entropy

MDPI

Article

PAC-Bayes Unleashed: Generalisation Bounds with Unbounded Losses

Maxime Haddouche [1], Benjamin Guedj [2,3,*], Omar Rivasplata [2] and John Shawe-Taylor [2]

[1] ENS Paris-Saclay, 91190 Gif-sur-Yvette, France; maxime.haddouche@ens-paris-saclay.fr
[2] Centre for Artificial Intelligence, Department of Computer Science, University College London, London WC1V 6LJ, UK; o.rivasplata@cs.ucl.ac.uk (O.R.); j.shawe-taylor@ucl.ac.uk (J.S.-T.)
[3] Inria, Lille–Nord Europe Research Centre and Inria London Programme, 59800 Lille, France
* Correspondence: b.guedj@ucl.ac.uk

Abstract: We present new PAC-Bayesian generalisation bounds for learning problems with unbounded loss functions. This extends the relevance and applicability of the PAC-Bayes learning framework, where most of the existing literature focuses on supervised learning problems with a bounded loss function (typically assumed to take values in the interval [0;1]). In order to relax this classical assumption, we propose to allow the range of the loss to depend on each predictor. This relaxation is captured by our new notion of *HYPothesis-dependent rangE* (HYPE). Based on this, we derive a novel PAC-Bayesian generalisation bound for unbounded loss functions, and we instantiate it on a linear regression problem. To make our theory usable by the largest audience possible, we include discussions on actual computation, practicality and limitations of our assumptions.

Keywords: statistical learning theory; PAC-Bayes; generalisation bounds

Citation: Haddouche, M.; Guedj, B.; Rivasplata, O.; Shawe-Taylor, J. PAC-Bayes Unleashed: Generalisation Bounds with Unbounded Losses. *Entropy* **2021**, *23*, 1330. https://doi.org/10.3390/e23101330

Academic Editor: Boris Ryabko

Received: 22 August 2021
Accepted: 25 September 2021
Published: 12 October 2021

1. Introduction

Since its emergence in the late 1990s, the PAC-Bayes theory (see the seminal works of [1–3], the recent survey by [4] and work by [5]) has been a powerful tool to obtain generalisation bounds and to derive efficient learning algorithms. Generalisation bounds are helpful for understanding how a learning algorithm may perform on future similar batches of data. While the classical generalization bounds typically address the performance of individual predictors from a given hypothesis class, PAC-Bayes bounds typically address a randomized predictor defined by a distribution over the hypothesis class.

PAC-Bayes bounds were originally meant for binary classification problems [6–8], but the literature now includes many contributions involving any bounded loss function (without loss of generality, with values in $[0;1]$), not just the binary loss. Our goal is to provide new PAC-Bayes bounds that are valid for unbounded loss functions, and thus extend the usability of PAC-Bayes to a much larger class of learning problems. To do so, we reformulate the general PAC-Bayes theorem of [9] and use it as basic building block to derive our new PAC-Bayes bound.

Some ways to circumvent the bounded range assumption on the losses have been explored in the recent literature. For instance, one approach consists of assuming a tail decay rate on the loss, such as sub-gaussian or sub-exponential tails [10,11]; however, this approach requires the knowledge of additional parameters. Some other works have also looked into the analysis for heavy-tailed losses, e.g., ref. [12] proposed a polynomial moment-dependent bound with f-divergences, while [13] devised an exponential bound that assumes the second (uncentered) moment of the loss is bounded by a constant (with a truncated risk estimator, as recalled in Section 4 below). A somewhat related approach was explored by [14], who do not assume boundedness of the loss, but instead control higher-order moments of the generalization gap through the Efron-Stein variance proxy. See also [5].

We investigate a different route here. We introduce the *HYPothesis-dependent rangE* (HYPE) condition, which means that the loss is upper-bounded by a term that depends on the chosen predictor (but does not depend on the data). Thus, effectively, the loss may have an arbitrarily large range. The HYPE condition allows us to derive an upper bound on the exponential moment of a suitably chosen functional, which, combined with the general PAC-Bayes theorem, leads to our new PAC-Bayes bound. To illustrate it, we instantiate the new bound on a linear regression problem, which additionally serves the purpose of illustrating that our HYPE condition is easy to verify in practice, given an explicit formulation of the loss function. In particular, we shall see in the linear regression setting that a mere use of the triangle inequality is enough to check the HYPE condition. The technical assumptions on which our results are based are comparable to those of the classical PAC-Bayes bounds; we state them in full detail, with discussions, for the sake of clarity and to make our work accessible.

Our contributions are twofold. (i) We propose PAC-Bayesian bounds holding with unbounded loss functions, therefore overcoming a limitation of the mainstream PAC-Bayesian literature for which a bounded loss is usually assumed. (ii) We analyse the bound, its implications, limitations of our assumptions, and their usability by practitioners. We hope this will extend the PAC-Bayes framework into a widely usable tool for a significantly wider range of problems, such as unbounded regression or reinforcement learning problems with unbounded rewards.

Outline. Section 2 introduces our notation and definition of the HYPE condition and provides a general PAC-Bayesian bound, which is valid for any learning problem complying with a mild assumption. For the sake of completeness, we present how our approach (designed for the unbounded case) behaves in the bounded case (Section 3). This section is not the core of our work, but rather serves as a safety check and particularises our bound to more classical PAC-Bayesian assumptions. We also provide numerical experiments. Section 4 introduces the notion of *softening functions* and particularises Section 2's PAC-Bayesian bound. In particular, we make explicit all terms in the right-hand side. Section 5.1 extends our results to linear regression (which has been studied from the perspective of PAC-Bayes in the literature, most recently by [15]). We also experimentally illustrate the behaviour of our bound. Finally, Section 6 presents, in detail, related works and Section 7 contains all proofs of the original claims we make in the paper.

2. Framework and Preliminary Results

The learning problem is specified by three variables $(\mathcal{H}, \mathcal{Z}, \ell)$ consisting of a set $\mathcal{H}$ of predictors, the data space $\mathcal{Z}$, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$.

For a given positive integer m, we consider size-m datasets. The space of all possible datasets of this fixed size is $\mathcal{S} = \mathcal{Z}^m$; an arbitrary element of this space is $s = (z_1, \ldots, z_m)$. We denote S as a random dataset: $S = (Z_1, \ldots, Z_m)$ where the random data points Z_i are independent and sampled from the same distribution μ over $\mathcal{Z}$. We call μ the data-generating distribution. The assumption that the Z_i's are *independent and identically distributed* is typically called the i.i.d. data assumption. It means that the random sample S (of size m) has distribution $\mu^{\otimes m}$ which is the product of m copies of μ.

For any predictor $h \in \mathcal{H}$, we define the *empirical risk* of h over a sample s, denoted $R_s(h)$, and the *theoretical risk* of h, denoted $R(h)$, as:

$$R_s(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) \qquad \text{and} \qquad R(h) = \mathbb{E}_\mu[\ell(h, Z)]$$

respectively, where $\mathbb{E}_\mu[\ell(h, Z)]$ denotes the expectation with respect to $Z \sim \mu$. Finally, we define the *risk gap* $\Delta_s(h) = R(h) - R_s(h)$ for any $h \in \mathcal{H}$ and $s \in \mathcal{S}$. Often, $\Delta_s(h)$ is referred to as the generalisation gap.

Notice that for a random dataset S, the empirical risk $R_S(h)$ is random, with expected value $\mathbb{E}_{\mu^{\otimes m}}[R_S(h)] = R(h)$, where $\mathbb{E}_{\mu^{\otimes m}}$ the expectation under the distribution of the random sample S.

In general, $\mathbb{E}_\mu[\cdot]$ denotes an expectation under the distribution μ. When we want to emphasize the role of the random variable $Z \sim \mu$ we write $\mathbb{E}_Z[\cdot]$ or $\mathbb{E}_{Z\sim\mu}[\cdot]$ instead of $\mathbb{E}_\mu[\cdot]$. We use a similar convention for expectations related to any other distributions and random quantities. We now introduce the key concept to our analysis.

Definition 1. (HYPE). *A loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ is said to satisfy the **hypothesis-dependent range** (`HYPE`) condition if there exists a function $K : \mathcal{H} \to \mathbb{R}^+\backslash\{0\}$ such that $\sup_{z\in\mathcal{Z}} \ell(h, z) \leq K(h)$ for every predictor h. We then say that ℓ is `HYPE`(K) compliant.*

Let $\mathcal{M}_1^+(\mathcal{H})$ be the set of probability distributions on $\mathcal{H}$. We assume that all considered probability measures on $\mathcal{H}$ are defined on a fixed σ-algebra over $\mathcal{H}$, while the notation $\mathcal{M}_1^+(\mathcal{H})$ hides the σ-algebra, for simplicity. For $P, P' \in \mathcal{M}_1^+(\mathcal{H})$, the notation $P' \ll P$ indicates that P' is absolutely continuous with respect to P (i.e., $P'(A) = 0$ if $P(A) = 0$ for measurable $A \subset \mathcal{H}$). We write $P' \sim P$ to indicate that $P' \ll P$ and $P \ll P'$, i.e., these two distributions are absolutely continuous with respect to each other.

We now recall a result from Germain et al. [9]. Note that while implicit in many PAC-Bayes works (including theirs), we make it explicit that both the prior P and the posterior Q must be absolutely continuous with respect to each other. We discuss this restriction below.

Theorem 1. (Adapted from [9], Theorem 2.1.) *For any $P \in \mathcal{M}_1^+(\mathcal{H})$ with no dependency on data, for any function $F : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$, define the exponential moment:*

$$\chi := \mathbb{E}_S \mathbb{E}_{h\sim P} \left[e^{F(R_S(h), R(h))} \right].$$

If F is convex, then for any $\delta \in [0; 1]$, with probability of at least $1 - \delta$ over random samples S, simultaneously for all $Q \in \mathcal{M}_1^+(\mathcal{H})$ such that $Q \sim P$ we have:

$$F\big(\mathbb{E}_{h\sim Q}[R_S(h)], \mathbb{E}_{h\sim Q}[R(h)]\big) \leq \mathrm{KL}(Q||P) + \log\Big(\frac{\chi}{\delta}\Big).$$

The proof is deferred to Section 7.1. Note that the proof in [9] requires that $P \ll Q$, although it is not explicitly stated; we highlight this in our own proof. While $Q \ll P$ is classical and necessary for the $\mathrm{KL}(Q||P)$ to be meaningful, $P \ll Q$ appears to be more restrictive. In particular, we have to choose Q such that it has the exact same support as P (e.g., choosing a Gaussian and a truncated Gaussian is not possible). However, we can still apply our theorem when P and Q belong to the same parametric family of distributions, e.g., both 'full-support' Gaussian or Laplace distributions, but these are just two examples and there are many others.

Note that Alquier et al. [10] (Theorem 4.1) adapted a result from Catoni [8], which only requires $Q \ll P$. This comes at the expense of what Alquier et al. [10] (Definition 2.3) called a *Hoeffding's assumption*, which means that the exponential moment χ is assumed to be bounded by a function depending only on the hyperparameters (such as the dataset size m or parameters given by Hoeffding's assumption). Our analysis does not require this assumption, which might prove restrictive in practice.

Theorem 1 may be seen as a basis to recover many classical PAC-Bayesian bounds. For instance, $F(x, y) = 2m(x - y)^2$, recovers McAllester's bound as recalled in [4] (Theorem 1). To get a usable bound, the outstanding task is to bound the exponential moment χ. Note that a previous attempt has been made in [11], as described in Section 6.1 below. Furthermore, under the assumption that the distribution P has no dependency on the data, we may swap the order of integration in the exponential moment thanks to Fubini-Tonelli's theorem and the positiveness of the exponential:

$$\chi = \mathbb{E}_{h\sim P} \mathbb{E}_S \left[e^{F(R_S(h), R(h))} \right].$$

This is the starting point for the way that the exponential moment was handled in several works in the PAC-Bayes literature. Essentially, for a fixed h, one may upper-bound the innermost expectation (with respect to S) using standard exponential moment inequalities.

In this work, we will use Theorem 1 with $F(x,y) = m^\alpha D(x,y)$, where $\alpha > 0$, and $D : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ is a convex function. In this case, the high-probability inequality of the theorem takes the form:

$$D\big(\mathbb{E}_{h\sim Q}[R_S(h)], \mathbb{E}_{h\sim Q}[R(h)]\big) \leq \frac{1}{m^\alpha}\left(\mathrm{KL}(Q||P) + \log\left(\frac{1}{\delta}\mathbb{E}_{h\sim P}\mathbb{E}_S\, e^{m^\alpha D(R_S(h),R(h))}\right)\right). \quad (1)$$

Our goal is to control $\mathbb{E}_S\, e^{m^\alpha D(R_S(h),R(h))}$ for a fixed h, when $D(x,y) = y - x$. This will readily give us control on the exponential moment χ. To do so, we propose the following theorem:

Theorem 2. *Let $h \in \mathcal{H}$ be a fixed predictor and $\alpha \in \mathbb{R}$. If the loss function ℓ is $\texttt{HYPE}(K)$ compliant, then for $\Delta_S(h) = R(h) - R_S(h)$ we have:*

$$\mathbb{E}_S\left[e^{m^\alpha \Delta_S(h)}\right] \leq \exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right).$$

Proof. Let $h \in \mathcal{H}$. Then:

$$\begin{aligned}\mathbb{E}_S\left[e^{m^\alpha \Delta_S(h)}\right] &= \mathbb{E}\left[\exp\left(m^{\alpha-1}\sum_{i=1}^m (l(h,Z_i) - R(h))\right)\right] \\ &= \mathbb{E}\left[\prod_{i=1}^m \exp\Big(m^{\alpha-1}(\ell(h,Z_i) - R(h))\Big)\right] \\ &= \prod_{i=1}^m \mathbb{E}\Big[\exp\Big(m^{\alpha-1}(\ell(h,Z_i) - R(h))\Big)\Big].\end{aligned}$$

We now apply Hoeffding's lemma, for any $i \in \{1..m\}$, the random (in Z_i) variable $\ell(h,Z_i) - R(h)$ is centered, taking values in $[-K(h); K(h)]$, so that:

$$\mathbb{E}\Big[\exp\Big(m^{\alpha-1}(\ell(h,Z_i) - R(h))\Big)\Big] \leq \exp\left(m^{2\alpha-2}\frac{4K(h)^2}{8}\right)$$

and finally:

$$\mathbb{E}_S\left[e^{m^\alpha \Delta_S(h)}\right] \leq \prod_{i=1}^m \exp\left(m^{2\alpha-2}\frac{4K(h)^2}{8}\right) = \exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right).$$

□

The strength of this result lies in the fact that $\frac{K(h)^2}{m^{1-2\alpha}}$, is a decreasing factor in m, when $\alpha \leq 1/2$, and more generally, one can control how fast the exponential moment will explode when m grows by the choice of the hyperparameter α.

For convenient cross-referencing, we state the following rewriting of Theorem 1.

Theorem 3. *Let the loss ℓ be $\texttt{HYPE}(K)$ compliant. For any $P \in \mathcal{M}_1^+(\mathcal{H})$ with no data dependency, for any $\alpha \in \mathbb{R}$ and for any $\delta \in [0;1]$, with probability of at least $1-\delta$ over size-m random samples S, simultaneously for all Q such that $Q \sim P$ we have:*

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}[R_S(h)] + \frac{1}{m^\alpha}\left(\mathrm{KL}(Q||P) + \log\frac{\mathbb{E}_{h\sim P}\left[\exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right)\right]}{\delta}\right).$$

Proof. We first apply Theorem 1 with $F(x,y) = m^\alpha(y-x)$. More precisely, we use Equation (1) with $D(x,y) = y - x$. We then conclude with Theorem 2. □

3. Safety Check: The Bounded Loss Case

3.1. Theoretical Results

At this stage, the reader might wonder whether this new approach allows for the recovery of known results in the bounded case: the answer is yes.

In this section, we study the case where ℓ is bounded by some constant $C \in \mathbb{R}^+ \setminus \{0\}$. In other words, we consider the case that $\sup_h \sup_z \ell(h,z) \leq C$. We provide a bound, valid for any choice of "priors" P and "posteriors" Q such that $P \sim Q$, which is an immediate corollary of Theorem 3.

Proposition 1. *Let ℓ be $\texttt{HYPE}(K)$ compliant, with $K(h) = C$ constant, and let $\alpha \in \mathbb{R}$. Let $P \in \mathcal{M}_1^+(\mathcal{H})$ be a distribution with no data dependency. Then, for any $\delta \in [0;1]$, with probability of at least $1-\delta$ over random m-samples S, simultaneously for all $Q \in \mathcal{M}_1^+(\mathcal{H})$ such that $Q \sim P$ we have:*

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}[R_S(h)] + \frac{\mathrm{KL}(Q||P) + \log(1/\delta)}{m^\alpha} + \frac{C^2}{2m^{1-\alpha}}.$$

Remark 1. *We provide Proposition 1 to evaluate the robustness of our approach. For instance, by comparing it with the PAC-Bayesian bound found in Germain et al. [11]. This discussion can be found in Section 6.1, where the bound from Germain et al. [11] is presented in detail.*

Remark 2. *At first glance, a naive remark: in order to control the rate of convergence of all the terms of the bound in Proposition 1 (as is often the case in classical PAC-Bayesian bounds), then the only case of interest is in fact $\alpha = \frac{1}{2}$. However, one could notice that the factor C^2 is not optimisable, while the KL is. In this way, if it appears that C^2 is too big, in practice, one wants to have the ability to attenuate its influence as much as possible and this may lead us to consider $\alpha < 1/2$. The following lemma answers this question.*

Lemma 1. *For any given $K_1 > 0$, the function $f_{K_1}(\alpha) := \frac{K_1}{m^\alpha} + \frac{C^2}{m^{1-\alpha}}$ reaches its minimum at*

$$\alpha_0 = \frac{1}{2} + \frac{1}{2\log(m)} \log\left(\frac{2K_1}{C^2}\right).$$

Proof. The explicit calculus of the f'_{K_1} and the resolution of $f'_{K_1}(\alpha) = 0$ provides the result. □

Remark 3. *Lemma 1 indicates that with a fixed "prior" P and "posterior" Q, taking $K_1 = \mathrm{KL}(Q||P) + \log(1/\delta)$, gives the optimised value of the bound in Proposition 1. We numerically show in Section 3.2 (first experiment there) that optimising α leads to significantly better results.*

Now the only remaining question is how to optimise the KL divergence. To do so, we may need to fix an "informed prior" to minimise the KL divergence with an interesting posterior. This idea has been studied by [16,17] and, more recently, by Mhammedi et al. [18], Rivasplata et al. [5], among others. We will adapt it to our problem in the simplest way.

We now introduce some additional notation. For a sample $s = (z_1, \ldots, z_m)$ and $k \in \{1..m\}$, we define $s_{\leq k} := \{z_1, \ldots, z_k\}$ and $s_{>k} := \{z_{k+1}, \ldots, z_m\}$. Then, similarly, for a random sample S, we have the splits $S_{\leq k}$ and $S_{>k}$.

Proposition 2. *Let ℓ be $\texttt{HYPE}(K)$ compliant, with constant $K(h) = C$, and $\alpha_1, \alpha_2 \in \mathbb{R}$. Consider any "priors" $P_1 \in \mathcal{M}_1^+(\mathcal{H})$ (possibly dependent on $S_{>m/2}$) and $P_2 \in \mathcal{M}_1^+(\mathcal{H})$ (possibly dependent*

on $S_{\leq m/2}$). Then, for any $\delta \in [0;1]$, with probability of at least $1-\delta$ over random size-m samples S, simultaneously for all $Q \in \mathcal{M}_1^+(\mathcal{H})$ such that $Q \sim P_1$ and $Q \sim P_2$ we have:

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}[R_S(h)] + \frac{1}{2}\left(\frac{\mathrm{KL}(Q||P_1)+\log(2/\delta)}{(m/2)^{\alpha_1}} + \frac{C^2}{2(m/2)^{1-\alpha_1}}\right) + \frac{1}{2}\left(\frac{\mathrm{KL}(Q||P_2)+\log(2/\delta)}{(m/2)^{\alpha_2}} + \frac{C^2}{2(m/2)^{1-\alpha_2}}\right).$$

Proof. Let P_1, P_2, Q be as stated in Proposition 2. We first notice that by using Proposition 1 on the two halves of the sample, we obtain, with a probability of at least $1-\delta/2$:

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}\left[\frac{1}{m/2}\sum_{i=1}^{m/2}\ell(h,Z_i)\right] + \frac{\mathrm{KL}(Q||P_1)+\log(2/\delta)}{(m/2)^{\alpha_1}} + \frac{C^2}{2(m/2)^{1-\alpha_1}}$$

and also with probability at least $1-\delta/2$:

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}\left[\frac{1}{m/2}\sum_{i=1}^{m/2}\ell(h,Z_{m/2+i})\right] + \frac{\mathrm{KL}(Q||P_2)+\log(2/\delta)}{(m/2)^{\alpha_2}} + \frac{C^2}{2(m/2)^{1-\alpha_2}}.$$

Hence, with a probability of at least $1-\delta$, both inequalities hold, and the result follows by adding them and dividing by 2. □

Remark 4. *One can notice that the main difference between Proposition 2 and Proposition 1 lies in the implicit PAC-Bayesian paradigm that our priors must not depend on the data. With this last proposition, we implicitly allow P_1 to depend on $S_{>m/2}$ and P_2 on $S_{\leq m/2}$, which can in practice lead to far more accurate priors. We numerically show this fact in Section 3.2's second experiment. Note that this idea is not new and has been studied, for instance, in [19] for the specific case of SVMs.*

3.2. Numerical Experiments

Our experimental framework has been inspired by the work of [18].

Settings. We generate synthetic data for classification, and we are using the 0–1 loss. The data space is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{0,1\}$ with $d \in \mathbb{N}$. The set of predictors $\mathcal{H}$ is parameterised with d-dimensional 'weight' vectors: $\mathcal{H} = \{h_w : \mathcal{X} \to \mathcal{Y} \mid w \in \mathbb{R}^d\}$. For simplicity, we identify h_w with w and we also identify the space $\mathcal{H}$, with the weight space $\mathcal{W} = \mathbb{R}^d$. For $z = (x,y) \in \mathcal{Z}$ and $w \in \mathcal{W}$, we define the loss as $\ell(w,z) := |\mathbb{1}\{\phi(w^\top x) > 1/2\} - y|$, where $\phi(r) = \frac{1}{1+e^{-r}}$. We want to learn an optimised predictor given a dataset $S = (Z_i)_{i=1..m}$ where $Z_i = (X_i, Y_i)$. To do so, we use *regularised logistic regression* and compute:

$$\hat{w}(S) := \arg\min_{w\in\mathcal{W}} \lambda\frac{||w||^2}{2} - \frac{1}{m}\sum_{i=1}^{m} y_i \log\Big(\phi(w^\top x_i)\Big) + (1-y_i)\log\Big(1-\phi(w^\top x_i)\Big) \qquad (2)$$

where λ is a fixed regularisation parameter.

We also restrict the probability distributions (over $\mathcal{W} = \mathbb{R}^d$), considered for this learning problem. We consider the Gaussian distribution $\mathcal{N}(w, \sigma^2 I_d)$ with centre $w \in \mathbb{R}^d$ and diagonal covariance $\sigma^2 I_d \in \mathbb{R}^{d\times d}$ with $\sigma^2 > 0$.

Parameters. We set $\delta = 0.05, \lambda = 0.01$. We approximately solve Equation (2) by using the `minimize` function of the optimisation module in Python, with the Powell method. To approximate gaussian expectations, we use Monte-Carlo sampling.

Synthetic data. We generate synthetic data for $d = 10$ according to the following process: for a fixed sample size m, we draw $X_1, ..., X_m$ under the multivariate Gaussian distribution $\mathcal{N}(0, I_d)$ and for each i we compute the label if X_i as: $Y_i = \mathbb{1}\{\phi(w^{*\top} x_i) > 1/2\}$ where w^* is the vector formed by the d first digits of the number π.

Normalisation trick. Given the predictors shape, we notice that for any $w \in \mathcal{W}$:

$$\mathbb{1}\{\phi(w^\top x) > 1/2\} = 1 \quad \Leftrightarrow \quad \frac{1}{1+\exp(-w^\top x)} > \frac{1}{2} \quad \Leftrightarrow \quad w^\top x < 0.$$

Thus, the value of the prediction is exclusively determined by the sign of the inner product, and this quantity is definitely not influenced by the norm of the vector. Then, for any sample S, we call the **normalisation trick** the fact of considering $\hat{w}(S)/||\hat{w}(S)||$ instead of $\hat{w}(S)$ in our calculations. This process will not deteriorate the quality of the prediction and will considerably enhance the value of the KL divergence.

3.2.1. First experiment

Our goal here is to highlight the point discussed in Remark 2, e.g., the influence of the parameter α in Proposition 1. We arbitrarily fix $\sigma_0^2 = 1/2$, and define our *naive prior* as $P_0 = \mathcal{N}(0, \sigma_0^2 I_d)$. For a fixed dataset S, we define our posterior as $P(S) := \mathcal{N}(\hat{h}(S), \sigma^2 I_d)$, with $\sigma^2 \in \{1/2, \ldots, 1/2^J\}$ (for $J = \log_2(m)$) such that it is minimising the bound among candidates. We computed two curves: first, Proposition 1 with $\alpha = 1/2$ second, Proposition 1 again with α equals to the value proposed in Lemma 1. Notice that to compute this last bound, we first optimised our choice of posterior with $\alpha = 1/2$ and then optimised α, to be consistent with Lemma 1. Indeed, we proved this lemma by assuming that the KL divergence was already fixed, hence our optimisation process is in two steps. Note that we chose to apply the normalisation trick here, we then obtained the left curve of Figure 1.

Discussion. From this curve, we formulate several remarks. First, we remark on this specific case, our theorem provides a tight result in practice (with an error rate lesser than 10% for the bound with optimised alpha). Second, we can now confirm that choosing an optimised α leads to a tighter bound. In further studies, it will be relevant to adjust α with regards to the different terms of our bound instead of looking for an identical convergence rate for all terms.

3.2.2. Second Experiment

We now study Proposition 2 to see if an informed prior effectively provides a tighter bound than a naive one. We will use the notations introduced in Proposition 2. For a dataset S, we define $w_1(S) = w(S_{>m/2})$ as the vector resulting from the optimisation of Equation (2) on $S_{>m/2}$. Similarly, we define $w_2(S) := w(S_{\leq m/2})$. We arbitrarily fix $\sigma_0^2 = 1/2$, and define our *informed priors* as: $P_1 = \mathcal{N}(w_1(S), \sigma_0^2 I_d)$ and $P_2 = \mathcal{N}(w_2(S), \sigma_0^2 I_d)$. Finally, we define our posterior as $P(S) := \mathcal{N}(\hat{w}(S), \sigma^2 I_d)$, with $\sigma^2 \in \{1/2, ..., 1/2^J\}$ (for $J = \log_2(m)$) with σ^2 optimising the bound among the same candidate than the first experiment. We computed two curves: first, Proposition 1 with α optimised accordingly to Lemma 1 secondly, Proposition 2 with α_1, α_2 optimised as well, and informed priors as defined above. We chose to not apply the normalisation trick here, we then obtained the right curve of Figure 1.

Discussion. It is clear, that with this framework, having an informed prior is a powerful tool to enhance the quality of our bound. Notice that we voluntarily chose to not apply the normalisation trick here. The reason is that this trick appears to be too powerful in practice, and applying it leads to counterproductive results; to highlight our point: the bound without informed prior would be tighter than the one with informed prior. Furthermore, this trick is linked to the specific structure of our problem and is not valid for any classification problem. Thus, the idea of providing informed priors remains an interesting tool for most cases.

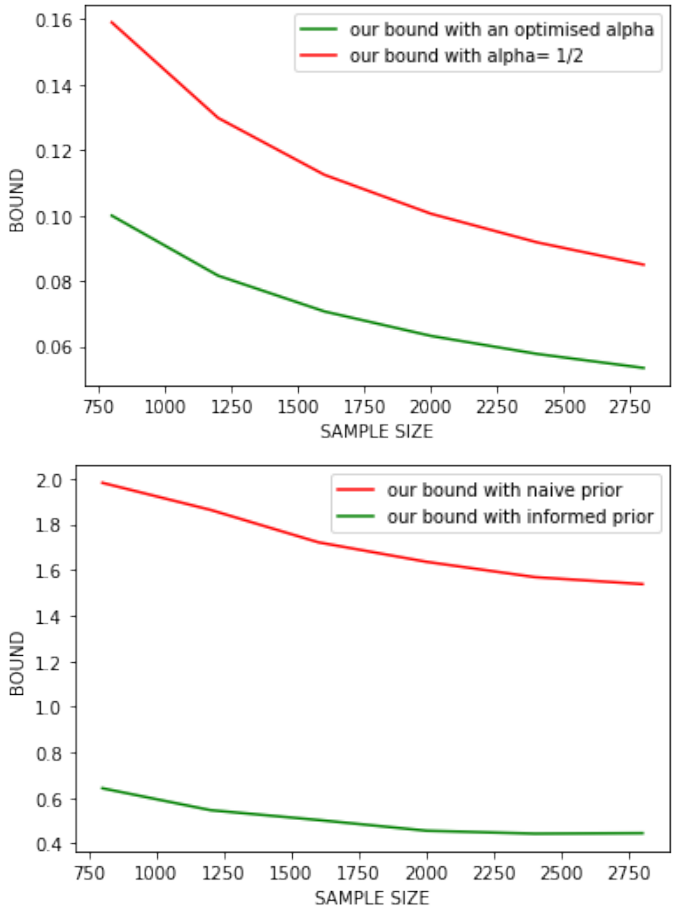

Figure 1. Above, result of the first experiment which highlight the importance of optimising α. Below, result of the second experiment which show how effective an informed prior is.

4. PAC Bayesian Bounds with Smoothed Estimator

We now move on to control the right-hand side term in Theorem 3 when K is not constant. A first step is to consider a transformed estimate of the risk, inspired by the truncated estimator from [20], also used in [21], and more recently in [13]. The following is inspired by the results of [13], which we summarise in Section 6.

The idea is to modify the estimator $R_S(h)$ for any h by introducing a threshold t and a function ψ which will attenuate the influence of the empirical losses $(\ell(h, Z_i))_{i=1..m}$ that exceed t.

Definition 2. ψ-risks. *For every $t > 0$, $\psi : \mathbb{R}^+ \to \mathbb{R}^+$, for any $h \in \mathcal{H}$, we define the empirical ψ-risk $R_{S,\psi,t}$ and the theoretical ψ-risk $R_{\psi,t}$ as follows:*

$$R_{S,\psi,t}(h) := \frac{t}{m} \sum_{i=1}^{m} \psi\left(\frac{\ell(h, Z_i)}{t}\right) \quad \text{and} \quad R_{\psi,t}(h) = \mathbb{E}_\mu\left[t\, \psi\left(\frac{\ell(h, Z)}{t}\right)\right]$$

where $Z \sim \mu$. Notice that $\mathbb{E}_S\left[R_{S,\psi,t}(h)\right] = R_{\psi,t}(h)$.

We now focus on what we call *softening functions*, i.e., functions that will temper high values of the loss function ℓ.

Definition 3. (Softening function). *We say that $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is a softening function if:*

- $\forall x \in [0; 1], \psi(x) = x$,
- *ψ is non-decreasing,*
- $\forall x \geq 1, \psi(x) \leq x$.

We let $\mathcal{F}$ denote the set of all softening functions.

Remark 5. *Notice that those three assumptions ensure that ψ is continuous at 1. For instance, the functions $f : x \mapsto x\mathbb{1}\{x \leq 1\} + \mathbb{1}\{x > 1\}$ and $g : x \mapsto x\mathbb{1}\{x \leq 1\} + (2\sqrt{x} - 1)\mathbb{1}\{x > 1\}$ are in $\mathcal{F}$. In Section 6 we compare these softening functions and those used by Holland [13].*

Using $\psi \in \mathcal{F}$, for a fixed threshold $t > 0$, the softened loss function $t\psi\left(\frac{\ell(h,z)}{t}\right)$ verifies for any $h \in \mathcal{H}, z \in \mathcal{Z}$:

$$t\ \psi\left(\frac{\ell(h,z)}{t}\right) \leq t\ \psi\left(\frac{K(h)}{t}\right)$$

because ψ is non-decreasing. In this way, the exponential moment in Theorem 3 can be far more controllable. The trade-off lies in the fact that softening ℓ (instead of taking directly ℓ) will deteriorate our ability to distinguish between two bad predictions when both of them are greater than t. For instance, if we choose $\psi \in \mathcal{F}$ such as $\psi = 1$ on $[1; +\infty)$ and $t > 0$, if $\psi(\ell(h,z)/t) = 1$ for a certain pair (h,z), then we cannot tell how far $\ell(h,z)$ is from t and we only can affirm that $\ell(h,z) \geq t$.

We now move on to the following lemma, which controls the shortfall between $\mathbb{E}_{h\sim Q}[R(h)]$ and $\mathbb{E}_{h\sim Q}[R_{\psi,t}(h)]$ for all $Q \in \mathcal{M}_1^+(\mathcal{H})$, for a given ψ and $t > 0$. To do that, we assume that K admits a finite moment under any posterior distribution:

$$\forall Q \in \mathcal{M}_1^+(\mathcal{H}),\ \ \mathbb{E}_{h\sim Q}[K(h)] < +\infty. \tag{3}$$

For instance, in the case of $\mathcal{H}$ identified with a weight space $\mathcal{W} = \mathbb{R}^N$, and if K is polynomial in $||w||$ (where $||.||$ denotes the Euclidean norm), then this assumption holds if we consider Gaussian priors and posteriors.

Lemma 2. *Assume that Equation (3) holds, and let $\psi \in \mathcal{F}$, $Q \in \mathcal{M}_1^+(\mathcal{H}), t > 0$. We have:*

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}[R_{\psi,t}(h)] + \mathbb{E}_{h\sim Q}[K(h)\mathbb{1}\{K(h) \geq t\}].$$

Proof. Let $\psi \in \mathcal{F}$, $Q \in \mathcal{M}_1^+(\mathcal{H}), t > 0$. We have, for $h \in \mathcal{H}$:

$$\begin{aligned} &R(h) - R_{\psi,t}(h) \\ &\quad = \mathbb{E}_{Z\sim\mu}\left[\ell(h,Z) - t\psi\left(\frac{\ell(h,Z)}{t}\right)\right] \end{aligned}$$

and using that $\forall x \in [0,1], \psi(x) = x$,

$$= \mathbb{E}_{Z\sim\mu}\left[\left(\ell(h,Z) - t\psi\left(\frac{\ell(h,Z)}{t}\right)\right)\mathbb{1}\{\ell(h,Z) \geq t\}\right]$$

while using that $\ell(h,z) \leq K(h)$,

$$= \mathbb{E}_{Z\sim\mu}\left[\left(\ell(h,Z) - t\psi\left(\frac{\ell(h,Z)}{t}\right)\right)\mathbb{1}\{\ell(h,Z) \geq t\}\mathbb{1}\{K(h) \geq t\}\right]$$

and continuing:

$$\begin{aligned} &\leq \mathbb{E}_{Z\sim\mu}[\ell(h,Z)\mathbb{1}\{\ell(h,Z) \geq t\}]\mathbb{1}\{K(h) \geq t\} && (\psi \geq 0) \\ &\leq K(h)\mathbb{P}_{Z\sim\mu}\{\ell(h,Z) \geq t\}\mathbb{1}\{K(h) \geq t\} && (\ \ell(h,Z) \leq K(h)) \end{aligned}$$

Finally, by crudely bounding the probability by 1, we get:

$$R(h) \leq R_{\psi,t}(h) + K(h)\mathbb{1}\{K(h) \geq t\}.$$

Hence the result by integrating over $\mathcal{H}$ with respect to Q. □

Finally we present the following theorem, which provides a PAC-Bayesian inequality bounding the theoretical risk by the empirical ψ-risk for $\psi \in \mathcal{F}$.

Theorem 4. *Let ℓ be $\texttt{HYPE}(K)$ compliant, and assume K satisfies Equation (3). Then for any $P \in \mathcal{M}_1^+(\mathcal{H})$ with no data dependency, for any $\alpha \in \mathbb{R}$, for any $\psi \in \mathcal{F}$ and for any $\delta \in [0;1]$, with probability of at least $1-\delta$ over size-m random samples S, simultaneously for all Q such that $Q \sim P$ we have:*

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}\left[R_{S,\psi,t}(h)\right] + \mathbb{E}_{h\sim Q}[K(h)\mathbb{1}\{K(h) \geq t\}] + \frac{\mathrm{KL}(Q||P) + \log\left(\frac{1}{\delta}\right)}{m^\alpha} + \frac{1}{m^\alpha}\log\left(\mathbb{E}_{h\sim P}\left[\exp\left(\frac{t^2}{2m^{1-2\alpha}}\psi\left(\frac{K(h)}{t}\right)^2\right)\right]\right).$$

Proof. Let $\psi \in \mathcal{F}$, we define the ψ-loss:

$$\ell_2(h,z) = t\psi\left(\frac{\ell(h,z)}{t}\right).$$

Since ψ is non decreasing, we have for all $(h,z) \in \mathcal{H}\times\mathcal{Z}$:

$$\ell_2(h,z) \leq t\psi\left(\frac{K(h)}{t}\right) := K_2(h).$$

Thus, we apply Theorem 3 to the learning problem defined with ℓ_2: for any α and $\delta \in (0,1)$, with probability at least $1-\delta$ over size-m random samples S, simultaneously for all Q such that $Q \sim P$ we have:

$$\mathbb{E}_{h\sim Q}\left[R_{\psi,t}(h)\right] \leq \mathbb{E}_{h\sim Q}\left[R_{S,\psi,t}(h)\right] + \frac{\mathrm{KL}(Q||P) + \log\left(\frac{1}{\delta}\right)}{m^\alpha} + \frac{1}{m^\alpha}\log\left(\mathbb{E}_{h\sim P}\left[\exp\left(\frac{K_2(h)^2}{2m^{1-2\alpha}}\right)\right]\right).$$

We then add $\mathbb{E}_{h\sim Q}[K(h)\mathbb{1}\{K(h)\geq t\}]$ on both sides of the latter inequality and apply Lemma 2. □

Remark 6. *Notice that the function $\psi : x \mapsto x\mathbb{1}\{x \leq 1\} + \mathbb{1}\{x > 1\}$ is such that for any given prior P we have $\mathbb{E}_{h\sim P}\left[\exp\left(\frac{t^2}{2m^{1-2\alpha}}\psi\left(\frac{K(h)}{t}\right)^2\right)\right] < +\infty$. So the exponential moment can be controlled with a good choice of ψ. Thus the strength of Theorem 4 is to provide a PAC-Bayesian bound valid for any set of posterior measures verifying Equation (3). The choice of ψ minimising the bound is still an open problem.*

5. The Linear Regression Problem

5.1. Theoretical Result

We now focus on the celebrated linear regression problem and see how our theory translates to that particular learning problem. We assume that the data is a size-m random sample $S = (Z_i)_{i=1..m}$ where the Z_i are i.i.d. drawn from the distribution μ, and $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^N$, $Y_i \in \mathbb{R}$.

Our goal here is to find the most accurate predictor h_w (with $w \in \mathbb{R}^N$), with respect to the loss function $\ell(h_w, z) = |\langle w, x\rangle - y|$, where $z = (x, y)$. We will make the following mild assumption: there exists $B, C \in \mathbb{R}+\backslash\{0\}$ such that for all $z = (x, y)$ drawn under μ:

$$||x|| \leq B \text{ and } |y| \leq C$$

where $||.||$ is the norm associated to the classical inner product of $\mathbb{R}^N$. Under this assumption we note that for all $z = (x, y)$ drawn according to μ, we have:

$$\ell(h_w, z) = |\langle w, x\rangle - y| \leq |\langle w, x\rangle| + |y] \leq ||w||.||x|| + |y| \leq B||w|| + C.$$

Thus we define $K(h_w) = B||w|| + C$ for $w \in \mathbb{R}^N$. If we first restrict ourselves to the framework of Section 2, we want to use Theorem 3 and doing so, our goal is to bound $\xi := \mathbb{E}_{w\sim P}\left[\exp\left(\frac{K(w)^2}{2m^{1-2\alpha}}\right)\right]$. The shape of K invites us to consider a Gaussian prior. Indeed, we notice that if $P = \mathcal{N}(0, \sigma^2 \boldsymbol{I}_N)$ with $0 < \sigma^2 < \frac{m^{1-2\alpha}}{B^2}$, then $\xi < +\infty$. Notice that we cannot take just any Gaussian prior, however with a small α, the condition $0 < \sigma^2 < \frac{m^{1-2\alpha}}{B^2}$ may become quite loose. Thus, we have the following:

Theorem 5. *Let $\alpha \in \mathbb{R}$ and $N \geq 6$. Assume that the loss ℓ is $\texttt{HYPE}(K)$ compliant with $K(h) = B||h|| + C$, with $B > 0, C \geq 0$. For a prior distribution, consider any Gaussian $P = \mathcal{N}(0, \sigma^2 \boldsymbol{I}_N)$ with $\sigma^2 = t\frac{m^{1-2\alpha}}{B^2}$, $0 < t < 1$. Then, for any $\delta \in [0; 1]$, with probability of at least $1 - \delta$ over size-m random samples S, simultaneously for all $Q \in \mathcal{M}_1^+(\mathcal{H})$ such that $P \sim Q$ we have:*

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}[R_S(h)] + \frac{\mathrm{KL}(Q||P) + \log(2/\delta)}{m^\alpha} + \frac{C^2}{2m^{1-\alpha}}\left(1 + f(t)^{-1}\right) + \frac{N}{m^\alpha}\left(\log\left(1 + \left(\frac{C}{\sqrt{2f(t)m^{1-2\alpha}}}\right)\right) + \log\left(\frac{1}{\sqrt{1-t}}\right)\right)$$

where $f(t) = \frac{1-t}{t}$.

The proof is deferred to Section 7.2. To compare our result with those found in the literature, we can fix $\alpha = 1/2$. Doing so, we lose the dependency in m for the choice of the variance of the prior (which now only depends on B), but we recover the classic decreasing factor $1/\sqrt{m}$.

Remark 7. *Notice that for now we did not use Section 4, even if we could (because K is polynomial in $||w||$ and we consider Gaussian priors and posteriors, so Equation (3) is satisfied). Doing so, we obtained a bound which appears to depend linearly on the dimension N. In practice, N may be too big, and in this case, introducing an adapted softening function ψ (one can think for instance of $\psi(x) = x\mathbb{1}\{x \leq 1\} + \mathbb{1}\{x > 1\}$) is a powerful tool to attenuate the weight of the exponential moment. This also extends the class of authorised Gaussian priors by avoidance, to stick with a variance $\sigma^2 = t\frac{m^{1-2\alpha}}{B^2}$, $0 < t < 1$.*

5.2. Numerical Experiment

5.2.1. Setting

In this section we apply Theorem 5 on a concrete linear regression problem. The situation is as follows: we want to approximate the function $f(x) = \sqrt{\langle w^*, x\rangle}$, where $w^* \in \mathbb{R}^d$. We assume that $\mathcal{W} = [-c, c]^d$ so that w^* lies in an hypercube centred at 0 of half-side $c > 0$, i.e., the set $\{(w_i)_{i=1,\ldots,d} \mid \forall i, |w_i| \leq c\}$. Doing so we have $||w^*|| \leq c\sqrt{d}$.

Furthermore, we assume that input data are drawn inside a hypercube of half-side $e > 0$, i.e., $\mathcal{X} = [-e, e]^d$. Doing so we have for any data x, $||x|| \leq e\sqrt{d}$.

For any data $x \in \mathbb{R}^d$, we define $y = f(x)$. As before, we identify the hypothesis set $\mathcal{H}$ with the weight space $\mathcal{W} = \mathbb{R}^d$. As described in Section 5.1, we set $\ell(h_w, x, y) = |\langle w, x\rangle - y|$. We then remark that for any (w, x, y):

$$\begin{aligned}\ell(h_w, x, y) &\leq |\langle w, x\rangle| + |y| \leq ||w||||x|| + |\sqrt{\langle w^*, x\rangle}| \\ &\leq e\sqrt{d}||w|| + \sqrt{||w^*||.||x||} \leq e\sqrt{d}||w|| + \sqrt{c\sqrt{d}.e\sqrt{d}} \\ &\leq e\sqrt{d}||w|| + \sqrt{cde}.\end{aligned}$$

Then we can define $B = e\sqrt{d}$ and $C = \sqrt{cde}$ to apply Theorem 5. We restrict (as before) the class of distributions over $\mathcal{W}$ to be d-dimensional Gaussians:

$$\left\{\mathcal{N}(w, \sigma^2 I_d) \mid w \in \mathcal{H}, \sigma^2 \in \mathbb{R}^+\right\},$$

which is the set of candidate distributions for this learning problem. Recall that in practice, given a fixed $\alpha \in \mathbb{R}$, we are only allowed to consider priors such that their variance $\sigma^2 \in \left]0; \frac{m^{1-2\alpha}}{B^2}\right[$. We want to learn an optimised predictor (posterior) given a random dataset $S = ((X_i, Y_i))_{i=1,\ldots,m}$. To do so, we consider synthetic data.

5.2.2. Synthetic Data

We draw w^* under a Gaussian (with mean 0 and standard deviation equal to 5) truncated to the hypercube centered at 0 of the half-side $c > 0$. We generate synthetic data according to the following process: for a fixed sample size m, we draw $X_1, \ldots, X_m$ under a Gaussian (with mean 0 and standard deviation equal to 5) truncated to the hypercube centered at 0 of the half-side $e > 0$.

5.2.3. Experiment

First, we fix $c = e = 10$. Our goal here is to obtain a generalisation bound on our problem. We fix arbitrarily, for a fixed $\alpha \in \mathbb{R}$, $t_0 = 1/2$ and $\sigma_0^2 = t_0 \frac{m^{1-2\alpha}}{B^2}$ and we define our *naive prior* as $P_0 = \mathcal{N}(0, \sigma_0^2 I_d)$. For a given dataset S, we define our posterior as $Q(S) := \mathcal{N}(\hat{w}(S), \sigma^2 I_d)$, with $\sigma^2 \in \{\sigma_0^2/2, ..., \sigma_0^2/2^J\}$ ($J = \log_2(m)$), such that it is minimising the bound among candidates. Note that all the previously defined parameters are dependent on α, which is why we choose $\alpha \in \{i/\texttt{step} \mid 0 \leq i \leq \texttt{step}\}$ for `step` a fixed integer (in practice `step` = 8 or 16) and we take the value of α minimising the bound among the candidates as well. Figure 2 contains two figures, one with $d = 10$, the other with $d = 50$. On each figure are computed the right-hand side term in Theorem 5 with an optimised α for each step.

5.2.4. Discussion

To the the best of our knowledge, this is the first attempt to numerically compute PAC-Bayes bounds for unbounded problems, making it impossible to compare to other results. We stress, however, that obtaining numerical values for the bound without assuming a bounded loss is a significant first step. Furthermore, we consider a rather hard problem: f is not linear, so we cannot rely on a linear approximation fitting perfectly data, and the larger the dimension, the larger the error, as illustrated by Figure 2. Thus, for any posterior Q, the quantity $\mathbb{E}_{h\sim Q}[R(h)]$ is potentially large in practice and our bound might not be tight. Finally, notice that optimising α (instead of taking $\alpha = 1/2$ to recover a classic convergence rate) leads to a significantly better bound. A numerical example of this assertion is presented in Section 3.2. We aim to conduct further studies to consider the convergence rate as an hyperparameter to optimise, rather than selecting the same rate for all terms in the bound.

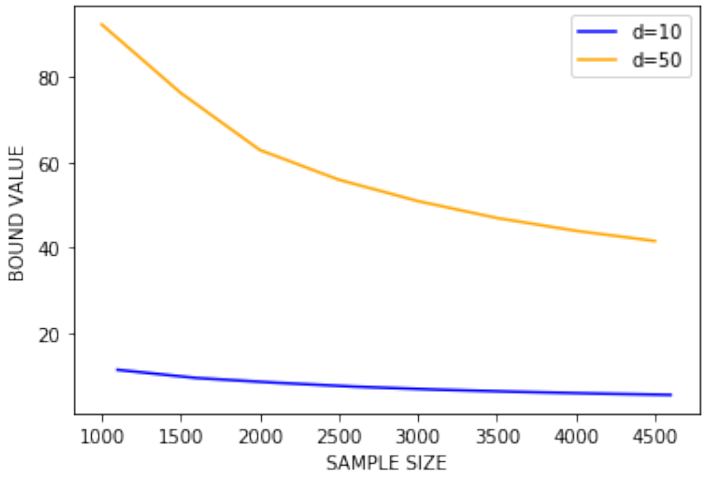

Figure 2. Evaluation of the right hand side in Theorem 5 with $d = 10$ and $d = 50$.

6. Existing Work

6.1. Germain et al., 2016

In Germain et al. [11] (Section 4), a PAC-Bayesian bound has been provided for all *sub-gamma* losses with a variance t^2 and scale parameter $c > 0$, under a data distribution μ and a prior P, i.e. losses such that for every $\lambda \in \left(0, \frac{1}{c}\right)$ the following is satisfied:

$$\log\left(\frac{1}{\delta}\mathbb{E}_{h\sim P}\mathbb{E}_S\, e^{\lambda(R(h)-R_S(h))}\right) \leq \frac{t^2}{c^2}(-\log(1-c\lambda)-\lambda c) \leq \frac{\lambda^2 t^2}{2(1-c\lambda)}.$$

Note that a sub-gamma loss (with regards to μ and P) is potentially unbounded. Germain et al. then propose the following PAC-Bayesian bound:

Theorem 6. *Ref. [11]. If the loss ℓ is sub-gamma with a variance t^2 and scale parameter c, under the data distribution μ and a fixed prior $P \in \mathcal{H}$, then for any $\delta \in [0;1]$, with probability $1-\delta$ over size-m random samples, simultaneously for all $Q \ll P$ we have:*

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}[R_S(h)] + \frac{\mathrm{KL}(Q||P) + \log(1/\delta)}{m} + \frac{t^2}{2(1-c)}.$$

Theorem 6 will be quoted several times in this paper given that it is a concrete PAC Bayesian bound provided with the will to overcome the constraint of a bounded loss. It is also one of the only one found in the literature.

Can we apply this theorem to the bounded case? The answer is yes: we remark that thanks to Hoeffding's lemma, if ℓ is bounded by $C > 0$, then for any $h \in \mathcal{H}$ it holds that $R_S(h) - R(h) \in [-C, C]$ almost surely. So, $\forall \lambda \in \mathbb{R}$, $\log \mathbb{E}_{z\sim\mu}\left[e^{\lambda(R(h)-R_S(h)}\right] \leq \frac{\lambda^2 C^2}{2}$. Therefore, for any prior P, we have:

$$\log \mathbb{E}_{h\sim P}\mathbb{E}_{z\sim\mu}\left[e^{\lambda(R(h)-R_S(h)}\right] \leq \frac{\lambda^2 C^2}{2}.$$

Thus, ℓ is sub-gamma with variance C^2 and scale parameter 0. Then, Theorem 6 can be applied with $t^2 = C^2$, $c = 0$.

Comparison with Proposition 1. We remark that by taking $K = C$ and $\alpha = 1$ in Proposition 1, we are recovering Theorem 6. However, our approach allows us to say that if

we can obtain a more precise form of K such that $\forall h \in \mathcal{H}$, $K(h) \leq C$ and K is non-constant, Theorem 3, will ensure that:

$$\frac{1}{m^\alpha} \log\left(\mathbb{E}_{h\sim P}\left[\exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right)\right]\right) \leq \frac{C^2}{2m^{1-\alpha}}.$$

Thus, having precise information on the behavior of the loss function ℓ, with regards to the predictor h, allows us to obtain a tighter control of the exponential moment, and hence a tighter bound.

Remark 8. *We can see that Theorem 6 cannot control the factor $C^2/2$. However, Ref. [11] remarked on this apparent weakness and partially corrected this issue [11] (Section 4, Equations (13) and (14)). Indeed, they proposed to balance the influence of m between the different terms of the PAC-Bayes bound by providing the same convergence rate in $1/\sqrt{m}$ to all terms.*

We can then see Proposition 1 as a proper generalisation of Germain et al. [11] (Section 4, Equations (13) and (14)). Indeed, our bound exhibits properly the influence of the parameter α. Thus, we understand (and Lemma 1 proves it) that the choice of α deserves a study in itself in the way it is now a parameter of our optimisation problem. This fact has already been highlighted in Alquier et al. [10] (Theorem 4.1) (where $\lambda := m^\alpha$).

6.2. Holland, 2019

In [13], Holland proposed a PAC Bayesian inequality with unbounded loss. For that, he introduced a function ψ verifying a few specific conditions, different to those used in Section 4 to define our set of softening functions. Indeed, he considered a function ψ such that:

- ψ is bounded,
- ψ is non decreasing,
- it exists $b > 0$ such that for all $u \in \mathbb{R}$:

$$-\log\left(1 - u + \frac{u^2}{b}\right) \leq \psi(u) \leq \log\left(1 + u + \frac{u^2}{b}\right). \tag{4}$$

We remark that, as Holland did, we supposed that our softening functions are non-decreasing. We chose softening functions to be equal to the identity function ($x \mapsto x$) on $[0, 1]$, which is quite restrictive. However, we are imposing softening functions to be lesser than the identity on $[1, +\infty)$; whereas, Holland supposed ψ to be bounded and satisfy Equation (4). A concrete example of such a function ψ, lies in the piecewise polynomial function of Catoni and Giulini [21], defined by:

$$\psi(u) = \begin{cases} -2\sqrt{2}/3 & \text{if } u \leq -\sqrt{2} \\ u - u^3/6 & \text{if } u \in [-2\sqrt{2}/3, 2\sqrt{2}/3] \\ 2\sqrt{2}/3 & \text{otherwise.} \end{cases}$$

As in Section 4, we are considering the ψ-empirical risk $R_{S,\psi,t}$ for any $t > 0$. Holland provided his theorem given the fact the following assumptions are realised:

- Bounds on lower-order moments. For all $h \in \mathcal{H}$, we have $\mathbb{E}_{Z\sim\mu}[\ell(h, Z)^2] \leq M_2 < +\infty$ and $\mathbb{E}_{Z\sim\mu}[\ell(h, Z)^3] \leq M_3 < +\infty$.
- Bounds on the risk. For all $h \in \mathcal{H}$, we suppose $R(h) \leq \sqrt{mM_2/(4\log(\delta^{-1}))}$.
- Large enough confidence, we require $\delta \leq e^{-1/9}$.

Now we can state Holland's theorem.

Theorem 7. *Ref. [13]. Let P be a prior distribution on model $\mathcal{H}$. Let the three assumptions listed above hold. Setting $t^2 = mM_2/(2\log(\delta^{-1}))$, then for any $\delta \in [0;1]$, with probability of at least $1-\delta$ over the random draw of the size-m sample S, simultaneously for all Q it holds that:*

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}\left[R_{S,\psi,t}(h)\right] + \frac{1}{\sqrt{m}}\left(\mathrm{KL}(Q||P) + \frac{1}{2}\log\left(\frac{8\pi M_2}{\delta^2}\right) - 1\right) + \frac{1}{\sqrt{m}}\nu^*(\mathcal{H}) + O\left(\frac{1}{m}\right)$$

where:

$$\nu^*(\mathcal{H}) := \frac{\mathbb{E}_{h\sim P}\left[\exp\left(\sqrt{m}(R(h) - R_{S,\psi,t}(h))\right)\right]}{\mathbb{E}_{h\sim P}\left[\exp\left(R(h) - R_{S,\psi,t}(h)\right)\right]}.$$

7. Proofs

7.1. Proof of Theorem 1

Proof. Let $F : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}$ be a convex function, P a fixed prior, and $\delta \in [0,1]$. Since $\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]$ is a nonnegative random variable, we know that, by Markov's inequality, for any $h \in \mathcal{H}$:

$$\mathbb{P}\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right] > \frac{1}{\delta}\mathbb{E}_S\ \mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right) \leq \delta.$$

So with probability of at least $1-\delta$, we have:

$$\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right] \leq \frac{1}{\delta}\mathbb{E}_S\ \mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right] = \frac{\chi}{\delta}.$$

Applying the log function on each side of this inequality gives us with probability of at least $1-\delta$ over samples S:

$$\log\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right) \leq \log\left(\frac{\chi}{\delta}\right).$$

We now rename $A := \log\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right)$.

Furthermore, if we denote by $\frac{dQ}{dP}$ the Radon-Nikodym derivative of Q with respect to P when $Q \ll P$, we then have, for all Q such that $Q \sim P$:

$$\begin{aligned} A &= \log\left(\mathbb{E}_{h\sim Q}\left[\frac{dP}{dQ}e^{F(R_S(h),R(h))}\right]\right) \\ &= \log\left(\mathbb{E}_{h\sim Q}\left[\left(\frac{dQ}{dP}\right)^{-1}e^{F(R_S(h),R(h))}\right]\right) \qquad \left(\tfrac{dP}{dQ} = \left(\tfrac{dQ}{dP}\right)^{-1}\right) \end{aligned}$$

and by concavity of log and Jensen's inequality,

$$\begin{aligned} &\geq -\mathbb{E}_{h\sim Q}\left[\log\left(\frac{dQ}{dP}\right)\right] + \mathbb{E}_{h\sim Q}[F(R_S(h),R(h))] \\ &= -\mathrm{KL}(Q||P) + \mathbb{E}_{h\sim Q}[F(R_S(h),R(h))] \end{aligned}$$

while by convexity of F with Jensen's inequality,

$$\geq -\mathrm{KL}(Q||P) + F\left(\mathbb{E}_{h\sim Q}[R_S(h)], \mathbb{E}_{h\sim Q}[R(h)]\right).$$

Hence, for Q such that $Q \sim P$,

$$F\big(\mathbb{E}_{h\sim Q}[R_S(h)], \mathbb{E}_{h\sim Q}[R(h)]\big) \leq \mathrm{KL}(Q||P) + A.$$

So with probability $1-\delta$, for Q such that $Q \sim P$,

$$F\big(\mathbb{E}_{h\sim Q}[R_S(h)], \mathbb{E}_{h\sim Q}[R(h)]\big) \leq \mathrm{KL}(Q||P) + \log\Big(\frac{\chi}{\delta}\Big).$$

This completes the proof of Theorem 1. □

7.2. Proof of Theorem 5

We first provide a technical property. Recall that:

$$\xi = \mathbb{E}_{h\sim P}\left[\exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right)\right].$$

Proposition 3. *Let $\alpha \in \mathbb{R}$. Suppose the loss ℓ is $\mathtt{HYPE}(K)$ compliant with $K(h) = B||h|| + C$, with $B > 0$, $C \geq 0$. Then, for any Gaussian prior $P = \mathcal{N}(0, \sigma^2 \boldsymbol{I}_N)$ with $\sigma^2 = t\frac{m^{1-2\alpha}}{B^2}$, $0 < t < 1$ and $N \geq 6$ we have:*

$$\xi \leq 2\exp\left(\frac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right)\frac{1}{\left(\sqrt{1-t}\right)^N}\left(1+\left(\frac{C}{\sqrt{2f(t)m^{1-2\alpha}}}\right)\right)^{N-1}$$

with $f(t) = \frac{1-t}{t}$.

Proof. We recall that $\sigma^2 = t\frac{m^{1-2\alpha}}{B^2}$. By expliciting the expectation and $K(h)$ we thus obtain:

$$\begin{aligned}
\xi &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \int_{h\in\mathbb{R}^N} \exp\left(\frac{(B||h||+C)^2}{2m^{1-2\alpha}} - \frac{||h||^2B^2}{2tm^{1-2\alpha}}\right)dh \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \int_{h\in\mathbb{R}^N} \exp\left(-\frac{1}{2m^{1-2\alpha}}\left(f(t)B^2||h||^2 - 2BC||h|| - C^2\right)\right)dh \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \int_{h\in\mathbb{R}^N} \exp\left(-\frac{B^2f(t)}{2m^{1-2\alpha}}\left(||h||^2 - \frac{2C||h||}{Bf(t)} - \frac{C^2}{B^2f(t)}\right)\right)dh \\
&= \exp\left(\frac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right)\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\int_{h\in\mathbb{R}^N}\exp\left(-\frac{B^2f(t)}{2m^{1-2\alpha}}\left(||h|| - \frac{C}{Bf(t)}\right)^2\right)dh.
\end{aligned}$$

We will use the spherical coordinates in N-dimensional Euclidean space given in [22]:

$$\varphi : (h_1, ..., h_N) \to (r, \varphi_1, ..., \varphi_{N-1})$$

where especially $r = ||h||$ and also the Jacobian of ϕ is given by:

$$d^NV = r^{N-1}\prod_{k=1}^{N-2}\sin^k(\varphi_{N-1-k}) = r^{N-1}d_{S^{N-1}}V.$$

Let us also precise that as given in Blumenson [22] (page 66), we have that the surface of the sphere of radius 1 in N-dimensional space is:

$$\int_{\varphi_1,\ldots,\varphi_{N-1}} d_{S^{N-1}}V\, d\varphi_1 \ldots d\varphi_{N-1} = \frac{2\sqrt{\pi}^N}{\Gamma\left(\frac{N}{2}\right)}$$

where Γ is the Gamma function defined as:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt \quad \text{for } x > -1.$$

Then, if we set:

$$A := \int_{h \in \mathbb{R}^N} \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} \left(||h|| - \frac{C}{Bf(t)} \right)^2 \right) dh$$

we obtain by a change of variable:

$$\begin{aligned} A &= \int_{r,\varphi_1,\ldots,\varphi_{N-1}} \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} \left(r - \frac{C}{Bf(t)} \right)^2 \right) d^N V \, dr d\varphi_1 \ldots d\varphi_{N-1} \\ &= \left(\frac{2\sqrt{\pi}^N}{\Gamma\left(\frac{N}{2}\right)} \right) \int_{r=0}^{+\infty} \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} \left(r - \frac{C}{Bf(t)} \right)^2 \right) r^{N-1} dr \\ &= \left(\frac{2\sqrt{\pi}^N}{\Gamma\left(\frac{N}{2}\right)} \right) \int_{r=-\frac{C}{Bf(t)}}^{+\infty} \left(r + \frac{C}{Bf(t)} \right)^{N-1} \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} r^2 \right) dr \\ &= \left(\frac{2\sqrt{\pi}^N}{\Gamma\left(\frac{N}{2}\right)} \right) \sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{Bf(t)} \right)^{N-k-1} \int_{r=-\frac{C}{Bf(t)}}^{+\infty} r^k \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} r^2 \right) dr. \end{aligned}$$

We fix a random variable X such that:

$$X \sim \mathcal{N}\left(0, \frac{m^{1-2\alpha}}{B^2 (f(t)} \right).$$

We then have for any k positive integer, if k is even:

$$\begin{aligned} \int_{r=-\frac{C}{Bf(t)}}^{+\infty} r^k \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} r^2 \right) dr &\leq \int_{r=-\infty}^{+\infty} r^k \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} r^2 \right) dr \\ &\leq \sqrt{2\pi \frac{m^{1-2\alpha}}{B^2 f(t)}} \mathbb{E}[|X|^k]. \end{aligned}$$

And if k is odd:

$$\begin{aligned} \int_{r=-\frac{C}{Bf(t)}}^{+\infty} r^k \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} r^2 \right) dr &\leq \int_{r=0}^{+\infty} r^k \exp\left(-\frac{B^2 f(t)}{2m^{1-2\alpha}} r^2 \right) dr \\ &\leq \sqrt{2\pi \frac{m^{1-2\alpha}}{B^2 f(t)}} \mathbb{E}[|X|^k \mathbb{1}(X \geq 0)] \\ &\leq \sqrt{2\pi \frac{m^{1-2\alpha}}{B^2 f(t)}} \mathbb{E}[|X|^k]. \end{aligned}$$

So we have:

$$A \leq \left(\frac{2\sqrt{\pi}^N}{\Gamma\left(\frac{N}{2}\right)} \right) \sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{Bf(t)} \right)^{N-k-1} \sqrt{2\pi \frac{m^{1-2\alpha}}{B^2 f(t)}} \mathbb{E}[|X|^k].$$

As precised in [23], we have for any k:

$$\mathbb{E}[|X|^k] = \left(\sqrt{\frac{m^{1-2\alpha}}{B^2 f(t)}} \right)^k 2^{k/2} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi}}.$$

So finally:

$$A \leq 2\sqrt{\pi}^N \sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{Bf(t)}\right)^{N-k-1} \left(\sqrt{\frac{2m^{1-2\alpha}}{B^2 f(t)}}\right)^{k+1} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{N}{2}\right)}.$$

Lemma 3. *If $N \geq 6$, then:*

$$\max_{k=0..N-1} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} = 1.$$

Proof. As precised in the introduction of Srinivasan and Zvengrowski [24], Gauss [25] (page 147) proved that on the interval $[x_0, +\infty)$ where $x_0 \in [1.46, 1.47]$, Γ is a monotonic increasing function. So, for $N-1 \geq k \geq 2, \Gamma(\frac{k+1}{2}) \leq \Gamma(\frac{N}{2})$. And because $\Gamma(1/2) = \sqrt{\pi}, \Gamma(1) = 1$, we have:

$$\max_{k=0..N-1} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} = \max\left(\frac{\sqrt{\pi}}{\Gamma\left(\frac{N}{2}\right)}, \frac{\Gamma\left(\frac{N-1+1}{2}\right)}{\Gamma\left(\frac{N}{2}\right)}\right) = \max\left(\frac{\sqrt{\pi}}{\Gamma\left(\frac{N}{2}\right)}, 1\right)$$

Because $N \geq 6$, and Γ is monotone and increasing on $[3; +\infty]$, we have $\Gamma(N/2) \geq \Gamma(3) \geq \sqrt{\pi}$. Hence the result. □

Using Lemma 3 allows us to write:

$$A \leq 2\sqrt{\pi}^N \sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{Bf(t)}\right)^{N-k-1} \left(\sqrt{\frac{2m^{1-2\alpha}}{B^2 f(t)}}\right)^{k+1}.$$

We recall that $\sigma^2 = t\frac{m^{1-2\alpha}}{B^2}$ and $f(t) = \frac{1-t}{t}$. Then we can write:

$$A \leq 2\sqrt{\pi}^N \sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{Bf(t)}\right)^{N-k-1} \left(\sqrt{\frac{2\sigma^2}{1-t}}\right)^{k+1}.$$

We now conclude with the final bound on ξ:

$$\begin{aligned}
\xi &\leq \exp\left(\tfrac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right) \tfrac{1}{(\sqrt{2\pi\sigma^2})^N} A \\
&\leq \exp\left(\tfrac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right) \tfrac{1}{(\sqrt{2\pi\sigma^2})^N} 2\sqrt{\pi}^N \textstyle\sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{Bf(t)}\right)^{N-k-1} \left(\sqrt{\frac{2\sigma^2}{1-t}}\right)^{k+1} \\
&\leq 2\exp\left(\tfrac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right) \textstyle\sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{Bf(t)}\right)^{N-k-1} \left(\sqrt{\frac{1}{1-t}}\right)^{k+1} \left(\sqrt{\frac{B^2}{2tm^{1-2\alpha}}}\right)^{N-k-1} \\
&\leq 2\exp\left(\tfrac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right) \textstyle\sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C\sqrt{t}}{(1-t)\sqrt{2m^{1-2\alpha}}}\right)^{N-k-1} \left(\sqrt{\frac{1}{1-t}}\right)^{k+1} \\
&\leq 2\frac{\exp\left(\frac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right)}{\left(\sqrt{1-t}\right)^N} \textstyle\sum_{k=0}^{N-1} \binom{N-1}{k} \left(\frac{C}{\sqrt{2f(t)m^{1-2\alpha}}}\right)^{N-k-1} \\
&\leq 2\frac{\exp\left(\frac{C^2}{2m^{1-2\alpha}f(t)}(1+f(t))\right)}{\left(\sqrt{1-t}\right)^N} \left(1 + \left(\frac{C}{\sqrt{2f(t)m^{1-2\alpha}}}\right)\right)^{N-1}.
\end{aligned}$$

This completes the proof of Proposition 3. □

Proof of Theorem 5. We combine Theorem 3 with Proposition 3. We also upper-bound $N-1$ by N. □

Author Contributions: Conceptualization, M.H., B.G. and J.S.-T.; Formal analysis, M.H., B.G. and O.R.; Project administration, B.G.; Supervision, B.G.; Writing—original draft, M.H., B.G. and O.R.; Writing—review and editing, M.H., B.G., O.R. and J.S.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the U.S. Army Research Laboratory and the U. S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1. BG acknowledges partial support from the French National Agency for Research, grants ANR-18-CE40-0016-01 and ANR-18-CE23-0015-02.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shawe-Taylor, J.; Williamson, R.C. A PAC analysis of a Bayes estimator. In Proceedings of the 10th Annual Conference on Computational Learning Theory, Nashville, TN, USA, 6–9 July 1997; ACM: New York, NY, USA, 1997; pp. 2–9.
2. McAllester, D.A. Some PAC-Bayesian theorems. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; ACM: New York, NY, USA, 1998; pp. 230–234.
3. McAllester, D.A. PAC-Bayesian model averaging. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory, Santa Cruz, CA, USA, 7–9 July 1999; ACM: New York, NY, USA, 1999; pp. 164–170.
4. Guedj, B. A Primer on PAC-Bayesian Learning. *arXiv* **2019**, arXiv:stat.ML/1901.05353.
5. Rivasplata, O.; Kuzborskij, I.; Szepesvári, C.; Shawe-Taylor, J. PAC-Bayes Analysis Beyond the Usual Bounds. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, Online, 6–12 December 2020.
6. Seeger, M. PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification. *J. Mach. Learn. Res.* **2002**, *3*, 233–269.
7. Langford, J. Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.* **2005**, *6*, 273–306.
8. Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*; Institute of Mathematical Statistics: Waite Hill, OH, USA, 2007.
9. Germain, P.; Lacasse, A.; Laviolette, F.; Marchand, M. PAC-Bayesian Learning of Linear Classifiers. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 353–360.
10. Alquier, P.; Ridgway, J.; Chopin, N. On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **2016**, *17*, 1–41.
11. Germain, P.; Bach, F.; Lacoste, A.; Lacoste-Julien, S. PAC-Bayesian Theory Meets Bayesian Inference. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2016; pp. 1884–1892.
12. Alquier, P.; Guedj, B. Simpler PAC-Bayesian bounds for hostile data. *Mach. Learn.* **2018**, *107*, 887–902. [CrossRef]
13. Holland, M. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 2715–2724.
14. Kuzborskij, I.; Szepesvári, C. Efron-Stein PAC-Bayesian Inequalities. *arXiv* **2019**, arXiv:1909.01931.
15. Shalaeva, V.; Fakhrizadeh Esfahani, A.; Germain, P.; Petreczky, M. Improved PAC-Bayesian Bounds for Linear Regression. In Proceedings of the AAAI 2020—Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
16. Lever, G.; Laviolette, F.; Shawe-Taylor, J. Distribution-Dependent PAC-Bayes Priors. In *Algorithmic Learning Theory*; Hutter, M., Stephan, F., Vovk, V., Zeugmann, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 119–133.
17. Lever, G.; Laviolette, F.; Shawe-Taylor, J. Tighter PAC-Bayes Bounds through Distribution-Dependent Priors. *Theor. Comput. Sci.* **2013**, *473*, 4–28. [CrossRef]
18. Mhammedi, Z.; Grünwald, P.; Guedj, B. PAC-Bayes Un-Expected Bernstein Inequality. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 12202–12213.
19. Parrado-Hernández, E.; Ambroladze, A.; Shawe-Taylor, J.; Sun, S. PAC-Bayes bounds with data dependent priors. *J. Mach. Learn. Res.* **2012**, *13*, 3507–3531.
20. Catoni, O. Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. H. Poincaré Probab. Statist.* **2012**, *48*, 1148–1185. [CrossRef]

21. Catoni, O.; Giulini, I. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv* **2017**, arXiv:math.ST/1712.02747.
22. Blumenson, L.E. A Derivation of n-Dimensional Spherical Coordinates. *Am. Math. Mon.* **1960**, *67*, 63–66. [CrossRef]
23. Winkelbauer, A. Moments and Absolute Moments of the Normal Distribution. *arXiv* **2012**, arXiv:math.ST/1209.4340.
24. Srinivasan, G.K.; Zvengrowski, P. On the Horizontal Monotonicity of $|\Gamma(s)|$. *Can. Math. Bull.* **2011**, *54*, 538–543. [CrossRef]
25. Gauss, C.F. Disquisitiones Generales Circa Seriem Infinitam (reprint). In *Werke*; Cambridge University Press: Cambridge, UK, 2011; Volume 3.