
PAC-Bayes-Chernoff bounds for unbounded losses

Ioar Casado

Machine Learning Group

Basque Center for Applied Mathematics (BCAM)

icasado@bcamath.org

Luis A. Ortega

Machine Learning Group

Computer Science Dept. - EPS.

Universidad Autónoma de Madrid

luis.ortega@uam.es

Aritz Pérez

Machine Learning Group

Basque Center for Applied Mathematics (BCAM)

aperez@bcamath.org

Andrés R. Masegosa

Department of Computer Science

Aalborg University

arma@cs.aau.dk

Abstract

We introduce a new PAC-Bayes oracle bound for unbounded losses that extends Cramér-Chernoff bounds to the PAC-Bayesian setting. The proof technique relies on controlling the tails of certain random variables involving the Cramér transform of the loss. Our approach naturally leverages properties of Cramér-Chernoff bounds, such as exact optimization of the free parameter in many PAC-Bayes bounds. We highlight several applications of the main theorem. Firstly, we show that our bound recovers and generalizes previous results. Additionally, our approach allows working with richer assumptions that result in more informative and potentially tighter bounds. In this direction, we provide a general bound under a new *model-dependent* assumption from which we obtain bounds based on parameter norms and log-Sobolev inequalities. Notably, many of these bounds can be minimized to obtain distributions beyond the Gibbs posterior and provide novel theoretical coverage to existing regularization techniques.

1 Introduction

PAC-Bayes theory provides powerful tools to analyze the generalization performance of randomized learning algorithms—for an introduction to the subject see the recent surveys of [Guedj \(2019\)](#); [Alquier \(2024\)](#) and [Hellström et al. \(2023\)](#). Let Θ be a model class parametrized by $\theta \in \mathbb{R}^d$. With a small abuse of notation, $\theta \in \Theta$ will represent both a model and its parameters. Instead of learning a single model, we consider the set of probability measures over our class, $\mathcal{M}_1(\Theta)$, and aim to find the optimal distribution $\rho^* \in \mathcal{M}_1(\Theta)$. This approach is more robust than finding the single best model in Θ , and is tightly related to Bayesian and ensemble methods. The learning algorithm infers this distribution from a sequence of n training data points $D = \{\mathbf{x}_i\}_{i=1}^n$, which are assumed to be i.i.d. sampled from an unknown distribution $\nu(\mathbf{x})$ with support in $\mathcal{X} \subseteq \mathbb{R}^k$.

Given a *loss function* $\ell : \Theta \times \mathcal{X} \rightarrow \mathbb{R}_+$, bounding the gap between the *population risk* $L(\theta) := \mathbb{E}_\nu[\ell(\theta, X)]$ and the *empirical risk* $\hat{L}(\theta, D) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, \mathbf{x}_i)$ of individual models is the standard approach in statistical learning theory. In contrast, PAC-Bayes theory provides high-probability bounds over the *population Gibbs risk* $\mathbb{E}_\rho[L(\theta)]$ in terms of the *empirical Gibbs risk* $\mathbb{E}_\rho[\hat{L}(\theta, D)]$ and an extra term measuring the dependence of ρ to the dataset D . This second term involves an information measure—usually the Kullback-Leibler divergence $KL(\rho|\pi)$ —between the data

dependent *posterior*¹ $\rho \in \mathcal{M}_1(\Theta)$ and a *prior* $\pi \in \mathcal{M}_1(\Theta)$, chosen before observing the data. These bounds hold simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$, hence minimizing them with respect to ρ provides an appealing approach to derive new learning algorithms with theoretically sound guarantees.

The foundational papers on PAC-Bayes theory (Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999; Seeger, 2002) worked with classification problems under bounded losses, usually the zero-one loss. This framework was significantly extended in Catoni (2007), who introduced some of the first bounds for unbounded losses. McAllester's bound (McAllester, 2003) is one of the most representative results for bounded losses: for any $\pi \in \mathcal{M}_1(\Theta)$ independent of D and every $\delta \in (0, 1)$, we have

$$\mathbb{E}_\rho[L(\theta)] \leq \mathbb{E}_\rho[\hat{L}(D, \theta)] + \sqrt{\frac{KL(\rho|\pi) + \log \frac{2\sqrt{n}}{\delta}}{2n}}, \quad (1)$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$, where the above inequality holds with probability no less than $1 - \delta$ over the choice of $D \sim \nu^n$. Another significant example under the bounded loss assumption is the Langford-Seeger-Maurer bound —after (Langford and Seeger, 2001; Seeger, 2002; Maurer, 2004)—. Under the same conditions as above,

$$kl\left(\mathbb{E}_\rho[\hat{L}(D, \theta)], \mathbb{E}_\rho[L(\theta)]\right) \leq \frac{KL(\rho|\pi) + \log \frac{2\sqrt{n}}{\delta}}{n}, \quad (2)$$

where kl is the so-called binary- kl distance, defined as $kl(a, b) := a \ln \frac{a}{b} + (1 - a) \ln \frac{1-a}{1-b}$.

These bounds illustrate typical trade-offs in PAC-Bayes theory. In (1), the relation between the empirical and the population risk is easy to interpret because the expected loss is bounded by the empirical loss plus a complexity term. More crucially, the right-hand side of the bound can be directly minimized with respect to the posterior ρ , resulting in the Gibbs posterior (Guedj, 2019). Conversely, (2) is known to be tighter than (1), but it is not straightforward to minimize because it requires inverting $kl(\mathbb{E}_\rho[\hat{L}(D, \theta)], \cdot)$ —several techniques deal with this issue (Thiemann et al., 2017; Reeb et al., 2018)—.

With the trade-offs among explainability, tightness, and generality in mind, the PAC-Bayes community has come up with novel bounds with applications in virtually every area of machine learning, ranging from the study of particular algorithms —linear regression (Alquier and Lounici, 2011; Germain et al., 2016), matrix factorization (Alquier and Guedj, 2017), kernel PCA (Haddouche et al., 2020), ensembles (Masegosa et al., 2020; Wu et al., 2021; Ortega et al., 2022) or Bayesian inference (Germain et al., 2016; Masegosa, 2020)— and generic versions of PAC-Bayes theorems (Bégin et al., 2016; Rivasplata et al., 2020) to the study of the generalization of deep neural networks (Dziugaitė and Roy, 2017; Rivasplata et al., 2019).

Such applications often require relaxing the assumptions of the classical PAC-Bayesian framework, such as data-independent priors, i.i.d. data or bounded losses. Here we are interested in the case of PAC-Bayes bounds for unbounded losses —such as the squared error loss and the log-loss—.

PAC-Bayes bounds for unbounded losses

The main challenge of working with unbounded losses is that one needs to deal with an exponential moment term which cannot be easily bounded without specific assumptions about the tails of the loss. The following theorems, fundamental starting points for many works on PAC-Bayes theory over unbounded losses, illustrate this point.

Theorem 1 ((Alquier et al., 2016; Germain et al., 2016)). *Let $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$ and any $\lambda > 0$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\theta)] \leq \mathbb{E}_\rho[\hat{L}(D, \theta)] + \frac{1}{\lambda n} \left[KL(\rho|\pi) + \log \frac{f_{\pi, \nu}(\lambda)}{\delta} \right],$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$. Here $f_{\pi, \nu}(\lambda) := \mathbb{E}_\pi \mathbb{E}_{\nu^n} \left[e^{\lambda n (L(\theta) - \hat{L}(D, \theta))} \right]$.

¹In PAC-Bayes theory the term *posterior* is used to emphasize that $\rho \in \mathcal{M}_1(\Theta)$ depends on training data, and does not necessarily refer to the *Bayesian posterior*. We refer the reader to Germain et al. (2016) for a discussion of the relation between PAC-Bayesian and Bayesian methods.

This is an *oracle* bound, because $f_{\pi,\nu}(\lambda)$ depends on the data generating distribution ν^n . To obtain empirical bounds from the theorem above, the exponential term $f_{\pi,\nu}$ is usually bounded by making the appropriate assumptions on the tails of the loss, such as the Hoeffding assumption (Alquier et al., 2016), sub-Gaussianity (Alquier and Guedj, 2018; Xu and Raginsky, 2017), sub-gamma (Germain et al., 2016) or sub-exponential (Catoni, 2004). See Section 5 of Alquier (2024) for an overview.

Many of these assumptions are generalized by the notion that the *cumulant generating function* (CGF) of the (centered) loss, $\Lambda_\theta(\lambda)$, exists and is bounded (Banerjee and Montúfar, 2021; Rodríguez-Gálvez et al., 2024). Remember that we say $\Lambda_\theta(\lambda)$ exists if it is bounded in some interval $[0, b)$ with $b > 0$.

Definition 2 (Bounded CGF). A loss function ℓ has bounded CGF if for all $\theta \in \Theta$, there is a convex and continuously differentiable function $\psi : [0, b) \rightarrow \mathbb{R}_+$ such that $\psi(0) = \psi'(0) = 0$ and

$$\Lambda_\theta(\lambda) := \log \mathbb{E}_\nu \left[e^{\lambda(L(\theta) - \ell(\mathbf{x}, \theta))} \right] \leq \psi(\lambda) \text{ for all } \lambda \in [0, b). \quad (3)$$

We will say that a loss function ℓ is ψ -bounded if it satisfies the above assumption under the function ψ . In this setup, Banerjee and Montúfar (2021) obtain the following PAC-Bayes bound under the Bounded CGF assumption:

Theorem 3. *Let ℓ be a loss with ψ -bounded CGF and $\pi \in \mathcal{M}_1(\Theta)$ any prior independent of D . Then, for any $\delta \in (0, 1)$ and any $\lambda \in (0, b)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\theta)] \leq \mathbb{E}_\rho[\hat{L}(D, \theta)] + \frac{KL(\rho|\pi) + \log \frac{1}{\delta}}{\lambda n} + \frac{\psi(\lambda)}{\lambda},$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Theorems 1 and 3 illustrate a pervasive problem of many PAC-Bayes bounds for unbounded losses: they often depend on a free parameter $\lambda > 0$ —see for example (Alquier et al., 2016; Hellström and Durisi, 2020; Guedj and Pujol, 2021; Banerjee and Montúfar, 2021; Haddouche et al., 2021)—. The choice of this free parameter is crucial for the tightness of the bounds, and it cannot be directly optimized because its choice is prior to the draw of data, while the optimal λ would be data-dependent. The standard approach is to optimize λ over a grid using union-bound arguments, but the resulting λ is not guaranteed to be optimal. Seldin et al. (2012) carefully design the grid so that the optimal λ is inside its range, but their work only deals with bounded losses. See the discussion in Section 2.1.4 of Alquier (2024) for an overview.

Recently, Rodríguez-Gálvez et al. (2024) improved the union-bound approach for the bounded CGF scenario using a convenient partition of the event space. However, their optimization remains approximate. Hellström and Durisi (2021) could circumvent this problem for the particular case of sub-Gaussian losses, but the *exact* optimization of λ for the general case remains an open question. A positive result in this direction would lift the restriction of having to optimize λ over restricted grids or using more complex approaches.

The use of the bounded CGF assumption entails another potential drawback: both uniform control of the CGF — $\Lambda_\theta(\lambda) \leq \psi(\lambda)$ for every $\theta \in \Theta$ — and integrability assumptions as in Alquier et al. (2016) — $\mathbb{E}_\pi[\Lambda_\theta(\lambda)] \leq C$ for some constant C —, necessarily drop information about the different concentration properties of individual models. For example, Masegosa and Ortega (2023) show that within the class of models defined by the weights of common neural networks, the behavior of their CGFs —and hence of their Cramér transforms, which control generalization error via Cramér-Chernoff bound— varies significantly. As Figure 1 shows, uniformly bounding the CGFs can result in loose *worst-case* bounds that ignore the properties of the models that interest us most.

The bounds of Seldin et al. (2012) and Haddouche et al. (2021) partially address this issue including average variances and model-dependent range functions to their bounds, which account for the fact that different models have different CGFs. However, the former only applies to bounded losses, while the latter cannot be exploited to obtain better posteriors because their model-dependent function only impacts the prior. Outside the PAC-Bayes framework, Jiao et al. (2017) generalized the bounded CGF assumption, but their result only applies to finite model sets.

Exploiting these differences among models within the same class could be an important line of research in PAC-Bayes theory. This approach would provide more informative generalization bounds, paving also the way for the design of better posteriors. This is one of the main appeals of our work.

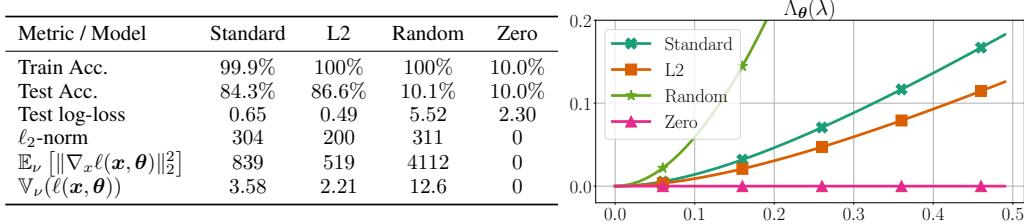


Figure 1: **Models with very different CGFs coexist within the same model class.** On the left, we display several metrics for InceptionV3 models trained on CIFAR10 without regularization (**Standard**) and with L2 regularization (**L2**). **Random** refers to a model learned over randomly reshuffled labels and **Zero** refers to a model where all the weights are equal to zero. For each model, the metrics include train and test accuracy, test log-loss, ℓ_2 -norm of the parameters of the model, the variance of the log-loss function, denoted $\mathbb{V}_\nu (\ell(\mathbf{x}, \boldsymbol{\theta}))$, and the expected norm of the input-gradients, denoted $\mathbb{E}_\nu [\|\nabla_x \ell(\mathbf{x}, \boldsymbol{\theta})\|_2^2]$. On the right, we display the estimated CGFs of each model, following [Masegosa and Ortega \(2023\)](#). Note how models with smaller variance $\mathbb{V}(\ell(\mathbf{x}, \boldsymbol{\theta}))$, ℓ_2 -norm or input-gradient norm $\mathbb{E}_\nu [\|\nabla_x \ell(\mathbf{x}, \boldsymbol{\theta})\|_2^2]$ have smaller CGFs. Bounds derived from Theorem 7 naturally exploit these differences. Experimental details in Appendix C.

Overview and Contributions

The main contribution of this paper is Theorem 7, a novel (oracle) PAC-Bayes bound for unbounded losses which extends the classic Cramér-Chernoff bound to the PAC-Bayesian setting. The theorem is introduced in Section 3, while Section 2 contains the necessary prerequisites. As far as we know, the proof technique based on Lemma 6 is also novel and can be of independent interest.

We discuss the applications of our main theorem in sections 4 and 5. First, we show that our bound allows *exact* optimization of the free parameter λ incurring in a $\log n$ penalty without resorting to union-bound approaches. In the case of bounded CGFs, we recover bounds which are optimal up to that logarithmic term. We also show that versions of many well-known bounds —such as the Langford-Seeger bound— can be recovered from Theorem 7.

In Section 5, we show how our main theorem provides a general framework to deal with richer assumptions that result in novel, more informative, and potentially tighter bounds. In Theorem 11, we generalize the bounded CGF assumption so that there is a different bounding function, $\psi(\boldsymbol{\theta}, \cdot)$, for the CGF of each model. We illustrate this idea in three cases: generalized sub-Gaussian losses, norm-based regularization, and input-gradient regularization based on log-Sobolev inequalities. Remarkably, the bounds in Section 5 can be minimized with respect to $\rho \in \mathcal{M}_1(\Theta)$, resulting in optimal posteriors beyond Gibbs' and opening the door to the design of novel algorithms.

Appendix A contains the proofs not included in the paper.

2 Preliminaries

In this section we introduce the necessary prerequisites in order to prove our main theorem. Its proof relies in controlling the concentration properties of each model in Θ using their Cramér transform.

Definition 4. Let $I \subseteq \mathbb{R}$ be an interval and $f : I \rightarrow \mathbb{R}$ a convex function. The *Legendre transform* of f is defined as

$$f^*(a) := \sup_{\lambda \in I} \{\lambda a - f(\lambda)\}, \quad \forall a \in \mathbb{R}. \quad (4)$$

Following this definition, the Legendre transform of the CGF of a model $\boldsymbol{\theta} \in \Theta$ is known as its *Cramér transform*:

$$\Lambda_{\boldsymbol{\theta}}^*(a) := \sup_{\lambda \in [0, b)} \{\lambda a - \Lambda_{\boldsymbol{\theta}}(\lambda)\}, \quad \forall a \in \mathbb{R}. \quad (5)$$

Throughout the paper, we will use the abbreviation $gen(\boldsymbol{\theta}, D)$ to denote the *generalization gap* $L(\boldsymbol{\theta}) - \hat{L}(D, \boldsymbol{\theta})$, in order to make the notation more compact. Cramér's transform provides non-

asymptotic bounds on the right tail of $gen(\boldsymbol{\theta}, D)$ via Cramér-Chernoff's theorem —see Section 2.2 of [Boucheron et al. \(2013\)](#) or Section 2.2 of [Dembo and Zeitouni \(2009\)](#)—.

Theorem 5 (Cramér-Chernoff). *For any $\boldsymbol{\theta} \in \Theta$ and $a \in \mathbb{R}$,*

$$\mathbb{P}_{\nu^n}(gen(\boldsymbol{\theta}, D) \geq a) \leq e^{-n\Lambda_{\boldsymbol{\theta}}^*(a)}. \quad (6)$$

Furthermore, the inequality is asymptotically tight up to exponential factors.

Importantly, Cramér-Chernoff's bound can be inverted to establish high-probability generalization bounds: for any $\delta \in (0, 1)$,

$$\mathbb{P}_{\nu^n}(gen(\boldsymbol{\theta}, D) \leq (\Lambda_{\boldsymbol{\theta}}^*)^{-1}\left(\frac{1}{n} \log \frac{1}{\delta}\right)) \geq 1 - \delta. \quad (7)$$

where $(\Lambda_{\boldsymbol{\theta}}^*)^{-1}$ is the inverse of the Cramér transform. We cannot directly use Theorem 5 to obtain PAC-Bayes bounds, because we need a bound which is uniform for every model. The following lemma will be our main technical tool for this purpose.

Lemma 6. *For any $\boldsymbol{\theta} \in \Theta$ and $c \geq 0$, we have*

$$\mathbb{P}_{\nu^n}(n\Lambda_{\boldsymbol{\theta}}^*(gen(\boldsymbol{\theta}, D)) \geq c) \leq \mathbb{P}_{X \sim \exp(1)}(X \geq c).$$

This way of controlling the survival function of $\Lambda_{\boldsymbol{\theta}}^*(gen(\boldsymbol{\theta}, D))$ for every $\boldsymbol{\theta} \in \Theta$ will allow us to bound an exponential moment term in our main theorem.

3 PAC-Bayes-Chernoff bound

As we hinted above, instead of directly introducing boundedness conditions on the loss or its CGF, we stay in the realm of oracle bounds, aiming to provide a flexible starting point for diverse applications. A key element of our theoretical approach is the averaging of the CGFs with respect to a posterior distribution. For any posterior distribution $\rho \in \mathcal{M}_1(\Theta)$, we may consider the expectation of the CGF, $\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]$, as in [Jiao et al. \(2017\)](#). In analogy to the standard definition, we define the *Cramér transform of a posterior distribution* ρ as the following function:

$$\Lambda_{\rho}^*(a) := \sup_{\lambda \in [0, b)} \{\lambda a - \mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]\}, \quad a \in \mathbb{R}. \quad (8)$$

Since the CGFs $\Lambda_{\boldsymbol{\theta}}(\lambda)$ are convex and continuously differentiable with respect to λ , their expectation $\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]$ retains the same properties. Hence according to Lemma 2.4 in [Boucheron et al. \(2013\)](#), the (generalized) inverse of Λ_{ρ}^* exists and can be written as

$$(\Lambda_{\rho}^*)^{-1}(s) = \inf_{\lambda \in [0, b)} \left\{ \frac{s + \mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]}{\lambda} \right\}. \quad (9)$$

With these definitions in hand, we are ready to introduce our main result, a novel (oracle) PAC-Bayes bound for unbounded losses:

Theorem 7 (PAC-Bayes-Chernoff bound). *Let $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_{\rho}[L(\boldsymbol{\theta})] \leq \mathbb{E}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \inf_{\lambda \in [0, b)} \left\{ \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]}{\lambda} \right\}$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

The above result gives an oracle PAC-Bayes analogue to the Cramér-Chernoff's bound of equation (7), because the second term of the right hand side of the inequality corresponds with the inverse of Λ_{ρ}^* , as described in equation (9). Theorem 7 shows that bounds like the one given in Theorem 3 can hold simultaneously for every $\lambda > 0$ —and hence optimized with respect to λ — by paying a $\log(n)$ penalty, virtually the same as if we were optimizing over a uniform grid of size n using union-bound arguments. Although the dependence in union bound arguments can be improved to $\log(\log(n))$ ([Alquier, 2024](#)), our optimization is exact.

Furthermore, Theorem 7 also shows that the posterior minimizing its right-hand-side is involved in a three-way trade-off: firstly, ρ must explain the training data due to $\mathbb{E}_\rho[\hat{L}(D, \theta)]$; secondly, it must be close to the prior due to the KL term $KL(\rho|\pi)$; and lastly, it must place its density in models with a lower CGF due to the $\mathbb{E}_\rho[\Lambda_\theta(\lambda)]$ term. We note that the first two elements are standard on most PAC-Bayesian bounds, but the role of the CGF in defining an optimal posterior ρ is novel compared to previous bounds. We explore the implications of this fact in Section 5.

4 Relation with previous bounds

We first relate Theorem 7 to previous bounds. As a first application, we show how some well-known PAC-Bayes bounds can be recovered from ours. When the distribution of the loss is known, we can often compute its Cramér transform. This happens to be the case with the zero-one loss, where we recover Langford-Seeger's bound (Seeger, 2002, Theorem 1).

Corollary 8. *Let ℓ be the $0 - 1$ loss and $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$kl\left(\mathbb{E}_\rho[\hat{L}(\theta, D)], \mathbb{E}_\rho[L(\theta)]\right) \leq \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1},$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

The dependence on n in Corollary 8 was further improved by Maurer (2004). However, our version is enough to illustrate the role played by Crámer transforms in obtaining tight PAC-Bayes bounds, which is a recent line of work by Foong et al. (2021) and Hellström and Guedj (2024).

When the loss is of bounded CGF —recall Definition 2—, Theorem 7 results in a generalization of Theorem 3, where the bound holds simultaneously for every $\lambda \in (0, b)$ with a $\log n$ penalty.

Corollary 9. *Let ℓ be a loss function with ψ -bounded CGF. Let $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\theta)] \leq \mathbb{E}_\rho[\hat{L}(D, \theta)] + \inf_{\lambda \in [0, b)} \left\{ \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{\lambda(n-1)} + \frac{\psi(\lambda)}{\lambda} \right\},$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Proof. Directly follows from the definition of ψ -bounded CGF and Theorem 7. \square

Corollary 9 is the first PAC-Bayes bound that allows exact optimization of the free parameter $\lambda \geq 0$ for the general case of losses with bounded CGF without resorting to union-bound approaches (Seldin et al., 2012; Rodríguez-Gálvez et al., 2024), which cannot guarantee an exact minimization. Although in the particular case of sub-Gaussian losses the $\log n$ penalty is worse than that in Hellström and Durisi (2021), when $KL(\rho|\pi) \geq \frac{\sqrt{6n/e}}{\pi} - 1$, Corollary 9 is tighter than the general Theorem 14 in Rodríguez-Gálvez et al. (2024). We instantiate Corollary 9 in the case of sub-Gaussian and sub-gamma losses in the Appendix B.

5 PAC-Bayes bounds under model-dependent assumptions

As discussed in the introduction, most boundedness conditions used to bound the exponential moment term in Theorem 7 discard information about the statistical properties of individual models. In this section, we show that the structure of Theorem 7 allows for a more fine-grained control of the CGFs, resulting in potentially tighter and more informative bounds. We start by generalizing Definition 2.

Definition 10 (Model-dependent bounded CGF). A loss function ℓ has model-dependent bounded CGF if for each $\theta \in \Theta$, there is a convex and continuously differentiable function $\psi(\theta, \lambda)$ such that $\psi(\theta, 0) = \psi'(\theta, 0) = 0$ and for any $\lambda \in [0, b)$,

$$\Lambda_\theta(\lambda) := \log \mathbb{E}_\nu \left[e^{\lambda(L(\theta) - \ell(\mathbf{x}, \theta))} \right] \leq \psi(\theta, \lambda). \quad (10)$$

As motivated in Figure 1 and in opposition to Definition 2, this new condition acknowledges the possibility of having different bounding functions, $\psi(\boldsymbol{\theta}, \lambda)$, for each $\Lambda_{\boldsymbol{\theta}}(\lambda)$.

Using this definition, we can easily exploit Theorem 7 to derive the following bound:

Theorem 11. *Let ℓ be a loss function satisfying Definition 10. Let $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_{\rho}[L(\boldsymbol{\theta})] \leq \mathbb{E}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \inf_{\lambda \in [0, b)} \left\{ \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_{\rho}[\psi(\boldsymbol{\theta}, \lambda)]}{\lambda} \right\}, \quad (11)$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Proof. The proof is analogue to that of Corollary 9. \square

This theorem can be understood as a PAC-Bayesian version of Theorem 2 in Jiao et al. (2017), where we allow infinite model classes. Note that if we tried to exploit this model-dependent CGF assumption on other oracle bounds, as the one shown in Theorem 1, we would end with an empirical bound where the $\psi(\boldsymbol{\theta}, \lambda)$ term would be *exponentially averaged* by the prior, $\ln \mathbb{E}_{\pi}[e^{\psi(\boldsymbol{\theta}, \lambda)}]$, instead of having the more amenable term $\mathbb{E}_{\rho}[\psi(\boldsymbol{\theta}, \lambda)]$, which directly impacts on the choice of the optimal posterior ρ^* .

As discussed in Section 3, the posterior distribution in Theorem 11 is involved in a three-way trade-off which has the potential to result in tighter bounds and the design of better posteriors. It is worth noting that the posterior minimizing the bound in Theorem 11 is not the standard Gibbs posterior.

Proposition 12. *If we fix some $\lambda > 0$, the bound in Theorem 11 can be minimized with respect to $\rho \in \mathcal{M}_1(\Theta)$. The optimal posterior is*

$$\rho^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp \left\{ -(n-1)\lambda \hat{L}(D, \boldsymbol{\theta}) - (n-1)\psi(\boldsymbol{\theta}, \lambda) \right\}. \quad (12)$$

Observe that under the posterior in Proposition 12, the *maximum a posteriori* (MAP) estimate is

$$\boldsymbol{\theta}_{\text{MAP}} = \arg \min_{\boldsymbol{\theta} \in \Theta} \left\{ \hat{L}(D, \boldsymbol{\theta}) + \frac{1}{\lambda} \psi(\boldsymbol{\theta}, \lambda) - \frac{1}{\lambda(n-1)} \ln \pi(\boldsymbol{\theta}) \right\}.$$

As we will illustrate below, the extra term, $\psi(\boldsymbol{\theta}, \lambda)$, can often be understood as a regularizer. In what remains, we exemplify the general recipe of Theorem 11 in several cases. To start providing concrete intuition, we instantiate Corollary 11 in the case of sub-Gaussian losses.

Generalized sub-Gaussian losses

It is well known that if X is a σ^2 -sub-Gaussian random variable, we have $\mathbb{V}(X) \leq \sigma^2$ (Arbel et al., 2020). In many cases, it might not be reasonable to bound $\mathbb{V}_{\nu}(\ell(\boldsymbol{\theta}, X)) \leq \sigma^2$ for every $\boldsymbol{\theta} \in \Theta$, because the variance of the loss function highly depends on the particular model, as illustrated in Figure 1. This is where Corollary 11 comes into play: we may assume that $\ell(\boldsymbol{\theta}, X)$ is $\sigma(\boldsymbol{\theta})^2$ -sub-Gaussian for each $\boldsymbol{\theta} \in \Theta$. In this case the variance proxy $\sigma(\boldsymbol{\theta})^2$ is specific for each model, leading to the following bound:

Corollary 13. *Assume the loss $\ell(\boldsymbol{\theta}, X)$ is $\sigma^2(\boldsymbol{\theta})$ -sub-Gaussian. Let $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_{\rho}[L(\boldsymbol{\theta})] \leq \mathbb{E}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{2\mathbb{E}_{\rho}[\sigma(\boldsymbol{\theta})^2] \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1}}, \quad (13)$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Proof. Use Theorem 11 and the fact that $\psi(\boldsymbol{\theta}, \lambda) = \frac{\lambda^2 \sigma^2(\boldsymbol{\theta})}{2}$. Then optimize λ . \square

This result generalizes sub-Gaussian PAC-Bayes bounds —Corollary 2 in Hellström and Durisi (2021) or Corollary 19— and shows that posteriors favoring models with small variance-proxy, $\sigma^2(\boldsymbol{\theta})$, generalize better. It is, therefore, potentially tighter than previous results, because the σ^2

factor in standard sub-Gaussian bounds is a *worst-case constant*, while Corollary 13 exploits the fact that some models have much smaller variance-proxy than others.

Analogous bounds can be straightforwardly derived for generalized sub-gamma or sub-exponential losses, but Theorem 11 is not limited to explicit tail assumptions on the loss. The following subsections explore model-dependent assumptions on $\Lambda_{\theta}(\lambda)$ that result in novel PAC-Bayes bounds involving well-known regularization techniques.

L2 regularization

We now introduce bounds based on the norm of the model parameters. For that purpose we use the following standard assumption in machine learning (Li and Orabona, 2019).

Assumption 1 (Parameter Lipschitz). The loss function $\ell(\mathbf{x}, \theta)$ is M -Lipschitz with respect to θ , that is, for any $\theta \in \Theta$ and any $\mathbf{x} \in \mathcal{X}$ we have $\|\nabla_{\theta} \ell(\mathbf{x}, \theta)\|_2^2 \leq M$.

If the model class is parametrized in such a way that the model with null parameter vector has null variance —i.e., $\mathbb{V}_{\nu}(\ell(\mathbf{x}, 0)) = 0$, which is the case of a neural net with null weights—, then, as shown in Masegosa and Ortega (2023), we can derive the following model-dependent bound for the CGF:

$$\Lambda_{\theta}(\lambda) \leq 2M\lambda^2\|\theta\|_2^2. \quad (14)$$

This case further illustrates the idea we motivated in the introduction: bounding the CGFs with model-dependent proxies for generalization. Figure 1 illustrates how models with smaller norm has increasingly smaller CGFs.

Using Equation (14) we obtain the following generalization bound penalizing model's L2-norm:

Corollary 14. *If Assumption 1 holds, then for any prior distribution $\pi \in \mathcal{M}_1(\Theta)$ independent of D and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws $D \sim \nu^n$,*

$$\mathbb{E}_{\rho}[L(\theta)] \leq \mathbb{E}_{\rho}[\hat{L}(D, \theta)] + \sqrt{2M\mathbb{E}_{\rho}[\|\theta\|_2^2] \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1}}, \quad (15)$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Proof. The result follows from using $\psi(\theta, \lambda) = 2M\lambda^2\|\theta\|_2^2$ in Theorem 11 and optimizing λ . \square

Corollary 14 shows that models with smaller parameter norms generalize better, and provides PAC-Bayesian certificates for norm-based regularization. Many other PAC-Bayesian bounds contains different kind of parameter norms (Germain et al., 2009, 2016; Neyshabur et al., 2017), but their parameter norm term is always introduced through the prior —e.g., using a zero-centered Gaussian prior distribution—. The novelty here is that this parameter norm term appears independently of the prior as a result of Theorem 11.

According to Proposition 12, the MAP estimate of the optimal posterior is, $\theta_{\text{MAP}} = \arg \min_{\theta} \{ \hat{L}(D, \theta) + \frac{2M}{\lambda} \|\theta\|_2^2 - \frac{1}{\lambda(n-1)} \ln \pi(\theta) \}$, which is the result of L2-regularization with tradeoff parameter $\frac{2M}{\lambda}$.

Gradient-based regularization

We finish this section providing novel bounds based on log-Sobolev inequalities that include a gradient term penalizing the sensitivity of models to small changes in the input data. First, let us simplify the notation for this section. We use $\|\nabla_{\mathbf{x}} \ell\|_2^2 := \mathbb{E}_{\nu} \|\nabla_{\mathbf{x}} \ell(\mathbf{x}, \theta)\|_2^2$ and $\|\hat{\nabla}_{\mathbf{x}} \ell\|_2^2 := \frac{1}{n} \sum_{i=1}^n \|\nabla_{\mathbf{x}} \ell(\mathbf{x}_i, \theta)\|_2^2$ to denote the expected and empirical (squared) gradient norms of the loss.

Including a gradient-based penalization is the idea behind input-gradient regularization (Varga et al., 2017), which minimizes an objective function of the form $\hat{L}(D, \theta) + \frac{1}{k} \|\hat{\nabla}_{\mathbf{x}} \ell\|_2^2$, where $k > 0$ is a trade-off parameter. This approach is often used to make models more robust against disturbances in input data and adversarial attacks (Ross and Doshi-Velez, 2018; Finlay and Oberman, 2021).

We make the connection between PAC-Bayes bounds and gradient norms assuming that ν and ℓ satisfy certain log-Sobolev inequality (Chafaï, 2004).

Assumption 2 (log-Sobolev). For any $\theta \in \Theta$ and any $\lambda > 0$, we have

$$\Lambda_\theta(\lambda) \leq \frac{C}{2} \lambda^2 \|\nabla_{\mathbf{x}} \ell\|_2^2, \text{ for some } C > 0.$$

For example, Assumption 2 holds when ν is strictly uniformly log-concave, as shown in Corollary 2.1 of [Chafaï \(2004\)](#)—this case includes the Gaussian and Weibull densities ([Saumard and Wellner, 2014](#))—. We also try to empirically verify Assumption 2 for certain class of neural networks in Appendix C.2. Assumption 2 is going to be our model-dependent bound on the CGF. We first provide a bound for expected gradients.

Theorem 15. *If Assumption 2 is satisfied, then for any prior distribution $\pi \in \mathcal{M}_1(\Theta)$ independent of D and any $\delta_1 \in (0, 1)$, with probability at least $1 - \delta_1$ over draws $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\theta)] \leq \mathbb{E}_\rho[\hat{L}(D, \theta)] + \sqrt{2C\mathbb{E}_\rho[\|\nabla_{\mathbf{x}} \ell\|_2^2] \frac{KL(\rho||\pi) + \log \frac{n}{\delta_1}}{n-1}} \quad (16)$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Proof. Follows from the application Assumption 2 to Theorem 11 and the optimization of λ . \square

This bound shows that—under certain regularity conditions— posteriors favoring models with smaller expected gradients, $\|\nabla_{\mathbf{x}} \ell\|_2^2$, generalize better, which is the heuristic behind input-gradient regularization ([Varga et al., 2017](#)). However, Theorem 15 is still an oracle bound. We can obtain a new, fully empirical one if we concatenate Corollary 15 with a PAC-Bayes concentration bound for $\mathbb{E}_\rho[\|\nabla_{\mathbf{x}} \ell\|_2^2]$. This can be done if we assume that the loss is Lipschitz w.r.t. the input-data.

Assumption 3 (Input-data Lipschitz). For any $\theta \in \Theta$ and for any $\mathbf{x} \in \mathcal{X}$, we have $\|\nabla_{\mathbf{x}} \ell(\mathbf{x}, \theta)\|_2^2 \leq L$.

Assumption 3 is satisfied in standard deep neural network architectures, and the Lipschitz constant can be efficiently estimated ([Virmaux and Scaman, 2018; Fazlyab et al., 2019](#)).

Theorem 16. *If Assumptions 2 and 3 are satisfied, then for any prior distribution $\pi \in \mathcal{M}_1(\Theta)$ independent of D and any $\delta \in (0, 1)$; with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\theta)] \leq \mathbb{E}_\rho[\hat{L}(D, \theta)] + \sqrt{2C\mathbb{E}_\rho[\|\hat{\nabla}_{\mathbf{x}} \ell\|_2^2] \cdot K(\rho, \pi, n, \delta) + \sqrt{2}CL \cdot K(\rho, \pi, n, \delta)^{\frac{3}{2}}}$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$, where $K(\rho, \pi, n, \delta) := \frac{KL(\rho||\pi) + \log \frac{2n}{\delta}}{n-1}$.

With minimal modifications, the result above also holds under the assumption that on-average gradients are bounded, $\|\nabla_{\mathbf{x}} \ell\|_2^2 \leq g$, or when $\|\nabla_{\mathbf{x}} \ell(\theta, X)\|_2^2$ are sub-Gaussian.

Similar bounds with input-gradients were first introduced by [Gat et al. \(2022\)](#) under the assumption that the underlying distribution of data was a mixture of Gaussians plus a technical “per-label loss balance” assumption. However, Theorem 16 is, as far as we know, the first empirical PAC-Bayes bound for input-gradients. Furthermore, in contrast with [Gat et al. \(2022\)](#), our bounds optimize the free parameter λ and explicitly relate the gradients with the generalization ability of the posterior $\rho \in \mathcal{M}_1(\Theta)$. The recent work of [Haddouche et al. \(2024\)](#) also includes gradient terms in their bounds, but they are gradients with respect to the model parameter, not input-gradients.

As for the optimal posterior ρ^* , if we minimize the bound in Theorem 15 for a fixed $\lambda > 0$, the MAP estimate in Proposition 12 is $\theta_{\text{MAP}} = \arg \min_\theta \{\hat{L}(D, \theta) + \lambda \frac{C}{2} \|\nabla_{\mathbf{x}} \ell\|_2^2 - \frac{1}{\lambda(n-1)} \ln \pi(\theta)\}$, which is the result of input-gradient regularization with trade-off parameter $\frac{\lambda C}{2}$.

In conclusion, the approach suggested by Theorem 11 not only provides novel insights—in the form of tight bounds or PAC-Bayesian interpretations of previously known algorithms—, it also hints towards the design of new regularized learning algorithms with solid theoretical guarantees.

6 Conclusion

We derived a novel PAC-Bayes oracle bound using basic properties of the Cramér transform —Theorem 7—. In general, our work aligns with very recent literature ([Rodríguez-Gálvez et al.,](#)

2024; Hellström et al., 2023) that highlights the importance of Cramér transforms in the quest for tight PAC-Bayes bounds. This bound has the potential to be a stepping stone in the development of novel, tighter empirical PAC-Bayes bounds for unbounded losses. Firstly, because it allows exact optimization of the free parameter $\lambda > 0$ without the need for more convoluted union-bound approaches. But, more relevantly, because it allows the introduction of flexible, fine-grained, model-dependent assumptions for bounding the CGF —Theorem 11— which results in optimal distributions beyond Gibbs’ posterior. The importance and wide applicability of this result have been illustrated with three model-dependent assumptions: generalized sub-Gaussian losses, bounds based on parameter norms, and input-gradients based on log-Sobolev inequalities. In the last case we introduce PAC-Bayes bounds including empirical input-gradients norms.

Limitations and future work

A limitation of our approach is that we are implicitly assuming that $\ell(\boldsymbol{\theta}, X)$ is light-tailed (equivalently, sub-exponential), as in every Cramér-Chernoff bound. This is only partially true. Although there are specific studies for heavy-tailed losses (Alquier and Guedj, 2018; Holland, 2019; Haddouche and Guedj, 2023; Chugg et al., 2023), we avoid this limitation because we are only interested in the right tail of $\text{gen}(\boldsymbol{\theta}, D)$ —that is, $\mathbb{P}_\nu(L(\boldsymbol{\theta}) - \hat{L}(D, \boldsymbol{\theta}) \geq a)$ for $a \geq 0$ —. This is done by defining the CGF only for $\lambda \geq 0$. In this way, as we show in Remark 17, the finiteness of $\mathbb{E}_\rho[L(\boldsymbol{\theta})]$ guarantees the existence of $\mathbb{E}_\rho[\Lambda_\boldsymbol{\theta}(\lambda)]$ in Theorem 7. This approach is motivated by the fact that most state-of-the-art models lie in the interpolating regime, hence $\mathbb{P}_\nu(\hat{L}(D, \boldsymbol{\theta}) - L(\boldsymbol{\theta}) \geq a)$ for $a \geq 0$ has less practical importance. However, in the case where one is looking for two-tailed bounds, our work is restricted to sub-exponential losses. See the discussion on Section 4.2.1 of Zhang et al. (2024).

Given the generality of Theorem 7, the dependence of the logarithmic penalty term in the bound may be suboptimal in certain cases, as discussed in Section 4. Going beyond Lemma 6 and improving this dependence is one of the main open tasks.

Our results provide a systematic approach for integrating general model-dependent CGF bounds into PAC-Bayes theory, and can open exciting new research lines: they can provide a PAC-Bayesian extension of the work of Masegosa and Ortega (2023), which studies the effects of different regularization techniques and the role of invariant architectures, data augmentation and over-parametrization in generalization. Since our bounds can be applied to unbounded losses such as the log-loss, it could be interesting to apply them in a Bayesian setting (Germain et al., 2016) in order to study the relation between marginal likelihood and generalization (Lotfi et al., 2022).

Acknowledgments and disclosure of funding

IC acknowledges financial support from the Basque Government through the IKUR program. LO acknowledges financial support from project PID2022-139856NB-I00 funded by MCIN/ AEI/ 10.13039/501100011033/FEDER, UE and from the Autonomous Community of Madrid (ELLIS Unit Madrid). AP acknowledges financial support by the Basque Government through the BERC 2022-2025 program and Elkartek program (KK-2023/00038), and by the Ministry of Science and Innovation: BCAM Severo Ochoa accreditation CEX2021-001142-S/MICIN/AEI/10.13039/501100011033. AM acknowledges financial support from Grant PID2022-139293NB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe.

References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Pierre Alquier and Benjamin Guedj. An oracle inequality for quasi-Bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics*, 26:55–67, 2017.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Pierre Alquier and Karim Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127 – 145, 2011.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Julyan Arbel, Olivier Marchal, and Hien D Nguyen. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability and Statistics*, 24:39–55, 2020.

Pradeep Kr Banerjee and Guido Montúfar. Information complexity and generalization bounds. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 676–681. IEEE, 2021.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2016.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, NY, 2006.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer Science & Business Media, 2004.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 1277 of *IMS Lecture Notes-Monograph Series*. 2007.

Djalil Chafaï. Entropies, convexity, and functional inequalities, on Φ -entropies and Φ -sobolev inequalities. *Journal of Mathematics of Kyoto University*, 44(2):325–363, 2004.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *Journal of Machine Learning Research*, 24(372):1–61, 2023.

Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer Science & Business Media, 2009.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Paul Embrechts and Marius Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77:423–432, 2013.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of Lipschitz constants for Deep Neural Networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Chris Finlay and Adam M Oberman. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, 3:100017, 2021.

Andrew Foong, Wessel Bruinsma, David Burt, and Richard Turner. How tight can PAC-Bayes be in the small data regime? *Advances in Neural Information Processing Systems*, 34:4093–4105, 2021.

Nina Gantert, Steven Soojin Kim, and Kavita Ramanan. Cramér’s theorem is atypical. In Gail Letzter, Kristin Lauter, Erin Chambers, Nancy Flouroy, Julia Elisenda Grigsby, Carla Martin, Kathleen Ryan, and Konstantina Trivisa, editors, *Advances in the Mathematical Sciences*, pages 253–270, Cham, 2016. Springer International Publishing.

Itai Gat, Yossi Adi, Alex Schwing, and Tamir Hazan. On the importance of gradient norm in PAC-Bayesian bounds. *Advances in Neural Information Processing Systems*, 35:16068–16081, 2022.

Pascal Germain, Alexandre Lacasse, Mario Marchand, Sara Shanian, and François Laviolette. From PAC-Bayes bounds to KL regularization. *Advances in Neural Information Processing Systems*, 22, 2009.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Benjamin Guedj and Louis Pujol. Still no free lunches: the price to pay for tighter PAC-Bayes bounds. *Entropy*, 23(11):1529, 2021.

Maxime Haddouche and Benjamin Guedj. PAC-Bayes generalization bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2023.

Maxime Haddouche, Benjamin Guedj, and John Shawe-Taylor. Upper and lower bounds on the performance of kernel PCA. *arXiv preprint arXiv:2012.10369*, 2020.

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

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A PAC-Bayesian link between generalisation and flat minima. *arXiv preprint arXiv:2402.08508*, 2024.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 2020.

Fredrik Hellström and Benjamin Guedj. Comparing comparators in generalization bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 73–81. PMLR, 2024.

Fredrik Hellström and Giuseppe Durisi. Corrections to “Generalization bounds via information density and conditional information density”. *IEEE Journal on Selected Areas in Information Theory*, 2(3):1072–1073, 2021.

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

Matthew Holland. PAC-Bayes under potentially heavy tails. *Advances in Neural Information Processing Systems*, 32, 2019.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479. IEEE, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Toronto, ON, Canada, 2009.

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

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, and Andrew Gordon Wilson. Bayesian model selection, the marginal likelihood, and generalization. In *International Conference on Machine Learning*, pages 14223–14247. PMLR, 2022.

Andrés Masegosa, Stephan Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. *Advances in Neural Information Processing Systems*, 33:5263–5273, 2020.

Andrés R Masegosa and Luis A Ortega. PAC-Chernoff Bounds: Understanding Generalization in the Interpolation Regime. *arXiv preprint arXiv:2306.10947*, 2023.

Andrés Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 33:5479–5491, 2020.

Andreas Maurer. A note on the PAC Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the 11th annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning theory*, pages 164–170. ACM, 1999.

David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Luis A Ortega, Rafael Cabañas, and Andres Masegosa. Diversity and generalization in neural network ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11720–11743. PMLR, 2022.

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.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems*, 33:16833–16845, 2020.

Borja Rodríguez-Gálvez, 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.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

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

Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-Bayes-Bernstein inequality for martingales and its application to multiarmed bandits. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 98–111. JMLR Workshop and Conference Proceedings, 2012.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the 10th annual conference on Computational Learning theory*, pages 2–9. ACM, 1997.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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

Dániel Varga, Adrián Csiszárík, and Zsolt Zombori. Gradient regularization improves accuracy of discriminative models. *arXiv preprint arXiv:1712.09936*, 2017.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Yi-Shan Wu, Andres Masegosa, Stephan Lorenzen, Christian Igel, and Yevgeny Seldin. Chebyshev-Cantelli PAC-Bayes-Bennett inequality for the weighted majority vote. *Advances in Neural Information Processing Systems*, 34:12625–12636, 2021.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017.

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.

A Proofs

We first gather some standard properties of CGFs and Cramér transforms.

Let $\mathcal{D}_{\Lambda_{\theta}} := \{\lambda \in \mathbb{R}_+ \mid \Lambda_{\theta}(\lambda) < \infty\}$. Under the assumption that there is some $\lambda > 0$ such that $\Lambda_{\theta}(\lambda) < \infty$, we have $\mathcal{D}_{\Lambda_{\theta}} = [0, b)$ for some $b > 0$. Furthermore, Λ_{θ} is strictly convex and \mathcal{C}^∞ on $(0, b)$ (Boucheron et al., 2013, Section 2.2).

In particular, since for any $\lambda > 0$ we have $\Lambda_{\theta}(\lambda) := \log \mathbb{E}_{\nu} [e^{\lambda(L(\theta) - \ell(x, \theta))}] \leq \log \mathbb{E}_{\nu} [e^{\lambda L(\theta)}] = \lambda L(\theta) < \infty$ because $\ell \geq 0$, we verify that $\mathcal{D}_{\Lambda_{\theta}} = [0, \infty)$ and that Λ_{θ} is strictly convex and \mathcal{C}^∞ on $(0, \infty)$.

Remark 17. Observe that the previous proof also shows that $\mathbb{E}_{\rho}[L(\theta)] < \infty$ is enough to guarantee that $\mathbb{E}_{\rho}[\Lambda_{\theta}(\lambda)]$ is finite for every $\lambda \geq 0$.

The following technical result will be important for the proof of Lemma 6. It combines the previous remarks with Lemma 2.2.5(c) of Dembo and Zeitouni (2009).

Lemma 18. $\Lambda_{\theta}^*(a) < \infty$ for any $a \in [0, L(\theta))$. Furthermore, Λ_{θ}^* is differentiable in $(0, L(\theta))$.

Proof. We start by proving the finiteness condition. Since Λ_{θ} is strictly convex on $(0, \infty)$, Λ'_{θ} is also strictly increasing on $(0, \infty)$. This means that Λ'_{θ} is a bijection from $(0, \infty)$ to $\left(\inf_{\lambda \in (0, \infty)} \Lambda'_{\theta}(\lambda), \sup_{\lambda \in (0, \infty)} \Lambda'_{\theta}(\lambda) \right)$. In other words, for every $a \in \left(\inf_{\lambda \in (0, \infty)} \Lambda'_{\theta}(\lambda), \sup_{\lambda \in (0, \infty)} \Lambda'_{\theta}(\lambda) \right)$, the equation $a = \Lambda'_{\theta}(\lambda)$ has a unique solution $\lambda_a \in (0, \infty)$.

Now consider the function $g(\lambda) = \lambda a - \Lambda_{\theta}(\lambda)$. The previous observation means that $g'(\lambda_a) = 0$, and since g is concave, $\Lambda_{\theta}^*(a) = \sup_{\lambda \in (0, \infty)} \{\lambda a - \Lambda_{\theta}(\lambda)\} = g(\lambda_a) < \infty$.

It only remains to prove that $\inf_{\lambda \in (0, \infty)} \Lambda'_{\theta}(\lambda) = 0$ and $\sup_{\lambda \in (0, \infty)} \Lambda'_{\theta}(\lambda) = L(\theta)$. The first part follows from the fact that $\Lambda'_{\theta}(0) = 0$. The second one is equivalent to proving

$$\lim_{\lambda \rightarrow \infty} \Lambda'_{\theta}(\lambda) = \text{ess sup}(L(\theta) - \ell(X, \theta)),$$

where, without loss of generality, we assume that $\text{ess inf } \ell(X, \theta) = 0$. This is also a standard property of CGFs (see, for example, [this proof](#)).

Finally, differentiability of Λ_{θ}^* follows from its finiteness in $[0, L(\theta))$ and Lemma 6(5) in Gantert et al. (2016). \square

With the tools we gathered above, we are ready to prove our main technical lemma:

Lemma 6. For any $\theta \in \Theta$ and $c \geq 0$, we have

$$\mathbb{P}_{\nu^n} \left(n\Lambda_{\theta}^*(\text{gen}(\theta, D)) \geq c \right) \leq \mathbb{P}_{X \sim \text{exp}(1)} (X \geq c).$$

Proof. The proof relies in the properties of generalized inverses. Consider

$$\mathbb{P}_{D \sim \nu^n} \left(n\Lambda_{\theta}^*(\text{gen}(\theta, D)) \geq a \right)$$

for $a \geq 0$. We will separately consider three cases:

Case 1: $a/n < \sup_{x \in (0, L(\theta))} \Lambda_{\theta}^*(x)$.

By Proposition 1(5) in Embrechts and Hofert (2013), we have

$$\mathbb{P}_{D \sim \nu^n} \left(n\Lambda_{\theta}^*(\text{gen}(\theta, D)) \geq a \right) \leq \mathbb{P}_{D \sim \nu^n} \left(\text{gen}(\theta, D) \geq (\Lambda_{\theta}^*)^{-1}(a/n) \right),$$

and using the Cramér-Chernoff bound on the right-hand side we obtain

$$\mathbb{P}_{D \sim \nu^n} \left(n\Lambda_{\theta}^*(\text{gen}(\theta, D)) \geq a \right) \leq e^{-n\Lambda_{\theta}^*((\Lambda_{\theta}^*)^{-1}(a/n))}.$$

This results in

$$\mathbb{P}_{D \sim \nu^n} \left(n \Lambda_{\theta}^*(gen(\theta, D)) \geq a \right) \leq e^{-a}$$

because Λ_{θ}^* is continuous (in fact differentiable, see Lemma 18) on a/n and we can apply Proposition 1(4) in [Embrechts and Hofert \(2013\)](#).

Case 2: $a/n = \sup_{x \in (0, L(\theta))} \Lambda_{\theta}^*(x)$.

$$\begin{aligned} \mathbb{P}_{D \sim \nu^n} \left(n \Lambda_{\theta}^*(gen(\theta, D)) \geq a \right) &= \mathbb{P}_{D \sim \nu^n} \left(gen(\theta, D) = L(\theta) \right) \\ &= \mathbb{P}_{X \sim \nu} \left(\ell(X, \theta) = 0 \right)^n \\ &= e^{-n \Lambda_{\theta}^*(L(\theta))}. \end{aligned}$$

If $\Lambda_{\theta}^*(L(\theta)) < \infty$, then $a/n = \Lambda_{\theta}^*(L(\theta))$ by the lower semi-continuity of Λ_{θ}^* , and the result holds. Otherwise it is trivially true.

Case 3: $a/n > \sup_{x \in (0, L(\theta))} \Lambda_{\theta}^*(x)$. In this case

$$\mathbb{P}_{D \sim \nu^n} \left(n \Lambda_{\theta}^*(gen(\theta, D)) \geq a \right) = \mathbb{P}_{D \sim \nu^n} \left(\Lambda_{\theta}^*(gen(\theta, D)) = \infty \right).$$

But $\Lambda_{\theta}^*(gen(\theta, D)) = \infty$ can only happen if $gen(\theta, D) = L(\theta)$ and $\Lambda_{\theta}^*(L(\theta)) = \infty$, and this means that $\mathbb{P}_{D \sim \nu^n} \left(gen(\theta, D) = L(\theta) \right) = 0$.

This concludes the proof. \square

Now we list every result whose proof is not included in the paper.

Theorem 7 (PAC-Bayes-Chernoff bound). *Let $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_{\rho} [L(\theta)] \leq \mathbb{E}_{\rho} [\hat{L}(D, \theta)] + \inf_{\lambda \in [0, b)} \left\{ \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_{\rho} [\Lambda_{\theta}(\lambda)]}{\lambda} \right\}$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Proof. For any posterior distribution $\rho \in \mathcal{M}_1(\Theta)$ and any positive $m < n$, consider $m \Lambda_{\rho}^*(gen(\rho, D))$, where $gen(\rho, D) := \mathbb{E}_{\rho} [L(\theta)] - \mathbb{E}_{\rho} [\hat{L}(D, \theta)]$. The function $\Lambda_{\rho}^*(\cdot)$ will play a role analogue to the convex comparator function in [Rivasplata et al. \(2020\)](#). Since $\sup_{\lambda} \mathbb{E} X_{\lambda} \leq \mathbb{E} \sup_{\lambda} X_{\lambda}$, it verifies that

$$m \Lambda_{\rho}^*(gen(\rho, D)) \leq m \mathbb{E}_{\rho} [\Lambda_{\theta}^*(gen(\theta, D))]. \quad (17)$$

Applying Donsker-Varadhan's change of measure ([Donsker and Varadhan, 1975](#)) to the right-hand side of the inequality we obtain

$$m \Lambda_{\rho}^*(gen(\rho, D)) \leq KL(\rho|\pi) + \log \mathbb{E}_{\pi} \left(e^{m \Lambda_{\theta}^*(gen(\theta, D))} \right). \quad (18)$$

We can now apply Markov's inequality to the random variable $\mathbb{E}_{\pi} \left(e^{m \Lambda_{\theta}^*(gen(\theta, D))} \right)$. Thus, with probability at least $1 - \delta$,

$$m \Lambda_{\rho}^*(gen(\rho, D)) \leq KL(\rho|\pi) + \log \frac{1}{\delta} + \log \mathbb{E}_{\nu^n} \mathbb{E}_{\pi} \left(e^{m \Lambda_{\theta}^*(gen(\theta, D))} \right). \quad (19)$$

Since π is data-independent, we can swap both expectations using Fubini's theorem, so that we need to bound $\mathbb{E}_{\nu^n} \left(e^{m \Lambda_{\theta}^*(gen(\theta, D))} \right)$ for any fixed $\theta \in \Theta$. Here is where Lemma 6 comes into play: we have that for any $c > 0$,

$$\mathbb{P}_{D \sim \nu^n} \left(n \Lambda_{\theta}^*(gen(\theta, D)) \geq \frac{n}{m} c \right) \leq \mathbb{P}_{X \sim \exp(1)} \left(X \geq \frac{n}{m} c \right).$$

Since $X \sim \exp(1)$, we get $kX \sim \exp(\frac{1}{k})$. Thus, multiplying by $\frac{m}{n}$,

$$\mathbb{P}_{D \sim \nu^n} (m \Lambda_{\theta}^*(gen(\theta, D)) \geq c) \leq \mathbb{P}_{X \sim \exp(\frac{n}{m})} (X \geq c). \quad (20)$$

which in turn results in

$$\mathbb{P}_{\nu^n} \left(e^{m\Lambda_{\theta}^*(gen(\theta, D))} \geq t \right) \leq \mathbb{P}_{\exp(\frac{n}{m})} \left(e^X \geq t \right) \quad (21)$$

for any $t \geq 1$. Finally, since $X \sim \exp(\frac{n}{m})$, we have $e^X \sim \text{Pareto}(\frac{n}{m}, 1)$. Thus, for any $t \geq 1$

$$\mathbb{P}_{\nu^n} \left(e^{m\Lambda_{\theta}^*(gen(\theta, D))} \geq t \right) \leq \mathbb{P}_{\text{Pareto}(\frac{n}{m}, 1)} \left(X \geq t \right). \quad (22)$$

Using that for any random variable Z with support $\Omega \subseteq \mathbb{R}_+$, its expectation can be written as

$$\mathbb{E}[Z] = \int_{\Omega} P(Z \geq z) dz, \quad (23)$$

we obtain the desired bound:

$$\begin{aligned} \mathbb{E}_{D \sim \nu^n} \left(e^{m\Lambda_{\theta}^*(gen(\theta, D))} \right) &= \int_1^{\infty} \mathbb{P}_{D \sim \nu^n} \left(e^{m\Lambda_{\theta}^*(gen(\theta, D))} \geq t \right) dt \\ &\leq \int_1^{\infty} \mathbb{P}_{X \sim \text{Pareto}(\frac{n}{m}, 1)} \left(X \geq t \right) dt \\ &= \mathbb{E}_{X \sim \text{Pareto}(\frac{n}{m}, 1)} (X) \\ &= \frac{\frac{n}{m}}{\frac{n}{m} - 1} = \frac{n}{n - m}. \end{aligned}$$

Observe how the condition $m < n$ is crucial because a $\text{Pareto}(1, 1)$ has no finite mean. In conclusion, with probability at least $1 - \delta$ we have

$$m\Lambda_{\rho}^*(gen(\rho, D)) \leq KL(\rho|\pi) + \log \frac{n}{n - m} + \log \frac{1}{\delta}. \quad (24)$$

Dividing by m , setting $m = n - 1$ and applying $(\Lambda_{\rho}^*)^{-1}(\cdot)$ in both sides concludes the proof. \square

Corollary 8. *Let ℓ be the 0–1 loss and $\pi \in \mathcal{M}_1(\Theta)$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$kl \left(\mathbb{E}_{\rho}[\hat{L}(\theta, D)], \mathbb{E}_{\rho}[L(\theta)] \right) \leq \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n - 1},$$

simultaneously for every $\rho \in \mathcal{M}_1(\Theta)$.

Proof. We know that $\ell(\theta, X) \sim \text{Bin}(L(\theta))$, hence following the approach in Section 2.2 of Boucheron et al. (2013), we obtain

$$\Lambda_{\theta}^*(a) = kl(L(\theta) - a | L(\theta)).$$

From the proof of Theorem 7 we have

$$\mathbb{E}_{\rho}[\Lambda_{\theta}^*(gen(\theta, D))] \leq \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n - 1}, \quad (25)$$

and the result follows from the convexity of kl and Jensen's inequality. \square

Proposition 12. *If we fix some $\lambda > 0$, the bound in Theorem 11 can be minimized with respect to $\rho \in \mathcal{M}_1(\Theta)$. The optimal posterior is*

$$\rho^*(\theta) \propto \pi(\theta) \exp \left\{ -(n - 1)\lambda \hat{L}(D, \theta) - (n - 1)\psi(\theta, \lambda) \right\}. \quad (12)$$

Proof. We can solve the constrained minimization problem using standard results from variational calculus —see Appendices D and E of Bishop (2006) for a succinct introduction—. We need to minimize the functional

$$\mathcal{B}_{\pi, \lambda}[\rho] := \mathbb{E}_{\rho}[\hat{L}(D, \theta)] + \frac{\mathbb{E}_{\rho}[\psi(\theta, \lambda)]}{\lambda} + \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{\lambda(n - 1)} + \gamma \left(\int_{\Theta} \rho(\theta) d\theta - 1 \right),$$

where $\gamma \geq 0$ is the Lagrange multiplier. For this purpose, we compute the functional derivative of $\mathcal{B}_{\pi,\lambda}[\rho]$ wrt ρ ,

$$\frac{\delta \mathcal{B}_{\pi,\lambda}}{\delta \rho} = \hat{L}(D, \boldsymbol{\theta}) + \frac{\psi(\boldsymbol{\theta}, \lambda)}{\lambda} + \frac{1}{\lambda(n-1)} \left(\log \frac{\rho}{\pi} + 1 \right) + \gamma,$$

and find $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ satisfying $\frac{\delta \mathcal{B}_{\pi,\lambda}}{\delta \rho} = 0$, which results in the desired ρ^* after straightforward algebraic manipulations. \square

Theorem 16. *If Assumptions 2 and 3 are satisfied, then for any prior distribution $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ independent of D and any $\delta \in (0, 1)$; with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\boldsymbol{\theta})] \leq \mathbb{E}_\rho[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{2C\mathbb{E}_\rho \left[\|\hat{\nabla}_{\mathbf{x}}\ell\|_2^2 \right] \cdot K(\rho, \pi, n, \delta)} + \sqrt{2CL \cdot K(\rho, \pi, n, \delta)^{\frac{3}{2}}}$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$, where $K(\rho, \pi, n, \delta) := \frac{KL(\rho|\pi) + \log \frac{2n}{\delta}}{n-1}$.

Proof. By Assumption 3, $\|\nabla_{\mathbf{x}}\ell(\boldsymbol{\theta}, \mathbf{x})\|_2^2 \leq L$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and any $\mathbf{x} \in \mathcal{X}$ —note that with this we can already obtain a bound that includes a model-dependent Lipschitz constant $L(\boldsymbol{\theta})$. In particular, $\|\nabla_{\mathbf{x}}\ell(\boldsymbol{\theta}, X)\|_2^2$ is $\frac{L^2}{4}$ -sub-Gaussian. Thus we can use Corollary 5.3 in Hellström et al. (2023) and obtain the following PAC-Bayes bound:

$$\mathbb{E}_\rho \left[\|\nabla_{\mathbf{x}}\ell\|_2^2 \right] \leq \mathbb{E}_\rho \left[\|\hat{\nabla}_{\mathbf{x}}\ell\|_2^2 \right] + \frac{L}{\sqrt{2}} \sqrt{\frac{KL(\rho|\pi) + \log \frac{\sqrt{n}}{\delta_2}}{n-1}}, \quad (26)$$

with probability at least $1 - \delta_2$.

Now taking $\delta_1 = \delta_2 = \frac{\delta}{2}$, the bound in Theorem 15 and the one in (26) hold simultaneously with probability at least $1 - \delta$. Finally, plugging (26) in Theorem 15 we obtain the desired result. \square

B PAC-Bayes bounds for losses with bounded CGF

Corollary 19. *Assume the loss is σ^2 -sub-Gaussian. Let $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\boldsymbol{\theta})] \leq \mathbb{E}_\rho[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{2\sigma^2 \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1}},$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Proof. Follows from the fact that $(\psi^*)^{-1}(s) = \sqrt{2\sigma^2 s}$ for σ^2 -sub-Gaussian random variables (Boucheron et al., 2013, Section 2). \square

Corollary 20. *Assume the loss is (σ^2, c) -sub-gamma. Let $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ be any prior independent of D . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,*

$$\mathbb{E}_\rho[L(\boldsymbol{\theta})] \leq \mathbb{E}_\rho[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{2\sigma^2 \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1}} + c \frac{KL(\rho|\pi) + \log \frac{n}{\delta}}{n-1},$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Proof. Follows from the fact that $(\psi^*)^{-1}(s) = \sqrt{2\sigma^2 s} + cs$ for (σ^2, c) -sub-gamma random variables (Boucheron et al., 2013, Section 2). \square

C Experimental details

C.1 Models and data

As for the experimental setting for Figure 1 and the verification of Assumption 2; we have used the small InceptionV3 architecture (Szegedy et al., 2016) used in Zhang et al. (2017), where the total number of parameters of the model is 1.814.106. We trained the model in the CIFAR10 dataset (Krizhevsky et al., 2009) with the default train/test split using SGD with momentum 0.9 and learning rate 0.01 with exponential decay of 0.95. All models are trained for 30.000 iterations of batches of size 200 or until the train loss is under 0.005. These settings are selected to ensure that the random label model converges to an interpolator. For ℓ_2 regularization, the multiplicative factor is 0.01. We include a Jupyter Notebook in the Supplementary Material with the code for reproducing our experiments and figures.

From the definition of the log-loss, we have

$$\Lambda_{\theta}(\lambda) = \ln \mathbb{E}_{\nu} \left[e^{\lambda(L(\theta) - \ell(\mathbf{y}, \mathbf{x}, \theta))} \right] = \ln \mathbb{E}_{\nu} [p(\mathbf{y}|\mathbf{x}, \theta)^{\lambda}] - \mathbb{E}_{\nu} [\ln p(\mathbf{y}|\mathbf{x}, \theta)^{\lambda}],$$

and we estimate the expectations using the test data:

$$\Lambda_{\theta}(\lambda) \approx \ln \left(\frac{1}{M} \sum_{(\mathbf{x}, \mathbf{y}) \in D^{test}} p(\mathbf{y}|\mathbf{x}, \theta)^{\lambda} \right) - \frac{1}{M} \sum_{(\mathbf{x}, \mathbf{y}) \in D^{test}} \ln p(\mathbf{y}|\mathbf{x}, \theta)^{\lambda}.$$

We also estimated $\mathbb{V}(\ell)$ and $\|\nabla_{\mathbf{x}}\ell\|_2$ using D^{test} .

C.2 Experimental backing for Assumption 2

We now experimentally evaluate Assumption 2 in the same setup as above. We need to show that there is certain $C > 0$ such that for any $\lambda > 0$, we have

$$\Lambda_{\theta}(\lambda) \leq \frac{C}{2} \lambda^2 \|\nabla_{\mathbf{x}}\ell\|_2^2,$$

for every $\theta \in \Theta$. Since in this case each θ represents a configuration of weights in the InceptionV3 architecture, we cannot strictly verify the assumption for every θ , however, we provide empirical backing for its feasibility.

We estimate $\Lambda_{\theta}(\lambda)$ and $\|\nabla_{\mathbf{x}}\ell\|_2^2$ with test data and plot $\frac{\Lambda_{\theta}(\lambda)}{0.5\lambda^2 \|\nabla_{\mathbf{x}}\ell\|_2^2}$ for the three different models introduced in Figure 1 (in the case of Zero the inequality trivially holds).

We plot the results in Figure 2, showing that the quantity of interest decreases with λ and that it is reasonable to expect Assumption 2 holding with $C < 1$. Observe that by a simple call to L'Hopital's rule, one can see that $\lim_{x \rightarrow 0^+} \frac{\Lambda_{\theta}(\lambda)}{0.5\lambda^2 \|\nabla_{\mathbf{x}}\ell\|_2^2} = \frac{\mathbb{V}(\ell)}{\|\nabla_{\mathbf{x}}\ell\|_2^2}$.

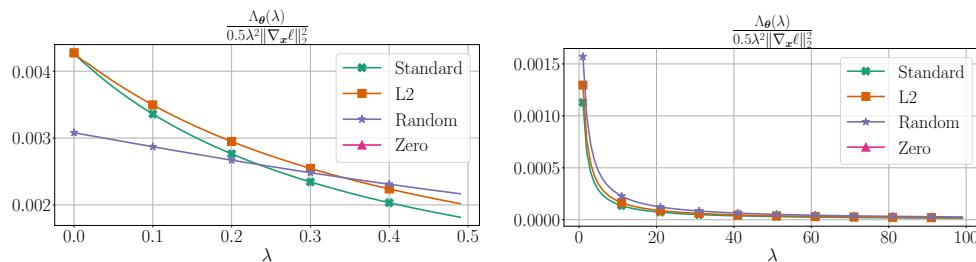


Figure 2: The plots of $\frac{\Lambda_{\theta}(\lambda)}{0.5\lambda^2 \|\nabla_{\mathbf{x}}\ell\|_2^2}$ in two different scales. In the case of these models, Assumption 2 holds with $C \approx 0.005$.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We mention and discuss bounds and techniques that are introduced later in the paper. Limitations of the assumptions are also mentioned.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Both during related work discussion and in the conclusions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Yes, complete proofs are included in Appendix A, and the assumptions are discussed and justified.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include our code in the supplementary material, and the details of the implementations are discussed in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: The same as above.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Yes, in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: Not necessary.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: No relevant issues in our work.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: This is a theoretical work aiming towards better understanding generalization, no negative societal impact. Potential positive impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.