Advances in PAC-Bayesian Certification of Deep Neural Networks: Tighter Closed-Form Inequalities and Optimization of Bounds on Non-Differentiable Losses

Diego García-Pérez[1] **Emilio Parrado-Hernández**[1] **John Shawe-Taylor**[2,3]

[1]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
[2]AI Centre, Dept. of Computer Science, University College London, UK
[3]IRCAI, Institute Jožef Stefan, Slovenia

Abstract

This paper presents three theoretical contributions that improve the usability of risk certificates for neural networks based on PAC-Bayes bounds. First, two bounds on the KL divergence between Bernoulli distributions enable the derivation of the tightest explicit bounds on the true risk of classifiers across different ranges of empirical risk. Then, a novel method to optimize bounds on non-differentiable objectives is introduced, allowing to optimize risk bounds directly on the 0-1 loss instead of resourcing to a surrogate differentiable loss, such as the bounded cross-entropy. These theoretical contributions are illustrated with an empirical evaluation on the MNIST and CIFAR-10 datasets. In fact, this paper presents the first non data-dependent generalization bounds on the 0-1 loss for neural networks fitted on CIFAR-10.

1 INTRODUCTION

A critical open problem in machine learning is certifying models' performance, particularly evaluating and guaranteeing their generalization capabilities. This is vital in regulated domains such as healthcare (e.g., diagnostic tools), finance (e.g., credit risk models), and autonomous systems (e.g., self-driving vehicles). The true risk of a classifier, defined as the probability of incorrectly classifying any test observation, would be an ideal certification. However, such risk turns out to be impossible to compute as in general the true underlying data distribution is unknown. The PAC-Bayesian framework offers a robust alternative to derive such certificates by producing non-trivial upper bounds on the true risk without requiring costly cross-validations or splitting data into training and testing subsets [Alquier, 2024].

The PAC-Bayes framework operates by defining two distributions over the hypothesis space: the prior, which must be independent from the data used to produce the certificate, and the posterior, which may depend on said data. The key to estimating the generalization capability lies in combining the empirical risk of the posterior with the quantification of the divergence between these two distributions. While this paper studies the generalization of the posterior, there are also "de-randomization" techniques that study the generalization of a sample from the posterior [Catoni, 2007, Blanchard and Fleuret, 2007, Viallard et al., 2024].

In bayesian learning, it is common to optimize objectives that can be decomposed into a performance term and a "prior" term. To obtain better performance, the weight of the prior term is frequently "attenuated", by introducing a coefficient Blundell et al. [2015] in the training target (see Guedj [2019] for an in-depth analysis). This paper argues that an extensive grid-search through different coefficient values can be replaced by a properly designed objective function.

PAC-Bayesian analyses have traditionally been effective with simpler models such as kernel methods [Parrado-Hernández et al., 2012], where the prior and posterior distributions can be characterized with relatively few parameters. Extending these techniques to certify deep neural networks (NNs) presents significant challenges due to the complexity and size of modern architectures [Dziugaite and Roy, 2017]. The state of the art PAC-Bayes bounds on the true risk of NNs are significantly weaker than those estimations obtained through traditional train/test split methods.

An approach that is not discussed in this paper, even though it produces models with strong performance and tight bounds, is the use of data-dependent priors [Parrado-Hernández et al., 2012, Dziugaite et al., 2021, Lotfi et al., 2022, Wu et al., 2024]. The reason for this is that when addressing more complex tasks such as classification on CIFAR-10, this approach becomes equivalent to a train-test split pipeline with the added complexity of the PAC-Bayes framework. This can be seen in the work of [Pérez-Ortiz et al., 2021], where such approach yields posteriors with

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1650–1662.

a negligible divergence from their priors and no increased performance.

Like most of the recent work in PAC-Bayes bounds in practice, this paper builds on Maurer's bound [Maurer, 2004]. Maurer's bound assumes a bounded loss function, and consequently most results in this line of research are constrained to bounded losses. Some exceptions are [Haddouche et al., 2021, Zhang et al., 2024].

In this context this paper aims at augmenting the usability of PAC-Bayes certificates for NNs with the following contributions:

- New relaxations of Maurer's bound [Maurer, 2004] that, to our knowledge, yield the tightest closed-form PAC-Bayesian risk bounds.
- A theoretical analysis of the KL-attenuation trick Blundell et al. [2015] that motivates a new method to optimize risk bounds on non-differentiable loss functions.
- A new training objective for SGD to optimize Maurer's bound when optimizing the bound on a 01 loss.

2 OBTAINING PAC-BAYESIAN RISK CERTIFICATES

Consider a set of n samples $S = \{(X_i, Y_i)\}_{i=1}^n$, where each $S_i = (X_i, Y_i)$ is an i.i.d. sample from an unknown distribution $Z := X \times Y$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \{0, 1\}$. The goal of our learning algorithm is to find a mapping $h : X \to Y$ from a hypotheses space $\mathcal{H}$ such that $h(X_t)$ is a good approximation to Y_t for any $(X_t, Y_t) \in Z$. Let us denote by Q_0 a data-independent prior distribution over $\mathcal{H}$, and by Q the posterior data-dependent distribution over $\mathcal{H}$ from which the learning algorithm selects the final h. The precision of each prediction made by h is measured with a bounded loss function $l : \mathcal{H} \times Z \to [0, 1]$, and the performance of h with the risk $L(h)$, defined as its expected loss over Z:

$$L(h) := \mathbb{E}_{z\sim Z}[l(h, z)]$$

The definition of risk can be extended to distributions over hypotheses: $L(Q) := \mathbb{E}_{h\sim Q}[L(h)]$. Analogously, the empirical risk of h w.r.t. S is the average of the loss over the elements of S: $\hat{L}_S(h) = \frac{1}{n}\sum_{i=1}^n l(h, S_i)$. Likewise, $\hat{L}_S(Q) := \mathbb{E}_{h\sim Q}[\hat{L}_S(h)]$.

The Kullback-Leibler divergence between two distributions is represented by $\mathrm{KL}(\cdot||\cdot)$. If these distributions are Bernoulli with means p and q, respectively, we will use:

$$\mathrm{kl}(p||q) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \quad p, q \in (0, 1)$$

The classical "PAC-Bayes-kl" bound with confidence $1 - \delta$ was introduced by Langford and Seeger [2001], and slightly improved by Maurer [2004]:

Theorem 2.1 (PAC-Bayes-kl)**.** *Let $S = \{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} Z$. For any distribution Q_0 independent from S, any distribution Q and a bounded loss function $0 \leq l \leq 1$, it holds with probability $1 - \delta$ that:*

$$\mathrm{kl}(\hat{L}_S(Q)||L(Q)) \leq \frac{\mathrm{KL}(Q||Q_0) + \log(\frac{2\sqrt{n}}{\delta})}{n} \quad (1)$$

For the remainder of the paper, K denotes the right-hand side of ineq. (1):

$$K := \frac{\mathrm{KL}(Q||Q_0) + \log(\frac{2\sqrt{n}}{\delta})}{n}$$

Since $\mathrm{kl}(\hat{L}_S(Q)||\hat{L}_S(Q)) = 0$, we assume $0 < \hat{L}_S(Q) \leq L(Q) < 1$ when discussing any upper bound on $L(Q)$ without loss of generality.

Although $L(Q)$ cannot be computed without access to Z, Theorem 2.1 gives an upper bound on $L(Q)$. However, because the bound is not explicit on $L(Q)$, it may be bothersome in practice. A common workaround to obtain an explicit bound on $L(Q)$ is to find a lower bound to the left side of (1) and solve for $L(Q)$. Another workaround is to consider the following inequality equivalent to (1)

$$L(Q) \leq \mathrm{kl}^{-1}(\hat{L}_S(Q), K) \quad (2)$$

where kl^{-1} is precisely the function that makes both inequalities equivalent, that is, $\mathrm{kl}^{-1}(p, k) = \sup\{q \in [p, 1) : \mathrm{kl}(p||q) \leq k\}$. In fact, due to the convexity and non-negativity of the divergences, $\mathrm{kl}(p||q) = 0 \iff p = q$ implies $\mathrm{kl}(p||q)$ is strictly increasing for $q > p$. This means $\mathrm{kl}^{-1}(p, k) := q$ such that $\mathrm{kl}(p||q) \leq k$ and $q \geq p$ is equivalent and well-defined. The inversion of the binary KL of Seeger [2002], $\mathrm{kl}^{-1}(p, k) := \Delta$ such that $\mathrm{kl}(p||p + \Delta) \leq k$ and $\Delta \geq 0$, while not consistent with the inverse of a function, simplifies the proofs in this paper. The second workaround involves upper-bounding the right-hand side of (2) [Alquier, 2024, Hellström et al., 2025]. Moreover, this explicit bound can then be used to optimize Theorem 2.1 with respect to Q. This is possible because, although an analytical expression for $\mathrm{kl}^{-1}(p, k)$ is not known, it can be approximated numerically.

2.1 RELAXING THE INEQUALITIES

This subsection reviews the most common relaxations of inequalities (1) and (2). The review is non-exhaustive, and does not cover parametric relaxations such as the PAC-Bayes-λ bound [Thiemann et al., 2017], which is closely related to KL-modulation (see subsection 4), or inequalities that are looser in the < 1 regime such as the Bretagnole-Huber inequality [Bretagnolle and Huber, 1979, Canonne, 2022]. A key point here is that any lower bound for the left-hand side of ineq. (1) that admits a closed-form solution for

$L(Q)$ can be turned into an upper bound for the right-hand side of ineq. (2), and any upper bound for the right-hand side of (2) that admits a closed-form solution for K can be turned into a lower bound for the right-hand side of (1). Since we need a closed-form solution for $L(Q)$ but not for K, the second approach is more general than the first and should be the tool to derive even tighter bounds in the future. Each bound presented in this section admits a closed form for K, and both forms are provided.

2.1.1 Pinsker's inequality

Consider Pinsker's inequality, a classical lower bound on the KL divergence that applied to two Bernoulli distributions with means p and q yields $\mathrm{kl}(p||q) \geq 2(q-p)^2$. Applying this lower bound to (1) produces a version of the classic McAllester bound [McAllester, 1999] with improved constants, highlighting the potential of Maurer's bound [Dziugaite and Roy, 2017]:

$$L(Q) \leq \hat{L}_S(Q) + \sqrt{\frac{K}{2}} \tag{3}$$

This bound is equivalent to upper bounding the right-hand side of (2) by $\mathrm{kl}^{-1}(p,k) \leq p + \sqrt{k/2}$.

2.1.2 Refined Pinsker's inequality

Applying a refinement of the Pinsker inequality, $\mathrm{kl}(p||q) \geq (q-p)^2/(2q)$, to the left-hand side of (1) and solving for $L(Q)$ yields the PAC-Bayes-quadratic (PBQ) bound [Rivasplata et al., 2019]:

$$L(Q) \leq \left(\sqrt{\hat{L}_S(Q) + \frac{K}{2}} + \sqrt{\frac{K}{2}}\right)^2 \tag{4}$$

This bound is tighter than (3) when and only when $L(Q) < 1/4$, which is often the case in practice. In fact, the objective functions derived from (4) produce much better results than those derived from (3) [Pérez-Ortiz et al., 2021].

Notice (4) is equivalent to upper bounding the right-hand side of (2) by $\mathrm{kl}^{-1}(p,k) \leq p + k + \sqrt{2pk + k^2} = (\sqrt{p + k/2} + \sqrt{k/2})^2$.

2.1.3 Tolstikhin and Seldin's (TS) bound

Tolstikhin and Seldin [2013] produce a PAC-Bayesian bound by applying $\mathrm{kl}^{-1}(p,k) \leq p + 2k + \sqrt{2pk}$ to the right-hand side of (2):

$$L(Q) \leq \hat{L}_S(Q) + 2K + \sqrt{2\hat{L}_S(Q)K} \tag{5}$$

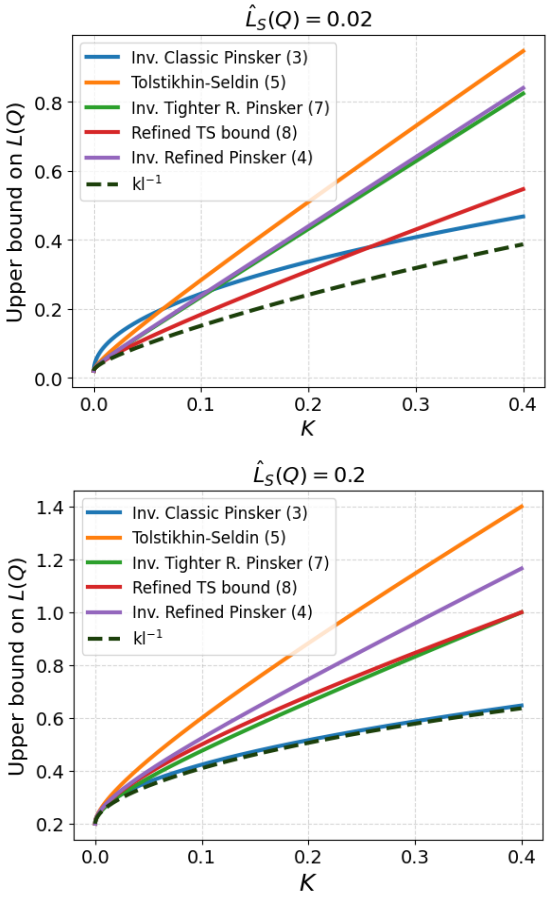

Figure 1: Different upper bounds on the true risk of the posterior over the hypothesis, $L(Q)$, as a function of the right hand side of (1), K, for a small (top) and a large (bottom) empirical risk $\hat{L}_S(Q)$. The dark green curve represents Maurer's bound, of which the rest are relaxations of, so the closer to the dark green curve the better the relaxation. The bottom plot shows that for larger values of the empirical risk the classic Pinsker bound is extremely tight.

This is equivalent to lower-bounding the left-hand side of (1) by $\mathrm{kl}(p||q) \geq (2q - p - \sqrt{4qp - 3p^2})/4$.

It has been described as "better" [Alquier, 2024] and "more refined" [Hellström et al., 2025] than (3), the bound derived from the classic Pinsker inequality. However, this tighter behavior strongly depends on the values of $\hat{L}_S(Q)$ and K.

2.2 GRADIENT OF MAURER'S BOUND

Closed-form approximations to the inverse of the KL enable the minimization of the bound on $L(Q)$ through gradient descent. This way, one can fit NNs with specific objective functions that optimize any of the risk certificates expressed by these bounds [Pérez-Ortiz et al., 2021]. However, it is possible to directly calculate the gradient of Maurer's bound

on $L(Q)$ with respect to θ, the parameters that define Q.

Let q be the bound on $L(Q)$ given in (1), for an empirical loss $p := \hat{L}_S(Q)$ and a right-hand term K:

$$\nabla_\theta q = \xi \left(\left(\log \frac{1-p}{1-q} - \log \frac{p}{q} \right) \nabla_\theta p + \nabla_\theta K \right) \tag{6}$$

where all the gradients are computed *w.r.t.* θ, and

$$\xi = \left(\frac{1-p}{1-q} - \frac{p}{q} \right)^{-1}$$

This and similar results, although useful in practice, have eluded mainstream adoption in PAC-Bayes, being rediscovered independently multiple times in the literature. To our knowledge, it was first introduced in Equation (5) of [Germain et al., 2009] in the context of linear classifiers. Then, Reeb et al. [2018] rediscovered it in the context of gaussian processes. Although the method had direct application to the PAC-Bayes-by-Backprop line of research Rivasplata et al. [2019], it was not introduced in it until Rodriguez-Galvez et al. [2024].

Hoping these results see widespread adoption, the method used during the writing of this paper to independently derive them is provided in Appendix B.

3 TIGHTER PAC-BAYES BOUNDS

The first contribution of the paper is a new, universally tighter bound than the one in (4).

Theorem 3.1. *Let* $f(p,q) := \frac{(q-p)^2}{2q(1-p)}$. *Then*

$$kl(p||q) \geq f(p,q) \qquad 0 < p \leq q < 1$$

Proof. See Appendix A. □

The new $1-p$ term makes the improvement over (4) uniform in $\{(p,q) : 0 < p \leq q < 1\}$. Now, applying the new bound to (1) and solving for $L(Q)$ yields the Tighter Refined Pinsker (TRP) bound:

$$L(Q) \leq \left(\sqrt{\hat{L}_S(Q) + \frac{K(1-\hat{L}_S(Q))}{2}} + \sqrt{\frac{K(1-\hat{L}_S(Q))}{2}} \right)^2 \tag{7}$$

This is equivalent to upper bounding the right-hand side of (2) by $\text{kl}^{-1}(p,k) \leq p + (1-p)k + \sqrt{2p(1-p)k + k^2(1-p)^2} = (\sqrt{p + k(1-p)/2} + \sqrt{(1-p)k/2})^2$.

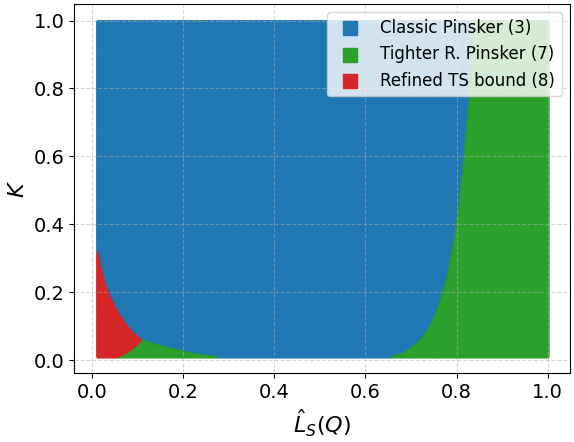

Figure 2: Map with the tightest bound (out of all the bounds studied in the paper) for different values of K (right-hand side of (1)) and $\hat{L}_S(Q)$.

Figure 1 shows that the TRP bound is universally tighter than PBQ, and that this difference is most noticeable for larger values of $\hat{L}_S(Q)$.

The second contribution of the paper is a new, universally tighter bound than (4) and (5), and tighter than (7) for smaller ranges of the empirical risk $\hat{L}_S(Q)$:

$$L(Q) \leq \hat{L}_S(Q) + K + \sqrt{2\hat{L}_S(Q)K} \tag{8}$$

This bound, which we call Refined Tolstikhin-Seldin (RTS) bound, can be obtained by lower-bounding the left-hand side of (1) by $\text{kl}(p||q) \geq q - \sqrt{2qp - p^2}$.

Theorem 3.2. *Let* $f(p,q) := q - \sqrt{2qp - p^2}$. *Then,*

$$kl(p||q) \geq f(p,q) \qquad 0 < p \leq q < 1$$

Proof. See Appendix A. □

Figure 1 shows that for a larger value of the empirical risk, the classic bound (3) is significantly tighter than the other bounds analyzed in the paper, and closely follows the inverse of the KL. When the empirical risk is smaller, RTS is the tightest for small values of K and the classic bound (3) is the tightest for increasing values of K. With respect to the bounds proposed in the paper, Figure 1 shows that they are universally tighter than their counterparts: RTS is tighter than TS bound and TRP is tighter than PBQ. This can also be easily seen by comparing their bounds on kl^{-1}. Eq. (7) is similar to (4) but with some terms multiplied by a constant which is less than 1, and when compared to Eq. (8), Eq. (4) has an extra K^2 term inside the square root, and Eq. (5) has a $2K$ term instead of just K.

Moreover, Figure 2 shows that just three of the bounds analyzed in the paper suffice to cover the whole range of

values of K and $\hat{L}_S(Q)$: The classic Pinsker relaxation of (3) covers the largest area while the bounds introduced in this paper are the tightest for small values of K and $\hat{L}_S(Q)$, as well as for large values of $\hat{L}_S(Q)$.

4 ON THE OPTIMIZATION OF NON-DIFFERENTIABLE FUNCTIONS

The reason why a better approximation $f(\hat{L}_S(Q), K) \approx \text{kl}^{-1}(\hat{L}_S(Q), K)$ produces a better risk certificate through gradient descent is not that the error $|f - \text{kl}^{-1}|$ is lower, but that the direction of the gradient of the bound on $L(Q)$ with respect to θ deviates less from the one derived directly from Maurer's bound. This gradient can be regarded as a weighted sum of the gradients of $\hat{L}_S(Q)$ and K (of course the coefficients of said weighted sum would depend heavily on the values of $\hat{L}_S(Q)$ and K). Then, the role of our approximation f will be to estimate the ratio between both coefficients of the weighted average. In those cases where f consistently over or underestimates this ratio, corrections should be applied.

This idea of correcting the relative weights of empirical loss and KL is not new. It has been used many times to derive risk certificates in the Bayes-by-backpropagation literature [Blundell et al., 2015, Pérez-Ortiz et al., 2021]. However, the motivations given in these works to justify this correction are fundamentally empirical. The strength of the correction is commonly optimized through grid search, or just fixed to a value that "seems to work well". We argue that an unbiased objective function does not need an arbitrary "KL-attenuating coefficient" (inequalities in Sections 2 and 3 are biased estimators of Maurer's bound by nature, as they are bounds), and if one seeks to optimize a function with unknown gradient through a surrogate function with known gradient, then all efforts should be devoted to the estimation of the relationship between both gradients, rather than to the optimization of the scaling of the KL term in the gradient of the surrogate function.

Problem (6) expresses the gradient of a training objective as a function of q, p, ∇p and ∇K. The terms q, p and ∇K are computable, but we require the empirical risk p to be differentiable for ∇p to exist. This is also a requisite in the traditional machine learning setting. A common procedure to optimize a non-differentiable loss, like the 0-1 loss l^{01}, is to assume that the minima of the non-differentiable loss will align with those of a differentiable one, like the cross-entropy loss l^{xe}, and optimize for the latter expecting to optimize for the former. However, this is not the case with PAC-Bayes bounds.

To work around this issue, consider the following setting: let l^t be a non-differentiable loss function whose bound on the true risk $L^t(Q)$ we seek to optimize, and l^d be a differentiable loss function. Let us denote by $\hat{L}^t_S$ and $\hat{L}^d_S$ the empirical risks corresponding to l^t and l^d, respectively. To use l^d as a surrogate loss, one must build a differentiable function $r(\hat{L}^d_S) \approx \hat{L}^t_S$. Now, applying the chain rule to (6) yields the following gradient ∇q_t for loss l^t:

$$\nabla q_t \approx \xi_t \left(\left(\log \frac{1-p_t}{1-q_t} - \log \frac{p_t}{q_t} \right) r'(p_d) \nabla p_d + \nabla K \right) \tag{9}$$

where subscripts d and t assigns the corresponding term to the implicit function for loss l^d or l^t, respectively, and

$$\xi_t = \left(\frac{1-p_t}{1-q_t} - \frac{p_t}{q_t} \right)^{-1}$$

This bypasses the need to introduce a KL-modulating coefficient $\eta > 0$ in the surrogate objective and optimize it through grid-search. In general, each minimum in the true objective is also a minimum in the surrogate objective for exactly one value of η. Consider you have found an arbitrarily good estimator r in equation (9). For a fixed $\theta \in \Theta$, the gradient of the arbitrarily good resulting objective function can be computed as $c_L \nabla \hat{L}^d_S + c_K \nabla K$. Now, consider the gradient of a different objective function evaluated at the same θ, $b_L \nabla \hat{L}^d_S + b_K \nabla K$. Since b_L, b_K, c_L and c_K are strictly positive, for any $\theta \in \Theta$, there exists a KL-attenuating coefficient $\eta > 0$ that verifies $c_L \nabla \hat{L}^d_S + c_K \nabla K \propto b_L \nabla \hat{L}^d_S + \eta b_K \nabla K$, which is $\eta = c_K b_L / (c_L b_K)$. Consequently, for any θ^* that is a local minimum of the arbitrarily good objective function, there is a value of η that makes θ^* a minimum of the KL-modulated objective function. Therefore, applying the KL modulating method with the optimal η to a surrogate function is sufficient to find any good local minimum, including the best possible solution.

This method is commonly known as "KL-attenuating", because in the literature the optimal value of η is generally less than 1. However, the experiments in Section 5 show that the optimal value of η to minimize a risk certificate may be greater than one, so we rename it "KL modulating".

5 EXPERIMENTS

This section gives some insights on the quality of the theoretical contributions of the paper by showing their performance in the certification of NNs trained on MNIST [Deng, 2012, CC3.0 License] and CIFAR-10 [Krizhevsky and Hinton, 2009, MIT License]. [1]

In all experiments, the prior distribution on the classifiers, Q_0, is the product of the individual priors on each weight of the network. Each individual prior is a Gaussian $\mathcal{G}(m, \sigma^2)$ where σ is explored in $[0.02, 0.08]$ (every prior has the same

[1] Code to replicate all experiments in this paper is available at `github.com/Diegogpcm/pacbayesgradients`.

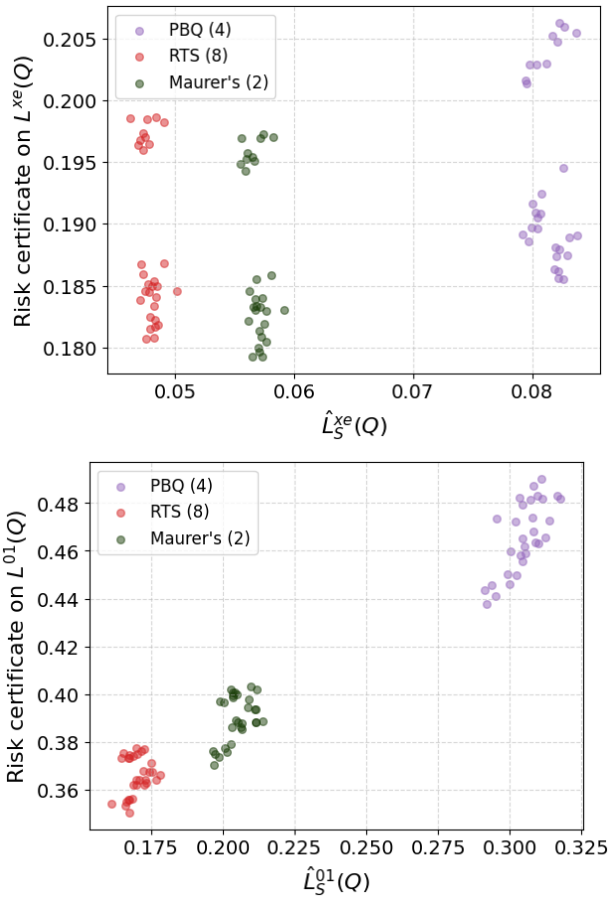

Figure 3: Risk certificates on cross-entropy loss (top) and 0-1 loss (bottom) on different experiment runs on the MNIST dataset. Color indicates training objective. The training objective functions work with L^{xe} in both plots, which explains why the RTS bound outperforms the stronger bypassed Maurer's bound in the bottom plot.

σ in a given experiment run) and m is randomly sampled from $\mathcal{G}(0, 1/n_{in})$, where n_{in} is the dimension of the input to the corresponding layer.

The posterior Q is a Gaussian with diagonal covariance matrix (i. e., the parameters are independent). The mean and covariance of Q are initialized to be those of Q_0 and optimized through 100 epochs of stochastic gradient descent on the corresponding objective function. Training is done by drawing a fresh sample from Q on each batch, performing inference and backpropagating as usual. We evaluate three of these objective functions:

- f_{pbq}, the PBQ bound of (4), that serves as baseline for the proposals of this work.
- f_{rts}, the RTS bound of (8).
- f_{mb}, resulting from the application of the procedure described in (6) to the Maurer's bound of (1).

These objectives optimize the risk bound on the bounded cross-entropy loss (see Pérez-Ortiz et al. [2021]) with class probabilities restricted to being larger than $1e-5$. These objectives are sometimes modified according to Section 4 to optimize the 01-loss instead. In such cases, they are presented as $\tilde{f}$. We do not include the results for f_{trp} because they are similar to f_{pbq}, and f_{pbq} serves as state-of-the-art baseline. The results for f_{trp} can be found in the supplementary material.

All risk certificates in the results are obtained with a confidence factor of $\delta = 0.025$. Moreover, the empirical risks in the bounds are estimated using Monte Carlo with 150000 models sampled from Q, following the same procedure as Pérez-Ortiz et al. [2021].

The provided confidence intervals are calculated with 10 different seeds for the random number generator. Reported values are formatted as $\mu \pm 2\sigma$, where μ represents the mean of the results and σ the standard deviation. Notice these confidence intervals have nothing to do with the confidence in the bound itself, which is already set to be $1-\delta = 0.975$. The relative size of these intervals also gives insight on the sensitivity of the final value of the bound to randomized elements of the learning algorithm, such as the prior initialization and the ordering of the samples during training. We also compute expected losses for the Gibbs classifier $x \mapsto h(x), h \sim Q$, the deterministic classifier $x \mapsto h(x), h = \mathbb{E}[Q]$ and the ensemble classifier $x \mapsto \frac{1}{n}\Sigma_{i=1}^{100} h_i(x), h_i \sim Q$. All experiment runs are done on NVIDIA RTX 4090 GPUs and each run takes a few minutes.

5.1 MNIST

The first set of experiments consists of fitting 3-layer MLPs with 600 neurons in each hidden layer on the MNIST dataset, and is a reproduction of the experiments in Pérez-Ortiz et al. [2021] and Rodriguez-Galvez et al. [2024] with the new RTS objective added.

Table 1 shows that f_{rts} achieves the tightest certificate for the 0-1 loss, and also serves as core model for the most accurate classifiers.

Figure 3 shows that the proposed RTS bound and the bypass procedure of (6) achieve risk certificates significantly tighter than the PBQ bound. This is particularly marked when the risk certificate involves the 0-1 loss but the objective functions work with the cross-entropy loss (bottom plot). With respect to the performance of the contributions of this paper, Figure 3 and Table 1 show that the a priori stronger theoretical result corresponding to f_{mb} achieves marginally better results than f_{rts} in the bound of L^{xe} while losing against it in every other metric. This is not a surprise; Section 4 shows that if the gradient of the bound on $L(Q)$ is expressed as $c_L \nabla \hat{L}_S^{xe} + c_K \nabla K$, f_{pbq} overestimates the ratio c_K/c_L in Maurer's bound, while f_{rts} underestimates

Table 1: Results for the three objective functions used to train the MLP with the MNIST data. The first five rows show information about the tightness of the risk certificate of the posterior Q, namely the bound for the cross-entropy loss (L^{xe}), the bound for the 0-1 loss (L^{01}), plus the corresponding empirical risks ($\hat{L}_S^{01}(Q)$ and $\hat{L}_S^{xe}(Q)$) and penalty (KL/n) used for their computation. Next rows show the accuracy, measured as the l^{xe} or the l^{01} averaged over the test set, of 4 classification strategies implemented with each model (Gibbs classifier, deterministic classifier, ensemble and average of the prior).

-	Metric	f_{pbq}	f_{rts}	f_{mb}
Risk bound	L^{xe}	$0.173_{\pm 0.002}$	$0.169_{\pm 0.002}$	$\mathbf{0.167}_{\pm 0.003}$
	L^{01}	$0.430_{\pm 0.014}$	$\mathbf{0.339}_{\pm 0.009}$	$0.361_{\pm 0.011}$
	KL/n	$0.032_{\pm 0.001}$	$0.065_{\pm 0.001}$	$0.053_{\pm 0.001}$
Emp. R bound	$\hat{L}_S^{xe}(Q)$	$0.095_{\pm 0.002}$	$\mathbf{0.059}_{\pm 0.001}$	$0.069_{\pm 0.001}$
	$\hat{L}_S^{01}(Q)$	$0.324_{\pm 0.012}$	$\mathbf{0.192}_{\pm 0.007}$	$0.227_{\pm 0.010}$
Gibbs	xe loss	$0.082_{\pm 0.001}$	$\mathbf{0.048}_{\pm 0.001}$	$0.057_{\pm 0.001}$
	01 loss	$0.299_{\pm 0.010}$	$\mathbf{0.168}_{\pm 0.008}$	$0.202_{\pm 0.010}$
$\mathbb{E}(Q)$	xe loss	$0.048_{\pm 0.006}$	$\mathbf{0.027}_{\pm 0.008}$	$0.032_{\pm 0.007}$
	01 loss	$0.145_{\pm 0.010}$	$\mathbf{0.097}_{\pm 0.006}$	$0.109_{\pm 0.007}$
Ensemble	xe loss	$0.0020_{\pm 0.0002}$	$\mathbf{0.0012}_{\pm 0.0002}$	$0.0014_{\pm 0.0002}$
	01 loss	$0.150_{\pm 0.013}$	$\mathbf{0.097}_{\pm 0.006}$	$0.110_{\pm 0.008}$
$\mathbb{E}(Q_0)$	01 loss		$0.899_{\pm 0.043}$	

it, effectively applying the "KL-attenuating trick". The consequence of this can be clearly seen in the bottom plot of Figure 3, where experiment runs that use f_{rts} as objective attain a better bound than f_{mb}, showing that the optimal objective for minimizing the bound on l^{xe} is not optimal for minizing the bound on l^{01}. This illustrates the goal of Section 4, which is designing the optimal objective.

5.1.1 Optimizing for accuracy

The objective functions used to obtain the results in Table 1 had a bounded version of the cross-entropy loss as optimization target. As discussed in Section 4, to optimize the bound for the 0-1 loss we must estimate $\hat{L}_S^{01}$ as a function of $\hat{L}_S^{xe}$. Figure 4 makes a strong case to choose r to be linear, and use the approximation $m\hat{L}_S^{xe} \approx \hat{L}_S^{01}$ in (9). A reasonable value for m is estimated by dividing a rolling average of L_{01} by a rolling average of L_{xe}. We denote by $\tilde{f}_{pbq}$, $\tilde{f}_{rts}$ and $\tilde{f}_{mb}$ the objective functions resulting by applying this procedure to f_{pbq}, f_{rts} and f_{mb}, respectively. As expected, Table 2 shows improved risk certificates for the 0-1 loss, and worsened risk certificates for the bounded cross-entropy loss when the objective functions pursue the optimization of the 0-1 loss. Just like in the previous experiment, $\tilde{f}_{rts}$ obtains better empirical performance due to a weaker penalty on KL divergence.

Figure 5 shows that when applying the correction method to optimize for accuracy, the objective corresponding to the proposed stronger theoretical result, $\tilde{f}_{mb}$, obtains slightly tighter risk certificates than the other proposal, $\tilde{f}_{rts}$, and significantly tighter than the baseline $\tilde{f}_{pbq}$.

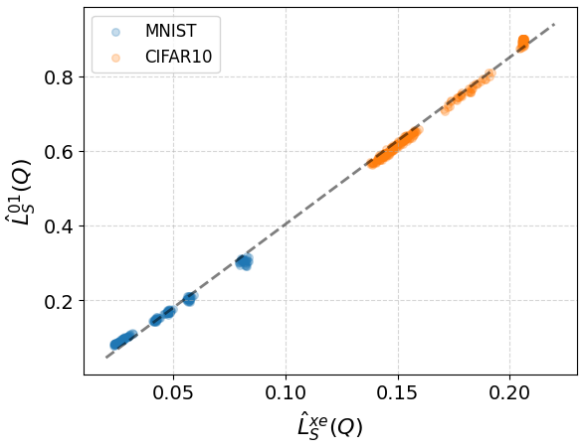

Figure 4: Observed values of $\hat{L}_S^{01}$ plotted against the value of $\hat{L}_S^{xe}$ for the same experiment run in both MNIST and CIFAR-10. An approximately linear relationship is strongly implied.

5.1.2 Applying the KL-modulating method

This set of experiments illustrates the capabilities of the KL-modulating method introduced in Section 4 to align the gradient of any objective function with that of the 0-1 loss. Table 3 shows that the three objective functions end up achieving very similar values for both risk certificates, especially for the one built on the 0-1 loss.

5.2 CIFAR-10

This set of experiments reproduces those in Section 5.1.1 on the CIFAR-10 dataset. Results in Table 4 show that the

Table 2: Results for the three modified objective functions used to train the MLP with the MNIST data. The objective functions are modified to directly optimize L^{01} through the application of the procedure in (9). The objective function based on Maurer's bound obtains a better bound on l^{01}, which is the target objective in this experiment.

-	Metric	$\tilde{f}_{pbq}$	$\tilde{f}_{rts}$	$\tilde{f}_{mb}$
Risk bound	L^{xe}	$\mathbf{0.182}_{\pm 0.006}$	$0.221_{\pm 0.003}$	$0.205_{\pm 0.003}$
	L^{01}	$0.345_{\pm 0.021}$	$0.329_{\pm 0.005}$	$\mathbf{0.325}_{\pm 0.006}$
	KL/n	$0.083_{\pm 0.001}$	$0.164_{\pm 0.003}$	$0.136_{\pm 0.003}$
Emp. R bound	$\hat{L}_S^{xe}(Q)$	$0.055_{\pm 0.004}$	$\mathbf{0.032}_{\pm 0.001}$	$0.037_{\pm 0.001}$
	$\hat{L}_S^{01}(Q)$	$0.177_{\pm 0.018}$	$\mathbf{0.099}_{\pm 0.003}$	$0.116_{\pm 0.003}$
Gibbs	xe loss	$0.044_{\pm 0.004}$	$\mathbf{0.024}_{\pm 0.001}$	$0.028_{\pm 0.001}$
	01 loss	$0.154_{\pm 0.019}$	$\mathbf{0.085}_{\pm 0.003}$	$0.100_{\pm 0.004}$
$\mathbb{E}(Q)$	xe loss	$0.025_{\pm 0.009}$	$\mathbf{0.012}_{\pm 0.007}$	$0.014_{\pm 0.007}$
	01 loss	$0.092_{\pm 0.007}$	$\mathbf{0.045}_{\pm 0.003}$	$0.053_{\pm 0.003}$
Ensemble	xe loss	$0.0011_{\pm 0.0003}$	$\mathbf{0.0006}_{\pm 0.0002}$	$0.0007_{\pm 0.0002}$
	01 loss	$0.098_{\pm 0.017}$	$\mathbf{0.047}_{\pm 0.004}$	$0.056_{\pm 0.004}$
$\mathbb{E}(Q_0)$	01 loss		$0.899_{\pm 0.043}$	

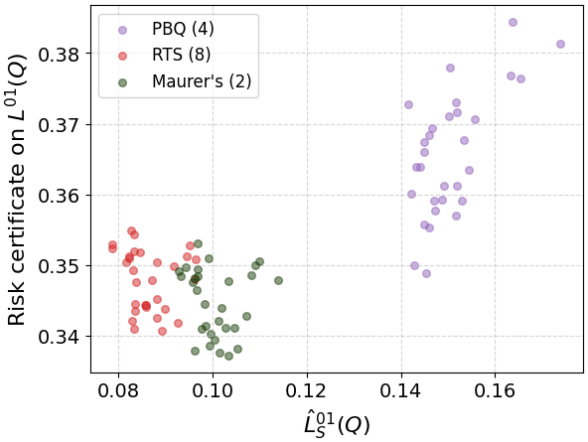

Figure 5: Risk certificates computed for each MLP (one MLP per objective function and explored length scale of the prior Q_0) using the 0-1 loss. The training objective functions work with L^{01}.

Table 3: The same results as in Table 1 after introducing an extensive grid-search over a KL-modulating coefficient. As predicted, similar risk certificates on L^{01} are attained with every method.

-	Metric	$\tilde{f}_{pbq}$	$\tilde{f}_{rts}$	$\tilde{f}_{mb}$
Risk bound	L^{xe}	0.186	0.191	0.190
	L^{01}	0.315	0.315	0.315
	KL/n	0.107	0.117	0.114

objective functions introduced in this paper combined with our method for optimizing the risk bound on accuracy produce non-trivial bounds on performance on the CIFAR-10 dataset, which is to our knowledge a first in PAC-Bayes and NNs. Deeper NNs (9+ layers) such as the ones used on CIFAR-10 in Pérez-Ortiz et al. [2021] produce vacuous bounds without data-dependent priors. In our experiments, we use shallower models consisting of 4 and 5 layers. Results for the unmodified objective functions are not provided for the CIFAR-10 dataset as they are all vacuous.

The results show a large gap in performance between optimizing directly against Maurer's bound and optimizing with surrogates. This gap grows even larger for deeper models, which might be explained by the error of the gradient accumulating throughout a larger amount of backpropagation steps. Another interesting result is that, while on the MNIST dataset the $\tilde{f}_{rts}$ objective yielded a posterior with higher KL divergence due to a lower KL-penalizing coefficient, on CIFAR-10 the KL divergence is higher for the $\tilde{f}_{mb}$ posterior.

The gap between the generalization error bound and the true generalization error remains large. A very precise estimate of the generalization error can be seen in Table 4 as the difference between the bound on the empirical loss and the Gibbs loss over the test set. The gap then hovers around a full order of magnitude.

Additionally, in the experiments in this section as well as Section 5.1.1, the variance of the bound on the risk is generally significantly smaller than the variance of the loss of the prior. This highlights the stability of the learning algorithm, which is a desirable property, but also indicates there is room for improvement, as it shows that good minima are distributed somewhat evenly throughout the parametric space and the KL penalty restricts the search to a very small region.

6 CONCLUSIONS

This paper has introduced two new explicit PAC-Bayes bounds on the true risk of binary classifiers and has demon-

Table 4: Results for the three modified objective functions used to train CNNs of varying sizes on CIFAR-10. The objective functions are again modified to directly optimize L^{01} through the application of the procedure in (9). Larger models give vacuous bounds and a KL of 0. The 01 loss of the prior, $l^{01}(Q_0)$, is also provided.

Model layers	# params	Metric	$\tilde{f}_{pbq}$	$\tilde{f}_{rts}$	$\tilde{f}_{mb}$
4	9946	Bound on L^{01}	$0.841_{\pm 0.101}$	$0.745_{\pm 0.011}$	$\mathbf{0.738}_{\pm 0.007}$
		Bound on $\hat{L}_S^{01}$	$0.820_{\pm 0.146}$	$0.648_{\pm 0.015}$	$\mathbf{0.610}_{\pm 0.012}$
		KL$/n$	$0.006_{\pm 0.009}$	$0.032_{\pm 0.002}$	$0.051_{\pm 0.004}$
		Gibbs 01 loss	$0.802_{\pm 0.158}$	$0.625_{\pm 0.017}$	$\mathbf{0.586}_{\pm 0.016}$
		$\mathbb{E}(Q)$ 01 loss	$0.751_{\pm 0.207}$	$0.549_{\pm 0.020}$	$\mathbf{0.507}_{\pm 0.019}$
		Ensemble 01 loss	$0.750_{\pm 0.231}$	$0.548_{\pm 0.025}$	$\mathbf{0.507}_{\pm 0.022}$
		$L^{01}(\mathbb{E}[Q_0])$		$0.906_{\pm 0.031}$	
4	315722	Bound on L^{01}	$0.823_{\pm 0.023}$	$0.773_{\pm 0.007}$	$\mathbf{0.758}_{\pm 0.005}$
		Bound on $\hat{L}_S^{01}$	$0.783_{\pm 0.040}$	$0.659_{\pm 0.012}$	$\mathbf{0.598}_{\pm 0.011}$
		KL$/n$	$0.012_{\pm 0.006}$	$0.045_{\pm 0.005}$	$0.077_{\pm 0.007}$
		Gibbs 01 loss	$0.762_{\pm 0.035}$	$0.638_{\pm 0.019}$	$\mathbf{0.577}_{\pm 0.014}$
		$\mathbb{E}(Q)$ 01 loss	$0.709_{\pm 0.062}$	$0.562_{\pm 0.018}$	$\mathbf{0.495}_{\pm 0.020}$
		Ensemble 01 loss	$0.679_{\pm 0.057}$	$0.555_{\pm 0.019}$	$\mathbf{0.486}_{\pm 0.020}$
		$L^{01}(\mathbb{E}[Q_0])$		$0.895_{\pm 0.018}$	
5	12266	Bound on L^{01}	>0.9	>0.9	$\mathbf{0.756}_{\pm 0.006}$
		Bound on $\hat{L}_S^{01}$	>0.9	>0.9	$\mathbf{0.624}_{\pm 0.010}$
		KL$/n$	0.000	0.000	$0.056_{\pm 0.005}$
		Gibbs 01 loss	$0.897_{\pm 0.009}$	$0.895_{\pm 0.011}$	$\mathbf{0.602}_{\pm 0.015}$
		$\mathbb{E}(Q)$ 01 loss	$0.894_{\pm 0.025}$	$0.888_{\pm 0.044}$	$\mathbf{0.525}_{\pm 0.016}$
		Ensemble 01 loss	$0.899_{\pm 0.008}$	$0.897_{\pm 0.011}$	$\mathbf{0.524}_{\pm 0.016}$
		$L^{01}(\mathbb{E}[Q_0])$		$0.900_{\pm 0.008}$	

strated that they are tighter than the state of the art. These bounds have been used to design objective functions that enable the training of NNs with the optimization of the risk certificate as main goal. Additionally, the paper has introduced a procedure to optimize the bound for non-differentiable loss functions such as the 0-1 loss, which is tied to the KL-attenuating method used in the literature. These theoretical results have been successfully applied to the calculation of risk certificates for neural networks trained on MNIST and CIFAR-10. The latter is particularly clarifying with respect to the efficacy of the results, as the contribution in section 4 is necessary to derive the nonvacuous risk bounds.

The contributions in this paper are orthogonal to other PAC-Bayesian learning techniques such as data-dependent priors, and thus can be employed simultaneously. Notably, they may be applied in all settings where PAC-Bayesian training of neural networks has already found success, such as transfer learning and domain adaptation.

Acknowledgements

The authors acknowledge the research environment provided by ELLIS Unit Madrid. This work has been partially funded by the Comunidad Autónoma de Madrid under project IDEA-CM (TEC-2024/COM-89) and grant PEJ-2024-AI/COM-32081 (2024 Ayudantes de Investigación program), and by the Agencia Estatal de Investigación of Spain through projects COBA (PID2023-146684NB-I00) and GENICOM (PID2023-148856OA-I00).

References

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024. URL `https://doi.org/10.1561/2200000100`.

Gilles Blanchard and François Fleuret. Occam's Hammer. In *Proceedings of the 20th Annual Conference on Learning Theory COLT*, pages 112–126. Springer, 2007.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International conference on machine learning*, pages 1613–1622. PMLR, 2015.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Clément L Canonne. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198*, 2022.

Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, page 1–163, 2007. URL http://dx.doi.org/10.1214/074921707000000391.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI*, 2017.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes bounds. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics AISTATS*, pages 604–612. PMLR, 2021.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 353–360, 2009.

Benjamin Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, 2019. URL https://arxiv.org/abs/1901.05353.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. PAC-Bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10): 1330, 2021.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and PAC-Bayes. *Foundations and Trends in Machine Learning*, 18(1):1–223, 2025.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

John Langford and Matthias Seeger. *Bounds for averaging classifiers*. Technical Report. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2001.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

Andreas Maurer. A Note on the PAC Bayesian Theorem. *CoRR*, cs.LG/0411099, 2004. URL http://arxiv.org/abs/cs.LG/0411099.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational learning theory COLT*, pages 164–170, 1999.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes bounds with data dependent priors. *The Journal of Machine Learning Research*, 13(1):3507–3531, 2012.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22 (227):1–40, 2021.

David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning Gaussian Processes by minimizing PAC-Bayesian generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.

Omar Rivasplata, Vikram M Tankasali, and Csaba Szepesvari. PAC-Bayes with backprop. *arXiv preprint arXiv:1908.07380*, 2019.

Borja Rodriguez-Galvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110): 1–43, 2024.

Matthias Seeger. PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification. *Journal of Machine Learning Research*, 2002.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 466–492. PMLR, 2017.

Ilya O Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein Inequality. In *Advances in Neural Information Processing Systems 26*, pages 109–117, 2013.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of PAC-Bayesian bounds. *Machine Learning*, 113(2):519–604, 2024.

Yi-Shan Wu, Yijie Zhang, Badr-Eddine Chérief-Abdellatif, and Yevgeny Seldin. Recursive PAC-Bayes: A frequentist approach to sequential prior updates with no information loss. In *Advances in Neural Information Processing Systems 37*, pages 17947–17971, 2024.

Xitong Zhang, Avrajit Ghosh, Guangliang Liu, and Rongrong Wang. Improving generalization of complex models under unbounded loss using PAC-bayes bounds. *Transactions on Machine Learning Research*, 2024.

Advances in PAC-Bayesian Certification of Deep Neural Networks: Tighter Closed-Form Inequalities and Optimization of Bounds on Non-Differentiable Losses
(Supplementary Material)

Diego García-Pérez[1] **Emilio Parrado-Hernández**[1] **John Shawe-Taylor**[2,3]

[1]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain
[2]AI Centre, Dept. of Computer Science, University College London, UK
[3]IRCAI, Institute Jožef Stefan, Slovenia

A PROOFS

Theorem (3.1)**.** *Let* $f(p,q) := \frac{(q-p)^2}{2q(1-p)}$.

$$kl(p||q) \geq f(p,q) \qquad 0 < p \leq q < 1$$

Proof. Let $g_p(\Delta) = \text{kl}(p||p+\Delta) - f(p, p+\Delta)$, for $\Delta \in [0, 1-p)$. $g_p(0) = 0$, since $\text{kl}(p||p) = f(p,p) = 0$. We first show $g'_p(\Delta) \geq 0$:

$$\begin{aligned} g'_p(\Delta) &= \frac{1-p}{1-p-\Delta} - \frac{p}{p+\Delta} - \frac{\Delta}{(1-p)(p+\Delta)} + \frac{\Delta^2}{2(1-p)(p+\Delta)^2} \\ &= \frac{\Delta}{(1-p-\Delta)(p+\Delta)} - \frac{\Delta}{(1-p)(p+\Delta)} + \frac{\Delta^2}{2(1-p)(p+\Delta)^2} \\ &\geq \frac{\Delta}{(1-p)(p+\Delta)} - \frac{\Delta}{(1-p)(p+\Delta)} + \frac{\Delta^2}{2(1-p)(p+\Delta)^2} \\ &= \frac{\Delta^2}{2(1-p)(p+\Delta)^2} \geq 0 \end{aligned}$$

Since g is continuous in its domain, the Mean Value Theorem states that $g_p(\Delta) = \Delta g'_p(\xi) + g_p(0) = \Delta g'_p(\xi)$ for some $\xi \in [0, \Delta]$. However, $g'_p(\xi) \geq 0$ for all $\xi \in [0, \Delta]$, so necessarily $g_p(\Delta) \geq 0$.

Finally, it is clear that $g_p(\Delta) \geq 0$ implies $\text{kl}(p||q) \geq f(p,q)$ if $q = p + \Delta$, and that each point in the set $\{(p,q) : 0 < p \leq q < 1\}$ is contained in the domain of g.

□

Theorem (3.2)**.** *Let* $f(p,q) := q - \sqrt{2qp - p^2}$.

$$kl(p||q) \geq f(p,q) \qquad 0 < p \leq q < 1$$

Proof. Let $g_p(\Delta) = \text{kl}(p||p+\Delta) - f(p, p+\Delta)$, for $\Delta \in [0, 1-p)$. $g_p(0) = 0$, since $\text{kl}(p||p) = f(p,p) = 0$. We first show $g'_p(\Delta) \geq 0$:

$$\begin{aligned} g'_p(\Delta) &= \frac{1-p}{1-p-\Delta} - \frac{p}{p+\Delta} + \frac{p}{\sqrt{p^2+2p\Delta}} - 1 = 1 + \frac{\Delta}{1-p-\Delta} - \frac{p}{p+\Delta} + \frac{p}{\sqrt{p^2+2p\Delta}} - 1 \\ &\geq -\frac{p}{p+\Delta} + \frac{p}{\sqrt{p^2+2p\Delta}} = -\frac{p}{\sqrt{(p+\Delta)^2}} + \frac{p}{\sqrt{(p+\Delta)^2-\Delta^2}} \geq 0 \end{aligned}$$

And the rest of the proof follows exactly the end of the proof of Theorem 3.1.

□

B DERIVATION OF THE DIRECT OPTIMIZATION OF THE BOUND WITH IMPLICIT DIFFERENTIATION

Surrogate training objectives are necessary to optimize a function without known gradient, but this is not the case with Maurer's bound:

$$L(Q) \leq \text{kl}^{-1}\left(\hat{L}_S(Q), \frac{\text{KL}(Q||Q^0) + \log(\frac{2\sqrt{n}}{\delta})}{n}\right) \tag{10}$$

Let q be the bound on $L(Q)$ given in (1), for an empirical loss $p := \hat{L}_S(Q)$ and a right-hand term K. Bound q is the solution of the following equation:

$$\text{kl}(p||q) = K \qquad 0 < p \leq q < 1, K \geq 0$$

Let us define a function $f : \mathbb{R}^3 \longrightarrow \mathbb{R}$ that analyzes the space of solutions:

$$f(p, K, q) := \text{kl}(p||q) - K$$

Since f is continuously differentiable and invertible in its domain $\{(p, K, q) : 0 < p \leq q < 1, K \geq 0\}$, the Implicit Function Theorem states that there exists a continuously differentiable function g such that $f(p, K, g(p, K)) = 0$. The gradient of this implicit function can be computed with:

$$\nabla g = \left[\tfrac{\partial g}{\partial p}, \tfrac{\partial g}{\partial K}\right] = -\left(\frac{\partial f}{\partial q}\right)^{-1}\left[\tfrac{\partial f}{\partial p}, \tfrac{\partial f}{\partial K}\right]$$

Using the gradient descent algorithm to minimize q in the parametric space of Q means gradually updating Q according to the gradient of q with respect to Q. Let Θ be the parametric space of the problem we are trying to solve by gradient descent. In other words, each particular posterior Q is represented by a particular θ in Θ. Therefore the optimization of the bound with respect to Q demands the computation of the gradients with respect to θ by means of the chain rule:

$$\begin{aligned} \nabla_\theta g(p(\theta), K(\theta)) &= \nabla_\theta p(\theta)\frac{\partial g}{\partial p} + \nabla_\theta K(\theta)\frac{\partial g}{\partial K} \\ = -\left(\nabla_\theta p\frac{\partial f}{\partial p} + \nabla_\theta K\frac{\partial f}{\partial K}\right)\left(\frac{\partial f}{\partial q}\right)^{-1} &= -\xi\left(\nabla_\theta p\left(\log\frac{p}{q} - \log\frac{1-p}{1-q}\right) - \nabla_\theta K\right) \\ &= \xi\left(\left(\log\frac{1-p}{1-q} - \log\frac{p}{q}\right)\nabla_\theta p + \nabla_\theta K\right) \end{aligned} \tag{11}$$

where all the gradients are computed w.r.t θ and

$$\xi = \left(\frac{1-p}{1-q} - \frac{p}{q}\right)^{-1}$$

C REPRODUCIBILITY

Code, experiment outputs and instructions to reproduce the experiments are available at `github.com/Diegogpcm/pacbayesgradients`.