entropy

Article

Still No Free Lunches: The Price to Pay for Tighter PAC-Bayes Bounds

Benjamin Guedj [1,2,*] **and Louis Pujol** [3,*]

[1] Centre for Artificial Intelligence, Department of Computer Science, University College London, London WC1V 6LJ, UK
[2] Inria Lille—Nord Europe Research Centre and Inria London, 59800 Lille, France
[3] Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, CNRS, 91405 Orsay, France
[*] Correspondence: b.guedj@ucl.ac.uk (B.G.); louis.pujol@universite-paris-saclay.fr (L.P.)

Abstract: "No free lunch" results state the impossibility of obtaining meaningful bounds on the error of a learning algorithm without prior assumptions and modelling, which is more or less realistic for a given problem. Some models are "expensive" (strong assumptions, such as sub-Gaussian tails), others are "cheap" (simply finite variance). As it is well known, the more you pay, the more you get: in other words, the most expensive models yield the more interesting bounds. Recent advances in robust statistics have investigated procedures to obtain tight bounds while keeping the cost of assumptions minimal. The present paper explores and exhibits what the limits are for obtaining tight probably approximately correct (PAC)-Bayes bounds in a robust setting for cheap models.

Keywords: statistical learning theory; PAC-Bayes theory; no free lunch theorems

check for updates

Citation: Guedj, B.; Pujol, L. Still No Free Lunches: The Price to Pay for Tighter PAC-Bayes Bounds. *Entropy* **2021**, *23*, 1529. https://doi.org/10.3390/e23111529

Academic Editor: Wray Buntine

Received: 29 August 2021
Accepted: 3 November 2021
Published: 18 November 2021

1. Introduction

For the sake of clarity, we focus on the supervised learning problem. We collect a sequence of input–output pairs $(X_i, Y_i)_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$, which we assume to be N independent realisations of a random variable drawn from a distribution P on $\mathcal{X} \times \mathcal{Y}$. The overarching goal in statistics and machine learning is to select a hypothesis f over a space $\mathcal{F}$ which, given a new input x in $\mathcal{X}$, delivers an output $f(x)$ in $\mathcal{Y}$, hopefully close (in a certain sense) to the unknown true output y. The quality of f is assessed through a loss function ℓ which characterises the discrepancy between the true output y and its prediction $f(x)$, and we define a global notion of risk as

$$R(f) = \mathbb{E}_{(X,Y)\sim \mathrm{P}}[\ell(f(X), Y)].$$

The aim of machine learning is to find a good (in the sense of a low risk) hypothesis $f \in \mathcal{F}$. In the generalised Bayes setting, the learning algorithm does not output a single hypothesis but rather a *distribution* ρ over the hypotheses space $\mathcal{F}$ and the associated bounds are called PAC-Bayesian bounds (see [1] for a survey of the topic).

As many probabilistic bounds stated in the statistics and machine learning literature, PAC-Bayesian bounds (where PAC stands for probably approximately correct—see [2]) commonly requires strong assumptions to hold, such as sub-Gaussian behaviour of some random variables. These assumptions can be misleading when dealing with true data as they do not take into account some practical situations, such as outlier contamination. Many efforts have been made recently to keep tight generalisation bounds valid with a few set of assumptions about the underlying distribution: this is known as robust learning [see [3] for a survey of the topic].

In this work we explore the possibility to establish a connection between recent techniques introduced by robust machine learning and PAC-Bayesian generalisation bounds. The result of our work is negative as we were not able to prove a PAC-Bayes bound in a robust statistics setting. However, we found it useful to write down our findings in order to give the interested reader a review of material involved in both robust statistics and PAC-Bayes theory and present the fundamental issues we faced as we believe it to be useful to the community.

Organisation of the paper. We introduce an elementary example and set a basic notation to illustrate the problem of robustness in Section 2, before providing an overview of recent advances in robust statistics in Section 3, and briefly introduce the field of PAC-Bayes learning in Section 4. We then propose in Section 5 a detailed study of the structural limits which do not allow for PAC-Bayes bounds which are simultaneously tight without requiring strong assumptions. The paper closes with a discussion in Section 6.

2. About the "No Free Lunch" Results

A class of results in statistics is known as "no free lunch" statements [see [4], Chapter 7]. The "no free lunch" results typically state that if one does not consider the restrictions on the modelling of the data-generating process, one cannot obtain meaningful deviation bounds in a non-asymptotic regime. The well-known trade-off is that the more restrictive the assumptions, the tighter the bounds. Let us illustrate this classical phenomenon by a simple example.

Assume that we have a dataset consisting in N real observations $x_1, \dots, x_N \in \mathbb{R}$ and consider they are independent, identically distributed (iid) realisations of a random variable X. Our goal is to estimate the mean of X and build a confidence interval for this estimate. As a start, let us focus on the empirical mean, denoted by $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. As "no free lunch" results state, we have to consider a class of distributions to which the data-generating distribution P belongs.

2.1. Expensive and Cheap Models

If there is always a price to pay in order to derive insightful result, there is a variety of degrees of restrictions. In the remainder of the paper, we will focus on two classical models corresponding to a different level of demand on the random variables.

A first type of restriction we can make is an "expensive modelling". For $\sigma > 0$, let $\mathcal{P}^{\sigma}_{\text{expensive}}$ be the set of all real-valued random variables X satisfying:

$$\log(\mathbb{E}[\exp\{\lambda(X - \mathbb{E}[X])\}]) \leq \frac{\lambda^2\sigma^2}{2}.$$

This $\mathcal{P}^{\sigma}_{\text{expensive}}$ is the class of sub-Gaussian random variables with variance factor σ^2 [see [5] for a complete coverage of the topic]. We call this model "expensive" as this restriction is often considered unrealistic for real-life datasets and is hard or impossible to check in practice.

An alternative type of restriction is a "cheap modelling". For $\sigma > 0$, let $\mathcal{P}^{\sigma}_{\text{cheap}}$ be the set of real-valued random variables with a finite variance, upper bounded by σ^2. We call this model "cheap" as this is considerably less restrictive than the expensive one and is much more likely to hold in practice.

2.2. Confidence Interval for the Empirical Mean

Proposition 1 (Confidence intervals). *If we assume that $X \in \mathcal{P}^{\sigma}_{expensive}$, then for all $\delta \in (0, 1/2)$, the following random interval is a confidence interval for the mean of X at level $1 - \delta$:*

$$\left[\ \bar{x} \pm \frac{\sigma}{\sqrt{N}}\sqrt{2} \times \sqrt{2\log\left(\frac{1}{\delta}\right)}\ \right]. \tag{1}$$

If we assume that $X \in \mathcal{P}^{\sigma}_{cheap}$, then for all $\delta \in (0, 1)$, the following random interval is a confidence interval for the mean of X at level $1 - \delta$:

$$\left[\bar{x} \pm \frac{\sigma}{\sqrt{N}} \sqrt{\frac{1}{\delta}} \right]. \tag{2}$$

In the case of a cheap model, there is no hope to obtain a significantly tighter confidence interval with respect to δ if one uses the empirical mean [as proved in [6], Proposition 6.2].

Proof. To establish the first confidence interval (1), we first remark that if $X \in \mathcal{P}^{\sigma}_{\text{expensive}}$, then $\bar{x} \in \mathcal{P}^{\sigma/\sqrt{N}}_{\text{expensive}}$ and $\mathbb{E}[\bar{x}] = \mathbb{E}[X]$. So, applying Theorem 2.1 of [5] to $\bar{x} - \mathbb{E}[X]$ we obtain, for all $a > 0$:

$$\begin{aligned} \mathbb{P}(|\bar{x} - \mathbb{E}[X]| > a) &= \mathbb{P}(\bar{x} - \mathbb{E}[X] > a) + \mathbb{P}(\bar{x} - \mathbb{E}[X] < -a) \\ &\leq 2\max(\mathbb{P}(\bar{x} - \mathbb{E}[X] > a), \mathbb{P}(\bar{x} - \mathbb{E}[X] < -a)) \\ &\leq 2\exp\left(-\frac{Na^2}{2\sigma^2}\right). \end{aligned}$$

Setting $\delta = \exp\left(-\frac{Na^2}{2\sigma^2}\right)$ leads to the expected result. The second confidence interval (2) is obtained through Chebychev's inequality. $\mathbb{E}[\bar{x}] = \mathbb{E}[X]$ and as $X \in \mathcal{P}^{\sigma}_{\text{cheap}}$, $\text{Var}(\bar{x}) = \frac{\text{Var}(X)}{N} \leq \frac{\sigma^2}{N}$. So for all $a > 0$

$$\mathbb{P}(|\bar{x} - \mathbb{E}[X]| > a) \leq \frac{\sigma^2}{Na^2}.$$

Now, setting $\delta = \frac{\sigma^2}{Na^2}$ we get

$$\mathbb{P}\left(|\bar{x} - \mathbb{E}[X]| > \frac{\sigma}{\sqrt{N}} \sqrt{\frac{1}{\delta}} \right) \leq \delta.$$

□

Note that the dependence in δ is fairly different in both confidence intervals defined in (1) and (2): for fixed σ^2 and N, the $\sqrt{2} \times \sqrt{2\log(1/\delta)}$ regime (following the lunch metaphor, the "good lunch") is much more favourable than the $1/\sqrt{\delta}$ regime (the "bad lunch"). We illustrate this in Figure 1, where we plot $\sqrt{2} \times \sqrt{2\log(1/\delta)}$ and $1/\sqrt{\delta}$ as a function of $\delta \in (0, 1/2)$. We remark that for small values of δ, corresponding to a higher confidence level, the interval (1) will be much tighter than (2).

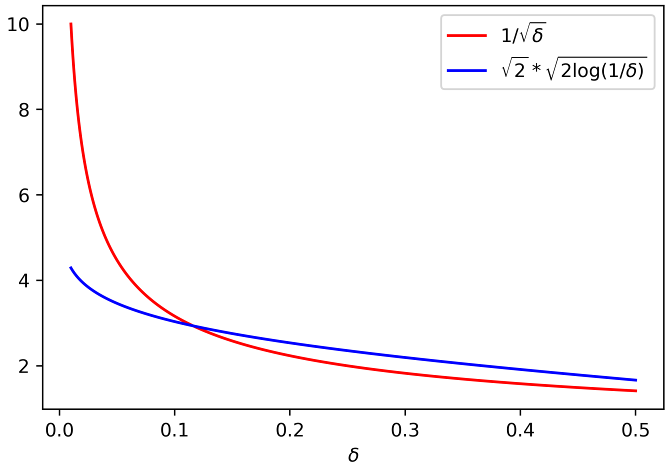

Figure 1. $\sqrt{2} \times \sqrt{2\log(1/\delta)}$ and $1/\sqrt{\delta}$ with respect to δ.

So, while it is clear that the best confidence interval requires more stringent assumptions, there have been attempts at relaxing those assumptions—or in other words, keeping equally good lunches at a cheaper cost.

3. Robust Statistics

Robust statistics address the following question: can we obtain tight bounds with minimal assumptions—or in other words, can we get a good cheap lunch? In the mean estimation case hinted in Section 2, the question becomes the following: if $\mathrm{P} \in \mathcal{P}^{\sigma}_{\text{cheap}}$, can we build a confidence interval at level $1-\delta$ with a size proportional to $\frac{\sigma}{\sqrt{N}}\sqrt{2\log(1/\delta)}$?

As mentioned above, there is no hope to achieve this goal with the empirical mean. Different alternative estimators have thus been considered in robust statistics, such as M-estimators [6] or median-of-means (MoM) estimators [see [7] for a recent survey, and references therein].

The key idea of MoM estimators is to achieve a compromise between the unbiased but non-robust empirical mean and the biased but robust median. As before, let us consider a sample of N real numbers $x_1, \dots, x_N$, assumed to be an iid sequence drawn from a distribution P. Let $K \leq N$ be a positive integer and assume for simplicity that K is a divisor of N. To compute the MoM estimator, the first step consists of dividing the sample $(x_1, \dots, x_N)$ into K non-overlapping blocks $B_1, \dots, B_K$, each of length N/K. For each block, we then compute the empirical mean

$$\bar{x}_{B_i} = \frac{K}{N}\sum_{j \in B_i} x_j.$$

The MoM estimator is defined as the median of those means:

$$\mathrm{MoM}_K(x_1 \dots, x_N) = \mathrm{median}\{\bar{x}_{B_1}, \dots, \bar{x}_{B_K}\}.$$

This estimator has the following nice property.

Proposition 2 ([7], Proposition 12). *Assume* $\mathrm{P} \in \mathcal{P}^{\sigma}_{cheap}$, *for* $\delta = \exp\left(-\frac{K}{8}\right)$,

$$\left[\mathrm{MOM}_K \pm \frac{\sigma}{\sqrt{N}} \times 4\sqrt{2\log\left(\frac{1}{\delta}\right)}\right] \tag{3}$$

is a confidence interval for the mean of X at the level $1-\delta$.

This property is quite encouraging, as for a cheap model we obtain a confidence interval similar, up to a numerical constant, to the best one (1) in Section 2. However, we also spot here an important limitation. The confidence interval (3) for MoM is only valid for the particular error threshold $\delta = \exp(-K/8)$, which depends on the number of blocks K (a parameter for the estimator MoM_K). The estimator must be changed each time we want to evaluate a different confidence level.

An ever more limiting feature is that the error threshold δ is constrained and cannot be set arbitrarily small, as in (1) or (2). Obviously, the number of blocks cannot exceed the sample size N, and the error threshold reaches its lowest tolerable value $\exp(-N/8)$. In other words, the interval defined in (3) can have confidence at most $1 - \exp(-N/8)$.

Is this strong limitation specific to MoM estimators? No, say [8], [Theorem 3.2 and following remark]. This limitation is universal; over the class $\mathcal{P}^{\sigma}_{\text{cheap}}$, there is no estimator $\hat{x}$ of the mean such that there exists a constant $L > 1$ such that

$$\left[\hat{x} \pm \frac{\sigma}{\sqrt{N}} \times L\sqrt{2\log\left(\frac{1}{\delta}\right)}\right]$$

is a confidence interval at level $1-\delta$ for δ lower than $e^{-\mathcal{O}(N)}$.

To sum up, a good and cheap lunch is possible, with the limitation that the bound is no longer valid for all confidence levels.

4. PAC-Bayes

We now briefly introduce the generalised Bayesian setting in machine learning, and the resulting generalisation bounds, the PAC-Bayesian bounds. PAC-Bayes is a sophisticated framework to derive new learning algorithms and obtain (often state-of-the-art) generalisation bounds, while maintaining probability distributions over hypotheses; as such, we are interested in studying how PAC-Bayes is compatible with good and cheap lunches. We refer the reader to [1,9] and the many references therein for recent surveys on PAC-Bayes including historical notes and main bounds. We focus on classical bounds from the PAC-Bayes literature, based on the empirical risk as a risk estimator—and we instantiate those bounds in two regimes matching the “expensive” and “cheap” models introduced in Section 2.

4.1. Notation

For any $f \in \mathcal{F}$, we define the empirical risk $R_N(f)$ as:

$$R_N(f) = \frac{1}{N}\sum_{i=1}^{N} \ell(f(X_i), Y_i).$$

In the following, we consider integrals over the hypotheses space $\mathcal{F}$. To keep the notation as compact as possible, we will write $\mu[g] = \int g \mathrm{d}\mu$ if μ is a measure over $\mathcal{F}$ and $g \in \mathcal{F}$ a μ-integrable function.

4.2. Generalised Bayes and PAC Bounds

The main advantage of PAC-Bayes over deterministic approaches which output single hypotheses (through optimisation of a particular criterion such as in model selection, etc.) is that the distributions allow us to capture uncertainty on hypotheses, and take into account correlations among possible hypotheses.

Denoting by ρ the posterior distribution, the quantity to control is:

$$\rho[R] = \int_{\mathcal{F}} R(f)\mathrm{d}\rho(f)$$

which is an aggregated risk over the class $\mathcal{F}$ and represents the expected risk if the predictor f is drawn from ρ for each new prediction. The distribution ρ is usually data-dependent and is referred to as a “posterior” distribution (by analogy with Bayesian statistics). We also fix a reference measure π over $\mathcal{F}$, called the “prior” (for similar reasons). We refer to [1,10] for in-depth discussions on the choice of the prior: a recent streamline of work has further investigated the choice of data-dependent priors [11–14].

The generalisation bounds associated to this setting are known as “PAC-Bayesian” bounds, where PAC stands for probably approximately correct. One important feature of PAC-Bayes bounds is that they hold true for any prior π and posterior ρ. In practice, bounds are optimised with respect to ρ and possibly π. In the following, we focus on establishing bounds for any choice of π and ρ and do not mean to optimise.

4.3. Notion of Divergence

An important notion used in PAC-Bayesian theory is the divergence between two probability distributions [see [15], for example, for a survey on divergences]. Let $\mathcal{E}$ be a measurable space and μ and ν two probability distributions on $\mathcal{E}$. Let f be a non-negative convex function defined on $\mathbb{R}_+$ such that $f(1)=0$, we define the f-divergence between μ and ν by

$$\mathcal{D}_f(\mu,\nu)=\begin{cases}\int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu & \text{if } \mu \ll \nu,\\ +\infty & \text{otherwise.}\end{cases}$$

Note that we also use the notation f to denote hypotheses elsewhere in the paper, but we believe the context to always be clear enough to avoid ambiguity.

Applying Jensen inequality, we have that $\mathcal{D}_f(\mu,\nu)$ is always non-negative and equal to zero if and only if $\mu=\nu$. The class of f-divergences includes many celebrated divergences, such as the Kullback–Leibler (KL) divergence, the reversed KL, the Hellinger distance, the total variation distance, χ^2-divergences, α-divergences, etc. Most PAC-Bayesian generalisation bounds involve the KL divergence.

A divergence can be thought of as a transport cost between two probability distributions. This interpretation will be useful for explaining PAC-Bayesian inequalities, where the divergence plays the role of a complexity term. In the following, we will just use two types of divergence. The first is the Kullback–Leibler divergence and corresponds to the choice $f(x)=x\log x$, which we denote it by

$$\mathrm{KL}(\mu,\nu)=\begin{cases}\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu & \text{if } \mu \ll \nu,\\ +\infty & \text{otherwise.}\end{cases}$$

The second is linked to Pearson's χ^2-divergence and corresponds to the choice $f(x)=x^2-1$. It is referred to as $\mathcal{D}_2$:

$$\mathcal{D}_2(\mu,\nu)=\begin{cases}\int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^2\mathrm{d}\nu\ -1 & \text{if } \mu \ll \nu,\\ +\infty & \text{otherwise.}\end{cases}$$

To illustrate the behaviour of these two divergences, consider the case where μ and ν are normal distributions on $\mathbb{R}^d$.

Proposition 3. *If $\mathcal{E}=\mathbb{R}^d$, $\mu=\mathcal{N}(a,I)$, and $\nu=\mathcal{N}(0,I)$ (where I stands for the $d\times d$ identity matrix), we have*

$$\begin{cases}\mathcal{D}_2(\mu,\nu)=e^{\|a\|^2}-1,\\ \mathrm{KL}(\mu,\nu)=\frac{1}{2}\|a\|^2.\end{cases}$$

Proof. We have:

$$\begin{cases}\mathrm{d}\mu(x)=\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}(x-a)^{\mathrm{T}}(x-a)\right)\mathrm{d}x,\\ \mathrm{d}\nu(x)=\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}x^{\mathrm{T}}x\right)\mathrm{d}x,\\ \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)=\exp\left(-\frac{1}{2}\left[-2x^{\mathrm{T}}a+a^{\mathrm{T}}a\right]\right)=\exp(-\|a\|^2/2)\exp(x^{\mathrm{T}}a).\end{cases}$$

Then:

$$\begin{aligned}\mathcal{D}_2(\mu,\nu)&=\exp\left(-\|a\|^2\right)\int\exp\left(2x^{\mathrm{T}}a\right)\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}x^{\mathrm{T}}x\right)\mathrm{d}x-1\\ &=\exp\left(-\|a\|^2\right)\int\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}x^{\mathrm{T}}x+2x^{\mathrm{T}}a\right)\mathrm{d}x-1\\ &=\exp\left(-\|a\|^2\right)\exp\left(2\|a\|^2\right)\int\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}(x-2a)^{\mathrm{T}}(x-2a)\right)\mathrm{d}x-1\\ &=e^{\|a\|^2}-1.\end{aligned}$$

And finally:

$$\begin{aligned}\mathrm{KL}(\mu,\nu) &= \int \left(-\frac{\|a\|^2}{2} + x^{\mathrm{T}}a\right)\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}(x-a)^{\mathrm{T}}(x-a)\right)\mathrm{d}x \\ &= -\frac{\|a\|^2}{2} + \int x^{\mathrm{T}}a\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}(x-a)^{\mathrm{T}}(x-a)\right)\mathrm{d}x \\ &= -\frac{\|a\|^2}{2} + \|a\|^2 = \frac{\|a\|^2}{2}.\end{aligned}$$

□

We therefore see that the divergence $\mathcal{D}_2$ penalises much more strongly the gap between the means of both distributions than the Kullback–Leibler divergence.

The following technical lemma involving the Kullback–Leibler divergence and a change of measure from posterior to prior distribution is pivotal in the PAC-Bayes literature:

Lemma 1 ([5–16], Corollary 4.15)**.** *Let g be a measurable function $g : \mathcal{F} \mapsto \mathbb{R}$ such that $\pi[e^g]$ is finite. Let π and ρ be respectively prior and posterior measures as defined in Section 4.1. The following inequality holds:*

$$\rho[g] \leq \log \pi[e^g] + \mathrm{KL}(\rho,\pi).$$

4.4. Expensive PAC-Bayesian Bound

The first PAC-Bayesian bound we present is called "expensive PAC-Bayesian bound" in the spirit of Section 2: it is obtained under a sub-Gaussian tails assumption. More precisely, we suppose here that for any $f \in \mathcal{F}$, the distribution of the random variable $\ell(f(X),Y)$ belongs to $\mathcal{P}^{\sigma}_{\text{expensive}}$, which means

$$\log \mathbb{E}[\exp\{\lambda(\ell(f(X),Y) - R(f))\}] \leq \frac{\lambda^2\sigma^2}{2}, \quad \forall \lambda \in \mathbb{R}.$$

In this setting, we have the following bound, close to the ones obtained by [10].

Proposition 4. *Assume that for any $f \in \mathcal{F}$, $\ell(f(X),Y) \in \mathcal{P}^{\sigma}_{\text{expensive}}$. For any prior π, posterior ρ, and any $\delta \in (0,1)$, the following inequality holds true with a probability greater than $1-\delta$:*

$$\rho[R] \leq \rho[R_N] + \frac{\sigma}{\sqrt{N}}\sqrt{2\left(\log\left(\frac{1}{\delta}\right) + \mathrm{KL}(\rho,\pi)\right)}.$$

Proof. The proof is decomposed in two steps. The first leverages Lemma 1. Let λ be a positive number and apply Lemma 1 to the function $\lambda(R - R_N)$:

$$\rho[R] \leq \rho[R_N] + \frac{1}{\lambda}\left(\log \pi\left[e^{\lambda(R-R_N)}\right] + \mathrm{KL}(\rho,\pi)\right).$$

The second step is to control the deviations of $\log \pi\left[e^{\lambda(R-R_N)}\right]$. With a probability $1-\delta$, we have, by Markov's inequality

$$\pi\left[e^{\lambda(R-R_N)}\right] \leq \frac{\mathbb{E}\left[\pi\left[e^{\lambda(R-R_N)}\right]\right]}{\delta}.$$

By Fubini's theorem, we can exchange the symbols $\mathbb{E}$ and π. Using the assumption $\mathcal{P}^{\sigma}_{\text{expensive}}$, we obtain with a probability greater than $1-\delta$

$$\pi\left[e^{\lambda(R-R_N)}\right] \leq \frac{\exp\{\lambda^2\sigma^2/2N\}}{\delta}.$$

Now, putting these results together and setting

$$\lambda = \frac{\sqrt{2N\left(\log\left(\frac{1}{\delta}\right) + \mathrm{KL}(\rho,\pi)\right)}}{\sigma}$$

we obtain the desired bound. □

A PAC-Bayesian inequality is a bound which treats the complexity in the following manner:

- At first, a global complexity measure is introduced with the change of measure and is characterised by the divergence term, measuring the price to switch from π (the reference distribution) to ρ (the posterior distribution on which all inference and prediction is based);
- Next, the stochastic assumption on the data-generating distribution is used to control $\pi\left[e^{\lambda(R-R_N)}\right]$ with high probability.

4.5. Cheap PAC-Bayesian Bounds

4.5.1. Using χ^2 Divergence

The vast majority of works in the PAC-Bayesian literature focuses on an expensive model. The main reason is that it includes the situation where the loss ℓ is bounded, a common (yet debatable) assumption in machine learning. The case where $\ell(f(X,Y)$ belongs to a cheap model has attracted far less attention; recently, ref. [17] have obtained the following bound.

Proposition 5 ([17], Theorem 1)**.** *Assume that for any $f \in \mathcal{F}$, $\ell(f(X),Y) \in \mathcal{P}^{\sigma}_{cheap}$. For any prior π, posterior ρ, and any $\delta \in (0,1)$, the following inequality holds true with a probability greater than $1-\delta$*

$$\rho[R] \leq \rho[R_N] + \frac{\sigma}{\sqrt{N}}\sqrt{\frac{\mathcal{D}_2(\rho,\pi)+1}{\delta}}.$$

The proof (see [17]) uses the same elementary ingredients as in the expensive case, replacing the Kullback–Leibler divergence by $\mathcal{D}_2$ and the dependence in δ moves from $\sqrt{2\log(1/\delta)}$ to $\frac{1}{\sqrt{\delta}}$. Note the correspondence between these two bounds and the confidence intervals introduced in Section 2.

4.5.2. Using Huber-Type Losses

With a different approach, ref. [18] obtained asymptotic PAC-Bayesian bounds for δ-dependent risk estimators based on the empirical mean of Huber-type influence functions. The author of [18] studied in a slightly more restrictive model than $\mathcal{P}_{\mathrm{cheap}}$, assuming in addition that the order 3 moment of $\ell(f(X),Y)$ is bounded for $f \in \mathcal{H}$. We rephrase here Theorem 9 of [18]: with a probability greater than $1-\delta$,

$$\rho[R] \leq \rho[\hat{R}_{\delta,N}] + \frac{1}{\sqrt{N}}\left(\mathrm{KL}(\rho,\pi) + \frac{\log(8\pi\sigma\delta^{-2})}{2} + \sigma + \pi^*_N(\mathcal{F}) - 1\right) + o\left(\frac{1}{N}\right),$$

where $\pi^*_N(\mathcal{F})$ is a term depending on the quality of the prior. In Remark 10, the author notes that assuming only finite moments for $\ell(f(X),Y)$, it is impossible in practice to choose a prior such that $\frac{\pi^*_N(\mathcal{F})}{\sqrt{N}}$ decreases at rate $1/\sqrt{N}$ or faster. Then, the dominant term necessarily converges at a slower rate than that of Proposition 4. However, this bounds leads to the definition of a robust PAC-Bayes estimator which proves efficient on simulated data (see Section 5 of [18]).

5. A Good Cheap Lunch: Towards a Robust PAC-Bayesian Bound?

If we take a closer look at the aforementioned PAC-Bayesian bounds from a robust statistics perspective, the following question arises: **can we obtain a PAC-Bayesian bound with a $\sqrt{\log(1/\delta)}$ dependence (possibly up to a numerical constant) in the confidence level with the cheap model?** In this section, we shed light on some structural issues. In the following, we assume the existence of $\sigma > 0$ such that for any $f \in \mathcal{F}$, $\ell(f(X), Y) \in \mathcal{P}^{\sigma}_{\text{cheap}}$.

5.1. A Necessary Condition

Let $\widehat{R}$ be an estimator of the risk (not necessarily the classical empirical risk). Here is a prototype of the inequality we are looking for: for any $\delta \in (0,1)$, with probability $1-\delta$

$$\rho[R] \leq \rho\left[\widehat{R}\right] + \frac{\sigma}{\sqrt{N}} \mathrm{A}(\rho, \pi, \delta),$$

where

$$\mathrm{A}(\rho, \pi, \delta) \underset{\delta \to 0}{=} \mathcal{O}\left(\sqrt{\log(1/\delta)}\right).$$

If we choose $\rho = \pi = \delta_{\{f\}}$ (Dirac mass in the single hypothesis f), the existence of such a PAC-Bayesian bound valid for all δ implies that

$$\left[\widehat{R}(f) \pm \frac{\sigma}{\sqrt{N}} \times c\sqrt{\log(1/\delta)}\right]$$

is a confidence interval for the risk $R(f)$ for any level $1-\delta$, where c is a constant.

Thus, a necessary condition for a PAC-Bayesian bound to be valid for all of the risk level δ is to have tight confidence intervals for any $f \in \mathcal{F}$.

However, as covered in Section 3, such estimators do not exist over the class $\mathcal{P}^{\sigma}_{\text{cheap}}$, and the possibility to derive a tight confidence interval is limited by the fact that the level δ must be greater that a positive constant of the form $e^{-\mathcal{O}(N)}$.

5.2. A δ-Dependent PAC-Bayesian Bound?

As a consequence, there is simply no hope for a robust PAC-Bayesian bound valid for any error threshold δ, for essentially the same reason which prevents it in the mean estimation case. The question we address now is the possibility of obtaining a robust PAC-Bayesian bound, with a dependence of magnitude $\sqrt{2\log(1/\delta)}$ (possibly up to a constant), with a possible limitation on the error threshold δ. In the following, we assume to have an estimator of the risk $\widehat{R}$ and an error threshold $\delta > 0$ such that there exists a constant $C > 0$ such that for any $f \in \mathcal{F}$,

$$\left[\widehat{R}(f) \pm \frac{\sigma}{\sqrt{N}} \times C\sqrt{\log(1/\delta)}\right]$$

is a confidence interval for $R(f)$ at level $1-\delta$. MoM is an example of such estimator. Let us stress that δ is fixed and cannot be used as a free parameter.

As seen above, a PAC-Bayesian bound proof proceeds in two steps:

- First, we use a convexity argument to control the target quantity $\rho[R - \widehat{R}]$ by an upper-bound involving a divergence term and a term of the form $g^{-1}\left(\pi\left[g(R-\widehat{R})\right]\right)$ where g is a non-negative, increasing, and convex function;
- Second, we control the term $\pi\left[g(R-\widehat{R})\right]$ in high probability, using Markov's inequality.

The first step does not require any use of a stochastic model on the data, and is always valid, regardless of whether we have a cheap or an expensive model. The second step uses the model and introduce the dependence in the error rate δ on the right-term of the bound: $g^{-1}(1/\delta)$. In the case of the "expensive bound", we had $g = \exp$, and the dependence was $\log(1/\delta)$, the final rate $\sqrt{\log(1/\delta)}$ was obtained by choosing a relevant value for λ.

Let us follow this scheme to obtain a robust PAC-Bayesian bound. The first step gives

$$\rho[R] \leq \rho[\widehat{R}] + \frac{1}{\lambda}\Big(\log \pi\Big[e^{\lambda(R-\widehat{R})}\Big] + \mathrm{KL}(\rho,\pi)\Big).$$

Our goal is now to control $\pi\Big[e^{\lambda(R-\widehat{R})}\Big]$ in high probability.

5.2.1. The Case $\pi = \delta_{\{f\}}$

Let us start with a very special case, where the prior is a Dirac mass on some hypothesis $f \in \mathcal{F}$. Then

$$\frac{1}{\lambda}\log \pi\Big[e^{\lambda(R-\widehat{R})}\Big] = R(f) - \widehat{R}(f).$$

Using how $\widehat{R}$ is defined, we can bound this quantity in the following way: with probability $1-\delta$,

$$R(f) - \widehat{R}(f) \leq \frac{\sigma}{\sqrt{N}} \times C\sqrt{\log(1/\delta)}.$$

Another way to formulate this result is to say that there exists an event $\mathcal{A}_f$ with a probability greater than $1-\delta$ such that for all $\omega \in \mathcal{A}_f$, the following holds true:

$$(R(f) - \widehat{R}(f,\omega)) \leq \frac{\sigma}{\sqrt{N}} \times C\sqrt{2\log(1/\delta)}.$$

In this example, we can control $\log \pi\Big[e^{\lambda(R-\widehat{R})}\Big]$ at the price of a maximal constraint on the choice of the posterior. Indeed, the only possible choice for ρ for the Kullback–Leibler $\mathrm{KL}(\rho,\pi)$ to make sense is $\rho = \pi = \delta_{\{f\}}$.

5.2.2. The Case $\pi = \alpha\delta_{\{f_1\}} + (1-\alpha)\delta_{\{f_2\}}$

Consider now a somewhat more sophisticated choice of prior which is a mixture of two Dirac masses in two distinct hypotheses. We do not fix the mixing proportion α and allow it to move freely between 0 and 1. The goal is to control the quantity

$$\pi\Big[e^{\lambda(R-\widehat{R})}\Big] = \alpha e^{\lambda(R(f_1)-\widehat{R}(f_1))} + (1-\alpha)e^{\lambda(R(f_2)-\widehat{R}(f_2))}.$$

More precisely, for all $\alpha \in (0,1)$, we want to find an event $\mathcal{A}_\alpha$ on which this quantity is under control. In view of the prior's structure, the only way to ensure such a control is to have $\mathcal{A}_\alpha \subset \mathcal{A}_{f_2} \cap \mathcal{A}_{f_2}$, where $\mathcal{A}_{f_1}$ (resp. $\mathcal{A}_{f_2}$) is the favourable event for the concentration of $\widehat{f_1}$ (resp. $\widehat{f_2}$) around its mean.

By the union bound, we have that with a probability greater than $1-2\delta$

$$\frac{1}{\lambda}\log \pi\Big[e^{\lambda(R-\widehat{R})}\Big] \leq \frac{\sigma}{\sqrt{N}} \times C\sqrt{\log(1/\delta)}.$$

We face a double problem here. As above, if we want the final bound to be non-vacuous, we have to ensure that $\mathrm{KL}(\rho,\pi)$ is finite, which restricts the support for the posterior to be included in the set $\{f_1, f_2\}$. In addition, the PAC-Bayesian bound holds with a probability greater than $1-2\delta$...

5.2.3. Limitation

... which hints at the fact that this will become $1-K\delta$ if the support for the prior contains K distinct hypotheses. If $K \geq 1/\delta$, the bound becomes vacuous. In particular, we cannot obtain a relevant bound using this approach in the situation where the cardinal of $\mathcal{F}$ is infinite (which is commonly the case in most PAC-Bayes works).

This limiting fact highlights that to derive PAC-Bayesian bounds, we cannot rely on the construction of confidence interval for all $R(f)$ for a fixed error threshold δ. The issue is that when we want to transfer this local property into a global one (valid for any mixture of hypotheses by the prior π), we cannot avoid a worst-case reasoning by the use of the union bound.

The established bounds in the PAC-Bayesian literature, both in cheap and expensive models, repeatedly use the fact that when we assume that for any $f \in \mathcal{F}$,

$$\log \mathbb{E}\left[e^{\lambda(R(f)-\ell(f(X),Y))}\right] \leq \frac{\lambda^2\sigma^2}{2}, \ \forall \lambda \in \mathbb{R}$$

or

$$\operatorname{var}(\ell(f(X),Y)) \leq \sigma^2,$$

we make an implicit assumption on the integrability of the tail of the distribution of $\ell(f(X),Y)$. This argument is crucial for the second step of the PAC-Bayesian proof because, by Fubini's theorem, it allows us to convert a local property (the tail distribution of each $\ell(f(X),Y)$) into a global one (the control of $\pi\left[e^{\lambda(R-R_N)}\right]$ or $\pi\left[(R-R_N))^2\right]$ in high probability).

5.3. Is That the End of the Story?

We have identified a structural limitation to derive a tight PAC-Bayesian bound in a cheap model. We make the case that we cannot replicate the PAC-Bayesian proof presented in Section 4. To conclude this section, we want to highlight the fact that, up to our knowledge, no proof of PAC-Bayesian bounds avoids these two steps (see, for example, the general presentation in [19]).

What if we try to avoid the change of the measure step and try to control directly $\rho[R] - \rho[\widehat{R}]$ in high probability? We remark that ρ can only be chosen with the information given by the observation of $\widehat{R}(f)$, where $f \in \mathcal{F}$. In particular, we cannot obtain any information of the concentration of each $\widehat{R}(f)$ around $R(f)$ as such knowledge requires to know the true risk. So, it seems that a direct control cannot avoid starting as a "worst-case" bound:

$$\rho[R] - \rho[\widehat{R}] \leq \sup_{f \in \mathcal{F}} \left\{ R(f) - \widehat{R}(f) \right\}.$$

Then, we have to control $\sup_{f\in\mathcal{F}}\left\{R(f)-\widehat{R}(f)\right\}$ in high probability (see [20] for a general presentation on such controls, and [7] for the recent results in the special case where $\widehat{R}$ is a MoM estimator). However, the obtained bound will take the following prototypic form:

$$\rho[R] \leq \rho[\widehat{R}] + \text{complexity term},$$

where the complexity term does not depend on the distribution ρ. Thus, the optimisation of the right term leads to choosing ρ as the Dirac mass in $\underset{f\in\mathcal{F}}{\arg\min}\, \widehat{R}(f)$.

So, the overall procedure amounts to a slightly modified empirical risk minimisation (where the empirical mean is replaced with any estimator of the risk), and will not fall into the category of generalised Bayesian approaches which take into account the uncertainty on hypotheses. Pretty much all the strengths of PAC-Bayes would then be lost.

6. Conclusions

The present paper contributes a better understanding of the profound structural reasons why good cheap lunches (tight bounds under minimal assumptions) are not possible with PAC-Bayes by walking gently through elementary examples.

From a theoretical perspective, PAC-Bayesian bounds requires too strong assumptions to adapt robust statistics results (where almost good lunches can be obtained for cheap models—with the limitation that the confidence level is constrained). The second step of the proof we have shown requires us to transform a local hypothesis, a control of some moments of $\ell(f(X),Y)$, into a global one, valid for all mixture of hypotheses by the prior π. As covered above, this transformation seems impossible.

To close on a more positive note after this negative result, let us stress that even if the conciliation of PAC-Bayes and robust statistics appears challenging, we believe that the recent ideas from robust statistics could be used in practical algorithms inspired by PAC-Bayes. In particular, we leave as an avenue for future work the empirical study of PAC-Bayesian posteriors (such as the Gibbs measure defined as $\rho \propto \exp(-\gamma\widehat{R})\pi$ for any inverse temperature $\gamma > 0$) where the risk estimator is not the empirical mean (as in most PAC-Bayes works) but rather a robust estimator, such as MoM.

Author Contributions: Conceptualization, B.G. and L.P.; Formal analysis, B.G. and L.P.; Supervision, B.G.; Writing—original draft, L.P.; Writing—review & editing, B.G. and L.P. All authors have read and agreed to the published version of the manuscript.

Funding: B.G. 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. B.G. 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. Guedj, B. A primer on PAC-Bayesian learning. *arXiv* **2019**, arXiv:1901.05353.
2. Valiant, L.G. A Theory of the Learnable. *Commun. ACM* **1984**, *27*, 1134–1142. [CrossRef]
3. Lecué, G.; Lerasle, M. Robust machine learning by median-of-means: Theory and practice. *Ann. Stat.* **2020**, *48*, 906–931. [CrossRef]
4. Devroye, L.; Györfi, L.; Lugosi, G. *A Probabilistic Theory of Pattern Recognition*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1996; Volume 31.
5. Boucheron, S.; Lugosi, G.; Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*; Oxford University Press: Oxford, UK, 2013.
6. Catoni, O. Challenging the empirical mean and empirical variance: A deviation study. *Ann. l'IHP Probabilités Stat.* **2012**, *48*, 1148–1185. [CrossRef]
7. Lerasle, M. Lecture Notes: Selected topics on robust statistical learning theory. *arXiv* **2019**, arXiv:1908.10761.
8. Devroye, L.; Lerasle, M.; Lugosi, G.; Oliveira, R.I. Sub-Gaussian mean estimators. *Ann. Stat.* **2016**, *44*, 2695–2725. [CrossRef]
9. Alquier, P. User-friendly introduction to PAC-Bayes bounds. *arXiv* **2021**, arXiv:2110.11216.
10. Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*; Lecture Notes-Monograph Series; IMS: Danbury, SC, USA, 2007.
11. Dziugaite, G.K.; Roy, D.M. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, 11–15 August 2017; Elidan, G., Kersting, K., Ihler, A.T., Eds.; AUAI Press: Montreal, QC, Canada, 2017.
12. Pérez-Ortiz, M.; Rivasplata, O.; Guedj, B.; Gleeson, M.; Zhang, J.; Shawe-Taylor, J.; Bober, M.; Kittler, J. Learning PAC-Bayes Priors for Probabilistic Neural Networks. *arXiv* **2021**, arXiv:2109.10304.
13. Pérez-Ortiz, M.; Rivasplata, O.; Shawe-Taylor, J.; Szepesvári, C. Tighter risk certificates for neural networks. *arXiv* **2020**, arXiv:2007.12911.
14. Dziugaite, G.K.; Hsu, K.; Gharbieh, W.; Arpino, G.; Roy, D. On the role of data in PAC-Bayes. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, Virtual Event, 13–15 April 2021; Banerjee, A., Fukumizu, K., Eds.; PMLR: New York, NY, USA, 2021; Volume 130, pp. 604–612.
15. Csiszár, I.; Shields, P.C. Information theory and statistics: A tutorial. In *Foundations and Trends® in Communications and Information Theory*; Now Publishers Inc.: Norwell, MA, USA, 2004; Volume 1, pp. 417–528.
16. Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158. [CrossRef]
17. Alquier, P.; Guedj, B. Simpler PAC-Bayesian bounds for hostile data. *Mach. Learn.* **2018**, *107*, 887–902. [CrossRef]

18. Holland, M.J. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems 32, Proceedings of the Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019*; Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc.: Montreal, QC, Canada, 2019; pp. 2711–2720.
19. Bégin, L.; Germain, P.; Laviolette, F.; Roy, J.F. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*; PMLR: New York, NY, USA, 2016; pp. 435–444.
20. Van der Vaart, A.W.; Wellner, J.A. Weak convergence. In *Weak Convergence and Empirical Processes*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 16–28.