^* into Equation (E9), we obtain

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

$$= \frac{1}{2} \left(D_2(\mathcal{Q}_S \| \mathcal{P}) + \ln \left[\mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \left(\frac{8\phi(h', S')^2}{\delta^3} \right) \right] \right),$$

which is the desired result. \square

Appendix F 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 $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$, for any measurable function $\phi: \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$, for any $\alpha > 1$, for any $\delta \in (0, 1]$, for any algorithm $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have*

$$\begin{aligned} \mathbb{P}_{S \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\forall \mathcal{P}_t \in \mathbf{P}, \frac{\alpha}{\alpha-1} \ln(\phi(h, S)) \leq D_\alpha(\mathcal{Q}_S \| \mathcal{P}) + \frac{\alpha}{\alpha-1} \ln \frac{2}{\delta} \right. \\ \left. + \ln \frac{2T}{\delta} + \ln \left(\mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', S')^{\frac{\alpha}{\alpha-1}} \right) \right) \right] \geq 1 - \delta, \end{aligned}$$

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

Proof The proof is mainly the same as Theorem 2. Indeed, we first derive the same equation as Equation (B5), we have

$$\begin{aligned} \mathbb{P}_{S \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\forall \mathcal{P} \in \mathcal{M}^*(\mathcal{H}), \frac{\alpha}{\alpha-1} \ln(\phi(h, S)) \leq D_\alpha(\mathcal{Q}_S \| \mathcal{P}) \right. \\ \left. + \frac{\alpha}{\alpha-1} \ln \frac{2}{\delta} + \ln \left(\mathbb{E}_{h' \sim \mathcal{P}} \left(\phi(h', S)^{\frac{\alpha}{\alpha-1}} \right) \right) \right] \geq 1 - \frac{\delta}{2}. \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}_{S \sim \mathcal{D}^m, h \sim \mathcal{Q}_S} \left[\ln \left(\mathbb{E}_{h' \sim \mathcal{P}_t} \left[\phi(h', S)^{\frac{\alpha}{\alpha-1}} \right] \right) \right. \\ \left. \leq \ln \frac{2T}{\delta} + \ln \left(\mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}_t} \left[\phi(h', S')^{\frac{\alpha}{\alpha-1}} \right] \right) \right] \geq 1 - \frac{\delta}{2T}. \end{aligned}$$

Finally, combining the $T + 1$ bounds with a union bound gives us the desired result. \square

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 set $\mathbf{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ of T priors on \mathcal{H} where $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$, for any algorithm $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, for any loss $\ell: \mathcal{H} \times \mathcal{Z} \rightarrow [0, 1]$, for any $\delta \in (0, 1]$, we have*

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

where $\text{kl}(a||b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$, $\mathcal{Q}_S = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$, and the hypothesis $h \sim \mathcal{Q}_S$ is parametrized by $\mathbf{w} + \boldsymbol{\epsilon}$.

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

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

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

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

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

Appendix G Proof of Corollary 7

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

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

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

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 $h \sim \mathcal{Q}_S$ with $\mathcal{Q}_S = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$.

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

$$\mathcal{Q}_S(h) = \left[\frac{1}{\sigma\sqrt{2\pi}} \right]^d \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{w} + \boldsymbol{\epsilon} - \mathbf{w}\|_2^2\right) = \left[\frac{1}{\sigma\sqrt{2\pi}} \right]^d \exp\left(-\frac{1}{2\sigma^2} \|\boldsymbol{\epsilon}\|_2^2\right)$$

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

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

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

\square

We can now prove Corollary 7.

Corollary 7 For any distribution \mathcal{D} on \mathcal{Z} , for any set \mathcal{H} , for any set $\mathbf{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ of T priors on \mathcal{H} where $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \{0, 1\}$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the learning sample $\mathcal{S} \sim \mathcal{D}^m$ and the hypothesis $h \sim \mathcal{Q}_{\mathcal{S}}$ parametrized by $\mathbf{w} + \epsilon$, we have $\forall \mathcal{P}_t \in \mathbf{P}$

$$\text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[\frac{\|\mathbf{w} + \epsilon - \mathbf{v}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \ln \frac{2T\sqrt{m}}{\delta} \right], \quad (7)$$

$$\forall b \in \mathbf{B}, \quad \text{kl}_+(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[\frac{b+1}{b} \left[\frac{\|\mathbf{w} + \epsilon - \mathbf{v}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} \right]_+ + \ln \frac{(b+1)T|\mathbf{B}|}{\delta} \right], \quad (8)$$

$$\forall c \in \mathbf{C}, \quad R_{\mathcal{D}}(h) \leq \frac{1 - \exp\left(-cR_{\mathcal{S}}(h) - \frac{1}{m} \left[\frac{\|\mathbf{w} + \epsilon - \mathbf{v}_t\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \ln \frac{T|\mathbf{C}|}{\delta} \right]\right)}{1 - e^{-c}}, \quad (9)$$

with $[x]_+ = \max(x, 0)$, and $\text{kl}_+(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) = \text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h))$ if $R_{\mathcal{S}}(h) < R_{\mathcal{D}}(h)$ and 0 otherwise. Moreover, $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ is a Gaussian noise such that $\mathbf{w} + \epsilon$ 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}, \mathbf{B} are two sets of hyperparameters fixed a priori.

Proof We will prove the three bounds separately.

Equation (7). We instantiate Theorem 1⁽ⁱ⁾ of Rivasplata et al. (2020) with $\phi(h, \mathcal{S}) = \exp[m\text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h))]$, 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 $h \sim \mathcal{Q}_{\mathcal{S}}$

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

From Maurer (2004), we upper-bound $\mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}_t} e^{m\text{kl}(R_{\mathcal{S}'}(h') \| R_{\mathcal{D}}(h'))}$ by $2\sqrt{m}$ and using Lemma 11 we rewrite the disintegrated KL divergence. Finally, a union-bound argument gives us the claim.

Equation (8). We apply $T|\mathbf{B}|$ times Proposition 3.1 of Blanchard and Fleuret (2007) with a confidence $\frac{\delta}{T|\mathbf{B}|}$ instead of δ . For each prior $\mathcal{P}_t \in \mathbf{P}$ and hyperparameters $b \in \mathbf{B}$, we have with probability at least $1 - \frac{\delta}{T|\mathbf{B}|}$ over the random choice of $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{Q}_{\mathcal{S}}$

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

From Lemma 11 and a union-bound argument, we obtain the claim.

Equation (9). 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 $h \sim \mathcal{Q}_{\mathcal{S}}$

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

From Lemma 11 and a union-bound argument, we obtain the claim. \square

Appendix H Proof of Corollary 8

Corollary 8 *For any distribution \mathcal{D} on \mathcal{Z} , for any \mathcal{H} , for any set $\mathbf{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ of T priors on \mathcal{H} where $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$, for any loss $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \{0, 1\}$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over $\mathcal{S} \sim \mathcal{D}^m$ and $\{h_1, \dots, h_n\} \sim \mathcal{Q}^n$, we have simultaneously $\forall \mathcal{P}_t \in \mathbf{P}$,*

$$\text{kl} \left(\mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h) \parallel \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h) \right) \leq \frac{1}{m} \left[\frac{\|\mathbf{w} - \mathbf{v}_t\|_2^2}{2\sigma^2} + \ln \frac{4T\sqrt{m}}{\delta} \right], \quad (10)$$

$$\text{and } \text{kl} \left(\frac{1}{n} \sum_{i=1}^n R_{\mathcal{S}}(h_i) \parallel \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h) \right) \leq \frac{1}{n} \ln \frac{4}{\delta}, \quad (11)$$

where $\mathcal{Q} = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$ and the hypothesis h sampled from \mathcal{Q} is parametrized by $\mathbf{w} + \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$.

Proof We instantiate Equation (3) (and apply Jensen's inequality on the left-hand 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} ,

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

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 $\text{KL}(\mathcal{Q} \parallel \mathcal{P}_t) = \frac{\|\mathbf{w} - \mathbf{v}_t\|_2^2}{2\sigma^2}$. 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}

$$\text{kl} \left(\mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{S}}(h) \parallel \mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h) \right) \leq \frac{1}{m} \left[\frac{\|\mathbf{w} - \mathbf{v}_t\|_2^2}{2\sigma^2} + \ln \frac{4T\sqrt{m}}{\delta} \right]. \quad (\text{Equation (10)})$$

Additionally, we obtained Equation (11) 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 (11) and (10) gives the claimed result. \square

Appendix I Evaluation and minimization of the bounds of Corollaries 6, 7, 8

This appendix presents more details on the optimization and the evaluation of the bounds.

I.1 Evaluation of the bounds

Note that, except for Equation (9), a generalization gap is upper-bounded instead of the true risk. Hence, to evaluate the bounds of the corollaries (except for Equation (9)) we use the invert binary kl divergence defined as

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

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

$$R_{\mathcal{D}}(h) \leq \underbrace{\text{kl}^{-1} \left(R_{\mathcal{S}}(h) \mid \frac{1}{m} \left[\frac{\|\mathbf{w} - \mathbf{v}_t\|_2^2}{\sigma^2} + \ln \frac{16T\sqrt{m}}{\delta^3} \right] \right)}_{\text{Corollary 6}},$$

$$R_{\mathcal{D}}(h) \leq \underbrace{\text{kl}^{-1} \left(R_{\mathcal{S}}(h) \mid \frac{1}{m} \left[\frac{\|\mathbf{w} + \boldsymbol{\epsilon} - \mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} + \ln \frac{2T\sqrt{m}}{\delta} \right] \right)}_{\text{Equation (7)}},$$

$$R_{\mathcal{D}}(h) \leq \underbrace{\text{kl}^{-1} \left(R_{\mathcal{S}}(h) \mid \frac{1}{m} \left[\frac{b+1}{b} \left[\frac{\|\mathbf{w} + \boldsymbol{\epsilon} - \mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2} \right]_+ + \ln \frac{(b+1)T|\mathbf{B}|}{\delta} \right] \right)}_{\text{Equation (8)}},$$

and $\mathbb{E}_{h \sim \mathcal{Q}} R_{\mathcal{D}}(h) \leq \underbrace{\text{kl}^{-1} \left(\spadesuit \mid \frac{1}{m} \left[\frac{\|\mathbf{w} - \mathbf{v}_t\|_2^2}{2\sigma^2} + \ln \frac{4T\sqrt{m}}{\delta} \right] \right)}_{\text{Corollary 8}},$

where $\spadesuit = \text{kl}^{-1} \left(\frac{1}{n} \sum_{i=1}^n R_{\mathcal{S}}(h_i) \mid \frac{1}{n} \ln \frac{4}{\delta} \right).$

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 $\text{kl}^{-1}(q|\psi)$ is performed by the bisection method. From this new formulation of the bounds, we can remark that the objective is to minimize the function $\text{kl}^{-1}(q|\psi)$ in order to minimize the true risk p . To do so, [Reeb, Doerr, Gerwinn, and Rakitsch \(2018\)](#) 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:

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

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

Note that these partial derivatives need the evaluation of $\text{kl}^{-1}(q|\psi)$ 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 automatically compute the derivatives of kl^{-1} with respect to the parameters.

I.2 Optimization of the bounds

The optimization of the bounds associated with the corollaries are presented in [Algorithm 2](#). This algorithm is divided in two steps: **1**) optimizing and choosing the prior \mathcal{P} (Line 6 to 28); and **2**) optimizing the posterior $\mathcal{Q}_{\mathcal{S}}$ (from Line 32 to 39).

In step **1**), the prior \mathcal{P}_t is obtained after the epoch $t \in \{1, \dots, T\}$ (line 16) by updating ω (parameterizing the prior \mathcal{P}_t) using a mini-batch gradient descent algorithm. For each epoch t and for each mini-batch $\mathcal{U} \subseteq \mathcal{S}_{\text{prior}}$ (Line 8 and 11), we sample a hypothesis h parameterized by $\omega + \epsilon$ (Line 12 and 13) and update ω with the gradient descent algorithm by minimizing the risk $R_{\mathcal{U}}(h)$ (Line 14).

After each epoch t , the prior \mathcal{P} is selected by early stopping on the learning sample \mathcal{S} . We first estimate the risk on \mathcal{S} (Line 19 to 23) by sampling $h \sim \mathcal{P}_t$ (Line 20 and 21) and computing the losses for each mini-batch \mathcal{U} . Then, we select the prior \mathcal{P}_t if it minimizes the risk (Line 24 to 27).

Given the prior \mathcal{P} , we learn a posterior $\mathcal{Q}_{\mathcal{S}}$ in step **2**) during T' epochs. For each epoch and each mini-batch $\mathcal{U} \subseteq \mathcal{S}$, we sample a hypothesis h associated with the weights $\omega + \epsilon$ (Line 38 and 39). At each iteration, the algorithm updates the weights ω (Line 39) by optimizing

$$\underbrace{\text{kl}^{-1}\left(R_{\mathcal{U}}(h) \left| \frac{1}{m} \left[\frac{\|\omega - \mathbf{v}_{t^*}\|_2^2}{\sigma^2} + \ln \frac{16T\sqrt{m}}{\delta^3} \right] \right. \right)}_{\text{Objective function for Corollary 6}}, \quad (\text{I10})$$

$$\underbrace{\text{kl}^{-1}\left(R_{\mathcal{U}}(h) \left| \frac{1}{m} \left[\frac{\|\omega + \epsilon - \mathbf{v}_{t^*}\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \ln \frac{2T\sqrt{m}}{\delta} \right] \right. \right)}_{\text{Objective function for Equation (7)}}, \quad (\text{I11})$$

$$\underbrace{\text{kl}^{-1}\left(R_{\mathcal{U}}(h) \left| \frac{1}{m} \left[\frac{b+1}{b} \left[\frac{\|\omega + \epsilon - \mathbf{v}_{t^*}\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} \right] + \ln \frac{(b+1)T|\mathbf{B}|}{\delta} \right] \right. \right)}_{\text{Objective function for Equation (8)}}, \quad (\text{I12})$$

$$\underbrace{\frac{1}{1 - e^{-c}} \left[1 - \exp \left(-cR_{\mathcal{U}}(h) - \frac{1}{m} \left[\frac{\|\omega + \epsilon - \mathbf{v}_{t^*}\|_2^2 - \|\epsilon\|_2^2}{2\sigma^2} + \ln \frac{T|\mathbf{C}|}{\delta} \right] \right) \right]}_{\text{Objective function for Equation (9)}}. \quad (\text{I13})$$

Note that, as stated in Section 5.3.3, $T' = 1$ for MNIST and FashionMNIST while $T' = 10$ for CIFAR-10 with a batch size of 32. Additionally, the loss is the bounded cross-entropy loss $\ell(h, (\mathbf{x}, y)) = -\frac{1}{Z} \ln(\Phi(h(\mathbf{x})[y]))$ of Dziugaite and Roy (2018) in the risk $R_{\mathcal{U}}(h)$. The update of the weights ω is done with the Adam optimizer (Kingma & Ba, 2015). Concerning the optimization of the hyperparameters $c \in \mathbf{C}$ and $b \in \mathbf{B}$ for Equations (8) and (9), we (a) initialize $b \in \mathbf{B}$ or $c \in \mathbf{C}$ with the one that performs best on the first mini-batch and (b) optimize by gradient descent the hyperparameter. To evaluate Equations (8) and (9), we take $b \in \mathbf{B}$ and $c \in \mathbf{C}$ that leads to the tightest bound.

Algorithm 2 Optimization of the bounds (Training Method)

```

1:           Optimizing the prior  $\mathcal{P}$  — Step 1) — Algorithm  $A_{\text{prior}}$ 
2:
3:  $\omega \leftarrow$  Initialize the weights  $\omega$ 
4:  $r^* \leftarrow +\infty$ 
5:  $t^* \leftarrow +\infty$ 
6: for each epoch  $t \leftarrow 1, \dots, T$  do
7:
8:   Optimizing the prior  $\mathcal{P}_t$ 
9:   for each mini-batch  $\mathcal{U} \subseteq \mathcal{S}_{\text{prior}}$  do
10:    Sample a noise  $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ 
11:     $h \leftarrow$  Hypothesis parameterized by  $\omega + \epsilon$ 
12:     $\omega \leftarrow$  Update  $\omega$  with  $R_{\mathcal{U}}(h)$ 
13:   end for
14:    $\mathcal{P}_t \leftarrow \mathcal{N}(\omega, \sigma^2 \mathbf{I}_d)$  where  $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$ 
15:
16:   Selecting the prior  $\mathcal{P}$ 
17:   for each mini-batch  $\mathcal{U} \subseteq \mathcal{S}$  do
18:    Sample a noise  $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ 
19:     $h \leftarrow$  Hypothesis parameterized by  $\mathbf{v}_t + \epsilon$ 
20:     $r \leftarrow r + \sum_{(\mathbf{x}, y) \in \mathcal{U}} \ell(h, (\mathbf{x}, y))$ 
21:   end for
22:   if  $r < r^*$  then
23:      $r^* \leftarrow r$ 
24:      $\mathcal{P} \leftarrow \mathcal{P}_t$ 
25:      $t^* \leftarrow t$ 
26:   end if
27: end for
28:
29:           Optimizing the posterior  $\mathcal{Q}_S$  — Step 2) — Algorithm  $A$ 
30:
31:  $\mathcal{Q}_S \leftarrow \mathcal{P} = \mathcal{N}(\omega, \sigma^2 \mathbf{I}_d) = \mathcal{N}(\mathbf{v}_{t^*}, \sigma^2 \mathbf{I}_d)$ 
32: for each epoch  $t' \leftarrow 1, \dots, T'$  do
33:   for each mini-batch  $\mathcal{U} \subseteq \mathcal{S}$  do
34:    Sample a noise  $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ 
35:     $h \leftarrow$  Hypothesis parameterized by  $\omega + \epsilon$ 
36:     $\omega \leftarrow$  Update  $\omega$  with either Equation (I10), (I11), (I12), or (I13)
37:   end for
38: end for
39: return  $\mathcal{Q}_S = \mathcal{N}(\omega, \sigma^2 \mathbf{I}_d) = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$  and  $\mathcal{P} = \mathcal{N}(\mathbf{v}_{t^*}, \sigma^2 \mathbf{I}_d)$ 

```

Appendix J 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.

J.1 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 measurable function $\phi : \mathcal{H} \times \mathcal{Z}^m \rightarrow [1, +\infty[$, for any $\delta \in (0, 1]$, for any deterministic algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have*

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

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 $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$

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

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

$$\begin{aligned} &\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \text{KL}(\mathcal{Q}_{\mathcal{S}} \| \mathcal{P}) + \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{Q}_{\mathcal{S}}} \ln \left[\frac{\mathcal{P}(h)}{\mathcal{Q}_{\mathcal{S}}(h)} \phi(h, \mathcal{S}) \right] \\ &\leq \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \text{KL}(\mathcal{Q}_{\mathcal{S}} \| \mathcal{P}) + \ln \left[\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{Q}_{\mathcal{S}}} \frac{\mathcal{P}(h)}{\mathcal{Q}_{\mathcal{S}}(h)} \phi(h, \mathcal{S}) \right] \\ &= \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \text{KL}(\mathcal{Q}_{\mathcal{S}} \| \mathcal{P}) + \ln \left[\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{P}} \phi(h, \mathcal{S}) \right]. \end{aligned} \tag{J14}$$

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

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} \left[\ln \phi(h, \mathcal{S}) \leq \frac{1}{\delta} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \mathbb{E}_{h \sim \mathcal{Q}_{\mathcal{S}}} \ln \phi(h, \mathcal{S}) \right] \geq 1 - \delta. \tag{J15}$$

Then, from Equations (J14) and (J15), we can deduce the stated result. \square

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 measurable function $\phi : \mathcal{H} \times \mathcal{Z}^m \rightarrow [1, +\infty[$, for any $\delta \in (0, 1]$, for any deterministic algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have

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

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

Proof Note that the mutual information is defined by $I(h; \mathcal{S}) = \min_{\mathcal{P} \in \mathcal{M}^*(\mathcal{H})} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \text{KL}(\mathcal{Q}_{\mathcal{S}} \| \mathcal{P})$. Hence, to prove Theorem 13, we have to instantiate Lemma 12 with the optimal prior, i.e., the prior \mathcal{P} which minimizes $\mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} \text{KL}(\mathcal{Q}_{\mathcal{S}} \| \mathcal{P})$. The optimal prior is well known (see, e.g., [Catoni, 2007](#); [Lever, Laviolette, & Shawe-Taylor, 2013](#)): for the sake of completeness, we derive it. First, we have

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

Hence,

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

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 probability at least $1 - \delta$ over $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{Q}_{\mathcal{S}}$

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

□

As you can remark, this bound is looser than Theorem 9, which is based on Sibson's mutual information. For example, when we instantiate this bound with $\phi(h, \mathcal{S}) = \exp[m \text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h))]$, 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).

J.2 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 measurable function $\phi : \mathcal{H} \times \mathcal{Z}^m$, for any $\alpha > 1$, for any $\delta \in (0, 1]$, for any deterministic algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have*

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

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

Proof Note that $\phi(h, \mathcal{S})$ is a non-negative random variable. From Markov's inequality, we have

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

Then, since both sides of the inequality are strictly positive, we take the logarithm to both sides of the equality and multiply by $\frac{\alpha}{\alpha-1} > 0$ to obtain

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

We develop the right-hand side of the inequality in the indicator function and make the expectation of the hypothesis over the distribution \mathcal{P} appear. We have for all priors $\mathcal{P} \in \mathcal{M}^*(\mathcal{H})$,

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

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

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

Since both sides of the inequality are positive, we take the logarithm. Moreover, we add $\ln(\frac{1}{\delta})$, and we multiply by $\frac{\alpha}{\alpha-1} > 0$ to both sides of the inequality. We have

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

Hence, we can deduce that

$$\mathbb{P}_{\mathcal{S} \sim \mathcal{D}^m, h \sim \mathcal{Q}_{\mathcal{S}}} \left[\forall \mathcal{P} \in \mathcal{M}^*(\mathcal{H}), \frac{\alpha}{\alpha-1} \ln(\phi(h, \mathcal{S})) \leq \frac{1}{\alpha-1} \ln \left(\mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} \left(\left[\frac{\mathcal{Q}_{\mathcal{S}'}(h')}{\mathcal{P}(h')} \right]^{\alpha} \right) \right) \right]$$

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

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

From Lemma 14, we prove Theorem 9.

Theorem 9 (Disintegrated Information-Theoretic Bound) *For any distribution \mathcal{D} on \mathcal{Z} , for any hypothesis set \mathcal{H} , for any measurable function $\phi : \mathcal{H} \times \mathcal{Z}^m \rightarrow \mathbb{R}_+^*$, for any $\alpha > 1$, for any $\delta \in (0, 1]$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, we have*

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

Proof Note that Sibson's mutual information is defined as $I_\alpha(h; S) = \min_{\mathcal{P} \in \mathcal{M}^*(\mathcal{H})} D_\alpha(\rho \parallel \pi)$. 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 $D_\alpha(\rho \parallel \pi)$. Actually, this optimal prior has a closed-form solution (Verdú, 2015). For the sake of completeness, we derive it. First, we have

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

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

From these equalities and using the fact that $D_\alpha(\mathcal{P}^* \parallel \mathcal{P})$ is minimal (*i.e.*, equal to zero) when $\mathcal{P}^* = \mathcal{P}$, we can deduce that

$$\operatorname{argmin}_{\mathcal{P} \in \mathcal{M}^*(\mathcal{H})} D_\alpha(\rho \parallel \pi)$$

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

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 probability at least $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{Q}_{\mathcal{S}}$

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

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

J.3 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 $\alpha \rightarrow 1^+$, with probability at least $1-\delta$ we have*

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

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

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

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

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

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

When $\alpha \rightarrow 1$.

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

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

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

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

Hence, we have

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

Finally, we can deduce that

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

When $\alpha \rightarrow +\infty$.

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

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

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

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

where $\|\gamma\|_{\alpha}$ is the L^{α} -norm of the function γ defined as $\gamma(h) = \frac{\mathcal{Q}_S(h)}{\mathcal{P}^*(h)}$. We have

$$\lim_{\alpha \rightarrow +\infty} \frac{\alpha-1}{\alpha} I_{\alpha}(h'; S') = \lim_{\alpha \rightarrow +\infty} \ln (\|\gamma\|_{\alpha}) = \ln \left(\lim_{\alpha \rightarrow +\infty} \|\gamma\|_{\alpha} \right)$$

$$= \ln(\|\gamma\|_\infty) = \ln\left(\operatorname{esssup}_{\mathcal{S} \in \mathcal{S}, h \in \mathcal{H}} \gamma(h)\right) = \ln\left(\operatorname{esssup}_{\mathcal{S} \in \mathcal{S}, h \in \mathcal{H}} \frac{\mathcal{Q}_{\mathcal{S}}(h)}{\mathcal{P}^*(h)}\right).$$

Finally, we can deduce that

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

□

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

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

$$2(R_{\mathcal{S}}(h) - R_{\mathcal{D}}(h))^2 \leq \frac{1}{m} \left[I_\alpha(h'; \mathcal{S}') + \ln \frac{2}{\delta^{\frac{\alpha}{\alpha-1}}} \right]. \quad (\text{J16})$$

Hence, we compare Equation (J16) 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 $\alpha > 1$, for any $\delta \in (0, 1]$, for any algorithm $A: \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \rightarrow \mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^m$ and $h \sim \mathcal{Q}_{\mathcal{S}}$, we have

$$\text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) \leq \frac{1}{m} \left[I_\alpha(h'; \mathcal{S}') + \ln \frac{2\sqrt{m}}{\delta^{\frac{\alpha}{\alpha-1}}} \right] \quad (\text{J17})$$

$$\text{and } 2(R_{\mathcal{S}}(h) - R_{\mathcal{D}}(h))^2 \leq \frac{1}{m} \left[I_\alpha(h'; \mathcal{S}') + \ln \frac{2\sqrt{m}}{\delta^{\frac{\alpha}{\alpha-1}}} \right]. \quad (\text{J18})$$

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

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

Then, from Maurer (2004), we upper-bound $\mathbb{E}_{\mathcal{S}' \sim \mathcal{D}^m} \mathbb{E}_{h' \sim \mathcal{P}} e^{m \text{kl}(R_{\mathcal{S}'}(h') \| R_{\mathcal{D}}(h'))}$ by $2\sqrt{m}$ to obtain Equation (J17). Finally, to obtain Equation (J18), we apply Pinsker’s inequality, i.e., $2(R_{\mathcal{S}}(h) - R_{\mathcal{D}}(h))^2 \leq \text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h))$ on Equation (J17). □

Equation (J18) is slightly looser than Equation (J16) since it involves an extra term of $\frac{1}{m} \ln \sqrt{m}$. However, Equation (J17) is tighter than Equation (J16) when $\text{kl}(R_{\mathcal{S}}(h) \| R_{\mathcal{D}}(h)) - 2(R_{\mathcal{S}}(h) - R_{\mathcal{D}}(h))^2 \geq \frac{1}{m} \ln \sqrt{m}$ (which becomes more frequent as m grows).

Appendix K Results presented in Section 5

This appendix presents the details of the results of Section 5. Tables K2 to K10 report empirical results for split ratios going from 0.0 to 0.9 presented in Figures 1 to 5. More precisely, we report the test risk $R_{\mathcal{T}}(h)$, the empirical risk $R_{\mathcal{S}}(h)$, the bound value (Bnd), and the divergence value associated with the network h sampled from the posterior $\mathcal{Q}_{\mathcal{S}}$ for each learning rate, variance, dataset, and bound type. Tables K11 to K13 report the performances of the prior before applying Step 2) outlined in Figures 4 and 5. In particular, we report the test risk $R_{\mathcal{T}}(h)$, the empirical risk $R_{\mathcal{S}}(h)$, the bound values of Corollary 6 and Equations (7), (8), (9) for each split ratio and variance.

Note that for the split 0.0, since Step 1) is skipped, the prior distribution \mathcal{P} is only initialized as introduced in Section 5.3.2. Note that in this case, $T = 1$ since we have only one prior. To do the same number of epochs compared to the other splits, we perform 11 epochs (instead of 1) for MNIST and Fashion-MNIST and 110 epochs (instead of 10) for CIFAR-10 during Step 2). The other parameters are not changed.

Table K2: Comparison of OURS, RIVASPLATA, BLANCHARD and CATONI based on the disintegrated bounds, and STOCHASTIC based on the randomized bounds learned with two learning rates $\text{lr} \in \{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_S(h)$), 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). More precisely, we report the mean \pm the standard deviation for 400 neural networks sampled from \mathcal{Q}_S for OURS, RIVASPLATA, BLANCHARD, and CATONI. We consider, in this table, that the split ratio is 0.1.

	$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div
h=10⁻⁶												
OURS	.035 ± .000	.044 ± .000	.039 ± .000	.022	.024 ± .000	.034 ± .000	.029 ± .000	2.122	.029 ± .002	.040 ± .002	.034 ± .002	12.754
BLANCHARD	.034 ± .000	.058 ± .002	.038 ± .000	90.876 ± 14.658	.024 ± .000	.033 ± .001	.030 ± .000	21.775 ± 6.848	.034 ± .002	.043 ± .002	.038 ± .002	3.949 ± 2.877
CATONI	.035 ± .000	.064 ± .001	.039 ± .000	139.665 ± 15.854	.024 ± .000	.036 ± .001	.030 ± .000	26.277 ± 7.490	.033 ± .002	.047 ± .002	.037 ± .002	4.067 ± 2.852
RIVASPLATA	.034 ± .000	.082 ± .001	.038 ± .000	104.880 ± 15.268	.024 ± .000	.036 ± .001	.029 ± .000	25.007 ± 7.187	.033 ± .002	.042 ± .002	.037 ± .002	4.06 ± 2.845
STOCHASTIC	—	.086	—	.311	—	.067	—	1.061	—	.074	—	6.377
Fashion												
OURS	.166 ± .001	.169 ± .000	.159 ± .000	.580	.157 ± .001	.160 ± .001	.150 ± .001	2.128	.160 ± .002	.161 ± .003	.151 ± .002	3.503
BLANCHARD	.165 ± .001	.192 ± .002	.159 ± .000	96.822 ± 14.116	.157 ± .001	.166 ± .002	.150 ± .001	21.592 ± 6.081	.163 ± .003	.162 ± .003	.159 ± .003	3.846 ± 2.660
CATONI	.165 ± .001	.190 ± .003	.159 ± .000	119.027 ± 15.938	.157 ± .001	.169 ± .002	.150 ± .001	26.363 ± 7.355	.162 ± .003	.161 ± .003	.159 ± .003	4.152 ± 2.945
RIVASPLATA	.165 ± .001	.183 ± .002	.158 ± .000	101.954 ± 14.463	.157 ± .001	.169 ± .002	.150 ± .001	23.098 ± 6.977	.162 ± .003	.161 ± .003	.159 ± .003	3.852 ± 2.798
STOCHASTIC	—	.227	—	.290	—	.216	—	1.064	—	.218	—	1.751
CEM4-10												
OURS	.479 ± .000	.487 ± .000	.479 ± .000	.052	.479 ± .000	.498 ± .000	.477 ± .000	.065	.458 ± .001	.479 ± .000	.468 ± .000	.299
BLANCHARD	.479 ± .000	.550 ± .003	.479 ± .000	27.644 ± 22.868	.479 ± .000	.529 ± .003	.477 ± .000	85.476 ± 12.781	.458 ± .001	.489 ± .003	.463 ± .000	24.608 ± 7.136
CATONI	.479 ± .000	.546 ± .005	.479 ± .000	269.855 ± 22.883	.479 ± .000	.511 ± .003	.477 ± .000	85.113 ± 12.806	.458 ± .001	.483 ± .002	.463 ± .000	25.453 ± 7.155
RIVASPLATA	.479 ± .000	.528 ± .002	.479 ± .000	27.588 ± 22.859	.479 ± .000	.511 ± .002	.477 ± .000	85.745 ± 13.357	.458 ± .001	.484 ± .002	.463 ± .001	25.051 ± 7.005
STOCHASTIC	—	.558	—	.026	—	.564	—	.032	—	.550	—	.150
h=10⁻⁴												
OURS	.035 ± .000	.048 ± .000	.039 ± .000	35.348	.024 ± .000	.037 ± .001	.029 ± .000	3.753	.022 ± .001	.042 ± .001	.027 ± .001	153.773
BLANCHARD	.032 ± .000	.442 ± .003	.039 ± .000	181.362 ± 14.449	.022 ± .000	.290 ± .003	.027 ± .000	3831.110 ± 84.274	.019 ± .001	.102 ± .002	.029 ± .001	1306.371 ± 51.396
CATONI	.035 ± .000	.362 ± .003	.039 ± .000	139.55734 ± 145.511	.024 ± .000	.134 ± .002	.029 ± .000	3841.248 ± 84.033	.027 ± .002	.084 ± .002	.032 ± .001	1535.287 ± 49.454
RIVASPLATA	.030 ± .000	.489 ± .002	.034 ± .000	142.23716 ± 151.137	.021 ± .000	.069 ± .009	.026 ± .000	86.836 ± 85.146	.018 ± .000	.072 ± .001	.022 ± .000	151.609 ± 93.400
STOCHASTIC	—	.084	—	17.674	—	.069	—	15.376	—	.072	—	76.887
Fashion												
OURS	.166 ± .001	.172 ± .000	.159 ± .000	13.884	.157 ± .001	.161 ± .001	.151 ± .001	16.513	.159 ± .002	.161 ± .002	.149 ± .002	2.344
BLANCHARD	.166 ± .001	.588 ± .003	.159 ± .000	1089.829 ± 137.125	.150 ± .001	.373 ± .003	.141 ± .001	37.441 ± 83.656	.155 ± .002	.174 ± .003	.140 ± .002	1221.062 ± 49.548
CATONI	.165 ± .001	.500 ± .003	.159 ± .000	11954.591 ± 141.463	.156 ± .001	.311 ± .002	.148 ± .001	3826.848 ± 86.111	.158 ± .002	.248 ± .003	.148 ± .002	1226.282 ± 5.332
RIVASPLATA	.158 ± .001	.459 ± .002	.151 ± .000	13541.128 ± 14.706	.149 ± .001	.302 ± .002	.140 ± .001	3878.145 ± 85.782	.154 ± .002	.230 ± .002	.144 ± .001	1244.035 ± 49.268
STOCHASTIC	—	.229	—	6.542	—	.219	—	8.257	—	.219	—	1.172
CEM4-10												
OURS	.479 ± .000	.489 ± .000	.479 ± .000	.482	.479 ± .000	.498 ± .000	.477 ± .000	.923	.458 ± .001	.480 ± .000	.468 ± .000	.498
BLANCHARD	.479 ± .000	.957 ± .001	.471 ± .000	2220.1035 ± 218.369	.479 ± .000	.854 ± .002	.477 ± .000	8766.571 ± 126.716	.457 ± .001	.699 ± .003	.461 ± .000	2758.075 ± 77.155
CATONI	.479 ± .000	.995 ± .001	.471 ± .000	26347.753 ± 225.908	.479 ± .000	.771 ± .002	.477 ± .000	8569.272 ± 124.834	.455 ± .001	.650 ± .002	.459 ± .000	3117.566 ± 75.178
RIVASPLATA	.479 ± .000	.915 ± .001	.471 ± .000	29489.241 ± 241.010	.479 ± .000	.769 ± .002	.477 ± .000	867.284 ± 126.038	.456 ± .001	.633 ± .002	.460 ± .000	2776.052 ± 72.901
STOCHASTIC	—	.559	—	2.441	—	.566	—	.437	—	.551	—	2.494

Table K3: Comparison of OURS, RIVASPLATA, BLANCHARD and CATONI based on the disintegrated bounds, and STOCHASTIC based on the randomized bounds learned with two learning rates $\text{lr} \in \{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_S(h)$), 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). More precisely, we report the mean \pm the standard deviation for 400 neural networks sampled from \mathcal{Q}_S for OURS, RIVASPLATA, BLANCHARD, and CATONI. We consider, in this table, that the split ratio is 0.2.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
		$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div
MNIST	OURS	.016 ± .000	.023 ± .000	.019 ± .000	.748	.014 ± .001	.020 ± .001	.016 ± .000	2.096	.019 ± .002	.024 ± .002	.020 ± .002	2.244
	BLANCHARD	.016 ± .000	.034 ± .001	.019 ± .000	97.590 ± 14.590	.015 ± .000	.029 ± .001	.019 ± .000	21.135 ± 6.514	.020 ± .002	.024 ± .002	.021 ± .002	3.71 ± .875
	CATONI	.016 ± .000	.034 ± .001	.019 ± .000	116.744 ± 15.447	.015 ± .000	.027 ± .002	.019 ± .000	23.135 ± 7.075	.020 ± .002	.024 ± .002	.021 ± .002	.410 ± .880
	RIVASPLATA	.016 ± .000	.029 ± .001	.019 ± .000	103.854 ± 14.728	.015 ± .000	.024 ± .001	.019 ± .000	24.063 ± 6.663	.020 ± .002	.024 ± .002	.021 ± .002	14.22
	STOCHASTIC	—	.092	—	.108	—	.047	—	.374	—	.063	—	1.048
Fashion	OURS	.165 ± .002	.169 ± .001	.157 ± .001	4.811	.145 ± .005	.153 ± .006	.139 ± .005	15.453	.160 ± .005	.166 ± .005	.155 ± .005	1.633
	BLANCHARD	.163 ± .002	.190 ± .003	.155 ± .001	96.264 ± 14.472	.152 ± .003	.165 ± .002	.147 ± .003	21.099 ± 6.507	.163 ± .006	.165 ± .006	.158 ± .006	.340 ± .885
	CATONI	.163 ± .002	.190 ± .004	.156 ± .001	121.542 ± 16.499	.150 ± .002	.158 ± .003	.144 ± .002	27.241 ± 7.318	.162 ± .005	.165 ± .005	.157 ± .005	.444 ± .968
	RIVASPLATA	.161 ± .001	.180 ± .002	.153 ± .001	106.403 ± 14.044	.150 ± .002	.158 ± .003	.145 ± .003	23.134 ± 7.064	.162 ± .006	.165 ± .005	.157 ± .005	.417 ± .928
	STOCHASTIC	—	.226	—	2.405	—	.207	—	5.428	—	.223	—	.816
CIFAR-10	OURS	.300 ± .000	.407 ± .000	.391 ± .000	0.040	.306 ± .001	.411 ± .000	.395 ± .000	1.555	.416 ± .002	.432 ± .002	.415 ± .001	.970
	BLANCHARD	.300 ± .000	.473 ± .004	.391 ± .000	271.616 ± 23.555	.404 ± .000	.445 ± .003	.398 ± .000	84.868 ± 13.050	.416 ± .002	.432 ± .002	.410 ± .001	4.496 ± 3.018
	CATONI	.300 ± .000	.473 ± .005	.391 ± .000	27.502 ± 23.371	.404 ± .000	.434 ± .003	.398 ± .000	84.848 ± 12.992	.415 ± .002	.431 ± .001	.415 ± .001	4.859 ± 3.176
	RIVASPLATA	.300 ± .000	.450 ± .002	.391 ± .000	271.700 ± 23.586	.403 ± .000	.433 ± .002	.398 ± .000	85.027 ± 13.047	.416 ± .002	.431 ± .002	.410 ± .001	4.610 ± 3.084
	STOCHASTIC	—	.477	—	.020	—	.482	—	.077	—	.503	—	.485
MNIST	OURS	.016 ± .000	.025 ± .000	.019 ± .000	14.490	.015 ± .000	.024 ± .000	.019 ± .000	8.583	.016 ± .001	.023 ± .001	.017 ± .001	25.556
	BLANCHARD	.016 ± .000	.430 ± .001	.018 ± .000	11465.062 ± 153.554	.014 ± .000	.290 ± .003	.018 ± .000	3799.912 ± 89.585	.016 ± .001	.019 ± .001	.016 ± .000	13.055
	CATONI	.016 ± .000	.435 ± .002	.019 ± .000	11564.106 ± 15.709	.015 ± .000	.149 ± .003	.019 ± .000	3528.942 ± 83.357	.016 ± .001	.019 ± .001	.016 ± .000	187.859 ± 48.790
	RIVASPLATA	.015 ± .000	.402 ± .002	.018 ± .000	11465.063 ± 163.364	.015 ± .000	.149 ± .003	.017 ± .000	4949.440 ± 82.512	.016 ± .001	.019 ± .001	.016 ± .000	248.708 ± 48.314
	STOCHASTIC	—	.085	—	7.246	—	.052	—	4.292	—	.068	—	6.528
Fashion	OURS	.165 ± .002	.172 ± .001	.157 ± .001	28.705	.141 ± .002	.156 ± .002	.137 ± .002	52.796	.131 ± .003	.147 ± .003	.126 ± .003	7.515
	BLANCHARD	.136 ± .001	.598 ± .003	.130 ± .001	11334.327 ± 145.083	.125 ± .001	.375 ± .003	.121 ± .001	3998.068 ± 88.992	.124 ± .001	.147 ± .003	.117 ± .001	126.184 ± 48.814
	CATONI	.162 ± .001	.525 ± .004	.154 ± .001	11965.068 ± 152.681	.141 ± .002	.380 ± .003	.137 ± .002	384.802 ± 84.123	.132 ± .003	.124 ± .004	.127 ± .002	1239.918 ± 49.594
	RIVASPLATA	.131 ± .001	.455 ± .002	.127 ± .001	1195.209 ± 155.390	.123 ± .001	.209 ± .002	.119 ± .001	4065.169 ± 89.793	.123 ± .003	.124 ± .002	.116 ± .001	1294.726 ± 49.874
	STOCHASTIC	—	.228	—	11.853	—	.209	—	26.368	—	.198	—	35.258
CIFAR-10	OURS	.300 ± .000	.411 ± .000	.391 ± .000	13.286	.306 ± .001	.414 ± .000	.398 ± .000	3.305	.415 ± .001	.433 ± .001	.415 ± .001	6.064
	BLANCHARD	.389 ± .000	.990 ± .000	.391 ± .000	74.247.764 ± 397.521	.403 ± .000	.829 ± .002	.397 ± .000	8851.324 ± 126.764	.405 ± .001	.621 ± .003	.394 ± .000	2738.066 ± 75.053
	CATONI	.300 ± .000	.990 ± .000	.391 ± .000	26.434.787 ± 228.500	.403 ± .000	.724 ± .003	.397 ± .000	8651.380 ± 50.413	.404 ± .001	.620 ± .003	.399 ± .000	4178.302 ± 9.315
	RIVASPLATA	.389 ± .000	.902 ± .001	.391 ± .000	61.647	.403 ± .000	.715 ± .002	.397 ± .000	897.893 ± 138.239	.394 ± .001	.578 ± .003	.395 ± .000	2741.257 ± 74.942
	STOCHASTIC	—	.480	—	6.643	—	.486	—	1.053	—	.483	—	1.568

Table K4: Comparison of OURS, RIVASPLATA, BLANCHARD and CATONI based on the disintegrated bounds, and STOCHASTIC based on the randomized bounds learned with two learning rates $\text{lr} \in \{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_S(h)$), the empirical risk ($R_S(h)$), 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). More precisely, we report the mean \pm the standard deviation for 400 neural networks sampled from \mathcal{Q}_S for OURS, RIVASPLATA, BLANCHARD, and CATONI. We consider, in this table, that the split ratio is 0.3.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
		$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div
MNIST	OURS	.012 \pm .000	.017 \pm .000	.013 \pm .000	.181	.009 \pm .000	.021 \pm .000	.011 \pm .000	.155	.012 \pm .000	.020 \pm .000	.016 \pm .000	1.655
	BLANCHARD	.012 \pm .000	.027 \pm .001	.013 \pm .000	93.915 \pm 14.110	.012 \pm .000	.023 \pm .001	.014 \pm .000	19.292 \pm 6.037	.012 \pm .000	.020 \pm .001	.015 \pm .000	3.023 \pm 2.430
	CATONI	.012 \pm .000	.025 \pm .001	.013 \pm .000	133.574 \pm 15.436	.012 \pm .000	.023 \pm .001	.014 \pm .000	22.347 \pm 6.377	.012 \pm .000	.020 \pm .001	.015 \pm .000	2.318 \pm 2.341
	RIVASPLATA	.012 \pm .000	.023 \pm .001	.013 \pm .000	96.392 \pm 14.330	.012 \pm .000	.023 \pm .001	.014 \pm .000	19.995 \pm 6.254	.012 \pm .000	.020 \pm .001	.015 \pm .000	3.024 \pm 2.446
	STOCHASTIC	—	.042	—	.091	—	.049	—	.077	—	.047	—	.082
Fashion	OURS	.126 \pm .000	.134 \pm .000	.124 \pm .000	.328	.126 \pm .001	.130 \pm .001	.119 \pm .001	1.092	.122 \pm .002	.126 \pm .002	.115 \pm .002	4.617
	BLANCHARD	.126 \pm .000	.157 \pm .003	.124 \pm .000	88.034 \pm 13.485	.126 \pm .001	.136 \pm .002	.120 \pm .001	18.852 \pm 6.115	.124 \pm .002	.127 \pm .002	.118 \pm .002	3.014 \pm 2.395
	CATONI	.126 \pm .000	.159 \pm .004	.124 \pm .000	114.259 \pm 15.300	.126 \pm .001	.139 \pm .002	.120 \pm .001	22.607 \pm 6.871	.124 \pm .002	.126 \pm .002	.118 \pm .002	3.109 \pm 2.513
	RIVASPLATA	.126 \pm .000	.148 \pm .002	.124 \pm .000	93.107 \pm 13.630	.126 \pm .001	.133 \pm .002	.120 \pm .001	19.724 \pm 6.320	.124 \pm .002	.126 \pm .002	.118 \pm .002	2.980 \pm 2.451
	STOCHASTIC	—	.187	—	.164	—	.182	—	.846	—	.178	—	2.309
CIFAR-10	OURS	.369 \pm .000	.375 \pm .000	.358 \pm .000	.028	.351 \pm .000	.368 \pm .000	.362 \pm .000	.041	.359 \pm .001	.377 \pm .000	.360 \pm .000	183
	BLANCHARD	.369 \pm .000	.446 \pm .004	.358 \pm .000	269.780 \pm 22.724	.351 \pm .000	.401 \pm .004	.362 \pm .000	84.113 \pm 12.530	.359 \pm .001	.388 \pm .003	.360 \pm .000	22.878 \pm 6.728
	CATONI	.369 \pm .000	.450 \pm .007	.358 \pm .000	269.843 \pm 24.225	.351 \pm .000	.390 \pm .004	.362 \pm .000	84.500 \pm 12.608	.359 \pm .001	.381 \pm .002	.360 \pm .000	23.567 \pm 7.181
	RIVASPLATA	.369 \pm .000	.421 \pm .003	.358 \pm .000	27.224 \pm 24.187	.351 \pm .000	.388 \pm .002	.362 \pm .000	84.250 \pm 13.274	.359 \pm .001	.382 \pm .002	.360 \pm .000	23.033 \pm 6.724
	STOCHASTIC	—	.445	—	.014	—	.438	—	.020	—	.447	—	.092
MNIST	OURS	.012 \pm .000	.019 \pm .000	.013 \pm .000	24.837	.012 \pm .000	.020 \pm .000	.014 \pm .000	12.358	.012 \pm .000	.021 \pm .000	.015 \pm .000	13.908
	BLANCHARD	.012 \pm .000	.407 \pm .004	.013 \pm .000	1183.223 \pm 154.992	.011 \pm .000	.211 \pm .003	.014 \pm .000	388.931 \pm 86.014	.010 \pm .000	.040 \pm .003	.014 \pm .000	121.397 \pm 51.944
	CATONI	.012 \pm .000	.339 \pm .002	.013 \pm .000	136.394 \pm 15.794	.012 \pm .000	.139 \pm .003	.014 \pm .000	35.839 \pm 8.155	.012 \pm .000	.070 \pm .002	.019 \pm .000	1215.305 \pm 51.783
	RIVASPLATA	.012 \pm .000	.409 \pm .003	.013 \pm .000	143.897 \pm 102.759	.011 \pm .000	.143 \pm .002	.014 \pm .000	40.739 \pm 9.947	.010 \pm .000	.041 \pm .001	.013 \pm .000	1215.638 \pm 49.952
	STOCHASTIC	—	.044	—	.12418	—	.046	—	.6179	—	.047	—	63.94
Fashion	OURS	.126 \pm .000	.137 \pm .000	.124 \pm .000	12.401	.125 \pm .001	.133 \pm .001	.119 \pm .001	14.631	.120 \pm .002	.128 \pm .002	.113 \pm .001	26.489
	BLANCHARD	.126 \pm .000	.692 \pm .003	.124 \pm .000	10558.872 \pm 130.107	.119 \pm .001	.383 \pm .004	.112 \pm .001	383.091 \pm 86.176	.113 \pm .001	.239 \pm .003	.106 \pm .001	1204.214 \pm 5.815
	CATONI	.126 \pm .000	.581 \pm .004	.124 \pm .000	11966.223 \pm 148.195	.125 \pm .001	.299 \pm .003	.118 \pm .001	382.806 \pm 85.864	.119 \pm .002	.209 \pm .002	.113 \pm .001	1225.310 \pm 48.900
	RIVASPLATA	.123 \pm .000	.458 \pm .003	.120 \pm .000	1209.156 \pm 143.319	.118 \pm .001	.287 \pm .002	.111 \pm .001	381.804 \pm 85.091	.112 \pm .001	.196 \pm .002	.105 \pm .001	126.956 \pm 49.255
	STOCHASTIC	—	.189	—	6.290	—	.184	—	7.316	—	.184	—	13.250
CIFAR-10	OURS	.369 \pm .000	.379 \pm .000	.358 \pm .000	11.657	.351 \pm .000	.369 \pm .000	.362 \pm .000	2.907	.359 \pm .001	.378 \pm .000	.360 \pm .000	2.616
	BLANCHARD	.369 \pm .000	.990 \pm .000	.358 \pm .000	40152.074 \pm 291.721	.351 \pm .000	.709 \pm .003	.351 \pm .000	8753.816 \pm 136.801	.358 \pm .001	.675 \pm .004	.359 \pm .000	2728.436 \pm 73.835
	CATONI	.369 \pm .000	.986 \pm .000	.358 \pm .000	24477.984 \pm 223.367	.351 \pm .000	.803 \pm .003	.351 \pm .000	8463.452 \pm 136.001	.357 \pm .001	.678 \pm .002	.357 \pm .000	3401.221 \pm 84.878
	RIVASPLATA	.369 \pm .000	.868 \pm .001	.358 \pm .000	24424.968 \pm 223.601	.351 \pm .000	.694 \pm .002	.351 \pm .000	8665.339 \pm 136.361	.358 \pm .001	.555 \pm .003	.358 \pm .000	274.051 \pm 74.784
	STOCHASTIC	—	.448	—	5.829	—	.439	—	1.134	—	.448	—	1.308

Table K5: Comparison of OURS, RIVASPLATA, BLANCHARD and CATONI based on the disintegrated bounds, and STOCHASTIC based on the randomized bounds learned with two learning rates $\text{lr} \in \{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_S(h)$), 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). More precisely, we report the mean \pm the standard deviation for 400 neural networks sampled from \mathcal{Q}_S for OURS, RIVASPLATA, BLANCHARD, and CATONI. We consider, in this table, that the split ratio is 0.4.

	$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$				
	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	
MNIST	OURS	.010 \pm .000	.017 \pm .000	.013 \pm .000	.134	.012 \pm .000	.015 \pm .000	.011 \pm .000	.255	.014 \pm .001	.020 \pm .001	.015 \pm .001	1.111
	BLANCHARD	.010 \pm .000	.028 \pm .001	.013 \pm .000	83.323 \pm 13.740	.012 \pm .000	.021 \pm .001	.014 \pm .000	16.799 \pm 5.702	.009 \pm .000	.014 \pm .001	.011 \pm .000	2.449 \pm 2.313
	CATONI	.010 \pm .000	.026 \pm .001	.013 \pm .000	109.202 \pm 15.654	.012 \pm .000	.020 \pm .002	.014 \pm .000	19.918 \pm 6.526	.009 \pm .000	.015 \pm .001	.011 \pm .000	2.486 \pm 2.362
	RIVASPLATA	.010 \pm .000	.024 \pm .001	.013 \pm .000	96.572 \pm 14.470	.012 \pm .000	.019 \pm .001	.014 \pm .000	16.092 \pm 5.882	.009 \pm .000	.014 \pm .000	.011 \pm .000	2.329 \pm 2.251
	STOCHASTIC	—	.046	—	.001	.012 \pm .000	.014	—	.009	—	.047	—	.117
Fashion	OURS	.118 \pm .001	.123 \pm .000	.112 \pm .000	.269	.113 \pm .001	.118 \pm .001	.107 \pm .001	.743	.117 \pm .002	.121 \pm .002	.110 \pm .002	2.600
	BLANCHARD	.118 \pm .001	.145 \pm .003	.112 \pm .000	82.403 \pm 13.230	.113 \pm .001	.123 \pm .002	.107 \pm .001	16.836 \pm 5.583	.119 \pm .002	.121 \pm .003	.112 \pm .003	2.644 \pm 2.369
	CATONI	.118 \pm .001	.151 \pm .004	.112 \pm .000	109.988 \pm 15.347	.113 \pm .001	.129 \pm .002	.107 \pm .001	19.889 \pm 6.689	.118 \pm .002	.120 \pm .003	.112 \pm .003	2.615 \pm 2.234
	RIVASPLATA	.118 \pm .001	.137 \pm .002	.112 \pm .000	87.804 \pm 13.640	.113 \pm .001	.129 \pm .002	.107 \pm .001	17.491 \pm 6.144	.118 \pm .002	.121 \pm .003	.112 \pm .003	2.549 \pm 2.175
	STOCHASTIC	—	.174	—	.135	.113 \pm .001	.168	—	.372	—	.172	—	1.300
CIFAR-10	OURS	.334 \pm .000	.346 \pm .000	.328 \pm .000	.025	.322 \pm .000	.334 \pm .000	.313 \pm .000	.050	.323 \pm .001	.334 \pm .000	.316 \pm .000	.160
	BLANCHARD	.334 \pm .000	.421 \pm .004	.328 \pm .000	269.875 \pm 23.982	.322 \pm .000	.364 \pm .004	.313 \pm .000	83.082 \pm 13.029	.323 \pm .001	.345 \pm .004	.316 \pm .000	21.614 \pm 6.670
	CATONI	.334 \pm .000	.433 \pm .008	.328 \pm .000	27.270 \pm 24.010	.322 \pm .000	.355 \pm .005	.313 \pm .000	84.148 \pm 13.578	.323 \pm .001	.339 \pm .002	.316 \pm .000	22.547 \pm 6.801
	RIVASPLATA	.334 \pm .000	.394 \pm .003	.328 \pm .000	27.133 \pm 24.109	.322 \pm .000	.351 \pm .003	.313 \pm .000	83.438 \pm 13.033	.323 \pm .001	.339 \pm .002	.316 \pm .000	21.688 \pm 6.718
	STOCHASTIC	—	.414	—	.013	.322 \pm .000	.389	—	.025	—	.403	—	.080
MNIST	OURS	.010 \pm .000	.019 \pm .000	.013 \pm .000	23.992	.012 \pm .000	.019 \pm .000	.014 \pm .000	7.767	.009 \pm .000	.015 \pm .000	.011 \pm .000	3.165
	BLANCHARD	.010 \pm .000	.009 \pm .001	.013 \pm .000	123.328 \pm 15.115	.012 \pm .000	.020 \pm .004	.014 \pm .000	38.840 \pm 94.218	.009 \pm .000	.006 \pm .003	.011 \pm .000	134.214 \pm 47.208
	CATONI	.010 \pm .000	.009 \pm .003	.013 \pm .000	191.598 \pm 194.180	.012 \pm .000	.019 \pm .003	.014 \pm .000	35.263 \pm 85.362	.009 \pm .000	.070 \pm .002	.011 \pm .000	121.772 \pm 49.984
	RIVASPLATA	.010 \pm .000	.016 \pm .003	.013 \pm .000	167.763 \pm 151.498	.012 \pm .000	.014 \pm .002	.014 \pm .000	34.136 \pm 84.542	.009 \pm .000	.015 \pm .001	.012 \pm .000	131.106 \pm 46.583
	STOCHASTIC	—	.046	—	11.996	.012 \pm .000	.046	—	.3884	—	.040	—	1.583
Fashion	OURS	.115 \pm .000	.127 \pm .000	.112 \pm .000	17.987	.113 \pm .001	.118 \pm .001	.107 \pm .001	6.361	.114 \pm .002	.123 \pm .002	.107 \pm .002	22.582
	BLANCHARD	.115 \pm .000	.169 \pm .004	.112 \pm .000	183.578 \pm 161.816	.110 \pm .001	.139 \pm .004	.104 \pm .000	32.856 \pm 94.279	.108 \pm .001	.124 \pm .004	.102 \pm .001	118.882 \pm 5.575
	CATONI	.115 \pm .001	.166 \pm .004	.112 \pm .000	1192.111 \pm 153.739	.113 \pm .001	.181 \pm .003	.107 \pm .000	38.224 \pm 85.225	.114 \pm .002	.108 \pm .003	.107 \pm .002	121.7.879 \pm 45.353
	RIVASPLATA	.114 \pm .000	.176 \pm .003	.109 \pm .000	120.629 \pm 140.549	.110 \pm .001	.192 \pm .003	.103 \pm .000	37.459 \pm 84.367	.106 \pm .001	.197 \pm .003	.101 \pm .001	122.005 \pm 51.052
	STOCHASTIC	—	.177	—	8.994	.110 \pm .000	.169	—	3.180	—	.172	—	11.291
CIFAR-10	OURS	.334 \pm .000	.350 \pm .000	.328 \pm .000	12.067	.322 \pm .000	.332 \pm .000	.313 \pm .000	4.172	.323 \pm .001	.336 \pm .000	.316 \pm .000	3.382
	BLANCHARD	.334 \pm .000	.977 \pm .001	.328 \pm .000	285.65 \pm 245.568	.322 \pm .000	.803 \pm .004	.313 \pm .000	87.979 \pm 126.804	.320 \pm .001	.343 \pm .001	.314 \pm .000	272.786 \pm 7.572
	CATONI	.334 \pm .000	.983 \pm .001	.328 \pm .000	241.36 \pm 211.963	.322 \pm .000	.694 \pm .004	.313 \pm .000	70.928 \pm 112.159	.321 \pm .001	.315 \pm .002	.314 \pm .000	237.763 \pm 65.952
	RIVASPLATA	.334 \pm .000	.922 \pm .001	.328 \pm .000	332.82 \pm 246.654	.322 \pm .000	.680 \pm .003	.312 \pm .000	84.03 \pm 128.894	.320 \pm .001	.327 \pm .003	.314 \pm .000	273.108 \pm 7.556
	STOCHASTIC	—	.417	—	6.033	.322 \pm .000	.400	—	2.086	—	.403	—	1.691

Table K6: Comparison of OURS, RIVASPLATA, BLANCHARD and CATONI based on the disintegrated bounds, and STOCHASTIC based on the randomized bounds learned with two learning rates $\text{lr} \in \{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_S(h)$), 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). More precisely, we report the mean \pm the standard deviation for 400 neural networks sampled from \mathcal{Q}_S for OURS, RIVASPLATA, BLANCHARD, and CATONI. We consider, in this table, that the split ratio is 0.5.

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
		$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div
MNIST	OURS	.008 ± .000	.015 ± .000	.010 ± .000	.084	.006 ± .000	.012 ± .000	.009 ± .000	.053	.008 ± .000	.014 ± .000	.010 ± .000	.179	.014 ± .001	.019 ± .001	.014 ± .001	.576
	BLANCHARD	.008 ± .000	.025 ± .001	.010 ± .000	81.167 ± 12.801	.006 ± .000	.014 ± .001	.009 ± .000	15.538 ± 5.438	.009 ± .000	.014 ± .001	.010 ± .000	2.140 ± 2.072	.019 ± .001	.018 ± .001	.015 ± .001	284 ± 6.49
	CATONI	.008 ± .000	.022 ± .001	.010 ± .000	14.068 ± 14.062	.006 ± .000	.013 ± .000	.009 ± .000	7.670 ± 5.963	.008 ± .000	.014 ± .001	.010 ± .000	1.132 ± 2.085	.019 ± .001	.018 ± .001	.015 ± .001	379 ± 6.80
	RIVASPLATA	.008 ± .000	.022 ± .001	.010 ± .000	84.851 ± 13.055	.006 ± .000	.013 ± .001	.009 ± .000	15.416 ± 5.394	.008 ± .000	.014 ± .001	.010 ± .000	2.865 ± 1.592	.019 ± .001	.018 ± .001	.015 ± .001	379 ± 6.85
	STOCHASTIC	—	.039	—	.042	—	.035	—	.026	—	.038	—	.080	—	.045	—	.288
Fashion	OURS	.106 ± .000	.113 ± .000	.101 ± .000	.133	.104 ± .001	.110 ± .000	.099 ± .000	.027	.108 ± .002	.112 ± .001	.101 ± .001	.903	.120 ± .004	.127 ± .003	.115 ± .003	.868
	BLANCHARD	.106 ± .000	.136 ± .003	.101 ± .000	77.573 ± 12.564	.104 ± .001	.115 ± .003	.099 ± .000	15.278 ± 5.599	.109 ± .002	.111 ± .002	.102 ± .001	2.153 ± 2.081	.122 ± .004	.126 ± .004	.117 ± .004	248 ± 7.15
	CATONI	.106 ± .000	.145 ± .005	.101 ± .000	104.356 ± 14.712	.104 ± .001	.112 ± .002	.099 ± .000	17.566 ± 5.996	.109 ± .002	.110 ± .001	.102 ± .001	2.217 ± 2.084	.122 ± .004	.126 ± .004	.117 ± .004	262 ± 6.99
	RIVASPLATA	.106 ± .000	.127 ± .002	.101 ± .000	82.150 ± 12.955	.104 ± .001	.112 ± .001	.099 ± .000	15.569 ± 5.629	.109 ± .002	.111 ± .001	.102 ± .001	2.178 ± 2.060	.122 ± .004	.126 ± .004	.117 ± .004	264 ± 7.04
	STOCHASTIC	—	.162	—	.066	—	.159	—	.164	—	.162	—	.451	—	.179	—	.434
CIFAR-10	OURS	.312 ± .000	.323 ± .000	.304 ± .000	.027	.281 ± .000	.304 ± .000	.285 ± .000	.085	.298 ± .001	.310 ± .000	.291 ± .000	.101	.315 ± .001	.329 ± .001	.309 ± .001	.368
	BLANCHARD	.312 ± .000	.405 ± .004	.304 ± .000	268.140 ± 22.835	.281 ± .000	.339 ± .004	.285 ± .000	8.690 ± 12.628	.298 ± .001	.320 ± .004	.291 ± .000	19.618 ± 6.249	.315 ± .001	.327 ± .003	.310 ± .001	3.213 ± 2.590
	CATONI	.312 ± .000	.428 ± .009	.304 ± .000	269.415 ± 22.884	.281 ± .000	.333 ± .005	.285 ± .000	83.414 ± 13.018	.298 ± .001	.314 ± .003	.291 ± .000	7.711 ± 6.481	.315 ± .001	.326 ± .001	.310 ± .001	3.273 ± 2.597
	RIVASPLATA	.312 ± .000	.375 ± .003	.304 ± .000	268.589 ± 22.845	.281 ± .000	.325 ± .003	.285 ± .000	81.532 ± 12.712	.298 ± .001	.315 ± .002	.291 ± .000	19.813 ± 6.288	.315 ± .001	.327 ± .002	.310 ± .001	3.233 ± 2.599
	STOCHASTIC	—	.391	—	.013	—	.370	—	.017	—	.377	—	.050	—	.397	—	.184
MNIST	OURS	.008 ± .000	.017 ± .000	.010 ± .000	29.983	.006 ± .000	.013 ± .000	.009 ± .000	3.162	.008 ± .000	.015 ± .000	.010 ± .000	1.418	.013 ± .001	.019 ± .001	.013 ± .001	12.231
	BLANCHARD	.008 ± .000	.574 ± .005	.010 ± .000	11804.556 ± 155.958	.006 ± .000	.256 ± .004	.009 ± .000	3828.115 ± 86.973	.008 ± .000	.018 ± .003	.010 ± .000	118.4777 ± 48.158	.019 ± .001	.022 ± .002	.011 ± .000	36.865 ± 28.054
	CATONI	.008 ± .000	.386 ± .003	.010 ± .000	1386.455 ± 15.722	.006 ± .000	.192 ± .002	.009 ± .000	352.9371 ± 53.072	.008 ± .000	.019 ± .002	.010 ± .000	12.6301 ± 48.751	.019 ± .001	.022 ± .002	.014 ± .001	383.277 ± 28.536
	RIVASPLATA	.008 ± .000	.362 ± .003	.010 ± .000	11365.971 ± 15.059	.006 ± .000	.185 ± .003	.009 ± .000	377.2597 ± 84.127	.008 ± .000	.069 ± .002	.010 ± .000	118.841 ± 5.943	.019 ± .001	.026 ± .001	.011 ± .000	369.079 ± 27.947
	STOCHASTIC	—	.041	—	14.966	—	.035	—	1.581	—	.039	—	.760	—	.049	—	6.116
Fashion	OURS	.106 ± .000	.114 ± .000	.101 ± .000	6.310	.102 ± .001	.113 ± .000	.099 ± .000	9.312	.106 ± .002	.115 ± .001	.100 ± .001	14.924	.113 ± .003	.126 ± .003	.110 ± .002	18.864
	BLANCHARD	.106 ± .000	.674 ± .004	.101 ± .000	10795.464 ± 148.426	.102 ± .001	.412 ± .004	.098 ± .000	3085.946 ± 82.481	.103 ± .001	.253 ± .004	.097 ± .001	1178.401 ± 48.359	.115 ± .003	.126 ± .003	.108 ± .002	338.697 ± 27.104
	CATONI	.106 ± .000	.623 ± .005	.101 ± .000	11971.564 ± 15.589	.104 ± .001	.321 ± .004	.099 ± .000	3825.370 ± 87.728	.107 ± .002	.208 ± .003	.100 ± .001	121.4976 ± 48.846	.116 ± .003	.184 ± .004	.111 ± .001	388.197 ± 27.580
	RIVASPLATA	.105 ± .000	.563 ± .003	.101 ± .000	11159.304 ± 15.540	.102 ± .000	.307 ± .003	.097 ± .000	381.075 ± 87.924	.102 ± .001	.201 ± .003	.096 ± .001	1201.832 ± 48.877	.112 ± .002	.161 ± .002	.107 ± .002	388.146 ± 27.482
	STOCHASTIC	—	.163	—	3.155	—	.161	—	4.656	—	.163	—	7.462	—	.176	—	9.182
CIFAR-10	OURS	.312 ± .000	.328 ± .000	.304 ± .000	1.802	.281 ± .000	.304 ± .000	.285 ± .000	1.802	.297 ± .001	.311 ± .000	.291 ± .000	2.056	.314 ± .001	.330 ± .001	.309 ± .001	4.782
	BLANCHARD	.312 ± .000	.990 ± .000	.304 ± .000	48007.471 ± 31.730	.280 ± .000	.825 ± .003	.284 ± .000	8824.774 ± 134.331	.296 ± .001	.417 ± .004	.290 ± .000	2724.775 ± 66.832	.303 ± .001	.457 ± .002	.303 ± .001	888.277 ± 41.530
	CATONI	.312 ± .000	.980 ± .000	.304 ± .000	21278.808 ± 207.839	.280 ± .000	.681 ± .004	.284 ± .000	6951.939 ± 118.540	.296 ± .001	.606 ± .001	.290 ± .000	2145.470 ± 63.045	.305 ± .001	.467 ± .002	.309 ± .001	103.494 ± 47.021
	RIVASPLATA	.312 ± .000	.964 ± .001	.304 ± .000	42834.626 ± 284.116	.280 ± .000	.690 ± .003	.284 ± .000	8075.531 ± 136.658	.296 ± .001	.521 ± .003	.290 ± .000	2718.415 ± 66.664	.307 ± .001	.434 ± .003	.301 ± .001	921.068 ± 42.158
	STOCHASTIC	—	.394	—	6.003	—	.371	—	.901	—	.378	—	1.028	—	.397	—	2.391

Table K8: Comparison of OURS, RIVASPLATA, BLANCHARD and CATONI based on the disintegrated bounds, and STOCHASTIC based on the randomized bounds learned with two learning rates $\text{lr} \in \{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_S(h)$), 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). More precisely, we report the mean \pm the standard deviation for 400 neural networks sampled from Q_S for OURS, RIVASPLATA, BLANCHARD, and CATONI. We consider, in this table, that the split ratio is 0.7.

	$\sigma^2 = 10^{-6}$					$\sigma^2 = 10^{-5}$					$\sigma^2 = 10^{-4}$					$\sigma^2 = 10^{-3}$						
	$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div		$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div		$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div		$R_{\mathcal{T}}(h)$	Bnd	$R_S(h)$	Div			
MNIST	OURS	.011 ± .000	.025 ± .000	.033 ± .000	.047	.010 ± .000	.018 ± .000	.022 ± .000	.125	.010 ± .000	.007 ± .000	.012 ± .000	.116	.010 ± .000	.018 ± .001	.012 ± .001	.132	.010 ± .000	.018 ± .001	.012 ± .001	.132	
	BLANCHARD	.011 ± .000	.032 ± .002	.033 ± .000	65.017 ± 11.069	.010 ± .000	.019 ± .001	.012 ± .000	1.819 ± 4.995	.010 ± .000	.007 ± .001	.012 ± .000	1.655 ± 1.635	.010 ± .000	.016 ± .001	.012 ± .001	.115 ± .560	.010 ± .000	.016 ± .001	.012 ± .001	.115 ± .560	
	CATONI	.011 ± .000	.028 ± .001	.033 ± .000	84.329 ± 13.023	.010 ± .000	.021 ± .000	.012 ± .000	11.910 ± 5.053	.010 ± .000	.017 ± .001	.012 ± .000	1.228 ± 1.637	.010 ± .000	.017 ± .001	.012 ± .001	.012 ± .001	.017 ± .001	.012 ± .001	.012 ± .001	.173 ± .512	
	RIVASPLATA	.011 ± .000	.026 ± .001	.033 ± .000	68.955 ± 11.606	.010 ± .000	.018 ± .001	.012 ± .000	1.637 ± 4.962	.010 ± .000	.015 ± .000	.012 ± .000	1.408 ± 1.639	.010 ± .000	.016 ± .001	.012 ± .001	.010 ± .001	.016 ± .001	.012 ± .001	.010 ± .001	.100 ± .529	
	STOCHASTIC	—	.044	—	.023	—	.043	—	.062	—	.042	—	—	.094	—	.043	—	.066	—	.043	—	.066
Fashion	OURS	.099 ± .000	.112 ± .000	.098 ± .000	.067	.107 ± .001	.115 ± .001	.100 ± .001	.542	.098 ± .002	.107 ± .001	.083 ± .001	.353	.108 ± .003	.117 ± .002	.102 ± .002	.312	.108 ± .003	.117 ± .002	.102 ± .002	.312	
	BLANCHARD	.099 ± .000	.138 ± .004	.098 ± .000	61.733 ± 1.862	.107 ± .001	.119 ± .003	.101 ± .001	1.651 ± 4.230	.099 ± .001	.104 ± .002	.084 ± .001	1.342 ± 1.664	.108 ± .003	.113 ± .003	.103 ± .002	1.44 ± .534	.108 ± .003	.113 ± .003	.103 ± .002	1.44 ± .534	
	CATONI	.099 ± .000	.155 ± .007	.098 ± .000	83.929 ± 12.212	.107 ± .001	.116 ± .003	.101 ± .001	11.543 ± 4.870	.099 ± .002	.103 ± .002	.094 ± .001	4.837 ± 1.594	.108 ± .003	.112 ± .003	.103 ± .002	1.55 ± .545	.108 ± .003	.112 ± .003	.103 ± .002	1.55 ± .545	
	RIVASPLATA	.099 ± .000	.128 ± .002	.098 ± .000	65.737 ± 11.733	.107 ± .001	.116 ± .002	.101 ± .001	1.958 ± 4.794	.099 ± .002	.105 ± .002	.094 ± .001	1.491 ± 1.618	.108 ± .003	.114 ± .003	.103 ± .002	1.55 ± .546	.108 ± .003	.114 ± .003	.103 ± .002	1.55 ± .546	
	STOCHASTIC	—	.161	—	.034	—	.164	—	.271	—	.155	—	.177	—	.166	—	.156	—	.166	—	.156	
CIFAR-10	OURS	.277 ± .000	.306 ± .000	.272 ± .000	.016	.266 ± .000	.281 ± .000	.257 ± .000	.022	.253 ± .001	.272 ± .000	.248 ± .000	.069	.236 ± .001	.258 ± .001	.235 ± .001	.118	.236 ± .001	.258 ± .001	.235 ± .001	.118	
	BLANCHARD	.277 ± .000	.399 ± .006	.272 ± .000	23.327	.266 ± .000	.322 ± .005	.257 ± .000	7.190 ± 11.685	.253 ± .001	.281 ± .005	.248 ± .000	5.214 ± 5.838	.236 ± .001	.255 ± .002	.235 ± .001	2.223 ± 2.016	.236 ± .001	.255 ± .002	.235 ± .001	2.223 ± 2.016	
	CATONI	.277 ± .000	.399 ± .002	.272 ± .000	269.048 ± 23.489	.266 ± .000	.328 ± .009	.257 ± .000	81.217 ± 13.476	.253 ± .001	.275 ± .004	.248 ± .000	16.576 ± 6.087	.236 ± .001	.253 ± .002	.235 ± .001	2.248 ± 1.982	.236 ± .001	.253 ± .002	.235 ± .001	2.248 ± 1.982	
	RIVASPLATA	.277 ± .000	.362 ± .004	.272 ± .000	258.908 ± 23.725	.266 ± .000	.305 ± .004	.257 ± .000	72.737 ± 12.750	.253 ± .001	.275 ± .003	.248 ± .000	15.342 ± 5.948	.236 ± .001	.255 ± .002	.235 ± .001	2.220 ± 2.180	.236 ± .001	.255 ± .002	.235 ± .001	2.220 ± 2.180	
	STOCHASTIC	—	.362	—	.008	—	.345	—	.011	—	.336	—	.034	—	.322	—	.059	—	.322	—	.059	

Table K11: Comparison of the bound values before performing Step 2) of our Training Method for OURS, RIVASPLATA, BLANCHARD and CATONI. More precisely, for each split and each variance $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, we report the mean \pm the standard deviation (for 400 neural networks sampled from \mathcal{P}) of the test risk ($R_{\mathcal{T}}(h)$), the empirical risk ($R_S(h)$), and the value of the bounds of Corollaries 6 and 7. We consider in this table that the dataset is MNIST.

Split	$R_{\mathcal{T}}(h)$	$R_S(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
0	.901 \pm .002	.901 \pm .002	.908 \pm .002	.906 \pm .002	.905 \pm .002	.906 \pm .002
1	.035 \pm .000	.039 \pm .000	.045 \pm .000	.043 \pm .000	.043 \pm .000	.042 \pm .000
2	.016 \pm .000	.019 \pm .000	.023 \pm .000	.022 \pm .000	.022 \pm .000	.022 \pm .000
3	.012 \pm .000	.013 \pm .000	.017 \pm .000	.016 \pm .000	.015 \pm .000	.015 \pm .000
4	.010 \pm .000	.013 \pm .000	.017 \pm .000	.016 \pm .000	.016 \pm .000	.016 \pm .000
5	.008 \pm .000	.010 \pm .000	.015 \pm .000	.013 \pm .000	.013 \pm .000	.014 \pm .000
6	.008 \pm .000	.010 \pm .000	.014 \pm .000	.013 \pm .000	.013 \pm .000	.014 \pm .000
7	.011 \pm .000	.013 \pm .000	.019 \pm .000	.017 \pm .000	.017 \pm .000	.018 \pm .000
8	.011 \pm .000	.013 \pm .000	.020 \pm .000	.018 \pm .000	.018 \pm .000	.020 \pm .000
9	.008 \pm .000	.009 \pm .000	.018 \pm .000	.015 \pm .000	.014 \pm .000	.015 \pm .000
0	.897 \pm .013	.897 \pm .012	.904 \pm .012	.902 \pm .012	.902 \pm .012	.903 \pm .012
1	.024 \pm .000	.030 \pm .001	.035 \pm .001	.034 \pm .001	.033 \pm .001	.033 \pm .001
2	.015 \pm .000	.019 \pm .000	.023 \pm .000	.022 \pm .000	.021 \pm .000	.021 \pm .000
3	.009 \pm .000	.011 \pm .000	.015 \pm .000	.014 \pm .000	.013 \pm .000	.013 \pm .000
4	.012 \pm .000	.014 \pm .000	.018 \pm .000	.017 \pm .000	.017 \pm .000	.017 \pm .000
5	.006 \pm .000	.009 \pm .000	.012 \pm .000	.011 \pm .000	.011 \pm .000	.012 \pm .000
6	.007 \pm .000	.009 \pm .000	.014 \pm .000	.013 \pm .000	.012 \pm .000	.013 \pm .000
7	.010 \pm .000	.012 \pm .000	.018 \pm .000	.016 \pm .000	.016 \pm .000	.017 \pm .000
8	.008 \pm .000	.010 \pm .000	.017 \pm .000	.015 \pm .000	.014 \pm .000	.017 \pm .000
9	.011 \pm .000	.010 \pm .000	.020 \pm .000	.017 \pm .000	.017 \pm .000	.018 \pm .000

Split	$R_{\mathcal{T}}(h)$	$R_S(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
0	.898 \pm .017	.898 \pm .017	.905 \pm .016	.903 \pm .016	.902 \pm .016	.903 \pm .016
1	.035 \pm .003	.039 \pm .002	.045 \pm .002	.044 \pm .002	.043 \pm .002	.043 \pm .002
2	.015 \pm .001	.016 \pm .001	.020 \pm .001	.019 \pm .001	.019 \pm .001	.019 \pm .001
3	.012 \pm .000	.016 \pm .000	.020 \pm .001	.019 \pm .001	.019 \pm .001	.019 \pm .001
4	.009 \pm .000	.011 \pm .000	.015 \pm .000	.014 \pm .000	.014 \pm .000	.014 \pm .000
5	.008 \pm .000	.010 \pm .000	.015 \pm .000	.013 \pm .000	.013 \pm .000	.014 \pm .000
6	.008 \pm .000	.009 \pm .000	.013 \pm .000	.012 \pm .000	.012 \pm .000	.013 \pm .000
7	.010 \pm .000	.012 \pm .000	.017 \pm .000	.016 \pm .000	.015 \pm .000	.016 \pm .000
8	.011 \pm .000	.011 \pm .000	.018 \pm .000	.016 \pm .000	.016 \pm .000	.018 \pm .000
9	.009 \pm .000	.009 \pm .000	.018 \pm .001	.015 \pm .001	.015 \pm .001	.016 \pm .001
0	.903 \pm .014	.902 \pm .014	.909 \pm .013	.907 \pm .013	.907 \pm .013	.907 \pm .013
1	.041 \pm .005	.045 \pm .005	.050 \pm .005	.049 \pm .005	.048 \pm .005	.048 \pm .005
2	.020 \pm .002	.022 \pm .002	.026 \pm .002	.025 \pm .002	.025 \pm .002	.024 \pm .002
3	.014 \pm .001	.015 \pm .001	.019 \pm .001	.018 \pm .001	.018 \pm .001	.018 \pm .001
4	.015 \pm .001	.016 \pm .001	.021 \pm .001	.020 \pm .001	.019 \pm .001	.019 \pm .001
5	.015 \pm .001	.015 \pm .001	.020 \pm .001	.019 \pm .001	.018 \pm .001	.018 \pm .001
6	.008 \pm .000	.010 \pm .000	.014 \pm .001	.013 \pm .001	.012 \pm .000	.013 \pm .000
7	.010 \pm .001	.012 \pm .001	.016 \pm .001	.016 \pm .001	.016 \pm .001	.017 \pm .001
8	.010 \pm .001	.010 \pm .001	.016 \pm .001	.014 \pm .001	.014 \pm .001	.016 \pm .001
9	.008 \pm .000	.009 \pm .001	.019 \pm .001	.016 \pm .001	.015 \pm .001	.017 \pm .001

Table K12: Comparison of the bound values before performing Step 2) of our Training Method for OURS, RIVASPLATA, BLANCHARD and CATONI. More precisely, for each split and each variance $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, we report the mean \pm the standard deviation (for 400 neural networks sampled from \mathcal{P}) of the test risk ($R_T(h)$), the empirical risk ($R_S(h)$), and the value of the bounds of Corollaries 6 and 7. We consider in this table that the dataset is Fashion-MNIST.

Split	$R_T(h)$	$R_S(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
.0	.970 \pm .028	.970 \pm .027	.972 \pm .025	.971 \pm .025	.971 \pm .026	.972 \pm .026
.1	.166 \pm .001	.159 \pm .000	.169 \pm .000	.167 \pm .000	.166 \pm .000	.167 \pm .000
.2	.168 \pm .002	.160 \pm .001	.170 \pm .001	.168 \pm .001	.167 \pm .001	.168 \pm .001
.3	.126 \pm .000	.124 \pm .000	.134 \pm .000	.132 \pm .000	.131 \pm .000	.131 \pm .000
.4	.118 \pm .001	.112 \pm .000	.123 \pm .000	.120 \pm .000	.119 \pm .000	.119 \pm .000
.5	.106 \pm .000	.101 \pm .000	.113 \pm .000	.110 \pm .000	.109 \pm .000	.109 \pm .000
.6	.109 \pm .000	.102 \pm .000	.115 \pm .000	.112 \pm .000	.110 \pm .000	.110 \pm .000
.7	.099 \pm .000	.098 \pm .000	.112 \pm .000	.109 \pm .000	.108 \pm .000	.107 \pm .000
.8	.103 \pm .000	.099 \pm .000	.117 \pm .000	.112 \pm .000	.111 \pm .000	.110 \pm .000
.9	.094 \pm .000	.089 \pm .000	.113 \pm .000	.107 \pm .000	.105 \pm .000	.106 \pm .000
.0	.945 \pm .038	.945 \pm .037	.949 \pm .035	.948 \pm .035	.948 \pm .036	.948 \pm .036
.1	.158 \pm .001	.151 \pm .001	.161 \pm .001	.159 \pm .001	.158 \pm .001	.159 \pm .001
.2	.157 \pm .003	.151 \pm .003	.162 \pm .003	.159 \pm .003	.158 \pm .003	.159 \pm .003
.3	.126 \pm .001	.121 \pm .001	.131 \pm .001	.128 \pm .001	.127 \pm .001	.128 \pm .001
.4	.114 \pm .001	.107 \pm .001	.118 \pm .001	.115 \pm .001	.114 \pm .001	.114 \pm .001
.5	.104 \pm .001	.099 \pm .000	.110 \pm .000	.108 \pm .000	.107 \pm .000	.106 \pm .000
.6	.115 \pm .001	.104 \pm .001	.117 \pm .001	.114 \pm .001	.113 \pm .001	.112 \pm .001
.7	.107 \pm .001	.101 \pm .001	.115 \pm .001	.111 \pm .001	.110 \pm .001	.109 \pm .001
.8	.098 \pm .001	.096 \pm .001	.114 \pm .001	.109 \pm .001	.108 \pm .001	.107 \pm .001
.9	.091 \pm .001	.095 \pm .001	.119 \pm .001	.113 \pm .001	.111 \pm .001	.112 \pm .001
.0	.912 \pm .027	.912 \pm .027	.918 \pm .026	.916 \pm .027	.916 \pm .027	.916 \pm .027
.1	.164 \pm .003	.154 \pm .003	.164 \pm .003	.162 \pm .003	.161 \pm .003	.162 \pm .004
.2	.164 \pm .009	.160 \pm .009	.170 \pm .010	.168 \pm .010	.167 \pm .010	.168 \pm .010
.3	.125 \pm .002	.119 \pm .002	.129 \pm .002	.126 \pm .002	.126 \pm .002	.126 \pm .002
.4	.119 \pm .003	.113 \pm .003	.124 \pm .003	.121 \pm .003	.120 \pm .003	.120 \pm .003
.5	.109 \pm .002	.102 \pm .001	.113 \pm .001	.110 \pm .001	.109 \pm .001	.109 \pm .001
.6	.102 \pm .001	.096 \pm .001	.109 \pm .001	.105 \pm .001	.105 \pm .001	.104 \pm .001
.7	.099 \pm .002	.094 \pm .001	.108 \pm .001	.104 \pm .001	.103 \pm .001	.102 \pm .001
.8	.104 \pm .001	.100 \pm .002	.118 \pm .002	.113 \pm .002	.112 \pm .002	.111 \pm .002
.9	.092 \pm .002	.089 \pm .001	.113 \pm .001	.107 \pm .001	.105 \pm .001	.106 \pm .001
.0	.899 \pm .026	.899 \pm .027	.906 \pm .026	.904 \pm .026	.904 \pm .026	.905 \pm .025
.1	.178 \pm .006	.170 \pm .006	.181 \pm .006	.178 \pm .006	.177 \pm .006	.179 \pm .006
.2	.164 \pm .006	.159 \pm .006	.169 \pm .006	.167 \pm .006	.166 \pm .006	.167 \pm .006
.3	.143 \pm .007	.138 \pm .007	.148 \pm .007	.146 \pm .007	.145 \pm .007	.145 \pm .007
.4	.133 \pm .005	.129 \pm .005	.140 \pm .005	.137 \pm .005	.137 \pm .005	.137 \pm .005
.5	.122 \pm .004	.117 \pm .004	.129 \pm .004	.126 \pm .004	.125 \pm .004	.125 \pm .004
.6	.111 \pm .003	.104 \pm .003	.117 \pm .003	.114 \pm .003	.113 \pm .003	.112 \pm .003
.7	.109 \pm .003	.103 \pm .003	.118 \pm .003	.114 \pm .003	.113 \pm .003	.112 \pm .003
.8	.108 \pm .004	.102 \pm .004	.120 \pm .004	.115 \pm .004	.114 \pm .004	.113 \pm .004
.9	.103 \pm .003	.099 \pm .002	.124 \pm .003	.118 \pm .003	.116 \pm .003	.116 \pm .003

Table K13: Comparison of the bound values before performing Step 2) of our Training Method for OURS, RIVASPLATA, BLANCHARD and CATONI. More precisely, for each split and each variance $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, we report the mean \pm the standard deviation (for 400 neural networks sampled from \mathcal{P}) of the test risk ($R_T(h)$), the empirical risk ($R_S(h)$), and the value of the bounds of Corollaries 6 and 7. We consider in this table that the dataset is CIFAR-10.

Split	$R_T(h)$	$R_S(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
.0	.899 ± .000	.899 ± .000	.906 ± .000	.904 ± .000	.903 ± .000	.904 ± .000
.1	.476 ± .000	.470 ± .000	.486 ± .000	.482 ± .000	.481 ± .000	.485 ± .000
.2	.390 ± .000	.389 ± .000	.406 ± .000	.402 ± .000	.401 ± .000	.404 ± .000
.3	.370 ± .000	.358 ± .000	.374 ± .000	.371 ± .000	.370 ± .000	.372 ± .000
.4	.334 ± .000	.328 ± .000	.346 ± .000	.342 ± .000	.341 ± .000	.342 ± .000
.5	.307 ± .000	.302 ± .000	.321 ± .000	.317 ± .000	.316 ± .000	.317 ± .000
.6	.274 ± .000	.276 ± .000	.297 ± .000	.293 ± .000	.291 ± .000	.291 ± .000
.7	.275 ± .000	.279 ± .000	.296 ± .000	.290 ± .000	.289 ± .000	.288 ± .000
.8	.249 ± .000	.237 ± .000	.265 ± .000	.259 ± .000	.257 ± .000	.256 ± .000
.9	.227 ± .000	.230 ± .000	.269 ± .000	.260 ± .000	.258 ± .000	.258 ± .000
.0	.899 ± .001	.899 ± .000	.906 ± .000	.904 ± .000	.904 ± .000	.904 ± .000
.1	.476 ± .000	.478 ± .000	.494 ± .000	.490 ± .000	.489 ± .000	.493 ± .000
.2	.403 ± .000	.398 ± .000	.414 ± .000	.410 ± .000	.409 ± .000	.412 ± .000
.3	.349 ± .000	.350 ± .000	.367 ± .000	.363 ± .000	.362 ± .000	.364 ± .000
.4	.322 ± .000	.313 ± .000	.330 ± .000	.327 ± .000	.326 ± .000	.327 ± .000
.5	.281 ± .000	.283 ± .000	.302 ± .000	.298 ± .000	.297 ± .000	.297 ± .000
.6	.290 ± .000	.286 ± .000	.307 ± .000	.303 ± .000	.301 ± .000	.301 ± .000
.7	.266 ± .000	.257 ± .000	.281 ± .000	.276 ± .000	.274 ± .000	.274 ± .000
.8	.247 ± .000	.243 ± .000	.271 ± .000	.265 ± .000	.263 ± .000	.262 ± .000
.9	.236 ± .000	.227 ± .000	.266 ± .000	.257 ± .000	.255 ± .000	.255 ± .000

Split	$R_T(h)$	$R_S(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
.0	.900 ± .004	.900 ± .003	.907 ± .003	.905 ± .003	.905 ± .003	.905 ± .003
.1	.458 ± .001	.464 ± .001	.479 ± .001	.476 ± .001	.475 ± .001	.478 ± .001
.2	.395 ± .001	.396 ± .000	.412 ± .000	.409 ± .000	.408 ± .000	.411 ± .000
.3	.361 ± .001	.361 ± .000	.378 ± .000	.375 ± .000	.373 ± .000	.376 ± .000
.4	.323 ± .001	.316 ± .000	.334 ± .000	.330 ± .000	.329 ± .000	.331 ± .000
.5	.296 ± .001	.291 ± .000	.310 ± .000	.306 ± .000	.304 ± .000	.305 ± .000
.6	.271 ± .001	.263 ± .000	.284 ± .000	.279 ± .000	.278 ± .000	.278 ± .000
.7	.253 ± .001	.246 ± .000	.270 ± .000	.265 ± .000	.263 ± .000	.262 ± .000
.8	.259 ± .001	.252 ± .001	.281 ± .001	.275 ± .001	.273 ± .001	.272 ± .001
.9	.217 ± .000	.216 ± .001	.255 ± .001	.246 ± .001	.243 ± .001	.244 ± .001
.0	.905 ± .012	.904 ± .012	.911 ± .011	.909 ± .011	.909 ± .011	.909 ± .011
.1	.479 ± .002	.480 ± .001	.496 ± .001	.493 ± .001	.491 ± .001	.495 ± .001
.2	.415 ± .002	.415 ± .001	.432 ± .001	.428 ± .001	.427 ± .001	.430 ± .001
.3	.417 ± .001	.416 ± .001	.434 ± .001	.430 ± .001	.429 ± .001	.431 ± .001
.4	.333 ± .001	.323 ± .001	.341 ± .001	.337 ± .001	.336 ± .001	.338 ± .001
.5	.316 ± .001	.311 ± .001	.331 ± .001	.327 ± .001	.325 ± .001	.326 ± .001
.6	.280 ± .001	.281 ± .001	.302 ± .001	.298 ± .001	.296 ± .001	.296 ± .001
.7	.239 ± .001	.234 ± .001	.257 ± .001	.252 ± .001	.250 ± .001	.250 ± .001
.8	.249 ± .001	.245 ± .001	.274 ± .001	.268 ± .001	.266 ± .001	.264 ± .001
.9	.233 ± .001	.232 ± .002	.272 ± .002	.263 ± .002	.260 ± .002	.260 ± .002