Machine Learning (2024) 113:519–604
https://doi.org/10.1007/s10994-023-06391-0

A general framework for the practical disintegration of PAC-Bayesian bounds

Paul Viallard[1] · **Pascal Germain**[2] · **Amaury Habrard**[3,4] · **Emilie Morvant**[3]

Received: 24 December 2021 / Revised: 8 February 2023 / Accepted: 16 August 2023 /
Published online: 11 October 2023

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

Editor: Krzysztof Dembczynski and Emilie Devijver.

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

✉ Paul Viallard
paul.viallard@inria.fr

Pascal Germain
pascal.germain@ift.ulaval.ca

Amaury Habrard
amaury.habrard@univ-st-etienne.fr

Emilie Morvant
emilie.morvant@univ-st-etienne.fr

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

[2] Département d'informatique et de génie logiciel, Université Laval, Quebec, Canada

[3] Laboratoire Hubert Curien UMR 5516, Institut d Optique Graduate School, CNRS, Université Jean Monnet Saint-Étienne, 42023 Saint-Étienne, France

[4] Institut Universitaire de France (IUF), Paris, France

1 Introduction

In statistical learning theory, PAC-Bayesian theory[1] (Shawe-Taylor & Williamson, 1997; McAllester, 1998) provides a powerful framework for analyzing the generalization ability of machine learning models such as linear classifiers (Germain et al., 2009), SVM (Ambroladze et al., 2006), or neural networks (Dziugaite & Roy, 2017; Pérez-Ortiz et al., 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 et al. (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 (Neyshabur et al., 2018; Nagarajan & Kolter, 2019). 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 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 et al. (2010, 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

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

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. (2010, 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 Sect. 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 Sect. 5. Before concluding in Sect. 7, we discuss in Sect. 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 measure[2] 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 functions $h : \mathcal{X}\to\mathcal{Y}$. The learner aims to find $h \in \mathcal{H}$ that assigns a label y to an input $\mathbf{x}$ as accurately as possible. Given an example z and a hypothesis h, we assess the quality of the prediction of h with a *loss function* $\ell : \mathcal{H}\times\mathcal{Z}\to[0, 1]$ evaluating to which extent the prediction is accurate. Given 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

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

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:

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

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.*, $\epsilon\left(\frac{1}{\delta}, \frac{1}{m}\right)$ 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:

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

Due to $\sup_{h\in\mathcal{H}}$, this bound can be seen as a *worst-case* analysis. Indeed, it means that the bound $\left|R_{\mathcal{D}}(h) - R_{\mathcal{S}}(h)\right| \leq \epsilon\left(\frac{1}{\delta}, \frac{1}{m}, \mathcal{H}\right)$ 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 $\epsilon(\frac{1}{\delta}, \frac{1}{m}, \mathcal{H}) < 1$. 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:

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

For example, this approach has been used by Hardt et al. (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:

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

A key notion is that the function $\varepsilon()$ 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 $h(\mathbf{x})$.

We recall below the classical PAC-Bayesian bounds in a general form as proposed by Germain et al. (2009); Bégin et al. (2016). The idea is to express the bound in terms of a generic function $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}_+^*$ that is meant to capture the 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 $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}_+^*$, for any $\delta \in (0,1]$ we have*

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

and

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

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

[3] The 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.

Note that Eq. (2) is more general than Eq. (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 Eq. (2), *(ii)* applying Jensen's inequality in order to move the expectation over $\mathcal{Q}$ in front of the logarithm, and *(iii)* taking the limit when α tends to 1. Note also the original bound statements of Germain et al. (2009); Bégin et al. (2016) are recovered by choosing a convex function $\Delta : [0,1]^2 \to \mathbb{R}$ 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 Eq. (1), or $\phi(h,\mathcal{S}) = \Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h))$ in Eq. (2), and then *(ii)* applying Jensen's inequality on the left-hand side of the in equation. 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]

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 $\ell : \mathcal{H}\times\mathcal{Z}\to[0,1]$ with $\phi(h,\mathcal{S}) = \exp\left(m\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h))\right)$ and $\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h)) = 2[R_{\mathcal{S}}(h) - R_{\mathcal{D}}(h)]^2$ we retrieve from Eq. (1) the bound proposed by McAllester (1998):

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

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

This bound illustrates the trade-off between the average empirical risk and $\epsilon\left(\frac{1}{\delta},\frac{1}{m},\mathcal{Q}\right) = \sqrt{\frac{1}{2m}(\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \ln\frac{2\sqrt{m}}{\delta})}$. More precisely, the higher m is, the lower $\epsilon\left(\frac{1}{\delta},\frac{1}{m},\mathcal{Q}\right)$ 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 Seeger (2002); Maurer (2004)'s bound that we retrieve with $\phi(h,\mathcal{S}) = \exp\left(m\,\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h))\right]\right)$ and $\Delta(R_{\mathcal{S}}(h), R_{\mathcal{D}}(h)) = \mathrm{kl}[R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h)]$:

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

where

$$\mathrm{kl}(q\|p) = q\ln\tfrac{q}{p} + (1-q)\ln\tfrac{1-q}{1-p} \tag{4}$$

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

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

$\mathcal{H}$). 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:

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

where $\mathcal{Q}_\mathcal{S}\triangleq A(\mathcal{S},\mathcal{P})$ with $A:\mathcal{Z}^m\times\mathcal{M}^*(\mathcal{H})\to\mathcal{M}(\mathcal{H})$ 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. (2010, 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-bounding[5] algorithm (Freund,

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

1998) with theoretical guarantees. As an illustration of the usefulness of our results, we provide, in Sect. 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 Eq. (2) our main result stated in Theorem 2 is a general bound involving the Rényi divergence $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ of order $\alpha > 1$.

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

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

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

Proof (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 $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$, we have, with probability at least $1-\frac{\delta}{2}$ on $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_\mathcal{S}$,

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

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. □

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 Sect. 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 Eq. (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 $\phi(h,\mathcal{S})=\exp(\frac{\alpha-1}{\alpha}m\,\mathrm{kl}(R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h))$ with $\mathrm{kl}(\cdot\|\cdot)$ defined by Eq. (4), the bound depends on m and converges as m increases (see Sect/ 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 $\alpha\to1^+$ and $\alpha\to+\infty$ showing that the bound converges when $\alpha\to1^+$ and $\alpha\to+\infty$.

Corollary 3 *Under the assumptions of Theorem 2, when $\alpha\to1^+$, we have*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_{\mathcal{S}}}\left(\ln\phi(h,\mathcal{S})\le\ln\frac{2}{\delta}+\ln\left[\operatorname{esssup}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right]\right)\ge1-\delta,$$

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

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_{\mathcal{S}}}\left(\ln\phi(h,\mathcal{S})\le\ln\operatorname{esssup}_{h'\in\mathcal{H}}\frac{\mathcal{Q}_{\mathcal{S}}(h')}{\mathcal{P}(h')}+\ln\left[\frac{4}{\delta^2}\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\right]\right)\ge1-\delta,$$

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

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

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

This corollary illustrates that the parameter α controls the trade-off between the Rényi divergence $D_\alpha(\mathcal{Q}_{\mathcal{S}}\|\mathcal{P})$ and $\ln\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right]$. Indeed, when $\alpha\to1^+$, the Rényi divergence vanishes while the other term converges toward $\ln\left[\operatorname{esssup}_{\mathcal{S}'\in\mathcal{Z},h'\in\mathcal{H}}\phi(h',\mathcal{S}')\right]$, roughly speaking the maximal value possible for the second term. On the other hand, when $\alpha\to+\infty$, the Rényi divergence increases and converges toward $\ln\operatorname{esssup}_{h'\in\mathcal{H}}\frac{\mathcal{Q}_{\mathcal{S}}(h')}{\mathcal{P}(h')}$ and the other term decreases toward $\ln\left[\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')\right]$.

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

For the sake of comparison, we recall in Eq. (6) the bound proposed by Rivasplata et al. (2010, 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 Catoni (2007); Blanchard and Fleuret (2007)) can be seen as a "disintegrated[6] KL divergence" depending only on the sampled $h\sim\mathcal{Q}_\mathcal{S}$. In contrast, our bound involves the Rényi divergence $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ between the prior $\mathcal{P}$ and the posterior $\mathcal{Q}_\mathcal{S}$, 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 Eq. 6). Nevertheless, our divergence term is the main advantage of our bound. Indeed, as confirmed by our experiments (Sect/ 5), our bound with $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ makes the learning procedure (in our self-bounding algorithm) more stable and efficient compared to the optimization of Eq. (6) with $\ln\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}$ 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 et al. (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 (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 $\phi : \mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}_+^*$, for any $\delta\in(0,1]$, for any algorithm $A : \mathcal{Z}^m\times\mathcal{M}^*(\mathcal{H})\to\mathcal{M}(\mathcal{H})$, we have*

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

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

Note that $e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}$ is closely related to the χ^2-distance. Indeed we have: $\chi^2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) \triangleq \mathbb{E}_{h\sim\mathcal{P}}\left[\frac{\mathcal{Q}_\mathcal{S}(h)}{\mathcal{P}(h)}\right]^2 - 1 = e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})} - 1$. An asset of Theorem 4 is the parameter λ controlling the trade-off between the exponentiated Rényi divergence $e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}$ and $\frac{1}{\delta^3}\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\phi(h',\mathcal{S}')^2$. Our bound is valid for all $\lambda>0$, thus, from a practical view, we can learn/tune the parameter λ to minimize the bound and control the possible numerical instability due to $e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}$. Indeed, if $D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ 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

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

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 $\phi : \mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}_+^*$, for any algorithm $A : \mathcal{Z}^m\times\mathcal{M}^*(\mathcal{H})\to\mathcal{M}(\mathcal{H})$, let*

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

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

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

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

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 over-parametrized, *i.e.*, $d\gg m$. We aim to learn the weights of the NN leading to the lowest true risk. Practitioners usually proceed by epochs[7] 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 $\forall t\in\{1,\ldots,T\}, \mathcal{P}_t = \mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_d)$ is a Gaussian distribution centered at $\mathbf{v}_t$ (the 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

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

$\mathcal{S}_{\text{prior}}$ such that $\mathcal{S}_{\text{prior}} \bigcap \mathcal{S}=\emptyset$, then Corollaries 6 and 7 will guide us to learn a posterior $\mathcal{Q}_{\mathcal{S}}=\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_d)$ from 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 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, Eq. (7) corresponds to Rivasplata et al. (2010)'s bound of Eq. (6), Eq. (8) to Blanchard and Fleuret (2007)'s one, and Eq. (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 $\mathbf{P}=\{\mathcal{P}_1, \ldots, \mathcal{P}_T\}$ of T priors on $\mathcal{H}$ where $\mathcal{P}_t=\mathcal{N}(\mathbf{v}_t, \sigma^2\mathbf{I}_d)$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{H}\times\mathcal{Z}\to[0,1]$, for any $\delta\in(0,1]$, we have*

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

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

Corollary 7 *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any set $\mathcal{H}$, for any set $\mathbf{P}=\{\mathcal{P}_1, \ldots, \mathcal{P}_T\}$ of T priors on $\mathcal{H}$ where $\mathcal{P}_t=\mathcal{N}(\mathbf{v}_t, \sigma^2\mathbf{I}_d)$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{H}\times\mathcal{Z}\to\{0,1\}$, 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 $\mathbf{w}+\boldsymbol{\epsilon}$, we have $\forall \mathcal{P}_t \in \mathbf{P}$*

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

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

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

with $[x]_+ = \max(x, 0)$, *and* $\mathrm{kl}_+(R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h)) = \mathrm{kl}(R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h))$ *if* $R_{\mathcal{S}}(h) < R_{\mathcal{D}}(h)$ *and* 0 *otherwise. Moreover, $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$ is a Gaussian noise such that $\mathbf{w}+\boldsymbol{\epsilon}$ are the weights of $h\sim\mathcal{Q}_{\mathcal{S}}$ with $\mathcal{Q}_{\mathcal{S}}=\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_d)$, 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

[8] Gaussian distributions have been first studied in PAC-Bayes in the context of linear classifiers (e.g., Ambroladze et al., 2006; Germain, Habrard, Laviolette, & Morvant, 2009; Germain et al., 2020), 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 Eq. (8) and Eq. (9); however, there is no closed-form solution for the expression of the minimum of this Eq. . 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 $\mathcal{Q}_\mathcal{S}$ leading to the lowest bound value. To obtain a tight bound, the divergence between one prior $\mathcal{P}_t\in\mathbf{P}$ and $\mathcal{Q}_\mathcal{S}$ must be low, *i.e.*, $\|\mathbf{w}-\mathbf{v}_t\|_2^2$ (or $\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2$) has to be 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 $\mathcal{Q}_\mathcal{S}$ with the algorithm A (defined *a priori*).

At first sight, the selection of the prior weights with $\mathcal{S}$ by early stopping may appear to be "cheating". However, this procedure can be seen as: **1)** first constructing $\mathbf{P}$ from the T "intermediate" NNs learned after each epoch from $\mathcal{S}_{\text{prior}}$, 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 $R_\mathcal{S}(h)$ 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 $\mathcal{Q}_\mathcal{S}$. 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 $\mathcal{Q}_\mathcal{S}$. Moreover, our training method to learn the prior differs greatly since **1)** we learn T NNs (*i.e.*, T priors) instead of only one, **2)** we fix the variance of the Gaussian in the posterior $\mathcal{Q}_\mathcal{S}$. Note that, as illustrated in Sect/ 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.

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

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.*, Langford and Caruana (2001); Dziugaite and Roy (2017, 2018); Zhou et al. (2019); Pérez-Ortiz et al. (2021)). In Corollary 8 below, we instantiate the bound of Eq. (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 $Q_S{=}\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_d)$ 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 $\mathbf{w}+\boldsymbol{\epsilon}\in\mathbb{R}^d$. Conversely, the stochastic NN associated with a posterior distribution $Q{=}\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_d)$ predicts the label of a given example by *(i)* first sampling h according to Q, *(ii)* then returning the label predicted by h. Thus, the risk of the stochastic NN is the expected risk value $\mathbb{E}_{h\sim Q}R_{\mathcal{D}}(h)$, where the expectation is taken over *all* h sampled from 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 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 $\mathbf{P}=\{P_1,\ldots,P_T\}$ of T priors on $\mathcal{H}$ where $P_t=\mathcal{N}(\mathbf{v}_t,\sigma^2\mathbf{I}_d)$, for any loss $\ell:\mathcal{H}\times\mathcal{Z}\to\{0,1\}$, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over $S\sim\mathcal{D}^m$ and $\{h_1,\ldots,h_n\}\sim Q^n$, we have simultaneously $\forall P_t\in\mathbf{P}$,*

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

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

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

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.*, $\mathrm{KL}(Q\|P){=}\frac{1}{2\sigma^2}\|\mathbf{w}{-}\mathbf{v}_t\|_2^2$), *(ii)* the bound holds for T priors thanks to a union-bound argument.

5 Experiments with neural networks[10]

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. (2010), Blanchard and Fleuret (2007) and Catoni (2007) (Corollary 7).

5.1 Training method

We follow our Training Method (Sect. 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 Eq. (7), as blanchard the one based on Eq. (8), and as catoni the one based on Eq. (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 et al., 2016) and because this loss is bounded between 0 and 1, which is required for the 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.

5.2 Optimization procedure in algorithms A and A_{prior}

[11] Let $\boldsymbol{\omega}$ 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 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ at each iteration of the optimizer. Then, we forward the examples of the mini-batch to the NN parametrized by the weights $\boldsymbol{\omega} + \boldsymbol{\epsilon}$, and we update $\boldsymbol{\omega}$ according to the bounded cross-entropy loss. Note that during phase **1)**, at the

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

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

end of each epoch t, $\mathcal{P}_t{=}\mathcal{N}(\boldsymbol{\omega}, \sigma^2\mathbf{I}_d){=}\mathcal{N}(\mathbf{v}_t, \sigma^2\mathbf{I}_d)$ and finally at the end of phase **2)** we have $\mathcal{Q}_S{=}\mathcal{N}(\boldsymbol{\omega}, \sigma^2\mathbf{I}_d){=}\mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_d)$.

5.3 Experimental setting

5.3.1 Datasets

We perform our experimental study on three datasets: MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 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 et al., 2015). The model is a 3-hidden layers convolutional network with 96 channels. We use 5×5 convolutions with a padding of size 1, and a stride of size 1 everywhere except on the second convolution where we use a stride of size 2. We adopt the Leaky ReLU activation functions after each convolution. Lastly, we use a global average pooling of size 8×8 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 et al. (2016) with 20 layers. The weights are initialized with Kaiming Normal initializer (He et al., 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 $\mathcal{P}_1, \ldots, \mathcal{P}_T$ 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 $\mathcal{P}_1, \ldots, \mathcal{P}_T$ are learned for $T = 100$ epochs, and the posterior 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 $\mathbf{C}{=}\left\{10^k | k{\in}\{-3, -2, \ldots, +3\}\right\}$. We try different values for $\sigma^2{\in}\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ associated with the disintegrated KL

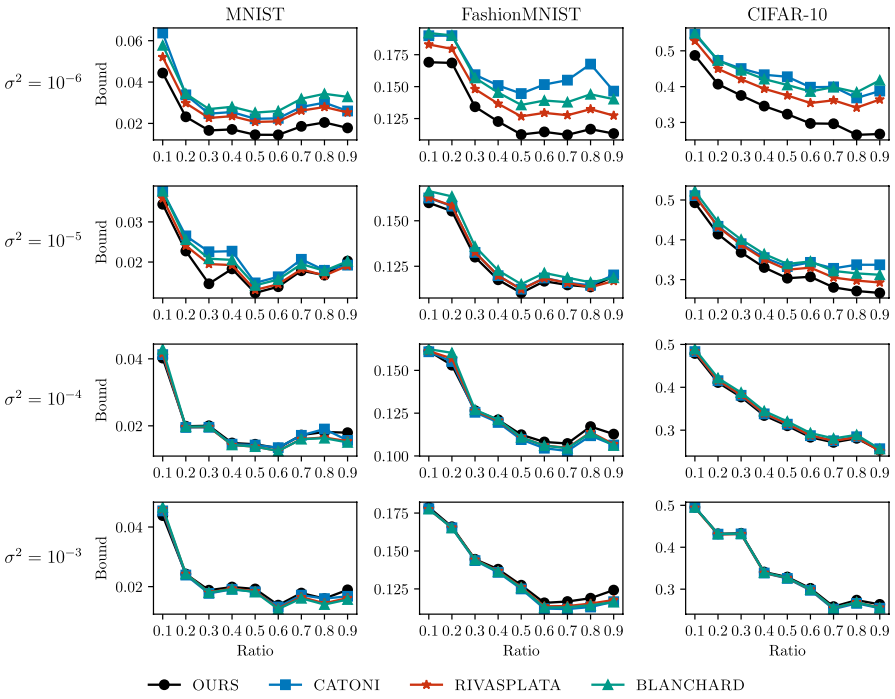

Fig. 1 Evolution of the bound values in terms of the split ratio. The x-axis represents the different split ratios, and the y-axis represents the bound values obtained after their optimization using our Training Method. Each row corresponds to a given variance σ^2, and each column corresponds to a dataset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

divergence $\ln \frac{Q_S(h)}{P(h)} = \frac{1}{2\sigma^2}(\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2)$, the "normal" Rényi divergence $D_2(Q\|P) = \frac{1}{\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ and the KL divergence $\mathrm{KL}(Q\|P) = \frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$. For all the figures, the values are averaged over 400 deterministic NNs sampled from 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 $\mathrm{KL}(Q\|P) = \frac{1}{2\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ 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 1. We first remark that the behavior is

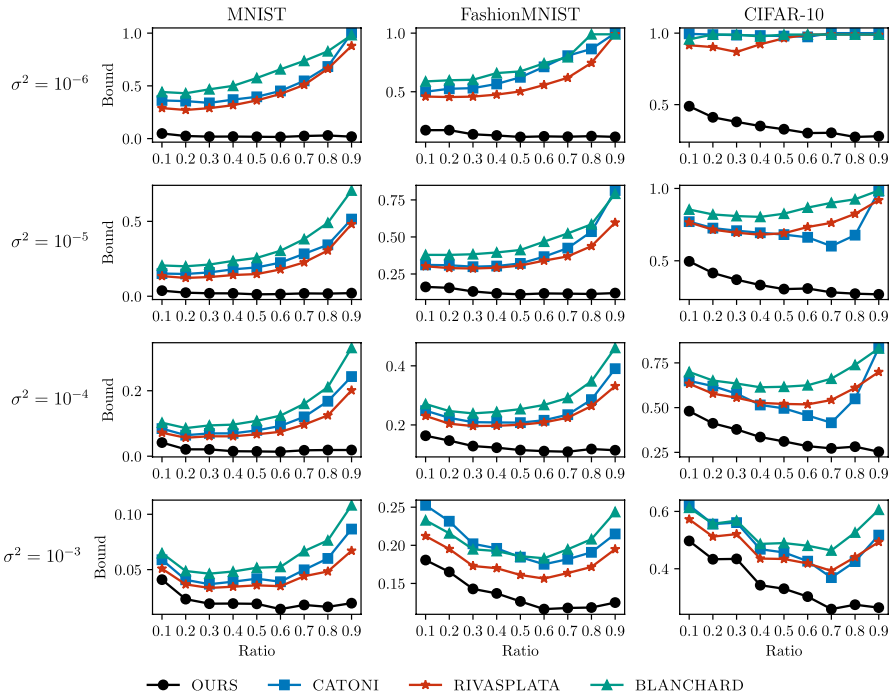

Fig. 2 Evolution of the bound values in terms of the split ratio. The x-axis represents the different split ratios, and the y-axis represents the bound values obtained after their optimization using our Training Method. Each row corresponds to a given variance σ^2, and each column corresponds to a dataset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

different for the two learning rates. On the one hand, for 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 2, 3, 4, 5, 6, 7, 8, 9, 10). 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 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 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.

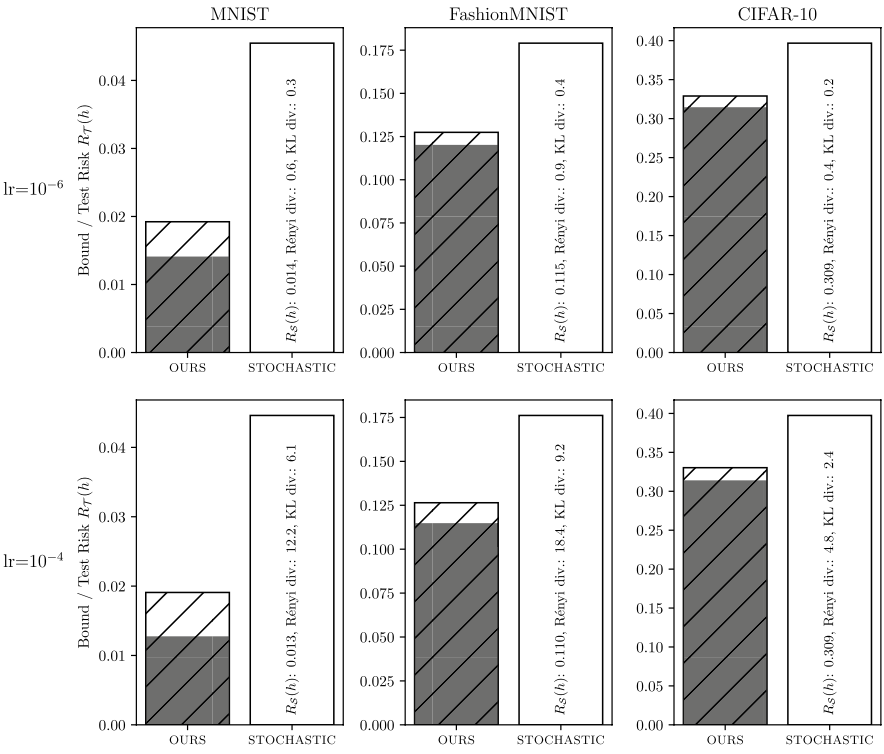

Fig. 3 The values of the PAC-Bayes bound (Corollary 8) and the values of the disintegrated bound (Corollary 6) for learning rates of 10^{-4} and 10^{-6}, and a split ratio is 0.5. The y-axis shows the values of the bounds (the hatched bar for ours (Corollary 6) and the white bar for stochastic (Corollary 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)

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 Fig. 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 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.

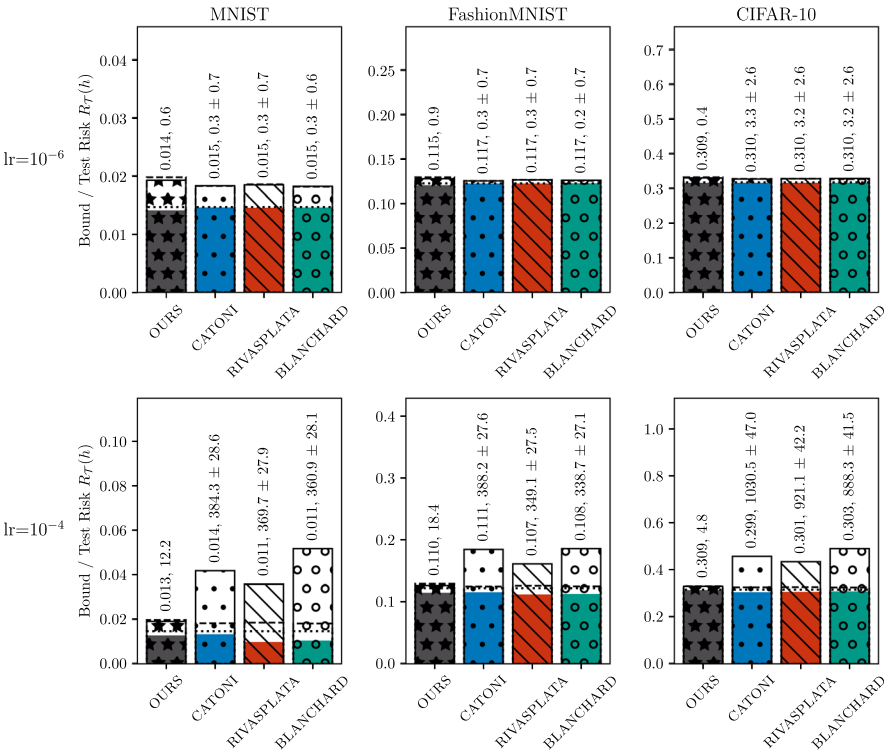

Fig. 4 The value of the bounds (hatched bars) and the test risks (colored bars) for Corollary 6 ("ours ") and Corollary 7 ("catoni ", "rivasplata " and "blanchard ") in two different settings, *i.e.*, with a learning rate of 10^{-6} and 10^{-4} and with split ratio of 0.5. We also plot the value of the 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 $R_{\mathcal{T}}(h)$. The empirical risk $R_{\mathcal{S}}(h)$ 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)

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 Fig. 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 $\mathcal{Q}_{\mathcal{S}}$). Moreover, we report above the bars the mean train risks $R_{\mathcal{S}}(h)$ and the mean/standard deviation divergence values obtained after Step **2)**, *i.e.*, the Rényi divergence $D_2(\mathcal{Q}_{\mathcal{S}}\|\mathcal{P}) = \frac{1}{\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ for ours and the disintegrated KL divergence $\ln \frac{\mathcal{Q}_{\mathcal{S}}(h)}{\mathcal{P}(h)} = \frac{1}{2\sigma^2}\left[\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2\right]$ for the others. First of all, we can remark that we observe two different behaviors for lr=10^{-4} and lr=10^{-6}. For 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 Fig. 4, we observe that the bound values and the test risks are close to the one associated with the prior distribution because the

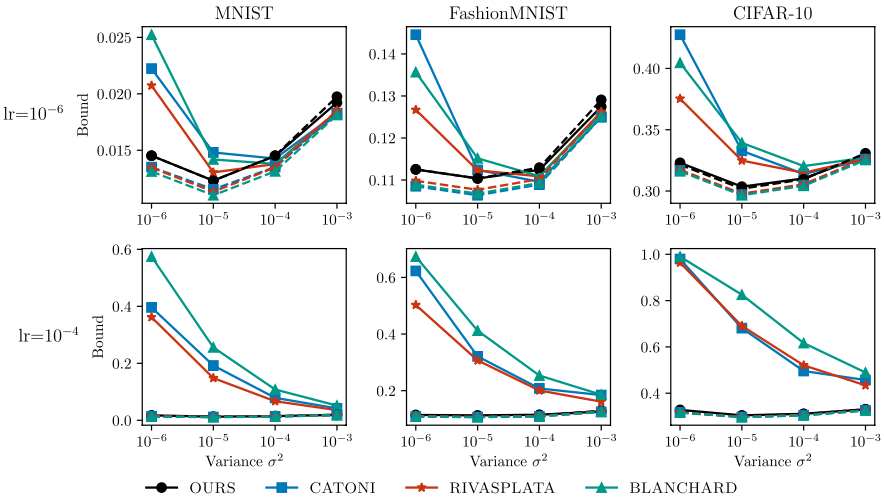

Fig. 5 We plot the evolution of the mean bound values (the plain lines) in terms of the variance σ^2 after optimizing the bounds with our Training Method. Moreover, we plot the mean bound values (the dashed lines) obtained before executing the Step **2)** of our Training Method. The variance is represented 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

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 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 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 lr$\in \{10^{-6}, 10^{-4}\}$, we report in Fig. 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 bound remains low, while the others increase drastically due to the explosion of the disintegrated KL divergence term (see Table 6 in Appendix K for more details). Concretely, the disintegrated KL divergence in Corollary 7 involves the noise $\boldsymbol{\epsilon}$ through $\frac{1}{2\sigma^2}\|\mathbf{w}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2$ compared to our divergence which is $\frac{1}{\sigma^2}\|\mathbf{w}-\mathbf{v}_t\|_2^2$ (without noise). Then, the sampled noise during the optimization procedure $\boldsymbol{\epsilon}$ 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 Corollarys 6 and 7 (Eq. 7). Roughly speaking, the divergence in Corollary 6 does not depend on the sampled hypothesis h (with weights $\boldsymbol{\omega}+\boldsymbol{\epsilon}$), while the divergence of Eq. (7) does. In consequence, the derivatives are less dependent on h for Corollary 6 than for Eq. (7). To be convinced of this, we propose to study the gradient with respect to the current mean vector $\boldsymbol{\omega}$. On the one hand, the gradient $\frac{\partial R_S(h)}{\partial \boldsymbol{\omega}}$ of the risk *w.r.t.* $\boldsymbol{\omega}$ is the same for both bounds; hence, the phenomenon cannot come from this derivative. On the other hand, the gradients of the divergence in Eq. (7) and Corollary 6 are respectively

$$\frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m}\left(\frac{\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2-\|\boldsymbol{\epsilon}\|_2^2}{2\sigma^2}\right)\right] = \frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m2\sigma^2}\|\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\|_2^2\right]$$
$$= \frac{1}{m\sigma^2}\left(\boldsymbol{\omega}+\boldsymbol{\epsilon}-\mathbf{v}_t\right) = \diamondsuit,$$

and

$$\frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m}\left(\frac{\|\boldsymbol{\omega}-\mathbf{v}_t\|_2^2}{\sigma^2}\right)\right] = \frac{\partial}{\partial \boldsymbol{\omega}}\left[\frac{1}{m\sigma^2}\|\boldsymbol{\omega}-\mathbf{v}_t\|_2^2\right]$$
$$= \frac{2}{m\sigma^2}\left(\boldsymbol{\omega}-\mathbf{v}_t\right) = \heartsuit.$$

From the two derivatives, we deduce that $\diamondsuit = \frac{1}{2}\heartsuit + \frac{1}{m\sigma^2}\boldsymbol{\epsilon}$. 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

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

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

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

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

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

This leads 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 $\phi : \mathcal{H}\times\mathcal{Z}^m \to \mathbb{R}_+^*$, for any $\alpha > 1$, for any $\delta \in (0,1]$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, we have*

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

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 $D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$ suggests that the learned posterior $\mathcal{Q}_\mathcal{S}$ should be close enough to the prior $\mathcal{P}$ to get a low bound. While in Theorem 9, the Sibson's mutual information $I_\alpha(h';\mathcal{S}')$ suggests that the random variable h has to be *not too much correlated* to $\mathcal{S}$. 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.

[12] We 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 Eq. (5), *i.e.*, when the derandomization step consists in *(i)* learning a posterior distribution $\mathcal{Q}_S$ on the classifiers set (given an algorithm, a learning sample S and a prior distribution $\mathcal{P}$) and *(ii)* sampling a hypothesis h from this posterior $\mathcal{Q}_S$. While our bound can be looser than the ones of Rivasplata et al. (2010); Blanchard and Fleuret (2007); Catoni (2007), 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 Rivasplata et al. (2010); Blanchard and Fleuret (2007); Catoni (2007) 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 Rivasplata et al. (2010); Blanchard and Fleuret (2007); Catoni (2007). 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 et al., 2021) or Transductive Learning (Bégin et al., 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.

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 $\phi : \mathcal{H} \times \mathcal{Z}^m \to \mathbb{R}_+^*$, for any $\delta \in (0,1]$ we have*

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

and

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

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

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

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

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

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

By merging Eqs. (A1) and (A2), we obtain Eq. (1).

The proof of Eq. (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

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

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

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

By merging Equations (A3) and (A4), Eq. (2) is obtained. □

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

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

where $\mathcal{Q}_\mathcal{S} \triangleq A(\mathcal{S},\mathcal{P})$ 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 $\mathcal{Q}_\mathcal{S} = A(\mathcal{S},\mathcal{P})$ 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

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

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

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

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

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

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

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

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:

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

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

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

From this inequality, we can deduce that

$$
\begin{aligned}
\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_{\mathcal{S}}}\Bigg[\forall\mathcal{P}\in\mathcal{M}^*(\mathcal{H}),\frac{\alpha}{\alpha-1}\ln\left(\phi(h,\mathcal{S})\right)\le D_{\alpha}(\mathcal{Q}_{\mathcal{S}}\|\mathcal{P})\\
+\frac{\alpha}{\alpha-1}\ln\frac{2}{\delta}+\ln\left(\mathbb{E}_{h'\sim\mathcal{P}}\left(\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\right)\Bigg]\ge1-\frac{\delta}{2}.
\end{aligned}
\tag{B5}
$$

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

$$
\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\mathbb{E}_{h'\sim\mathcal{P}}\left(\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\le\frac{2}{\delta}\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\left(\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\right]\ge1-\frac{\delta}{2}.
$$

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

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

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

$$
\begin{aligned}
&\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_{\mathcal{S}}}\left[\mathbb{E}_{h'\sim\mathcal{P}}\left(\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\le\frac{2}{\delta}\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\left(\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\right]\ge1-\frac{\delta}{2}\\
&\iff\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m,h\sim\mathcal{Q}_{\mathcal{S}}}\Bigg[\frac{\alpha}{\alpha-1}\ln\frac{2}{\delta}+\ln\left(\mathbb{E}_{h'\sim\mathcal{P}}\left(\phi(h',\mathcal{S})^{\frac{\alpha}{\alpha-1}}\right)\right)\le\frac{2\alpha-1}{\alpha-1}\ln\frac{2}{\delta}\\
&\qquad+\ln\left(\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\left(\phi(h',\mathcal{S}')^{\frac{\alpha}{\alpha-1}}\right)\right)\Bigg]\ge1-\frac{\delta}{2}.
\end{aligned}
\tag{B6}
$$

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

Appendix C: Proof of Corollary 3

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

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

when $\alpha \to +\infty$, *we have*

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

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

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

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

Proof Starting from Theorem 2 and rearranging, we have

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

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

When $\alpha \to 1$.

First, we have $\lim_{\alpha\to1^+} \frac{2\alpha-1}{\alpha} \ln \frac{2}{\delta} = \ln \frac{2}{\delta}$ and $\lim_{\alpha\to1^+} \frac{\alpha-1}{\alpha} D_\alpha(\mathcal{Q}_\mathcal{S}\|\mathcal{P}) = 0$.

Furthermore, note that

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

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

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

Hence, we have

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

Finally, we can deduce that

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

When $\alpha\to+\infty$.

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

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

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

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

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

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

Finally, we can deduce that

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

□

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*

$$\sqrt{\frac{a}{b}}=\operatorname{argmin}_{\lambda>0}\left(\frac{a}{\lambda}+\lambda b\right),\ \text{and}\ 2\sqrt{ab}=\min_{\lambda>0}\left(\frac{a}{\lambda}+\lambda b\right),$$
$$\text{and}\ \forall\lambda>0,\sqrt{ab}\le\frac{1}{2}\left(\frac{a}{\lambda}+\lambda b\right).$$

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. □

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 $\phi:\mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}_+^*$, for any $\delta\in(0,1]$, for any algorithm $A:\mathcal{Z}^m\times\mathcal{M}^*(\mathcal{H})\to\mathcal{M}(\mathcal{H})$, we have*

$$\mathbb{P}_{\substack{\mathcal{S}\sim\mathcal{D}^m,\\ h\sim\mathcal{Q}_\mathcal{S}}}\left(\forall\lambda>0,\ \ln\left(\phi(h,\mathcal{S})\right)\le\ln\left[\frac{\lambda}{2}e^{D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P})}+\frac{8}{2\lambda\delta^3}\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\left[\phi(h',\mathcal{S}')^2\right]\right]\right)\ge1-\delta,$$

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

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

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

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

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

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

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

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

Hence, we can deduce that

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

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

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

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

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

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,

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

Hence, we can deduce that

$$
\begin{aligned}
& \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m, h\sim\mathcal{Q}_\mathcal{S}}\Bigg[\forall\lambda{>}0,\ \ln\left(\frac{1}{2}\left[\lambda\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))+\frac{4}{\lambda\delta^2}\mathbb{E}_{h'\sim\mathcal{P}}\big(\phi(h',\mathcal{S})^2\big)\right]\right) \\
& \le \ln\left(\frac{1}{2}\left[\lambda\exp(D_2(\mathcal{Q}_\mathcal{S}\|\mathcal{P}))+\frac{8}{\lambda\delta^3}\mathbb{E}_{\mathcal{S}'\sim\mathcal{D}^m}\mathbb{E}_{h'\sim\mathcal{P}}\big(\phi(h',\mathcal{S}')^2\big)\right]\right)\Bigg] \ge 1-\tfrac{\delta}{2}.
\end{aligned} \tag{D8}
$$

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

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 $\phi : \mathcal{H}\times\mathcal{Z}^m\to\mathbb{R}_+^*$, for any algorithm $A : \mathcal{Z}^m\times\mathcal{M}^*(\mathcal{H})\to\mathcal{M}(\mathcal{H})$, let*

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

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

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

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

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

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

Since ln is a strictly increasing function, we have

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

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

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

which is the desired result. □

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

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

where $\mathcal{Q}_S\triangleq A(\mathcal{S},\mathcal{P})$ 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 Eq. (B5), we have

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

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

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

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

We now prove Corollary 6 from Theorem 2'.

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

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

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

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

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

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

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

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

Appendix G: Proof of Corollary 7

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

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

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

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

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

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

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

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

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

□

We can now prove Corollary 7.

Corollary 7 *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any set $\mathcal{H}$, for any set $\mathbf{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_T\}$ of T priors on $\mathcal{H}$ where $\mathcal{P}_t = \mathcal{N}(\mathbf{v}_t, \sigma^2\mathbf{I}_d)$, for any algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, for any loss $\ell : \mathcal{H}\times\mathcal{Z}\to[0,1]$, 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 $\mathbf{w}{+}\boldsymbol{\epsilon}$, we have $\forall\mathcal{P}_t \in \mathbf{P}$*

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

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

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

with $[x]_+ = \max(x, 0)$, and $\mathrm{kl}_+(R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h)) = \mathrm{kl}(R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h))$ if $R_{\mathcal{S}}(h) < R_{\mathcal{D}}(h)$ and 0 otherwise. Moreover, $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$ is a Gaussian noise such that $\mathbf{w}{+}\boldsymbol{\epsilon}$ are the weights of $h\sim\mathcal{Q}_{\mathcal{S}}$ with $\mathcal{Q}_{\mathcal{S}} = \mathcal{N}(\mathbf{w}, \sigma^2\mathbf{I}_d)$, 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. (2010) with $\phi(h,\mathcal{S}) = \exp\left[m\mathrm{kl}(R_{\mathcal{S}}(h)\|R_{\mathcal{D}}(h))\right]$, however, we apply the theorem T times for each prior $\mathcal{P}_t \in \mathbf{P}$ (with a confidence $\frac{\delta}{T}$ instead of δ). Hence, for each prior $\mathcal{P}_t \in \mathbf{P}$, we have with probability at least $1-\frac{\delta}{T}$ over the random choice of $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_{\mathcal{S}}$

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

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

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

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. □

Appendix H: Proof of Corollary 8

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

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

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

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

Proof We instantiate Eq. (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 $\mathcal{H}$,

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

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

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

Additionally, we obtained Eq. (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. □

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

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

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

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

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

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

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

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

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

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

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

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

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 chosing 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 $t \in \{1, \ldots, T\}$ (line 16) by updating $\boldsymbol{\omega}$ (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 $\boldsymbol{\omega} + \boldsymbol{\epsilon}$ (Line 12 and 13) and update $\boldsymbol{\omega}$ 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 $\boldsymbol{\omega} + \boldsymbol{\epsilon}$ (Line 38 and 39). At each iteration, the algorithm updates the weights $\boldsymbol{\omega}$ (Line 39) by optimizing

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

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

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

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

Note that, as stated in Sect. 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 $\boldsymbol{\omega}$ 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)

`Optimizing the prior` $\mathcal{P}$ — Step **1)** — Algorithm A_{prior}

$\boldsymbol{\omega} \leftarrow$ Initialize the weights $\boldsymbol{\omega}$
$r^* \leftarrow +\infty$
$t^* \leftarrow +\infty$
for each epoch $t \leftarrow 1, \ldots, T$ **do**

  `Optimizing the prior` $\mathcal{P}_t$
  for each mini-batch $\mathcal{U} \subseteq \mathcal{S}_{\text{prior}}$ **do**
    Sample a noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$
    $h \leftarrow$ Hypothesis parameterized by $\boldsymbol{\omega}+\boldsymbol{\epsilon}$
    $\boldsymbol{\omega} \leftarrow$ Update $\boldsymbol{\omega}$ with $R_{\mathcal{U}}(h)$
  end for
  $\mathcal{P}_t \leftarrow \mathcal{N}(\boldsymbol{\omega}, \sigma^2 \mathbf{I}_d)$ where $\mathcal{P}_t$=$\mathcal{N}(\mathbf{v}_t, \sigma^2 \mathbf{I}_d)$

  `Selecting the prior` $\mathcal{P}$
  for each mini-batch $\mathcal{U} \subseteq \mathcal{S}$ **do**
    Sample a noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$
    $h \leftarrow$ Hypothesis parameterized by $\mathbf{v}_t+\boldsymbol{\epsilon}$
    $r \leftarrow r + \sum_{(\mathbf{x},y)\in\mathcal{U}} \ell(h, (\mathbf{x}, y))$
  end for
  if $r < r^*$ **then**
    $r^* \leftarrow r$
    $\mathcal{P} \leftarrow \mathcal{P}_t$
    $t^* \leftarrow t$
  end if
end for

`Optimizing the posterior` $\mathcal{Q}_{\mathcal{S}}$ — Step **2)** — Algorithm A

$\mathcal{Q}_{\mathcal{S}} \leftarrow \mathcal{P} = \mathcal{N}(\boldsymbol{\omega}, \sigma^2 \mathbf{I}_d) = \mathcal{N}(\mathbf{v}_{t^*}, \sigma^2 \mathbf{I}_d)$
for each epoch $t' \leftarrow 1, \ldots, T'$ **do**
  for each mini-batch $\mathcal{U} \subseteq \mathcal{S}$ **do**
    Sample a noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$
    $h \leftarrow$ Hypothesis parameterized by $\boldsymbol{\omega}+\boldsymbol{\epsilon}$
    $\boldsymbol{\omega} \leftarrow$ Update $\boldsymbol{\omega}$ with either Equation (I10), (I11), (I12), or (I13)
  end for
end for
return $\mathcal{Q}_{\mathcal{S}} = \mathcal{N}(\boldsymbol{\omega}, \sigma^2 \mathbf{I}_d) = \mathcal{N}(\mathbf{w}, \sigma^2 \mathbf{I}_d)$ 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 $\phi : \mathcal{H} \times \mathcal{Z}^m \to [1, +\infty[$, for any $\delta \in (0, 1]$, for any deterministic algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, we have*

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

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

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

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

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

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

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

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

We are now ready to prove Theorem 13.

Theorem 13 *For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any hypothesis set $\mathcal{H}$, for any measurable function $\phi : \mathcal{H} \times \mathcal{Z}^m \to [1, +\infty[$, for any $\delta \in (0, 1]$, for any deterministic algorithm $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, 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 $I(h;\mathcal{S}) = \min_{\mathcal{P}\in\mathcal{M}^*(\mathcal{H})} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}\mathrm{KL}(\mathcal{Q}_\mathcal{S}\|\mathcal{P})$.

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

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

Hence,

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

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 $h \sim \mathcal{Q}_\mathcal{S}$

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

□

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 $\phi(h,\mathcal{S}) = \exp\left[m\mathrm{kl}(R_\mathcal{S}(h)\|R_\mathcal{D}(h))\right]$, 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 $A : \mathcal{Z}^m \times \mathcal{M}^*(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, we have*

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

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

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

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

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

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

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

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

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

Hence, we can deduce that

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

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

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

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

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

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

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

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

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

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 $h\sim\mathcal{Q}_{\mathcal{S}}$

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

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 $\alpha\to1^+$, with probability at least $1-\delta$ we have*

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

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

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

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

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

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

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

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

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

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

Hence, we have

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

Finally, we can deduce that

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

When $\alpha\to+\infty$.

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

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

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

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

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

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

Finally, we can deduce that

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

□

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

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 $h\sim\mathcal{Q}_{\mathcal{S}}$:

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

Hence, we compare Eq. (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 $A:\mathcal{Z}^m\times\mathcal{M}^*(\mathcal{H})\to\mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$ and $h\sim\mathcal{Q}_{\mathcal{S}}$, we have*

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

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

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

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

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

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

Appendix K: Results presented in Section 5

This appendix presents the details of the results of Sect. 5. Tables 2, 3, 4, 5, 6, 7, 8, 9, 10 report empirical results for split ratios going from 0.0 to 0.9 presented in Figs. 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 11, 12, 13 report the

Table 1 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.901 ± .002	.908 ± .002	.901 ± .002	.005	.897 ± .013	.904 ± .012	.897 ± .012	.009	.898 ± .017	.905 ± .016	.898 ± .016	.027	.902 ± .015	.908 ± .014	.901 ± .015	.671
	Blanchard	.901 ± .002	.926 ± .002	.901 ± .002	122.846 ± 15.952	.897 ± .013	.912 ± .012	.897 ± .013	39.350 ± 8.999	.898 ± .017	.907 ± .016	.898 ± .017	13.023 ± 4.818	.901 ± .015	.907 ± .014	.901 ± .014	3.041 ± 2.459
	Catoni	.901 ± .002	.926 ± .003	.901 ± .002	121.860 ± 15.930	.897 ± .013	.909 ± .012	.897 ± .013	38.552 ± 8.872	.898 ± .017	.905 ± .016	.898 ± .017	12.474 ± 4.774	.901 ± .014	.906 ± .013	.901 ± .014	3.088 ± 2.379
	Rivasplata	.901 ± .002	.920 ± .002	.901 ± .002	123.301 ± 15.941	.896 ± .014	.908 ± .012	.896 ±.013	39.195 ± 8.959	.897 ± .017	.905 ± .016	.897 ± .017	12.827 ± 4.858	.902 ± .015	.907 ± .014	.901 ± .015	3.232 ± 2.454
	Stochastic	–	.944	–	.002	–	.941	–	.004	–	.941	–	.014	–	.944	–	.336
Fashion	Ours	.970 ± .028	.972 ± .025	.970 ± .027	.016	.944 ± .038	.949 ± .035	.944 ± .037	.046	.910 ± .027	.917 ± .026	.910 ± .027	.140	.901 ± .026	.909 ± .025	.901 ± .026	1.255
	Blanchard	.970 ± .029	.978 ± .019	.970 ± .028	122.508 ± 16.085	.942 ± .038	.952 ± .032	.943 ± .038	39.957 ± 8.610	.910 ± .031	.919 ± .029	.910 ± .031	12.649 ± 4.846	.899 ± .028	.905 ± .027	.899 ± .028	3.206 ± 2.566
	Catoni	.970 ± .028	.983 ± .017	.970 ± .027	122.364 ± 15.860	.945 ± .038	.954 ± .036	.945 ± .037	38.555 ± 8.873	.912 ± .032	.919 ± .031	.912 ± .032	12.167 ± 4.762	.899 ± .027	.905 ± .026	.899 ± .027	3.122 ± 2.392
	Rivasplata	.970 ± .028	.977 ± .021	.971 ± .027	123.328 ± 15.929	.943 ± .038	.950 ± .033	.943 ± .038	39.300 ± 8.991	.908 ± .031	.916 ± .029	.908 ± .031	12.627 ± 4.890	.899 ± .028	.905 ± .027	.899 ± .028	3.591 ± 2.610
	Stochastic	–	.990	–	.008	–	.975	–	.023	–	.950	–	.070	–	.944	–	.627

Table 1 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
CIFAR-10	Ours	.899 ± .000	.907 ± .000	.899 ± .000	3.113	.896 ± .002	.914 ± .002	.894 ± .002	107.797	.826 ± .011	.885 ± .009	.825 ± .010	76.475	.786 ± .019	.851 ± .015	.788 ± .018	714.351
	Blanchard	.899 ± .000	.940 ± .001	.898 ± .000	314.983 ± 26.377	.888 ± .004	.927 ± .002	.885 ± .003	28.250 ± 25.255	.823 ± .010	.885 ± .008	.822 ± .010	422.401 ± 29.323	.798 ± .019	.856 ± .015	.799 ± .018	292.706 ± 25.318
	Catoni	.899 ± .000	.941 ± .000	.898 ± .000	285.415 ± 25.085	.894 ± .002	.930 ± .004	.892 ± .002	169.713 ± 19.543	.857 ± .010	.915 ± .009	.856 ± .010	273.554 ± 23.212	.815 ± .019	.864 ± .017	.816 ± .018	209.069 ± 21.230
	Rivas-plata	.899 ± .001	.930 ± .001	.898 ± .000	362.070 ± 28.420	.864 ± .004	.933 ± .002	.862 ± .004	1568.007 ± 55.492	.748 ± .010	.837 ± .007	.750 ± .009	1219.178 ± 49.610	.769 ± .018	.828 ± .015	.771 ± .017	526.068 ± 33.837
	Stochastic	–	.942	–	1.557	–	.945	–	53.898	–	.914	–	38.237	–	.884	–	357.175
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.901 ± .002	.909 ± .002	.901 ± .002	3.767	.896 ± .014	.904 ± .013	.896 ± .014	.835	.898 ± .016	.905 ± .015	.898 ± .016	1.062	.901 ± .015	.909 ± .014	.901 ± .015	6.022
	Blanchard	.900 ± .003	.990 ± .000	.900 ± .003	12004.196 ± 152.632	.894 ± .017	.986 ± .006	.894 ± .016	3837.785 ± 93.560	.888 ± .021	.957 ± .013	.888 ± .020	1221.198 ± 49.920	.898 ± .015	.939 ± .012	.897 ± .015	391.343 ± 28.182
	Catoni	.900 ± .003	.997 ± .002	.900 ± .003	5694.194 ± 102.906	.889 ± .020	.967 ± .012	.889 ± .019	3331.617 ± 78.945	.879 ± .025	.941 ± .016	.880 ± .025	1481.726 ± 53.973	.888 ± .023	.937 ± .015	.888 ± .023	567.893 ± 33.441
	Rivas-plata	.900 ± .004	.990 ± .000	.900 ± .003	1199.818 ± 152.557	.892 ± .017	.970 ± .009	.892 ± .016	3846.699 ± 84.643	.886 ± .020	.940 ± .015	.886 ± .020	1224.463 ± 49.970	.897 ± .018	.928 ± .015	.897 ± .018	393.757 ± 29.158
	Stochastic	–	.944	–	1.884	–	.940	–	.417	–	.941	–	.531	–	.944	–	3.011

Table 1 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
Fashion	Ours	.977 ± .024	.979 ± .021	.977 ± .023	3.926	.947 ± .038	.951 ± .035	.947 ± .038	1.623	.907 ± .030	.914 ± .029	.907 ± .030	2.947	.900 ± .026	.910 ± .025	.900 ± .026	15.978
	Blanchard	.984 ± .015	.990 ± .000	.984 ± .015	12019.121 ± 166.251	.912 ± .029	.988 ± .004	.911 ± .029	3846.861 ± 84.568	.883 ± .029	.953 ± .019	.883 ± .029	1232.645 ± 5.285	.403 ± .041	.648 ± .038	.399 ± .041	3853.231 ± 87.867
	Catoni	.983 ± .018	1.000 ± .000	.983 ± .017	5654.642 ± 114.040	.903 ± .021	.985 ± .012	.902 ± .021	4354.538 ± 94.427	.751 ± .033	.867 ± .023	.750 ± .033	2702.652 ± 76.863	.504 ± .041	.673 ± .037	.502 ± .041	3172.609 ± 78.698
	Rivas-plata	.983 ± .016	.990 ± .000	.983 ± .016	11976.720 ± 165.964	.905 ± .023	.975 ± .007	.905 ± .023	3855.872 ± 84.676	.855 ± .035	.916 ± .027	.855 ± .035	125.110 ± 51.837	.365 ± .032	.559 ± .032	.359 ± .033	4823.725 ± 103.813
	Stochastic	−	.990	−	1.963	−	.977	−	.812	−	.948	−	1.473	−	.944	−	7.989
CIFAR-10	Ours	.899 ± .000	.915 ± .000	.899 ± .000	63.416	.890 ± .003	.932 ± .003	.886 ± .003	68.353	.786 ± .011	.888 ± .008	.787 ± .010	2072.610	.769 ± .017	.859 ± .013	.770 ± .017	1406.824
	Blanchard	.869 ± .002	.990 ± .000	.866 ± .001	27237.938 ± 251.770	.813 ± .004	.990 ± .000	.812 ± .003	12052.733 ± 159.732	.697 ± .011	.920 ± .005	.700 ± .009	5137.799 ± 103.680	.674 ± .020	.861 ± .014	.675 ± .020	2814.450 ± 76.004
	Catoni	.928 ± .001	1.000 ± .000	.925 ± .001	2145276.795 ± 2095.160	.821 ± .002	1.000 ± .000	.821 ± .002	375019.277 ± 896.780	.689 ± .011	.870 ± .007	.692 ± .010	5292.535 ± 106.380	.629 ± .019	.805 ± .015	.628 ±.019	4159.131 ± 96.763
	Rivas-plata	.867 ± .002	.990 ± .000	.864 ± .001	35956.152 ± 268.304	.812 ± .004	.976 ± .001	.811 ± .003	12135.134 ± 157.621	.698 ± .010	.874 ± .006	.701 ± .009	5191.665 ± 102.712	.677 ± .020	.819 ± .015	.678 ± .019	2839.514 ± 81.432
	Stochastic	−	.947	−	31.708	−	.954	−	34.176	−	.908	−	1036.305	−	.886	−	703.412

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this figure, that the split ratio is 0.0

Table 2 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.035 ± .000	.044 ± .000	.039 ± .000	.622	.024 ± .000	.034 ± .000	.029 ± .000	2.122	.029 ± .002	.040 ± .002	.034 ± .002	12.754	.034 ± .004	.044 ± .004	.038 ± .004	7.303
	Blanchard	.034 ± .000	.058 ± .002	.038 ± .000	99.876 ± 14.858	.024 ± .000	.038 ± .001	.030 ± .000	21.775 ± 6.848	.034 ± .002	.043 ± .002	.038 ± .002	3.949 ± 2.877	.039 ± .005	.047 ± .005	.043 ± .005	.590 ± 1.085
	Catoni	.035 ± .000	.064 ± .001	.039 ± .000	119.663 ± 15.854	.024 ± .000	.038 ± .001	.030 ± .000	26.277 ± 7.490	.033 ± .002	.041 ± .002	.037 ± .002	4.067 ± 2.882	.038 ± .005	.045 ± .005	.042 ± .004	.759 ± 1.217
	Rivasplata	.034 ± .000	.052 ± .001	.038 ± .000	104.880 ± 15.268	.024 ± .000	.036 ± .001	.029 ± .000	23.007 ± 7.187	.033 ± .002	.042 ± .002	.037 ± .002	4.116 ± 2.845	.038 ± .005	.046 ± .004	.042 ± .004	.775 ± 1.231
	Stochastic	–	.080	–	.311	–	.067	–	1.061	–	.074	–	6.377	–	.079	–	3.651
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	.176 ± .006	.179 ± .006	.168 ± .005	1.268
	Blanchard	.165 ± .001	.192 ± .002	.159 ± .000	96.822 ± 14.116	.157 ± .001	.166 ± .002	.150 ± .001	21.592 ± 6.681	.163 ± .003	.162 ± .003	.153 ± .003	3.846 ± 2.660	.178 ± .005	.178 ± .005	.170 ± .005	.463 ± .954
	Catoni	.165 ± .001	.190 ± .003	.159 ± .000	119.927 ± 15.938	.157 ± .001	.163 ± .002	.150 ± .001	26.363 ± 7.355	.162 ± .003	.161 ± .003	.152 ± .003	4.152 ±2.945	.177 ± .006	.178 ± .006	.169 ± .006	.548 ± 1.032
	Rivasplata	.165 ± .001	.183 ± .002	.158 ± .000	101.954 ± 14.463	.157 ± .001	.163 ± .002	.150 ± .001	23.098 ± 6.977	.162 ± .003	.161 ± .003	.153 ± .003	3.852 ± 2.798	.177 ± .006	.177 ± .006	.169 ± .006	.516 ± .985
	Stochastic	–	.227	–	.290	–	.216	–	1.064	–	.218	–	1.751	–	.237	–	.634

Table 2 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
CIFAR-10	Ours	.479 ± .000	.487 ± .000	.472 ± .000	.052	.479 ± .000	.493 ± .000	.477 ± .000	.065	.458 ± .001	.479 ± .000	.463 ± .000	.299	.480 ± .002	.495 ± .001	.480 ± .001	.793
	Blanchard	.479 ± .000	.550 ± .003	.472 ± .000	27.644 ± 22.868	.479 ± .000	.522 ± .003	.477 ± .000	85.476 ± 12.781	.458 ± .001	.489 ± .003	.463 ± .000	24.608 ± 7.136	.481 ± .002	.495 ± .002	.480 ± .001	5.093 ± 3.299
	Catoni	.479 ± .000	.546 ± .005	.472 ± .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	.480 ± .002	.495 ± .001	.480 ± .001	5.468 ± 3.315
	Rivas-plata	.479 ± .000	.528 ± .002	.472 ± .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	.481 ± .002	.494 ± .001	.480 ± .001	5.155 ± 3.260
	Stochastic	–	.558	–	.026	–	.564	–	.032	–	.550	–	.150	–	.566	–	.397
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	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	.025 ± .002	.041 ± .002	.029 ± .002	97.840
	Blanchard	.032 ± .000	.442 ± .003	.036 ± .000	1181.482 ± 14.449	.022 ± .000	.206 ± .003	.027 ± .000	3851.110 ± 84.274	.019 ± .001	.102 ± .002	.023 ± .001	1306.371 ± 51.396	.024 ± .002	.065 ± .003	.027 ± .002	411.772 ± 29.458
	Catoni	.035 ± .000	.362 ± .003	.039 ± .000	11925.734 ± 145.511	.024 ± .000	.152 ± .002	.029 ± .000	3841.248 ± 84.033	.027 ± .002	.084 ± .002	.032 ± .001	1235.287 ± 49.454	.027 ± .002	.059 ± .003	.030 ± .002	403.300 ± 28.587
	Rivas-plata	.030 ± .000	.289 ± .002	.034 ± .000	12022.576 ± 151.157	.021 ± .000	.134 ± .002	.026 ± .000	3912.803 ± 85.146	.018 ± .000	.072 ± .001	.022 ± .000	1348.169 ± 53.400	.023 ± .002	.051 ± .002	.026 ± .001	424.971 ± 29.301

Table 2 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Stochastic	–	.084	–	17.674	–	.069	–	15.376	–	.072	–	76.887	–	.072	–	48.920
Fashion	Ours	.166 ± .001	.172 ± .000	.159 ± .000	13.084	.157 ± .001	.163 ± .001	.150 ± .001	16.513	.159 ± .002	.164 ± .002	.149 ± .002	2.344	.176 ± .005	.181 ± .005	.168 ± .005	11.331
	Blanchard	.160 ± .001	.588 ± .003	.153 ± .000	1089.829 ± 137.125	.150 ± .001	.379 ± .003	.141 ± .001	3744.491 ± 83.656	.155 ± .002	.271 ± .003	.145 ± .002	1221.062 ± 49.548	.173 ± .005	.233 ± .006	.165 ± .004	369.721 ± 27.211
	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	.174 ± .005	.252 ± .006	.166 ± .004	393.542 ± 27.890
	Rivas-plata	.158 ± .001	.459 ± .002	.151 ± .000	11541.128 ± 14.706	.149 ± .001	.302 ± .002	.140 ± .001	3878.145 ± 85.782	.154 ± .002	.230 ± .002	.144 ± .001	1244.035 ± 49.268	.172 ± .005	.212 ± .005	.164 ± .004	378.990 ± 27.559
	Stochastic	–	.229	–	6.542	–	.219	–	8.257	–	.219	–	1.172	–	.239	–	5.666
CIFAR-10	Ours	.479 ± .000	.489± .000	.472 ± .000	4.882	.479 ± .000	.496 ± .000	.477 ± .000	9.273	.458 ± .001	.480 ± .000	.463 ± .000	4.988	.480 ± .002	.497 ± .001	.479 ± .001	8.681
	Blanchard	.479 ± .000	.957 ± .001	.471 ± .000	22201.935 ± 218.369	.479 ± .000	.854 ± .002	.477 ± .000	8777.551 ± 125.716	.457 ± .001	.699 ± .003	.461 ± .000	2758.075 ± 77.155	.474 ± .001	.613 ± .003	.472 ± .001	903.948 ± 4.742
	Catoni	.479 ± .000	.995 ± .000	.471 ± .000	26347.736 ± 225.908	.479 ± .000	.771 ± .002	.477 ± .000	8566.272 ± 124.834	.455 ± .001	.650 ± .002	.459 ± .000	3117.566 ± 75.178	.468 ± .001	.621 ± .001	.466 ± .001	1481.520 ± 52.533
	Rivas-plata	.479 ± .000	.915 ± .001	.471 ± .000	29489.241 ± 241.010	.479 ± .000	.765 ± .002	.477 ± .000	867.264 ± 126.038	.456 ± .001	.633 ± .002	.460 ± .000	2776.052 ± 72.901	.472 ± .001	.572 ± .002	.470 ± .001	937.091 ± 42.116
	Stochastic	–	.559	–	2.441	–	.566	–	4.637	–	.551	–	2.494	–	.567	–	4.340

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.1

Table 3 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.016 ± .000	.023 ± .000	.019 ± .000	.336	.015 ± .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 ± 14.260	.015 ± .000	.026 ± .001	.019 ± .000	21.153 ± 6.514	.015 ± .001	.020 ± .001	.016 ± .001	3.362 ± 2.569	.020 ± .002	.024 ± .002	.021 ± .002	.371 ± .875
	Catoni	.016 ± .000	.034 ± .001	.019 ± .000	116.744 ± 15.447	.015 ± .000	.027 ± .002	.019 ± .000	24.135 ± 7.075	.015 ± .001	.020 ± .001	.016 ± .001	3.352 ± 2.667	.020 ± .002	.024 ± .002	.021 ± .002	.410 ± .890
	Rivasplata	.016 ± .000	.030 ± .001	.019 ± .000	101.334 ± 14.728	.015 ± .000	.024 ± .001	.019 ± .000	21.663 ± 6.603	.015 ± .001	.020 ± .001	.016 ± .001	3.409 ± 2.666	.020 ± .002	.024 ± .002	.021 ± .002	.446 ± .927
	Stochastic	–	.052	–	.168	–	.051	–	.374	–	.047	–	1.048	–	.053	–	1.122
Fashion	Ours	.165 ± .002	.169 ± .001	.157 ± .001	4.811	.148 ± .003	.155 ± .002	.143 ± .002	1.856	.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	.163 ± .003	.147 ± .003	21.099 ± 6.507	.155 ± .007	.160 ± .007	.151 ± .007	3.929 ± 2.841	.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	.151 ± .006	.155 ± .006	.146 ± .006	5.120 ± 3.150	.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	.153 ± .006	.157 ± .006	.148 ± .007	4.439 ± 2.924	.162 ± .006	.165 ± .005	.157 ± .005	.417 ± .928
	Stochastic	–	.226	–	2.405	–	.210	–	5.428	–	.207	–	7.727	–	.223	–	.816

Table 3 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
CIFAR-10	Ours	.390 ± .000	.407 ± .000	.391 ± .000	.040	.404 ± .000	.414 ± .000	.398 ± .000	.070	.396 ± .001	.411 ± .000	.395 ± .000	.155	.416 ± .002	.432 ± .001	.415 ± .001	.970
	Blanchard	.390 ± .000	.473 ± .004	.391 ± .000	271.616 ± 23.555	.404 ± .000	.445 ± .003	.398 ± .000	84.868 ± 13.050	.396 ± .001	.422 ± .003	.395 ± .000	23.962 ± 7.208	.416 ± .002	.432 ± .002	.416 ± .001	4.496 ± 3.018
	Catoni	.390 ± .000	.473 ± .006	.391 ± .000	27.502 ± 23.371	.404 ± .000	.434 ± .003	.398 ± .000	84.848 ± 12.992	.396 ± .001	.415 ± .002	.395 ± .000	24.505 ± 6.942	.416 ± .002	.431 ± .001	.415 ± .001	4.859 ± 3.176
	Rivas-plata	.390 ± .000	.450 ± .002	.391 ± .000	271.700 ± 23.586	.403 ± .000	.433 ± .002	.398 ± .000	85.027 ± 13.047	.396 ± .001	.416 ± .002	.395 ± .000	23.955 ± 7.093	.416 ± .002	.431 ± .002	.416 ± .001	4.610 ± 3.084
	Stochastic	–	.477	–	.020	–	.485	–	.035	–	.482	–	.077	–	.503	–	.485
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.016 ± .000	.025 ± .000	.019 ± .000	14.490	.015 ± .000	.024 ± .000	.019 ± .000	8.583	.014 ± .000	.021 ± .001	.016 ± .000	13.055	.016 ± .001	.023 ± .001	.017 ± .001	25.556
	Blanchard	.016 ± .000	.430 ± .004	.018 ± .000	11405.062 ± 153.554	.014 ± .000	.200 ± .003	.018 ± .000	3799.912 ± 89.585	.013 ± .000	.086 ± .002	.014 ± .000	1187.859 ± 48.700	.015 ± .001	.049 ± .002	.016 ± .001	38.983 ± 27.857
	Catoni	.016 ± .000	.355 ± .002	.019 ± .000	11954.106 ± 15.709	.015 ± .000	.149 ± .003	.019 ± .000	3828.342 ± 83.937	.014 ± .001	.064 ± .002	.016 ± .001	1218.708 ± 48.514	.017 ± .001	.041 ± .002	.018 ± .001	389.726 ± 29.076

Table 3 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Rivas-plata	.015 ± .000	.272 ± .002	.018 ± .000	1173.953 ± 149.364	.013 ± .000	.122 ± .002	.017 ± .000	3691.345 ± 82.512	.012 ± .000	.056 ± .001	.013 ± .000	1206.615 ± 5.381	.015 ± .001	.037 ± .001	.015 ± .001	391.881 ± 28.344
	Sto-chas-tic	–	.053	–	7.245	–	.052	–	4.292	–	.048	–	6.528	–	.051	–	12.778
Fashion	Ours	.165 ± .002	.172 ± .001	.157 ± .001	23.705	.141 ± .002	.156 ± .002	.137 ± .002	52.736	.131 ± .003	.147 ± .003	.126 ± .003	7.515	.156 ± .004	.165 ± .004	.151 ± .003	16.954
	Blan-chard	.136 ± .001	.598 ± .003	.130 ± .001	11334.327 ± 145.083	.125 ± .001	.379 ± .003	.121 ± .001	3998.068 ± 88.992	.124 ± .001	.247 ± .003	.117 ± .001	126.184 ± 48.814	.152 ± .003	.216 ± .004	.147 ± .003	364.531 ± 28.029
	Catoni	.162 ± .001	.525 ± .004	.154 ± .001	11965.668 ± 152.681	.141 ± .002	.309 ± .003	.137 ± .002	384.802 ± 84.123	.132 ± .003	.224 ± .004	.127 ± .002	1239.918 ± 49.594	.155 ± .004	.232 ± .005	.150 ± .004	394.607 ±28.146
	Rivas-plata	.131 ± .001	.455 ± .002	.127 ± .001	1193.209 ± 155.390	.123 ± .001	.290 ± .002	.119 ± .001	4005.169 ± 89.793	.123 ± .001	.204 ± .002	.116 ± .001	1294.726 ± 49.874	.152 ± .004	.195 ± .004	.146 ± .003	378.905 ± 27.422
	Sto-chas-tic	–	.228	–	11.853	–	.209	–	26.368	–	.198	–	35.258	–	.221	–	8.477
CIFAR-10	Ours	.390 ± .000	.411 ± .000	.391 ± .000	13.286	.404 ± .000	.415 ± .000	.398 ± .000	3.305	.396 ± .001	.412 ± .000	.395 ± .000	3.136	.415 ± .001	.433 ± .001	.415 ± .001	6.064
	Blan-chard	.389 ± .000	.990 ± .000	.391 ± .000	75424.764 ± 397.521	.403 ± .000	.820 ± .002	.397 ± .000	8815.324 ± 126.764	.395 ± .001	.651 ± .003	.394 ± .000	2738.066 ± 75.053	.408 ± .001	.557 ± .003	.405 ± .001	918.500 ± 42.347

Table 3 (continued)

	$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
Catoni	.390 ± .000	.990 ± .000	.391 ± .000	26434.787 ± 228.500	.403 ± .000	.726 ± .003	.397 ± .000	8651.380 ± 126.473	.394 ± .001	.620 ± .002	.393 ± .000	4178.302 ± 9.315	.401 ± .001	.556 ± .001	.396 ± .001	1462.235 ± 55.526
Rivas-plata	.389 ± .000	.902 ± .001	.391 ± .000	31497.669 ± 249.683	.403 ± .000	.715 ± .002	.397 ± .000	8707.893 ± 133.239	.394 ± .001	.578 ± .003	.393 ± .000	2741.257 ± 74.942	.405 ± .001	.512 ± .002	.402 ± .001	967.818 ± 43.629
Sto-chas-tic	–	.480	–	6.643	–	.486	–	1.653	–	.483	–	1.568	–	.503	–	3.032

We report the test risk ($R_{\mathcal{T}}(h)$), the bound alue (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divrgence (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 ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.2

Table 4 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.012 ± .000	.017 ± .000	.013 ± .000	.181	.009 ± .000	.015 ± .000	.011 ± .000	.155	.012 ± .000	.020 ± .000	.016 ± .000	1.655	.013 ± .001	.019 ± .001	.015 ± .001	.615
	Blanchard	.012 ± .000	.027 ± .001	.013 ± .000	93.915 ± 14.109	.012 ± .000	.021 ± .001	.014 ± .000	19.292 ± 6.037	.012 ± .000	.020 ± .001	.016 ± .000	3.023 ± 2.430	.014 ± .001	.018 ± .001	.015 ± .001	.368 ± .831
	Catoni	.012 ± .000	.025 ± .001	.013 ± .000	113.574 ± 15.436	.012 ± .000	.023 ± .002	.014 ± .000	22.347 ± 6.877	.012 ± .000	.020 ± .001	.016 ± .000	2.918 ± 2.341	.013 ± .001	.018 ± .001	.015 ± .001	.336 ± .807
	Rivasplata	.012 ± .000	.023 ± .001	.013 ± .000	96.392 ± 14.300	.012 ± .000	.020 ± .001	.014 ± .000	19.905 ± 6.254	.012 ± .000	.020 ± .001	.016 ± .000	2.931 ± 2.446	.013 ± .001	.018 ± .001	.015 ± .001	.355 ± .813
	Stochastic	–	.042	–	.091	–	.039	–	.077	–	.047	–	.827	–	.045	–	.308
Fashion	Ours	.126 ± .000	.134 ± .000	.124 ± .000	.328	.126 ± .001	.130 ± .001	.119 ± .001	1.692	.122 ± .002	.126 ± .002	.115 ± .002	4.617	.139 ± .005	.145 ± .005	.133 ± .005	2.425
	Blanchard	.126 ± .000	.157 ± .003	.124 ± .000	88.034 ± 13.485	.126 ± .001	.136 ± .002	.120 ± .001	18.852 ± 6.115	.124 ± .002	.127 ± .002	.118 ± .002	3.014 ± 2.395	.142 ± .006	.144 ± .006	.137 ± .006	.370 ± .819
	Catoni	.126 ± .000	.159 ± .004	.124 ± .000	114.259 ± 15.300	.126 ± .001	.133 ± .002	.120 ± .001	22.607 ± 6.871	.124 ± .002	.126 ± .002	.118 ± .002	3.100 ± 2.513	.141 ± .006	.144 ± .006	.136 ± .006	.390 ± .898
	Rivasplata	.126 ± .000	.148 ± .002	.124 ± .000	93.107 ± 13.630	.126 ± .001	.133 ± .002	.120 ± .001	19.724 ± 6.320	.124 ± .002	.126 ± .002	.118 ± .002	2.980 ± 2.451	.142 ± .006	.144 ± .006	.136 ± .006	.371 ± .869
	Stochastic	–	.187	–	.164	–	.182	–	.846	–	.178	–	2.309	–	.199	–	1.212
CIFAR-10	Ours	.369 ± .000	.375 ± .000	.358 ± .000	.028	.351 ± .000	.368 ± .000	.352 ± .000	.041	.359 ± .001	.377 ± .000	.360 ± .000	.183	.419 ± .001	.433 ± .001	.416 ± .001	.759
	Blanchard	.369 ± .000	.446 ± .004	.358 ± .000	269.789 ± 22.724	.351 ± .000	.401 ± .004	.352 ± .000	84.113 ± 12.530	.359 ± .001	.388 ± .003	.360 ± .000	22.878 ± 6.728	.419 ± .001	.432 ± .003	.416 ± .001	4.089 ± 2.818
	Catoni	.369 ± .000	.450 ± .007	.358 ± .000	269.843 ± 24.225	.351 ± .000	.390 ± .004	.352 ± .000	84.500 ± 12.608	.359 ± .001	.381 ± .002	.360 ± .000	23.567 ± 7.181	.419 ± .001	.432 ± .001	.416 ± .001	4.285 ± 2.942

Table 4 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Rivas-plata	.369 ± .000	.421 ± .003	.358 ± .000	27.224 ± 24.187	.351 ± .000	.388 ± .002	.352 ± .000	84.250 ± 13.274	.359 ± .001	.382 ± .002	.360 ± .000	23.053 ± 6.724	.419 ± .001	.431 ± .002	.416 ± .001	4.141 ± 2.985
	Stochas-tic	–	.445	–	.014	–	.438	–	.020	–	.447	–	.092	–	.504	–	.380
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.012 ± .000	.019 ± .000	.013 ± .000	24.837	.012 ± .000	.020 ± .000	.014 ± .000	12.358	.012 ± .000	.021 ± .000	.015 ± .000	13.908	.013 ± .001	.019 ± .001	.014 ± .001	16.179
	Blan-chard	.012 ± .000	.467 ± .004	.013 ± .000	11819.223 ± 154.992	.011 ± .000	.211 ± .003	.014 ± .000	3808.981 ± 86.014	.010 ± .000	.094 ± .003	.014 ± .000	121.397 ± 51.944	.012 ± .001	.046 ± .002	.013 ± .001	372.832 ± 26.602
	Catoni	.012 ± .000	.339 ± .002	.013 ± .000	1196.394 ± 15.704	.012 ± .000	.159 ± .003	.014 ± .000	3838.459 ± 88.155	.012 ± .000	.070 ± .002	.016 ± .000	1218.505 ± 51.783	.013 ± .001	.037 ± .001	.014 ± .001	386.824 ± 28.233
	Rivas-plata	.012 ± .000	.289 ± .003	.013 ± .000	1191.037 ± 152.759	.011 ± .000	.128 ± .002	.014 ± .000	3768.785 ± 9.947	.010 ± .000	.061 ± .001	.013 ± .000	1231.638 ± 49.362	.011 ± .001	.033 ± .001	.012 ± .000	382.225 ± 28.481
	Stochas-tic	–	.044	–	12.418	–	.046	–	6.179	–	.047	–	6.954	–	.045	–	8.089
Fashion	Ours	.126 ± .000	.137 ± .000	.124 ± .000	12.401	.125 ± .001	.132 ± .001	.119 ± .001	14.631	.120 ± .002	.128 ± .002	.113 ± .001	26.499	.133 ± .003	.143 ± .003	.127 ± .003	23.702
	Blan-chard	.123 ± .000	.602 ± .003	.121 ± .000	10558.872 ± 139.107	.119 ± .001	.383 ± .004	.112 ± .001	3893.091 ± 86.176	.113 ± .001	.239 ± .003	.106 ± .001	1204.211 ± 5.815	.132 ± .003	.195 ± .004	.125 ± .003	362.146 ± 27.801

Table 4 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Catoni	.126 ± .000	.531 ± .004	.124 ± .000	11966.223 ± 148.195	.125 ± .001	.299 ± .003	.118 ± .001	3829.806 ± 85.864	.119 ± .002	.209 ± .002	.113 ± .001	1225.310 ± 48.090	.134 ± .004	.202 ± .005	.127 ± .003	395.243 ± 29.182
	Rivas-plata	.123 ± .000	.458 ± .003	.120 ± .000	11209.156 ± 143.319	.118 ± .001	.287 ± .002	.111 ± .001	3815.804 ± 85.091	.112 ± .001	.196 ± .002	.105 ± .001	126.956 ± 49.255	.130 ± .003	.173 ± .004	.124 ± .003	376.904 ± 27.549
	Stochas-tic	–	.189	–	6.200	–	.184	–	7.316	–	.179	–	13.250	–	.195	–	11.851
CIFAR-10	Ours	.369 ± .000	.379 ± .000	.358 ± .000	11.657	.351 ± .000	.369 ± .000	.352 ± .000	2.267	.359 ± .001	.378 ± .000	.360 ± .000	2.616	.418 ± .001	.434 ± .001	.415 ± .001	5.675
	Blan-chard	.369 ± .000	.990 ± .000	.358 ± .000	40152.974 ± 291.721	.351 ± .000	.809 ± .003	.351 ± .000	8753.816 ± 136.801	.358 ± .001	.635 ± .004	.359 ± .000	2728.436 ± 73.835	.412 ± .001	.568 ± .004	.407 ± .001	91.026 ± 44.096
	Catoni	.369 ± .000	.986 ± .000	.358 ± .000	24477.984 ±223.367	.351 ± .000	.708 ± .003	.351 ± .000	8463.452 ± 135.001	.357 ± .001	.578 ± .002	.357 ± .000	3401.221 ± 84.878	.405 ± .001	.561 ± .002	.399 ± .001	1354.100 ± 51.315
	Rivas-plata	.369 ± .000	.868 ± .001	.358 ± .000	24424.968 ± 223.601	.351 ± .000	.694 ± .002	.351 ± .000	8665.339 ± 136.361	.358 ± .001	.555 ± .003	.358 ± .000	274.651 ± 74.784	.409 ± .001	.521 ± .003	.403 ± .001	955.211 ± 44.609
	Stochas-tic	–	.448	–	5.829	–	.439	–	1.134	–	.448	–	1.308	–	.504	–	2.838

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.3

Table 5 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.010 ± .000	.017 ± .000	.013 ± .000	.194	.012 ± .000	.018 ± .000	.014 ± .000	.138	.009 ± .000	.015 ± .000	.011 ± .000	.235	.014 ± .001	.020 ± .001	.015 ± .001	1.111
	Blanchard	.010 ± .000	.028 ± .001	.013 ± .000	88.323 ± 13.740	.012 ± .000	.021 ± .001	.014 ± .000	16.792 ± 5.702	.009 ± .000	.014 ± .001	.011 ± .000	2.449 ± 2.313	.014 ± .001	.019 ± .001	.016 ± .001	.244 ± .765
	Catoni	.010 ± .000	.026 ± .001	.013 ± .000	109.202 ± 15.634	.012 ± .000	.023 ± .002	.014 ± .000	19.918 ± 6.526	.009 ± .000	.015 ± .001	.011 ± .000	2.486 ± 2.362	.014 ± .001	.019 ± .001	.016 ± .001	.298 ± .762
	Rivasplata	.010 ± .000	.024 ± .001	.013 ± .000	91.872 ± 14.470	.012 ± .000	.019 ± .001	.014 ± .000	17.002 ± 5.882	.009 ± .000	.014 ± .000	.011 ± .000	2.529 ± 2.251	.014 ± .001	.019 ± .001	.016 ± .001	.308 ± .778
	Stochastic	–	.043	–	.097	–	.044	–	.069	–	.039	–	.117	–	.047	–	.555
Fashion	Ours	.118 ± .001	.123 ± .000	.112 ± .000	.269	.113 ± .001	.118 ± .001	.107 ± .001	.743	.117 ± .002	.121 ± .002	.110 ± .002	2.600	.131 ± .004	.138 ± .004	.126 ± .004	1.229
	Blanchard	.118 ± .001	.145 ± .003	.112 ± .000	82.403 ± 13.230	.113 ± .001	.123 ± .002	.107 ± .001	16.836 ± 5.583	.119 ± .002	.121 ± .003	.112 ± .003	2.641 ± 2.369	.133 ± .004	.136 ± .004	.128 ± .004	.297 ± .731
	Catoni	.118 ± .001	.151 ± .004	.112 ± .000	109.988 ± 15.347	.113 ± .001	.120 ± .002	.107 ± .001	19.889 ± 6.689	.118 ± .002	.120 ± .003	.112 ± .003	2.615 ± 2.234	.132 ± .004	.136 ± .004	.128 ± .004	.300 ± .811
	Rivasplata	.118 ± .001	.137 ± .002	.112 ± .000	87.804 ± 13.640	.113 ± .001	.120 ± .002	.107 ± .001	17.491 ± 6.144	.118 ± .002	.121 ± .003	.112 ± .003	2.549 ± 2.175	.133 ± .005	.137 ± .004	.128 ± .004	.322 ± .794
	Stochastic	–	.174	–	.135	–	.168	–	.372	–	.172	–	1.300	–	.191	–	.615
CIFAR-10	Ours	.334 ± .000	.346 ± .000	.328 ± .000	.025	.322 ± .000	.331 ± .000	.313 ± .000	.050	.323 ± .001	.334 ± .000	.316 ± .000	.160	.333 ± .001	.341 ± .001	.323 ± .001	.461

Table 5 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Blanchard	.334 ± .000	.421 ± .004	.328 ± .000	269.875 ± 23.982	.322 ± .000	.364 ± .004	.313 ± .000	83.082 ± 13.029	.323 ± .001	.345 ± .004	.316 ± .000	21.614 ± 6.670	.333 ± .001	.340 ± .002	.323 ± .001	3.630 ± 2.750
	Catoni	.334 ± .000	.433 ± .008	.328 ± .000	27.270 ± 24.201	.322 ± .000	.355 ± .005	.313 ± .000	84.148 ± 13.578	.323 ± .001	.338 ± .002	.316 ± .000	22.547 ± 6.801	.333 ± .001	.338 ± .001	.323 ± .001	3.831 ± 2.801
	Rivas-plata	.334 ± .000	.394 ± .003	.328 ± .000	27.133 ± 24.109	.322 ± .000	.351 ± .003	.313 ± .000	83.438 ± 13.033	.323 ± .001	.339 ± .002	.316 ± .000	21.688 ± 6.718	.333 ± .001	.339 ± .002	.323 ± .001	3.667 ± 2.757
	Stochastic	–	.414	–	.013	–	.399	–	.025	–	.403	–	.080	–	.409	–	.230
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.010 ± .000	.019 ± .000	.013 ± .000	23.992	.012 ± .000	.019 ± .000	.014 ± .000	7.767	.009 ± .000	.015 ± .000	.011 ± .000	3.165	.012 ± .001	.019 ± .001	.013 ± .001	18.413
	Blanchard	.010 ± .000	.500 ± .004	.013 ± .000	1123.328 ± 151.115	.012 ± .000	.236 ± .004	.014 ± .000	381.840 ± 94.218	.009 ± .000	.096 ± .003	.011 ± .000	1184.214 ± 47.208	.011 ± .001	.048 ± .002	.012 ± .001	363.194 ± 26.547
	Catoni	.010 ± .000	.369 ± .003	.013 ± .000	1191.598 ± 154.180	.012 ± .000	.180 ± .003	.014 ± .000	3826.581 ± 85.362	.009 ± .000	.070 ± .002	.011 ± .000	1217.723 ± 49.984	.012 ± .001	.039 ± .002	.014 ± .001	384.476 ± 29.126

Table 5 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Rivas-plata	.010 ± .000	.316 ± .003	.013 ± .000	11557.703 ± 151.498	.012 ± .000	.142 ± .002	.014 ± .000	3751.391 ± 84.542	.009 ± .000	.061 ± .002	.011 ± .000	1172.156 ± 46.933	.010 ± .001	.035 ± .001	.012 ± .001	373.003 ± 27.844
	Stochastic	–	.045	–	11.996	–	.045	–	3.884	–	.040	–	1.583	–	.045	–	9.207
Fashion	Ours	.118 ± .000	.127 ± .000	.112 ± .000	17.987	.113 ± .001	.119 ± .001	.107 ± .001	6.361	.114 ± .002	.123 ± .002	.107 ± .002	22.582	.125 ± .003	.137 ± .003	.122 ± .003	16.872
	Blanchard	.115 ± .001	.659 ± .004	.110 ± .000	11835.780 ± 161.816	.110 ± .001	.395 ± .004	.104 ± .000	3828.562 ± 94.279	.108 ± .001	.244 ± .004	.102 ± .001	1185.882 ± 5.575	.123 ± .003	.192 ± .004	.119 ± .002	346.265 ± 27.827
	Catoni	.118 ± .001	.566 ± .004	.112 ± .000	11921.114 ± 153.739	.113 ± .001	.304 ± .003	.107 ± .000	3822.647 ± 85.225	.114 ± .002	.208 ± .003	.107 ± .002	1217.879 ± 52.353	.125 ± .003	.196 ± .004	.121 ± .002	388.473 ± 29.475
	Rivas-plata	.114 ± .000	.476 ± .003	.109 ± .000	11206.239 ± 149.549	.110 ± .001	.292 ± .003	.103 ± .000	3745.930 ± 84.367	.106 ± .001	.197 ± .003	.101 ± .001	1229.005 ± 51.052	.122 ± .003	.170 ± .004	.118 ± .003	361.652 ± 28.452
	Stochastic	–	.177	–	8.994	–	.169	–	3.180	–	.172	–	11.291	–	.189	–	8.436
CIFAR-10	Ours	.334 ± .000	.350 ± .000	.328 ± .000	12.067	.322 ± .000	.332 ± .000	.313 ± .000	4.172	.323 ± .001	.336 ± .000	.316 ± .000	3.382	.332 ± .001	.343 ± .001	.322 ± .001	6.855
	Blanchard	.334 ± .000	.977 ± .001	.328 ± .000	28565.558 ± 245.568	.322 ± .000	.803 ± .003	.313 ± .000	8479.553 ± 126.804	.321 ± .001	.614 ± .004	.315 ± .000	2727.786 ± 7.572	.327 ± .001	.487 ± .004	.317 ± .001	887.578 ± 42.449

Table 5 (continued)

	$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
Catoni	.334 ± .000	.983 ± .000	.328 ± .000	24136.528 ± 211.963	.322 ± .000	.694 ± .004	.313 ± .000	7928.671 ± 122.159	.320 ± .001	.515 ± .002	.314 ± .000	237.703 ± 65.952	.323 ± .001	.468 ± .002	.312 ± .001	1157.073 ± 47.283
Rivas-plata	.334 ± .000	.922 ± .001	.328 ± .000	33282.032 ± 246.654	.322 ± .000	.680 ± .003	.312 ± .000	8493.458 ± 128.894	.320 ± .001	.527 ± .003	.314 ± .000	2739.108 ± 7.556	.325 ± .001	.436 ± .003	.314 ± .001	91.066 ± 43.389
Sto-chas-tic	–	.417	–	6.033	–	.400	–	2.086	–	.403	–	1.691	–	.410	–	3.427

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.4

Table 6 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	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	15.518 ± 5.438	.009 ± .000	.014 ± .001	.010 ± .000	2.140 ± 2.072	.015 ± .001	.018 ± .001	.015 ± .001	.284 ± .649
	Catoni	.008 ± .000	.022 ± .001	.010 ± .000	104.063 ± 14.662	.006 ± .000	.015 ± .000	.009 ± .000	17.676 ± 5.963	.008 ± .000	.014 ± .001	.010 ± .000	2.152 ± 2.085	.015 ± .001	.018 ± .001	.015 ± .001	.252 ± .680
	Rivasplata	.008 ± .000	.021 ± .001	.010 ± .000	84.581 ± 13.035	.006 ± .000	.013 ± .001	.009 ± .000	15.545 ± 5.594	.008 ± .000	.014 ± .000	.010 ± .000	2.185 ± 1.992	.015 ± .001	.018 ± .001	.015 ± .001	.276 ± .693
	Stochastic	–	.039	–	.042	–	.035	–	.026	–	.038	–	.090	–	.045	–	.288
Fashion	Ours	.106 ± .000	.113 ± .000	.101 ± .000	.133	.104 ± .001	.110 ± .000	.099 ± .000	.327	.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 ± .715
	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	.125 ± .004	.117 ± .004	.262 ± .699
	Rivasplata	.106 ± .000	.127 ± .002	.101 ± .000	82.150 ± 12.955	.104 ± .001	.112 ± .001	.099 ± .000	15.509 ± 5.629	.109 ± .002	.111 ± .001	.102 ± .001	2.178 ± 2.060	.122 ± .004	.126 ± .004	.117 ± .004	.264 ± .704
	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	.035	.298 ± .001	.310 ± .000	.291 ± .000	.101	.315 ± .001	.329 ± .001	.309 ± .001	.368

Table 6 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Blanchard	.312 ± .000	.405 ± .004	.304 ± .000	268.149 ± 22.835	.281 ± .000	.339 ± .004	.285 ± .000	8.690 ± 12.628	.298 ± .001	.320 ± .004	.291 ± .000	19.648 ± 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	2.711 ± 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
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.008 ± .000	.017 ± .000	.010 ± .000	29.993	.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	11894.556 ± 155.958	.006 ± .000	.256 ± .004	.009 ± .000	3826.515 ± 86.973	.008 ± .000	.108 ± .003	.010 ± .000	1184.777 ± 48.158	.010 ± .001	.052 ± .002	.011 ± .000	36.865 ± 28.054
	Catoni	.008 ± .000	.396 ± .003	.010 ± .000	11986.455 ± 15.722	.006 ± .000	.192 ± .002	.009 ± .000	3824.971 ± 85.072	.008 ± .000	.079 ± .002	.010 ± .000	1213.611 ± 48.751	.013 ± .001	.042 ± .002	.014 ± .001	384.275 ± 28.556
	Rivasplata	.008 ± .000	.362 ± .003	.010 ± .000	11905.971 ± 15.609	.006 ± .000	.148 ± .003	.009 ± .000	377.259 ± 84.127	.008 ± .000	.067 ± .002	.010 ± .000	118.841 ± 5.043	.010 ± .001	.036 ± .001	.011 ± .000	369.675 ± 27.947

Table 6 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Stochastic	–	.041	–	14.996	–	.035	–	1.581	–	.039	–	.709	–	.045	–	6.116
Fashion	Ours	.106 ± .000	.114 ± .000	.101 ± .000	6.310	.103 ± .001	.113 ± .000	.099 ± .000	9.312	.106 ± .002	.115 ± .001	.100 ± .001	14.924	.115 ± .003	.126 ± .003	.110 ± .002	18.364
	Blanchard	.105 ± .000	.674 ± .004	.101 ± .000	10795.464 ± 143.426	.102 ± .000	.412 ± .004	.098 ± .000	3685.940 ± 82.481	.103 ± .001	.253 ± .004	.097 ± .001	1178.401 ± 48.359	.113 ± .002	.186 ± .004	.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	1214.976 ± 48.846	.116 ± .003	.184 ± .004	.111 ± .003	388.197 ± 27.580
	Rivasplata	.105 ± .000	.503 ± .003	.100 ± .000	11139.304 ± 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 ± .003	.107 ± .002	349.146 ± 27.482
	Stochastic	–	.163	–	3.155	–	.161	–	4.656	–	.163	–	7.462	–	.176	–	9.182
CIFAR-10	Ours	.312 ± .000	.328 ± .000	.304 ± .000	12.006	.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	.617 ± .004	.290 ± .000	2723.775 ± 66.832	.309 ± .001	.490 ± .004	.303 ± .001	888.277 ± 41.530
	Catoni	.312 ± .000	.980 ± .000	.304 ± .000	21278.808 ± 207.839	.280 ± .000	.681 ± .004	.284 ± .000	6951.932 ± 118.540	.296 ± .001	.496 ± .003	.290 ± .000	2145.470 ± 6.045	.305 ± .001	.457 ± .002	.299 ± .001	103.494 ± 47.021

Table 6 (continued)

	$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
Rivas-plata	.312 ± .000	.964 ± .001	.304 ± .000	42834.626 ± 284.116	.280 ± .000	.690 ± .003	.284 ± .000	8675.531 ± 136.658	.296 ± .001	.521 ± .003	.290 ± .000	2718.415 ± 66.664	.307 ± .001	.434 ± .003	.301 ± .001	921.068 ± 42.158
Sto-chas-tic	–	.394	–	6.003	–	.371	–	.901	–	.378	–	1.028	–	.397	–	2.391

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.5

Table 7 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.008 ± .000	.014 ± .000	.010 ± .000	.040	.007 ± .000	.014 ± .000	.009 ± .000	.068	.008 ± .000	.013 ± .000	.009 ± .000	.092	.008 ± .000	.014 ± .001	.009 ± .000	.128
	Blanchard	.008 ± .000	.026 ± .002	.010 ± .000	75.043 ± 11.586	.007 ± .000	.016 ± .001	.009 ± .000	13.220 ± 4.956	.008 ± .000	.012 ± .001	.009 ± .000	1.774 ± 1.772	.008 ± .000	.012 ± .001	.009 ± .000	.190 ± .594
	Catoni	.008 ±.000	.022 ± .001	.010 ± .000	96.561 ± 13.980	.007 ± .000	.016 ± .000	.009 ± .000	15.107 ± 5.370	.008 ± .000	.013 ± .001	.009 ± .000	1.835 ± 1.837	.008 ± .000	.013 ± .000	.009 ± .000	.219 ± .619
	Rivasplata	.008 ± .000	.021 ± .001	.010 ± .000	76.898 ± 12.301	.007 ± .000	.014 ± .001	.009 ± .000	13.370 ± 4.931	.008 ± .000	.013 ± .000	.009 ± .000	1.695 ± 1.741	.008 ± .000	.013 ± .001	.009 ± .000	.183 ± .580
	Stochastic	–	.038	–	.020	–	.037	–	.034	–	.037	–	.046	–	.037	–	.064
Fashion	Ours	.109 ± .000	.115 ± .000	.102 ± .000	.128	.114 ± .001	.117 ± .001	.104 ± .001	.436	.101 ± .001	.108 ± .001	.096 ± .001	.452	.110 ± .003	.116 ± .003	.103 ± .003	.438
	Blanchard	.109 ± .000	.139 ± .003	.102 ± .000	7.878 ± 11.599	.114 ± .001	.121 ± .003	.104 ± .001	13.041 ± 5.012	.102 ± .001	.106 ± .002	.096 ± .001	1.840 ± 1.864	.111 ± .003	.113 ± .003	.104 ± .002	.184 ± .600
	Catoni	.109 ± .000	.152 ± .006	.102 ± .000	96.732 ± 13.464	.114 ± .001	.119 ± .002	.104 ± .001	15.103 ± 5.363	.102 ± .001	.105 ± .001	.096 ± .001	1.825 ± 1.886	.111 ± .003	.112 ± .003	.104 ± .003	.224 ± .610
	Rivasplata	.109 ± .000	.129 ± .002	.102 ± .000	75.029 ± 11.918	.114 ± .001	.118 ± .002	.104 ± .001	13.495 ± 5.112	.102 ± .001	.106 ± .001	.096 ± .001	1.798 ± 1.859	.111 ± .003	.114 ± .003	.104 ± .002	.219 ± .610
	Stochastic	–	.164	–	.064	–	.167	–	.218	–	.157	–	.226	–	.165	–	.219
CIFAR-10	Ours	.277 ± .000	.297 ± .000	.276 ± .000	.021	.288 ± .000	.307 ± .000	.286 ± .000	.027	.273 ± .001	.284 ± .000	.263 ± .000	.079	.281 ± .001	.302 ± .001	.281 ± .001	.227
	Blanchard	.277 ± .000	.386 ± .005	.276 ± .000	262.952 ± 24.385	.288 ± .000	.346 ± .005	.286 ± .000	76.609 ± 12.923	.273 ± .001	.293 ± .004	.263 ± .000	17.724 ± 6.241	.281 ± .001	.299 ± .002	.281 ± .001	2.580 ± 2.299

Table 7 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Catoni	.277 ± .000	.398 ± .001	.276 ± .000	268.083 ± 24.567	.288 ± .000	.343 ± .007	.286 ± .000	82.887 ± 13.493	.273 ± .001	.287 ± .003	.263 ± .000	18.978 ± 6.437	.281 ± .001	.297 ± .001	.281 ± .001	2.661 ± 2.317
	Rivas-plata	.277 ± .000	.354 ± .004	.276 ± .000	263.581 ± 24.435	.288 ± .000	.330 ± .003	.286 ± .000	77.488 ± 12.464	.273 ± .001	.288 ± .002	.263 ± .000	17.704 ± 5.927	.281 ± .001	.299 ± .002	.281 ± .001	2.619 ± 2.297
	Stochas-tic	–	.363	–	.010	–	.374	–	.014	–	.349	–	.040	–	.368	–	.113
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.008 ± .000	.016 ± .000	.010 ± .000	9.520	.007 ± .000	.014 ± .000	.009 ± .000	3.594	.008 ± .000	.014 ± .000	.009 ± .000	1.877	.008 ± .000	.014 ± .001	.009 ± .000	6.589
	Blanchard	.008 ± .000	.657 ± .005	.010 ± .000	1209.158 ± 157.539	.007 ± .000	.304 ± .005	.009 ± .000	3795.285 ± 88.141	.008 ± .000	.124 ± .004	.009 ± .000	1183.704 ± 5.113	.007 ± .000	.052 ± .002	.009 ± .000	347.860 ± 25.275
	Catoni	.008 ± .000	.452 ± .004	.010 ± .000	12032.708 ± 157.184	.007 ± .000	.225 ± .003	.009 ± .000	3834.246 ± 89.809	.008 ± .000	.093 ± .003	.009 ± .000	1225.575 ± 51.027	.007 ± .000	.039 ± .002	.008 ± .000	39.374 ± 26.987
	Rivasplata	.008 ± .000	.423 ± .004	.010 ± .000	11943.688 ± 156.365	.007 ± .000	.179 ± .003	.009 ± .000	3787.407 ± 87.968	.008 ± .000	.075 ± .002	.009 ± .000	1173.457 ± 49.846	.007 ± .000	.035 ± .002	.008 ± .000	348.717 ± 26.495
	Stochastic	–	.039	–	4.760	–	.038	–	1.797	–	.037	–	.938	–	.038	–	3.294

Table 7 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
Fashion	Ours	.109 ± .000	.119 ± .000	.102 ± .000	16.776	.114 ± .001	.119 ± .001	.104 ± .001	7.869	.101 ± .001	.111 ± .001	.095 ± .001	14.224	.109 ± .002	.116 ± .002	.101 ± .002	9.187
	Blanchard	.108 ± .000	.743 ± .004	.101 ± .000	11048.501 ± 146.969	.112 ± .001	.468 ± .005	.101 ± .001	3798.865 ± 87.270	.099 ± .001	.268 ± .005	.093 ± .001	1144.740 ± 49.199	.106 ± .002	.183 ± .004	.099 ± .002	328.466 ± 24.435
	Catoni	.109 ± .000	.712 ± .005	.102 ± .000	1191.096 ± 15.212	.114 ± .001	.367 ± .005	.104 ± .001	3831.104 ± 88.371	.101 ± .001	.216 ± .003	.095 ± .001	1221.392 ± 5.970	.108 ± .002	.175 ± .003	.101 ± .002	386.528 ± 26.498
	Rivasplata	.108 ± .000	.557 ± .003	.101 ± .000	11148.085 ± 145.818	.111 ± .001	.340 ± .003	.100 ± .001	3757.976 ± 83.965	.098 ± .001	.209 ± .003	.092 ± .001	1176.081 ± 49.829	.106 ± .002	.156 ± .003	.098 ± .002	34.716 ± 24.874
	Stochastic	–	.168	–	8.388	–	.168	–	3.935	–	.159	–	7.112	–	.165	–	4.594
CIFAR-10	Ours	.277 ± .000	.301 ± .000	.276 ± .000	8.466	.288 ± .000	.308 ± .000	.286 ± .000	2.415	.273 ± .001	.285 ± .000	.263 ± .000	2.256	.280 ± .001	.303 ± .001	.280 ± .001	2.747
	Blanchard	.277 ± .000	.990 ± .000	.276 ± .000	58878.209 ± 356.845	.288 ± .000	.868 ± .003	.286 ± .000	8858.838 ±134.545	.272 ± .001	.625 ± .005	.262 ± .000	2709.659 ± 76.197	.278 ± .001	.480 ± .005	.277 ± .001	86.940 ± 43.864
	Catoni	.277 ± .000	.974 ± .000	.276 ± .000	17581.286 ± 185.476	.288 ± .000	.662 ± .005	.286 ± .000	5118.582 ± 105.636	.272 ± .001	.456 ± .003	.262 ± .000	1548.107 ± 58.565	.277 ± .001	.426 ± .002	.274 ± .001	783.103 ± 41.593
	Rivasplata	.277 ± .000	.990 ± .000	.276 ± .000	82459.214 ± 398.763	.288 ± .000	.733 ± .003	.286 ± .000	8674.850 ± 13.468	.272 ± .001	.518 ± .004	.262 ± .000	2709.173 ± 77.205	.277 ± .001	.418 ± .004	.275 ± .001	874.307 ± 44.089
	Stochastic	–	.366	–	4.233	–	.374	–	1.207	–	.350	–	1.128	–	.369	–	1.374

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.6

Table 8 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	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	.019 ± .000	.013 ± .000	.047	.010 ± .000	.018 ± .000	.012 ± .000	.125	.010 ± .000	.017 ± .000	.012 ± .000	.116	.010 ± .001	.018 ± .001	.012 ± .001	.132
	Blanchard	.011 ± .000	.032 ± .002	.013 ± .000	65.017 ± 11.099	.010 ± .000	.019 ± .001	.012 ± .000	1.819 ± 4.995	.010 ± .000	.016 ± .001	.012 ± .000	1.551 ± 1.635	.010 ± .001	.016 ± .001	.012 ± .001	.115 ± .560
	Catoni	.011 ± .000	.028 ± .001	.013 ± .000	84.529 ± 13.023	.010 ± .000	.021 ± .000	.012 ± .000	11.910 ± 5.053	.010 ± .000	.017 ± .001	.012 ± .000	1.228 ± 1.637	.010 ± .001	.017 ± .001	.012 ± .001	.173 ± .512
	Rivasplata	.011 ± .000	.026 ± .001	.013 ± .000	68.055 ± 11.606	.010 ± .000	.018 ± .001	.012 ± .000	1.637 ± 4.962	.010 ± .000	.016 ± .000	.012 ± .000	1.408 ± 1.639	.010 ± .001	.016 ± .001	.012 ± .001	.160 ± .529
	Stochastic	–	.044	–	.023	–	.043	–	.062	–	.042	–	.058	–	.043	–	.066
Fashion	Ours	.099 ± .000	.112 ± .000	.098 ± .000	.067	.107 ± .001	.115 ± .001	.100 ± .001	.542	.098 ± .002	.107 ± .001	.093 ± .001	.353	.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	.094 ± .001	1.342 ± 1.664	.108 ± .003	.113 ± .003	.103 ± .002	.143 ± .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	1.437 ± 1.594	.108 ± .003	.112 ± .003	.103 ± .002	.153 ± .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	.155 ± .546
	Stochastic	–	.161	–	.034	–	.164	–	.271	–	.155	–	.177	–	.166	–	.156
CIFAR-10	Ours	.277 ± .000	.296 ± .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
	Blanchard	.277 ± .000	.399 ± .006	.272 ± .000	257.371 ± 23.327	.266 ± .000	.322 ± .005	.257 ± .000	7.190 ± 11.685	.253 ± .001	.281 ± .005	.248 ± .000	15.214 ± 5.838	.236 ± .001	.255 ± .002	.235 ± .001	2.223 ± 2.016

Table 8 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	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 ± 2.082
	Rivas-plata	.277 ± .000	.362 ± .004	.272 ± .000	258.993 ± 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
	Sto-chas-tic	–	.362	–	.008	–	.345	–	.011	–	.336	–	.034	–	.322	–	.059
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.011 ± .000	.025 ± .000	.013 ± .000	45.094	.010 ± .000	.019 ± .000	.012 ± .000	7.479	.010 ± .000	.018 ± .000	.012 ± .000	5.269	.010 ± .000	.018 ± .001	.011 ± .001	6.510
	Blanchard	.011 ± .000	.737 ± .004	.013 ± .000	11285.050 ± 147.363	.010 ± .000	.381 ± .006	.012 ± .000	3785.071 ± 85.889	.010 ± .000	.160 ± .005	.011 ± .000	1181.043 ± 46.219	.009 ± .000	.067 ± .003	.011 ± .001	34.267 ± 26.244
	Catoni	.011 ± .000	.547 ± .004	.013 ± .000	11965.668 ± 153.481	.010 ±.000	.283 ± .004	.012 ± .000	3811.642 ± 88.111	.010 ± .000	.120 ± .004	.011 ± .000	1212.373 ± 48.835	.009 ± .000	.050 ± .002	.010 ± .000	383.387 ± 27.059
	Rivas-plata	.011 ± .000	.509 ± .004	.013 ± .000	11555.623 ± 15.287	.010 ± .000	.226 ± .004	.012 ± .000	3695.054 ± 9.289	.009 ± .000	.096 ± .003	.011 ± .000	1171.892 ± 47.812	.009 ± .000	.044 ± .002	.010 ± .000	343.025 ± 25.804
	Stochastic	–	.050	–	22.547	–	.044	–	3.740	–	.043	–	2.634	–	.043	–	3.255
Fashion	Ours	.099 ± .000	.116 ± .000	.098 ± .000	11.922	.107 ± .001	.117 ± .001	.101 ± .001	6.556	.097 ± .001	.109 ± .001	.092 ± .001	9.235	.105 ± .002	.118 ± .002	.100 ± .002	1.362

Table 8 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Blanchard	.098 ± .000	.795 ± .004	.098 ± .000	10179.790 ± 138.889	.101 ± .001	.524 ± .006	.096 ± .001	3752.748 ± 9.952	.095 ± .001	.291 ± .005	.090 ± .001	1091.018 ± 47.577	.104 ± .002	.195 ± .005	.098 ± .002	309.857 ± 24.422
	Catoni	.099 ± .000	.808 ± .002	.098 ± .000	11999.071 ± 158.418	.107 ± .001	.425 ± .006	.100 ± .001	3817.800 ± 91.674	.098 ± .001	.235 ± .004	.093 ± .001	1216.042 ± 5.641	.106 ± .002	.182 ± .004	.101 ± .002	376.493 ± 27.018
	Rivas-plata	.098 ± .000	.619 ± .004	.097 ± .000	10768.160 ± 146.634	.099 ± .001	.369 ± .004	.094 ± .001	3565.270 ± 88.164	.094 ± .001	.224 ± .004	.089 ± .001	1137.876 ± 48.421	.103 ± .002	.164 ± .003	.097 ± .002	318.512 ± 24.741
	Stochastic	−	.164	−	5.961	−	.166	−	3.278	−	.156	−	4.618	−	.166	−	5.181
CIFAR-10	Ours	.277 ± .000	.303 ± .000	.272 ± .000	12.803	.266 ± .000	.282 ± .000	.257 ± .000	2.312	.253 ± .001	.272 ± .000	.248 ± .000	1.641	.236 ± .001	.259 ± .001	.235 ± .001	1.929
	Blanchard	.277 ± .000	.990 ± .000	.272 ± .000	2577.092 ± 236.075	.266 ± .000	.901 ± .003	.257 ± .000	8788.732 ± 134.680	.253 ± .001	.662 ± .005	.247 ± .000	2683.054 ± 73.139	.235 ± .001	.464 ± .006	.233 ± .001	85.586 ± 41.917
	Catoni	.277 ± .000	1.000 ± .000	.272 ± .000	177807.417 ± 546.892	.266 ± .000	.601 ± .005	.257 ± .000	331.757 ± 83.561	.253 ± .001	.416 ± .003	.247 ± .000	85.973 ± 4.961	.234 ± .001	.369 ± .003	.233 ± .001	485.863 ± 31.335
	Rivas-plata	.277 ± .000	.990 ± .000	.272 ± .000	48522.489 ± 309.735	.266 ± .000	.762 ± .003	.257 ± .000	850.968 ± 131.507	.252 ± .001	.542 ± .004	.247 ± .000	2696.074 ± 73.062	.234 ± .001	.393 ± .004	.232 ± .001	858.936 ± 41.972
	Stochastic	−	.366	−	6.401	−	.346	−	1.156	−	.336	−	.821	−	.322	−	.965

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.7

Table 9 Comparison of ours, rivasplata, blanchard andcatoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.011 ± .000	.020 ± .000	.013 ± .000	.064	.008 ± .000	.017 ± .000	.010 ± .000	.050	.011 ± .000	.018 ± .000	.011 ± .000	.112	.010 ± .001	.016 ± .001	.009 ± .001	.073
	Blan-chard	.011 ± .000	.034 ± .003	.013 ± .000	49.248 ± 1.541	.008 ± .000	.018 ± .001	.010 ± .000	8.031 ± 3.654	.011 ± .000	.016 ± .001	.011 ± .000	.810 ± 1.248	.010 ± .001	.014 ± .001	.010 ± .001	.102 ± .448
	Catoni	.011 ± .000	.030 ± .002	.013 ± .000	66.244 ± 11.961	.008 ± .000	.018 ± .001	.010 ± .000	8.685 ± 3.987	.011 ± .000	.019 ± .001	.011 ± .000	1.011 ± 1.283	.010 ± .001	.016 ± .001	.010 ± .001	.131 ± .422
	Rivas-plata	.011 ± .000	.028 ± .002	.013 ± .000	5.344 ± 1.600	.008 ± .000	.017 ± .001	.010 ± .000	7.757 ± 4.187	.011 ± .000	.017 ± .001	.011 ± .000	.861 ± 1.361	.010 ± .001	.014 ± .001	.010 ± .001	.090 ± .460
	Sto-chas-tic	–	.046	–	.032	–	.041	–	.025	–	.043	–	.056	–	.040	–	.037
Fashion	Ours	.103 ± .000	.117 ± .000	.099 ± .000	.068	.098 ± .001	.114 ± .001	.096 ± .001	.178	.104 ± .001	.117 ± .002	.099 ± .002	.587	.107 ± .004	.119 ± .004	.101 ± .003	.328
	Blan-chard	.103 ± .000	.144 ± .004	.099 ± .000	5.069 ± 9.537	.098 ± .001	.116 ± .003	.096 ± .001	8.105 ± 3.874	.104 ± .001	.113 ± .002	.100 ± .002	.990 ± 1.435	.109 ± .004	.114 ± .004	.102 ± .004	.102 ± .444
	Catoni	.103 ± .000	.168 ± .009	.099 ± .000	66.761 ± 1.939	.098 ± .001	.115 ± .004	.096 ± .001	8.698 ± 3.974	.104 ± .001	.112 ± .002	.100 ± .002	.934 ± 1.413	.109 ± .004	.113 ± .004	.102 ± .004	.100 ± .457
	Rivas-plata	.103 ± .000	.132 ± .003	.099 ± .000	52.096 ± 1.745	.098 ± .001	.113 ± .002	.096 ± .001	7.820 ± 4.154	.104 ± .001	.114 ± .002	.100 ± .002	.939 ± 1.417	.108 ± .004	.115 ± .004	.102 ± .004	.100 ± .464
	Sto-chas-tic	–	.165	–	.034	–	.162	–	.089	–	.166	–	.294	–	.168	–	.164
CIFAR-10	Ours	.249 ± .000	.265 ± .000	.237 ± .000	.014	.247 ± .000	.271 ± .000	.243 ± .000	.018	.259 ± .001	.281 ±.001	.252 ± .001	.055	.249 ± .001	.274 ± .001	.245 ± .001	.072
	Blan-chard	.249 ± .000	.384 ± .007	.237 ± .000	24.108 ± 22.114	.247 ± .000	.316 ± .006	.243 ± .000	59.096 ± 1.459	.259 ± .001	.289 ± .006	.252 ± .001	11.804 ± 5.001	.249 ± .001	.269 ± .003	.245 ± .001	1.578 ± 1.705

Table 9 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Catoni	.249 ± .000	.368 ± .003	.237 ± .000	27.284 ± 23.618	.247 ± .000	.337 ± .012	.243 ± .000	73.833 ± 11.692	.259 ± .001	.285 ± .005	.252 ± .001	12.808 ± 5.089	.249 ± .001	.266 ± .002	.245 ± .001	1.635 ± 1.773
	Rivas-plata	.249 ± .000	.341 ± .005	.237 ± .000	244.258 ± 22.339	.247 ± .000	.298 ± .004	.243 ± .000	61.907 ± 11.135	.259 ± .001	.283 ± .003	.252 ± .001	11.818 ± 4.923	.249 ± .001	.269 ± .002	.245 ± .001	1.629 ± 1.763
	Stochastic	–	.328	–	.007	–	.334	–	.009	–	.344	–	.028	–	.337	–	.036
		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.011 ± .000	.030 ± .000	.013 ± .000	53.875	.008 ± .000	.018 ± .000	.010 ± .000	4.369	.011 ± .000	.019 ± .000	.011 ± .000	5.063	.009 ± .001	.016 ± .001	.009 ± .001	4.854
	Blanchard	.011 ± .000	.828 ± .004	.013 ± .000	10014.066 ± 14.769	.008 ± .000	.491 ± .007	.010 ± .000	3707.758 ± 86.461	.011 ± .000	.211 ± .007	.011 ± .000	1151.660 ± 47.238	.009 ± .001	.076 ± .005	.009 ± .001	303.402 ± 25.072
	Catoni	.011 ± .000	.684 ± .004	.013 ± .000	12238.359 ± 158.595	.008 ± .000	.343 ± .005	.010 ± .000	3834.114 ± 88.516	.011 ± .000	.168 ± .006	.011 ± .000	121.777 ± 48.460	.009 ± .001	.060 ± .003	.008 ± .001	356.740 ± 25.649
	Rivas-plata	.011 ± .000	.662 ± .005	.013 ± .000	1207.265 ± 161.842	.008 ± .000	.305 ± .005	.010 ± .000	3785.930 ± 87.976	.011 ± .000	.125 ± .004	.010 ± .000	1141.437 ± 46.910	.009 ± .001	.048 ± .002	.008 ± .000	305.573 ± 23.629
	Stochastic	–	.055	–	26.937	–	.042	–	2.185	–	.044	–	2.532	–	.040	–	2.427

Table 9 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
Fashion	Ours	.102 ± .000	.121 ± .000	.099 ± .000	1.120	.098 ± .001	.115 ± .001	.096 ± .001	3.956	.102 ± .001	.118 ± .002	.098 ± .001	7.830	.103 ± .003	.118 ± .003	.097 ± .002	8.797
	Blanchard	.101 ± .000	.990 ± .000	.098 ± .000	27936.970 ± 235.840	.096 ± .001	.585 ± .007	.094 ± .001	321.105 ± 81.006	.098 ± .001	.348 ± .007	.094 ± .001	1045.641 ± 44.087	.101 ± .002	.208 ± .006	.095 ± .002	273.641 ± 24.046
	Catoni	.103 ± .000	.865 ± .002	.099 ± .000	12143.837 ± 161.857	.098 ± .001	.536 ± .008	.096 ± .001	3802.871 ± 87.750	.103 ± .001	.286 ± .006	.098 ± .001	1202.907 ± 47.928	.105 ± .003	.191 ± .005	.098 ± .003	354.246 ± 25.507
	Rivas-plata	.102 ± .000	.746 ± .004	.098 ± .000	11305.448 ± 149.693	.096 ± .001	.438 ± .005	.093 ± .001	3458.977 ± 83.715	.097 ± .001	.264 ± .004	.093 ± .001	1101.567 ± 44.816	.099 ± .002	.172 ± .004	.094 ± .002	285.588 ± 24.451
	Stochastic	–	.168	–	5.060	–	.163	–	1.978	–	.166	–	3.915	–	.166	–	4.399
CIFAR-10	Ours	.249 ± .000	.274 ± .000	.237 ± .000	14.083	.247 ± .000	.273 ± .000	.243 ± .000	1.770	.259 ± .001	.282 ± .001	.252 ± .001	1.098	.248 ± .001	.275 ± .001	.245 ± .001	1.461
	Blanchard	.249 ± .000	.990 ± .000	.237 ± .000	26575.507 ± 218.278	.247 ± .000	.925 ± .002	.243 ± .000	7135.143 ± 117.030	.259 ± .001	.739 ± .006	.251 ± .001	2581.211 ± 74.799	.247 ± .001	.526 ± .007	.243 ± .001	831.790 ± 4.592
	Catoni	.249 ± .000	1.000 ± .000	.237 ± .000	154168.585 ± 539.590	.247 ± .000	.677 ± .008	.243 ± .000	3148.174 ± 83.069	.259 ± .001	.549 ± .006	.252 ± .001	1735.530 ± 57.888	.248 ± .001	.425 ± .005	.244 ± .001	675.780 ± 38.306
	Rivas-plata	.249 ± .000	.990 ± .000	.237 ± .000	35062.089 ± 246.257	.247 ± .000	.824 ± .003	.243 ± .000	8092.236 ± 125.162	.259 ± .001	.610 ± .005	.251 ± .001	2652.857 ± 75.369	.247 ± .001	.441 ± .005	.242 ± .001	84.056 ± 4.952
	Stochastic	–	.334	–	7.041	–	.335	–	.885	–	.345	–	.549	–	.337	–	.731

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.8

Table 10 Comparison of ours, rivasplata, blanchard and catoni based on the disintegrated bounds, and stochastic based on the randomized bounds learned with two learning rates lr$\in\{10^{-4}, 10^{-6}\}$ and different variances $\sigma^2\in\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.008 ± .000	.018 ± .000	.008 ± .000	.029	.011 ± .000	.020 ± .000	.010 ± .000	.052	.009 ± .000	.018 ± .001	.009 ± .000	.059	.008 ± .000	.019 ± .001	.009 ± .001	.023
	Blanchard	.008 ± .000	.033 ± .004	.009 ± .000	35.446 ± 8.610	.011 ± .000	.020 ±.002	.010 ± .000	4.933 ± 2.958	.009 ± .000	.015 ± .001	.009 ± .000	.490 ± .960	.008 ± .001	.016 ± .001	.009 ± .001	.059 ± .299
	Catoni	.008 ± .000	.026 ± .002	.009 ± .000	41.267 ± 9.234	.011 ± .000	.019 ± .001	.010 ± .000	4.564 ± 3.263	.009 ± .000	.016 ± .001	.009 ± .000	.581 ± .989	.008 ± .001	.017 ± .001	.009 ± .001	.078 ± .320
	Rivasplata	.008 ± .000	.025 ± .002	.009 ± .000	35.856 ± 8.648	.011 ± .000	.019 ± .001	.010 ± .000	4.620 ± 2.983	.009 ± .000	.015 ± .001	.009 ± .000	.448 ± 1.045	.008 ± .000	.016 ± .001	.009 ± .001	.041 ± .330
	Stochastic	–	.041	–	.014	–	.045	–	.026	–	.042	–	.030	–	.043	–	.012
Fashion	Ours	.094 ± .000	.113 ± .000	.089 ± .000	.029	.091 ± .001	.119 ± .001	.095 ± .001	.107	.092 ± .002	.113 ± .001	.089 ± .001	.097	.103 ± .003	.124 ± .003	.099 ± .003	.045
	Blanchard	.094 ± .000	.140 ± .006	.089 ± .000	32.563 ± 8.007	.091 ± .001	.119 ± .004	.095 ± .001	4.567 ± 2.912	.092 ± .002	.106 ± .002	.089 ± .001	.468 ± 1.101	.104 ± .003	.116 ± .003	.099 ± .003	.063 ± .300
	Catoni	.094 ± .000	.146 ± .002	.089 ± .000	4.355 ± 9.121	.091 ± .001	.120 ± .005	.095 ± .001	4.895 ± 3.064	.092 ± .002	.106 ± .002	.089 ± .001	.473 ± 1.052	.103 ± .003	.117 ± .003	.099 ± .003	.079 ± .319
	Rivasplata	.094 ± .000	.127 ± .004	.089 ± .000	33.175 ± 8.710	.091 ± .001	.117 ± .002	.095 ± .001	4.774 ± 3.003	.092 ± .002	.107 ± .002	.089 ± .001	.479 ± .924	.103 ± .003	.118 ± .003	.099 ± .002	.045 ± .330
	Stochastic	–	.159	–	.015	–	.166	–	.053	–	.159	–	.048	–	.172	–	.023
CIFAR-10	Ours	.231 ± .000	.268 ± .000	.228 ± .000	.011	.235 ± .000	.267 ± .000	.227 ± .000	.009	.218 ± .001	.253 ± .001	.214 ± .001	.024	.231 ± .001	.264 ± .002	.224 ± .002	.036
	Blanchard	.231 ± .000	.418 ± .010	.228 ± .000	193.922 ± 19.216	.235 ± .000	.312 ± .009	.227 ± .000	39.705 ± 8.929	.218 ± .000	.256 ± .007	.214 ± .001	6.919 ± 3.722	.231 ± .001	.255 ± .003	.224 ± .002	.878 ± 1.248

Table 10 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-6}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Catoni	.231 ± .000	.388 ± .005	.228 ± .000	255.538 ± 22.306	.235 ± .000	.337 ± .003	.227 ± .000	53.736 ± 1.302	.218 ± .000	.257 ± .007	.214 ± .001	7.060 ± 3.626	.231 ± .001	.255 ± .003	.224 ± .002	.857 ± 1.264
	Rivas-plata	.231 ± .000	.364 ± .007	.228 ± .000	202.026 ± 19.688	.235 ± .000	.293 ± .006	.227 ± .000	42.458 ± 9.250	.218 ± .001	.251 ± .004	.214 ± .001	6.780 ± 3.575	.231 ± .001	.256 ± .002	.224 ± .002	.854 ± 1.275
	Stochas-tic	–	.328	–	.005	–	.327	–	.005	–	.312	–	.012	–	.324	–	.018
	lr=10^{-4}	$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
		$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
MNIST	Ours	.008 ± .000	.018 ± .000	.009 ± .000	2.107	.011 ± .000	.021 ± .000	.010 ± .000	1.329	.008 ± .000	.019 ± .001	.008 ± .000	3.598	.008 ± .001	.020 ± .001	.009 ± .001	4.216
	Blan-chard	.008 ± .000	.982 ± .001	.008 ± .000	11722.999 ± 157.452	.011 ± .000	.706 ± .008	.010 ± .000	3475.807 ± 77.708	.008 ± .000	.331 ± .011	.008 ± .000	1076.767 ± 46.059	.008 ± .000	.108 ± .008	.009 ± .001	242.819 ± 23.775
	Catoni	.008 ± .000	1.000 ± .000	.008 ± .000	60838.120 ± 346.289	.011 ± .000	.515 ± .007	.010 ± .000	3728.586 ± 86.803	.008 ± .000	.243 ± .007	.008 ± .000	1166.491 ± 48.086	.008 ± .000	.087 ± .006	.009 ± .001	277.823 ± 25.431
	Rivas-plata	.008 ± .000	.879 ± .003	.008 ± .000	12257.175 ± 152.738	.010 ± .000	.481 ± .007	.010 ± .000	3602.529 ± 78.717	.008 ± .000	.201 ± .007	.008 ± .000	1126.882 ± 47.430	.008 ± .001	.067 ± .004	.009 ± .001	242.366 ± 22.337
	Stochas-tic	–	.042	–	1.053	–	.045	–	.664	–	.042	–	1.799	–	.043	–	2.108
Fashion	Ours	.094 ± .000	.115 ± .000	.089 ± .000	2.501	.091 ± .001	.121 ± .001	.095 ± .001	2.925	.092 ± .002	.114 ± .001	.088 ± .001	3.069	.102 ± .002	.125 ± .003	.098 ± .002	3.159

Table 10 (continued)

		$\sigma^2 = 10^{-6}$				$\sigma^2 = 10^{-5}$				$\sigma^2 = 10^{-4}$				$\sigma^2 = 10^{-3}$			
	lr=10^{-4}	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div	$R_{\mathcal{T}}(h)$	Bnd	$R_{\mathcal{S}}(h)$	Div
	Blanchard	.094 ± .000	.990 ± .000	.089 ± .000	19455.864 ± 19.460	.089 ± .001	.792 ± .007	.093 ± .001	3402.546 ± 86.590	.090 ± .001	.461 ± .010	.087 ± .001	1002.861 ± 44.393	.102 ± .002	.244 ± .009	.098 ± .002	206.177 ± 2.051
	Catoni	.094 ± .000	1.000 ± .000	.089 ± .000	60888.029 ± 346.501	.091 ± .001	.813 ± .012	.095 ± .001	3756.375 ± 9.419	.092 ± .002	.390 ± .010	.089 ± .001	1161.884 ± 52.073	.103 ± .003	.215 ± .007	.099 ± .002	277.284 ± 25.479
	Rivasplata	.094 ± .000	.990 ± .000	.089 ± .000	27137.315 ± 227.934	.088 ± .001	.597 ± .007	.093 ± .001	3371.321 ± 86.352	.090 ± .001	.331 ± .007	.086 ± .001	1003.481 ± 48.362	.101 ± .002	.195 ± .006	.097 ± .002	207.442 ± 21.896
	Stochastic	–	.160	–	1.250	–	.167	–	1.463	–	.160	–	1.535	–	.172	–	1.579
CIFAR-10	Ours	.231 ± .000	.279 ± .000	.228 ± .000	12.925	.235 ± .000	.268 ± .000	.227 ± .000	1.371	.218 ± .001	.254 ± .001	.214 ± .001	.715	.231 ± .001	.264 ± .002	.224 ± .002	1.019
	Blanchard	.231 ± .000	.990 ± .000	.228 ± .000	26032.808 ± 222.475	.235 ± .000	.986 ± .001	.227 ± .000	6875.633 ± 112.137	.217 ± .000	.831 ± .006	.214 ± .001	2292.053 ± 68.347	.230 ± .001	.606 ± .010	.222 ± .001	76.644 ± 39.246
	Catoni	.231 ± .000	1.000 ± .000	.228 ± .000	17684.651 ± 576.711	.235 ± .000	.980 ± .000	.227 ± .000	8265.727 ± 123.941	.218 ± .000	.834 ± .011	.214 ± .001	2664.069 ± 73.915	.231 ± .001	.517 ± .009	.224 ± .002	85.593 ± 41.022
	Rivasplata	.231 ± .000	.988 ± .001	.228 ± .000	14284.846 ± 169.166	.235 ± .000	.919 ± .002	.227 ± .000	7121.350 ± 114.645	.217 ± .000	.699 ± .006	.213 ± .001	2502.412 ± 68.728	.229 ± .001	.494 ± .007	.221 ± .001	776.237 ± 39.540
	Stochastic	–	.335	–	6.462	–	.328	–	.685	–	.313	–	.358	–	.324	–	.510

We report the test risk ($R_{\mathcal{T}}(h)$), the bound value (Bnd), the empirical risk ($R_{\mathcal{S}}(h)$), and the divergence (Div) associated with each bound (the Rényi divergence for ours, the KL divergence for stochastic, and the disintegrated KL divergence for rivasplata, blanchard and catoni). More precisely, we report the mean ± the standard deviation for 400 neural networks sampled from $\mathcal{Q}_{\mathcal{S}}$ for ours, rivasplata, blanchard, and catoni. We consider, in this table, that the split ratio is 0.9

Table 11 Comparison of the bound values before performing Step **2)** of our Training Method for ours, rivasplata, blanchard and catoni

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
$\sigma^2 = 10^{-6}$	.0	.901 ± .002	.901 ± .002	.908 ± .002	.906 ± .002	.905 ± .002	.906 ± .002
	.1	.035 ± .000	.039 ± .000	.045 ± .000	.043 ± .000	.043 ± .000	.042 ± .000
	.2	.016 ±.000	.019 ± .000	.023 ± .000	.022 ± .000	.022 ± .000	.022 ± .000
	.3	.012 ± .000	.013 ± .000	.017 ± .000	.016 ± .000	.015 ± .000	.015 ± .000
	.4	.010 ± .000	.013 ± .000	.017 ± .000	.016 ± .000	.016 ± .000	.016 ± .000
	.5	.008 ± .000	.010 ± .000	.015 ± .000	.013 ± .000	.013 ± .000	.014 ± .000
	.6	.008 ± .000	.010 ± .000	.014 ± .000	.013 ± .000	.013 ± .000	.014 ± .000
	.7	.011 ± .000	.013 ± .000	.019 ± .000	.017 ± .000	.017 ± .000	.018 ± .000
	.8	.011 ± .000	.013 ± .000	.020 ± .000	.018 ± .000	.018 ± .000	.020 ± .000
	.9	.008 ± .000	.009 ± .000	.018 ± .000	.015 ± .000	.014 ± .000	.015 ± .000
$\sigma^2 = 10^{-5}$	.0	.897 ± .013	.897 ± .012	.904 ± .012	.902 ± .012	.902 ± .012	.903 ± .012
	.1	.024 ± .000	.030 ± .001	.035 ± .001	.034 ± .001	.033 ± .001	.033 ± .001
	.2	.015 ± .000	.019 ± .000	.023 ± .000	.022 ± .000	.021 ± .000	.021 ± .000
	.3	.009 ± .000	.011 ± .000	.015 ± .000	.014 ± .000	.013 ± .000	.013 ± .000
	.4	.012 ± .000	.014 ± .000	.018 ± .000	.017 ± .000	.017 ± .000	.017 ± .000
	.5	.006 ± .000	.009 ± .000	.012 ± .000	.011 ± .000	.011 ± .000	.012 ± .000
	.6	.007 ± .000	.009 ± .000	.014 ± .000	.013 ± .000	.012 ± .000	.013 ± .000
	.7	.010 ± .000	.012 ± .000	.018 ± .000	.016 ± .000	.016 ± .000	.017 ± .000
	.8	.008 ± .000	.010 ± .000	.017 ± .000	.015 ± .000	.014 ± .000	.017 ± .000
	.9	.011 ± .000	.010 ± .000	.020 ± .000	.017 ± .000	.017 ± .000	.018 ± .000
$\sigma^2 = 10^{-4}$	.0	.898 ± .017	.898 ±.017	.905 ± .016	.903 ± .016	.902 ± .016	.903 ± .016
	.1	.035 ± .003	.039 ± .002	.045 ± .002	.044 ± .002	.043 ± .002	.043 ± .002
	.2	.015 ± .001	.016 ± .001	.020 ± .001	.019 ± .001	.019 ± .001	.019 ± .001
	.3	.012 ± .000	.016 ± .000	.020 ± .001	.019 ± .001	.019 ± .001	.019 ± .001
	.4	.009 ± .000	.011 ± .000	.015 ± .000	.014 ± .000	.014 ± .000	.014 ± .000
	.5	.008 ± .000	.010 ± .000	.015 ± .000	.013 ± .000	.013 ± .000	.014 ± .000
	.6	.008 ± .000	.009 ± .000	.013 ± .000	.012 ± .000	.012 ± .000	.016 ± .000
	.7	.010 ± .000	.012 ± .000	.017 ± .000	.016 ± .000	.015 ± .000	.013 ± .000
	.8	.011 ± .000	.011 ± .000	.018 ± .000	.016 ± .000	.016 ± .000	.018 ± .000
	.9	.009 ± .000	.009 ± .000	.018 ± .001	.015 ± .001	.015 ± .001	.016 ± .001
$\sigma^2 = 10^{-3}$	.0	.903 ± .014	.902 ± .014	.909 ± .013	.907 ± .013	.907 ± .013	.907 ± .013
	.1	.041 ± .005	.045 ± .005	.050 ± .005	.049 ± .005	.048 ± .005	.048 ± .005
	.2	.020 ± .002	.022 ± .002	.026 ± .002	.025 ± .002	.025 ± .002	.024 ± .002
	.3	.014 ± .001	.015 ± .001	.019 ± .001	.018 ± .001	.018 ± .001	.018 ± .001
	.4	.015 ± .001	.016 ± .001	.021 ± .001	.020 ± .001	.019 ± .001	.019 ± .001
	.5	.015 ± .001	.015 ± .001	.020 ± .001	.019 ± .001	.018 ± .001	.018 ± .001
	.6	.008 ± .000	.010 ± .000	.014 ± .001	.013 ± .001	.012 ± .000	.013 ± .000
	.7	.010 ± .001	.012 ± .001	.018 ± .001	.016 ± .001	.016 ± .001	.017 ± .001
	.8	.010 ± .001	.010 ± .001	.016 ± .001	.014 ± .001	.014 ± .001	.016 ± .001
	.9	.008 ± .000	.009 ± .001	.019 ± .001	.016 ± .001	.015 ± .001	.017 ± .001

More precisely, for each split and each variance $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, we report the mean ± the standard deviation (for 400 neural networks sampled from $\mathcal{P}$) of the test risk ($R_{\mathcal{T}}(h)$), the empirical risk ($R_{\mathcal{S}}(h)$), and the value of the bounds of Corollaries 6 and 7. We consider in this table that the dataset is MNIST

Table 12 Comparison of the bound values before performing Step **2)** of our Training Method for ours, rivasplata, blanchard and catoni

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
$\sigma^2 = 10^{-6}$	.0	.970 ± .028	.970 ± .027	.972 ± .025	.971 ± .025	.971 ± .026	.972 ± .026
	.1	.166 ± .001	.159 ± .000	.169 ± .000	.167 ± .000	.166 ± .000	.167 ± .000
	.2	.168 ± .002	.160 ± .001	.170 ± .001	.168 ± .001	.167 ± .001	.168 ± .001
	.3	.126 ± .000	.124 ± .000	.134 ± .000	.132 ± .000	.131 ± .000	.131 ± .000
	.4	.118 ± .001	.112 ± .000	.123 ± .000	.120 ± .000	.119 ± .000	.119 ± .000
	.5	.106 ± .000	.101 ± .000	.113 ± .000	.110 ± .000	.109 ± .000	.109 ± .000
	.6	.109 ± .000	.102 ± .000	.115 ± .000	.112 ± .000	.110 ± .000	.110 ± .000
	.7	.099 ± .000	.098 ± .000	.112 ± .000	.109 ± .000	.108 ± .000	.107 ± .000
	.8	.103 ± .000	.099 ± .000	.117 ± .000	.112 ± .000	.111 ± .000	.110 ± .000
	.9	.094 ± .000	.089 ± .000	.113 ± .000	.107 ± .000	.105 ± .000	.106 ± .000
$\sigma^2 = 10^{-5}$	.0	.945 ± .038	.945 ± .037	.949 ± .035	.948 ± .035	.948 ± .036	.948 ± .036
	.1	.158 ± .001	.151 ± .001	.161 ± .001	.159 ± .001	.158 ± .001	.159 ± .001
	.2	.157 ± .003	.151 ± .003	.162 ± .003	.159 ± .003	.158 ± .003	.159 ± .003
	.3	.126 ± .001	.121 ± .001	.131 ± .001	.128 ± .001	.127 ± .001	.128 ± .001
	.4	.114 ± .001	.107 ± .001	.118 ± .001	.115 ± .001	.114 ± .001	.114 ± .001
	.5	.104 ± .001	.099 ± .000	.110 ± .000	.108 ± .000	.107 ± .000	.106 ± .000
	.6	.115 ± .001	.104 ± .001	.117 ± .001	.114 ± .001	.113 ± .001	.112 ± .001
	.7	.107 ± .001	.101 ± .001	.115 ± .001	.111 ± .001	.110 ± .001	.109 ± .001
	.8	.098 ± .001	.096 ± .001	.114 ± .001	.109 ± .001	.108 ± .001	.107 ± .001
	.9	.091 ± .001	.095 ± .001	.119 ± .001	.113 ± .001	.111 ± .001	.112 ± .001
$\sigma^2 = 10^{-4}$	.0	.912 ± .027	.912 ± .027	.918 ± .026	.916 ± .027	.916 ± .027	.916 ± .026
	.1	.164 ± .003	.154 ± .003	.164 ± .003	.162 ± .003	.161 ± .003	.162 ± .004
	.2	.164 ± .009	.160 ± .009	.170 ± .010	.168 ± .010	.167 ± .010	.168 ± .010
	.3	.125 ± .002	.119 ± .002	.129 ± .002	.126 ± .002	.126 ± .002	.126 ± .002
	.4	.119 ± .003	.113 ± .003	.124 ± .003	.121 ± .003	.120 ± .003	.120 ± .003
	.5	.109 ± .002	.102 ± .001	.113 ± .001	.110 ± .001	.109 ± .001	.109 ± .001
	.6	.102 ± .001	.096 ± .001	.109 ± .001	.105 ± .001	.105 ± .001	.104 ± .001
	.7	.099 ± .002	.094 ± .001	.108 ± .001	.104 ± .001	.103 ± .001	.102 ± .001
	.8	.104 ± .001	.100 ± .002	.118 ± .002	.113 ± .002	.112 ± .002	.111 ± .002
	.9	.092 ± .002	.089 ± .001	.113 ± .001	.107 ± .001	.105 ± .001	.106 ± .001
$\sigma^2 = 10^{-3}$	.0	.899 ± .026	.899 ± .027	.906 ± .026	.904 ± .026	.904 ± .026	.905 ± .025
	.1	.178 ± .006	.170 ± .006	.181 ± .006	.178 ± .006	.177 ± .006	.179 ± .006
	.2	.164 ± .006	.159 ± .006	.169 ± .006	.167 ± .006	.166 ± .006	.167 ± .006
	.3	.143 ± .007	.138 ± .007	.148 ± .007	.146 ± .007	.145 ± .007	.145 ± .007
	.4	.133 ± .005	.129 ± .005	.140 ± .005	.137 ± .005	.137 ± .005	.137 ± .005
	.5	.122 ± .004	.117 ± .004	.129 ± .004	.126 ± .004	.125 ± .004	.125 ± .004
	.6	.111 ± .003	.104 ± .003	.117 ± .003	.114 ± .003	.113 ± .003	.112 ± .003
	.7	.109 ± .003	.103 ± .003	.118 ± .003	.114 ± .003	.113 ± .003	.112 ± .003
	.8	.108 ± .004	.102 ± .004	.120 ± .004	.115 ± .004	.114 ± .004	.113 ± .004
	.9	.103 ± .003	.099 ± .002	.124 ± .003	.118 ± .003	.116 ± .003	.116 ± .003

More precisely, for each split and each variance $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, we report the mean ± the standard deviation (for 400 neural networks sampled from $\mathcal{P}$) of the test risk ($R_{\mathcal{T}}(h)$), the empirical risk ($R_{\mathcal{S}}(h)$), and the value of the bounds of Corollaries 6 and 7. We consider in this table that the dataset is Fashion-MNIST

Table 13 Comparison of the bound values before performing Step **2)** of our Training Method for ours, rivasplata, blanchard and catoni

	Split	$R_{\mathcal{T}}(h)$	$R_{\mathcal{S}}(h)$	Cor. 6	Eq. (7)	Eq. (8)	Eq. (9)
$\sigma^2 = 10^{-6}$	.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	.272 ± .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
$\sigma^2 = 10^{-5}$	.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
$\sigma^2 = 10^{-4}$	.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	.361 ± .000	.412 ± .000	.409 ± .000	.408 ± .000	.411 ± .000
	.3	.361 ± .001	.396 ± .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
$\sigma^2 = 10^{-3}$	.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	.260 ± .002	.264 ± .001
	.9	.233 ± .001	.232 ± .002	.272 ± .002	.263 ± .002	.266 ± .001	.260 ± .002

More precisely, for each split and each variance $\sigma^2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, we report the mean ± the standard deviation (for 400 neural networks sampled from $\mathcal{P}$) of the test risk ($R_{\mathcal{T}}(h)$), the empirical risk ($R_{\mathcal{S}}(h)$), and the value of the bounds of Corollaries 6 and 7. We consider in this table that the dataset is CIFAR-10

performances of the prior before applying Step **2)** outlined in Figs. 4 and 5. In particular, we report the test risk $R_{\mathcal{T}}(h)$, the empirical risk $R_S(h)$, 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 Sect. 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.

Acknowledgements 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.

Author contributions Conceptualization: PV, EM, PG, PAH; Formal analysis and investigation: PV; Software: PV; Writing—original draft preparation: PV; Writing—review and editing: PG, PAH; Funding acquisition: EM, PG, PAH; Supervision: EM, PG, PAH.

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.

Data availability Not applicable.

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

Declarations

Conflict of interest 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.

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. In: *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. In: *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 PACBayes. International conference on artificial intelligence and statistics (AISTATS) (Vol. 151, pp. 3709–3731). PMLR.

Blanchard, G., & Fleuret, F. (2007). Occam's hammer.In: *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. In: *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. In: *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*, 124–0131.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In: *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. In: *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.In: *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. In: *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. In: *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. In: *Annual conference on computational learning theory (COLT)* (pp. 230–234). ACM.

Nagarajan, V., & Kolter, Z. (2019). *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). PACBayes 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. In: *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. In: *International conference on learning representations (ICLR)*.

Thiemann, N., Igel, C., Wintenberger, O., & Seldin, Y. (2017). A strongly quasiconvex PAC-Bayesian bound. In: *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., Viallard, 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.