How Tight Can PAC-Bayes be in the Small Data Regime?

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we investigate the question: *Given a small number of datapoints, for example $N = 30$, how tight can PAC-Bayes and test set bounds be made?* For such small datasets, test set bounds adversely affect generalisation performance by discarding data. In this setting, PAC-Bayes bounds are especially attractive, due to their ability to use all the data to simultaneously learn a posterior and bound its generalisation risk. We focus on the case of i.i.d. data with a bounded loss and consider the generic PAC-Bayes theorem of Germain et al. [2009] and Bégin et al. [2016]. While their theorem is known to recover many existing PAC-Bayes bounds, it is unclear what the tightest bound derivable from their framework is. Surprisingly, we show that for a fixed learning algorithm and dataset, the tightest bound of this form coincides with the tightest bound of the more restrictive family of bounds considered in Catoni [2007]. In contrast, in the more natural case of distributions over datasets, we give examples (both analytic and numerical) showing that the family of bounds in Catoni [2007] can be suboptimal. Within the proof framework of Germain et al. [2009] and Bégin et al. [2016], we establish a lower bound on the best bound achievable in expectation, which recovers the Chernoff test set bound in the case when the posterior is equal to the prior. Finally, to illustrate how tight these bounds can potentially be, we study a synthetic one-dimensional classification task in which it is feasible to meta-learn both the prior and the form of the bound to obtain the tightest PAC-Bayes and test set bounds possible. We find that in this simple, controlled scenario, PAC-Bayes bounds are surprisingly competitive with comparable, commonly used Chernoff test set bounds. However, the sharpest test set bounds still lead to better guarantees on the generalisation error than the PAC-Bayes bounds we consider.

1 Introduction

Generalisation bounds are of both practical and theoretical importance. Practically, tight bounds provide certificates that algorithms will perform well on unseen data. Theoretically, the bounds and underlying proof techniques can help explain the phenomenon of learning. Among the tightest known bounds are *PAC-Bayes* [McAllester, 1999] and *test set* bounds [Langford, 2002]. In this paper, we investigate their numerical tightness when applied to small datasets ($N \approx 30$–60 datapoints). The comparison between PAC-Bayes and test set bounds is particularly interesting in this setting as one cannot discard data to compute a test set bound without significantly harming performance. PAC-Bayes on the other hand provides valid bounds while using all of the data for learning. The small data setting can also be quite different from the big data setting, as lower-order terms in PAC-Bayes bounds have a non-negligible contribution, and the detailed structure of the bound becomes important. Fortunately, we do not have to study each PAC-Bayes bound separately: remarkably, Germain et al. [2009] and Bégin et al. [2016] showed that a wide range of bounds can be obtained as special cases of a single *generic*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

PAC-Bayes theorem that captures the central ideas of many PAC-Bayes proofs. This theorem has a free parameter: it holds for any convex function Δ. By choosing Δ appropriately, one can recover the well-known bounds of Langford and Seeger [2001] and Catoni [2007]. We focus on two questions related to this set-up. *First*, *what is the tightest bound achievable by any convex function* Δ? An answer would characterise the limits of the generic PAC-Bayes theorem, and thereby of a wide range of bounds, by telling us how much improvement could be obtained before new ideas or assumptions are needed. *Second*, since test set bounds are the *de facto* standard for larger datasets, but PAC-Bayes has benefits when N is small, we ask: *in the small data regime, can PAC-Bayes be tighter than test set bounds?*

In Sec 3, Thm 4, we show that in the (artificial) case of a *fixed* dataset and learning algorithm, the tightest version of the generic PAC-Bayes theorem is obtained by one of the *Catoni bounds* [Catoni, 2007]. In the more realistic case of a stochastic dataset, we do not fully characterise the tightest bound, but in Cor 3 we *lower bound* the tightest bound achievable (in expectation) with any Δ. We also provide numerical evidence in Fig 2 that suggests this lower bound can in some cases be attained, by parameterising a generic convex function with a neural network. Interestingly, this lower bound coincides with removing a lower-order term from the Langford and Seeger [2001] bound, and relaxes to the well-known *Chernoff test set bound* when the PAC-Bayes posterior is equal to the prior.

In Sec 4, we investigate the tightness of PAC-Bayes and test set bounds in synthetic 1D classification. The goal of this experiment is to find out how tight the bounds could be made in principle. We use meta-learning to adapt all aspects of the bounds and algorithms to the task distribution. We find that, in this setting, PAC-Bayes can be competitive with the Chernoff test set bound, but is outperformed by the *binomial tail test set bound*, of which the Chernoff bound is a relaxation. This suggests that, for standard PAC-Bayes to be quantitatively competitive with the best test set bounds on small datasets, a new proof technique leading to bounds that gracefully relax to the binomial tail bound is required.

2 Background and Related Work

We consider supervised learning. Let X and Y denote the *input space* and *output space*, and let $Z = X \times Y$. Assume there is an (unknown) probability measure[1] D over Z, with the dataset $S \sim D^N$. Denote the *hypothesis space* by $\mathcal{H} \subseteq Y^X$. A learning algorithm is then a map $Z^N \to \mathcal{H}$. In PAC-Bayes, we also consider maps $Z^N \to \mathcal{M}_1(\mathcal{H})$, where $\mathcal{M}_1$ is the set of probability measures on its argument. The performance of a hypothesis $h \in \mathcal{H}$ is measured by a *loss function* $\ell\colon Z \times \mathcal{H} \to [0,1]$. The *(generalisation) risk* of h is $R_D(h) \coloneqq \mathbb{E}_{(x,y)\sim D}[\ell((x,y),h)]$ and its *empirical risk on* S is $R_S(h) \coloneqq \frac{1}{N}\sum_{(x,y)\in S} \ell((x,y),h)$. For $Q \in \mathcal{M}_1(\mathcal{H})$ its *(generalisation Gibbs) risk* is $\overline{R}_D(Q) \coloneqq \mathbb{E}_{h\sim Q}[R_D(h)]$ and its *empirical (Gibbs) risk* is $\overline{R}_S(Q) \coloneqq \mathbb{E}_{h\sim Q}[R_S(h)]$. In PAC-Bayes, we usually fix a *prior* $P \in \mathcal{M}_1(\mathcal{H})$, chosen without reference to S and learn a *posterior* $Q \in \mathcal{M}_1(\mathcal{H})$ which can depend on S. The *KL-divergence* between Q and P is defined as $\mathrm{KL}(Q\|P) = \int \log \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q$ if $Q \ll P$ and ∞ otherwise. Let $\mathcal{C}$ denote the set of proper, convex, lower semicontinuous (l.s.c.) functions $\mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$; if a convex function's domain is a subset of $\mathbb{R}^2$, extend it to all of $\mathbb{R}^2$ with the value $+\infty$. See App C for more details on convex analysis, which we use in Sec 3.

Test Set Bounds. *Test set bounds* rely on a subset of data which is not used to select the hypothesis, called a *test set* or *held-out set*. Let $S = S_{\text{train}} \cup S_{\text{test}}$, with $|S| = N$, $|S_{\text{train}}| = N_{\text{train}}$ and $|S_{\text{test}}| = N_{\text{test}}$. In Thms 1 and 2, we assume h is chosen independently of S_{test}. For a zero-one loss, where $\ell \in \{0,1\}$, $N_{\text{test}} R_{S_{\text{test}}}(h)$ is a binomial random variable with parameters $(N, R_D(h))$. This leads to the following simple bound, which, for $\ell \in \{0,1\}$, is tight among test set bounds:

Theorem 1 (Binomial tail test set bound, Langford, 2005, Theorem 3.3)**.**
Let $\overline{e}(M,k,\delta) \coloneqq \sup\left\{p : \delta \le \sum_{i=1}^{k} \binom{M}{i} p^i (1-p)^{M-i}\right\}$. *Then, for any* $h \in \mathcal{H}$ *and* $\ell \in \{0,1\}$,

$$\Pr\Big(R_D(h) \le \overline{e}(N_{\text{test}}, N_{\text{test}} R_{S_{\text{test}}}(h), \delta)\Big) \ge 1-\delta. \tag{1}$$

Often, looser bounds with a simpler form are applied. These can be obtained via the Chernoff method:

Theorem 2 (Chernoff test set bound, Langford, 2005, Corollary 3.7)**.**
For $q, p \in [0,1]$, *let* $\mathrm{kl}(q,p) \coloneqq q \log \frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$. *Then, for any* $h \in \mathcal{H}$ *and* $\ell \in [0,1]$,

$$\Pr\left(\mathrm{kl}(R_{S_{\text{test}}}(h), R_D(h)) \le \frac{1}{N_{\text{test}}} \log\frac{1}{\delta}\right) \ge 1-\delta. \tag{2}$$

[1] We will colloquially refer to measures on sets without specifying a σ-algebra. We implicitly assume functions are measurable with respect to the σ-algebras on which the relevant measures are defined.

PAC-Bayes Bounds. The PAC-Bayes approach bounds the generalisation Gibbs risk of *stochastic classifiers*, and does not require discarding data. Since the seminal paper of McAllester [1999], a large variety of PAC-Bayes bounds have been derived. Germain et al. [2009] and Bégin et al. [2016] prove a very general form of the PAC-Bayes theorem which encompasses many of these. Their proof technique consists of a series of inequalities shared by PAC-Bayes proofs (Jensen's, change of measure, Markov's, supremum over risk), and reveals their common structure. Thus understanding the properties of this generic theorem can give insight into many PAC-Bayes bounds at once:

Theorem 3 (Generic PAC-Bayes theorem, Bégin et al. [2016], Germain et al. [2009])**.**[2]
Fix $P \in \mathcal{M}_1(\mathcal{H})$, $\ell \in [0,1]$, $\delta \in (0,1)$, and Δ a proper, convex, l.s.c. function $[0,1]^2 \to \mathbb{R} \cup \{+\infty\}$. Then

$$\Pr\left((\forall Q)\ \Delta(\overline{R}_S(Q), \overline{R}_D(Q)) \leq \tfrac{1}{N}\left[\mathrm{KL}(Q\|P) + \log \tfrac{\mathcal{I}_\Delta(N)}{\delta}\right]\right) \geq 1-\delta, \tag{3}$$

where $\mathcal{I}_\Delta(N) \coloneqq \sup_{r\in[0,1]} \sum_{k=0}^{N} \binom{N}{k} r^k (1-r)^{N-k} e^{N\Delta(k/N,r)}$.

Remark 1. *We lose no generality in assuming $\Delta(q,\cdot)$ is monotonically increasing for all $q \in [0,1]$. Otherwise, we can define a Δ' that leads to at least as good an upper bound on $\overline{R}_D(Q)$. See App D.*

For completeness, we provide a proof of Thm 3 in App B. Following Bégin et al. [2016], we briefly recap some of the bounds that can be recovered as special cases (or looser versions) of Thm 3. Setting $\Delta(q,p) = C_\beta(q,p) \coloneqq -\log(1+p(e^{-\beta}-1)) - \beta q$ for $\beta > 0$, we recover the *Catoni bounds*:

Corollary 1 (Catoni, 2007, Theorem 1.2.6)**.** *For any $\beta > 0$,*

$$\Pr\left((\forall Q)\ \overline{R}_D(Q) \leq \tfrac{1}{1-e^{-\beta}}\left[1-\exp\left(-\beta\overline{R}_S(Q) - \tfrac{1}{N}\left(\mathrm{KL}(Q\|P) + \log\tfrac{1}{\delta}\right)\right)\right]\right) \geq 1-\delta. \tag{4}$$

This specifies a bound for every value of $\beta > 0$. If we instead choose $\Delta(q,p) = \mathrm{kl}(q,p)$, we obtain a slightly tighter version of Maurer's refinement of the Langford-Seeger "kl-bound":

Corollary 2 (Langford and Seeger, 2001, Theorem 3, Maurer, 2004)**.** *For $N \geq 8$,*

$$\Pr\left((\forall Q)\ \mathrm{kl}(\overline{R}_S(Q), \overline{R}_D(Q)) \leq \tfrac{1}{N}\left[\mathrm{KL}(Q\|P) + \log\tfrac{2\sqrt{N}}{\delta}\right]\right) \geq 1-\delta. \tag{5}$$

We refer to this as the *Maurer–Langford–Seeger* or *MLS bound*. It is very slightly looser than Thm 3 with $\Delta = \mathrm{kl}$, since Maurer [2004] upper bounds $\mathcal{I}_{\mathrm{kl}}(N)$ by $2\sqrt{N}$ using Stirling's formula. The Catoni and MLS bounds are among the tightest PAC-Bayes bounds known and have been applied in settings where numerical tightness is key, such as obtaining generalisation bounds for stochastic neural networks [Dziugaite and Roy, 2017; Zhou et al., 2019]. Many other bounds can be obtained by loosening these bounds. Applying Pinsker's inequality $\mathrm{kl}(q,p) \geq 2(q-p)^2$ to Eq (5) yields the "square-root" version of the PAC-Bayes theorem [McAllester, 2003]. The "PAC-Bayes-λ" [Thiemann et al., 2017] and "PAC-Bayes-quadratic" bounds [Rivasplata et al., 2019] can be derived as loosened versions of the MLS bound using the inequality $\mathrm{kl}(q,p) \geq (q-p)^2/(2p)$, valid for $q < p$. The "linear" bound in McAllester [2013] can be derived by loosening the Catoni bound using: $C_\beta(q,p) \leq A \implies p \leq \frac{1}{1-\beta/2}(q + \frac{1}{\beta}A)$, which is valid for $\beta \in (0,2)$.

How Tight Are PAC-Bayes Bounds? A fundamental question we can ask about a generalisation bound is how tight it is, and whether it can be tightened. Comparing the MLS and Chernoff test set bounds when $Q = P$ (so the PAC-Bayes bound essentially becomes a test set bound) shows they are identical except for a $\log(2\sqrt{N})/N$ on the RHS of the MLS bound. Whether this term (or similar discrepancies between PAC-Bayes and *Occam bounds* [Langford, 2002, Cor 4.6.2]; see App A) can be removed has been an open question since Langford [2002, Problem 6.1.2]. Maurer [2004] reduced this term to its current value, improving on work by Langford and Seeger [2001]. Interestingly, Germain et al. [2009, Prop 2.1] shows that the expression obtained by dropping $\log(2\sqrt{N})/N$ from the MLS bound is identical to that obtained by minimising the Catoni bound with respect to β, *illegally without a union bound*. The Chernoff test set bound is itself a looser version of the binomial tail bound, raising the question of whether a PAC-Bayes bound can be found that reduces to the binomial tail bound when $Q = P$. We provide new insights into these problems in Sec 3.

Researchers have also compared PAC-Bayes bounds numerically on actual learning problems. Langford [2005] and Germain et al. [2009] were able to obtain reasonable guarantees on small datasets.

[2] We state a simpler version of their result WLOG, absorbing a free parameter into the function Δ.

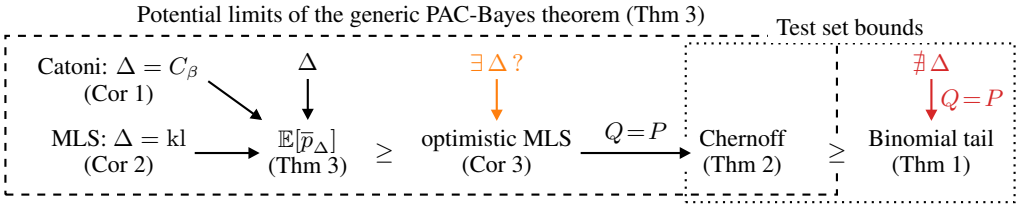


Figure 1: **Illustration of the relationship between various PAC-Bayes and test set bounds**; see Sec 3. It is unclear if there always exists a Δ that recovers the optimistic MLS bound (and hence the Chernoff bound when $Q = P$; Open Prob. 1), but there certainly does not exist a Δ that recovers the Binomial tail bound when $Q = P$.

However, Langford [2005] found that on datasets with $N \approx 145$, PAC-Bayes was outperformed by test set bounds. Langford and Caruana [2002], Dziugaite and Roy [2017] and Pérez-Ortiz et al. [2020] provide non-vacuous bounds for neural networks using PAC-Bayes. Even so, Dziugaite et al. [2021] states that tighter bounds would be obtained using a test set instead. In Sec 4 we find that if the bounds and learning algorithms are optimised for a task distribution, PAC-Bayes can be tight enough to compete with the Chernoff test set bound, but not the binomial tail test set bound.

3 Characterising the Limits of the Generic PAC-Bayes Proof Technique

This section establishes our main theoretical contributions, which characterise the limits of the generic PAC-Bayes theorem (Thm 3). For a convex $\Delta \in \mathcal{C}$, Thm 3 gives a high-probability upper bound on $\Delta(\overline{R}_S(Q), \overline{R}_D(Q))$. Define $B[f, y] \coloneqq \sup\{p \in [0,1] : f(p) \leq y\}$ for $f\colon [0,1] \to \mathbb{R}$ and $y \in \mathbb{R}$, where the supremum of the empty set is taken to be 1. This upper bound can be "inverted" to obtain a high-probability upper bound on $\overline{R}_D(Q)$: with probability at least $1 - \delta$, for all $Q \in \mathcal{M}_1(\mathcal{H})$,

$$\overline{R}_D(Q) \leq \overline{p}_\Delta \quad \text{where} \quad \overline{p}_\Delta \coloneqq B\big[\Delta(\overline{R}_S(Q), \cdot\,), \tfrac{1}{N}\big(\mathrm{KL}(Q\|P) + \log \tfrac{\mathcal{I}_\Delta(N)}{\delta}\big)\big]. \tag{6}$$

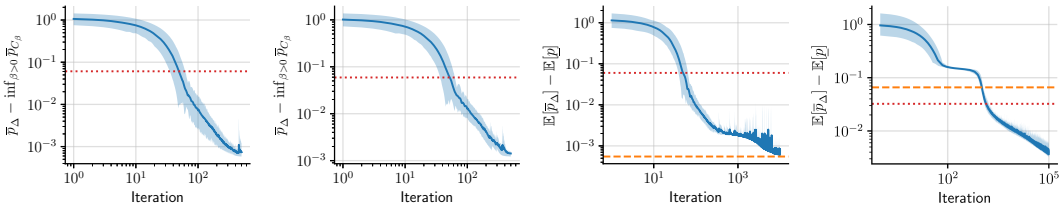
Since (6) holds for all $\Delta \in \mathcal{C}$, a natural question is: *Which Δ minimises the RHS of* (6)*?* This would characterise how tight, numerically, PAC-Bayes theorems can be made without introducing ideas beyond those needed to prove the bounds stated in Sec 2. Before considering the general case of a stochastic dataset, we first characterise the optimal Δ in the simplified scenario of a fixed dataset S and posterior Q (Thm 4). This is artificial because it is not valid to choose Δ based on S. Next, from Thm 4, we derive a lower bound on the best possible generic PAC-Bayes bound (in expectation) in the more realistic stochastic case (Cor 3). We then connect this lower bound to various existing PAC-Bayes and test set bounds. An overview is shown in Fig 1. We now state our first result.

Theorem 4. *For a fixed dataset S and any $Q, P \in \mathcal{M}_1(\mathcal{H})$, the tightest Catoni bound is as tight as the tightest bound possible within the generic PAC-Bayes theorem (Thm 3). Formally, let $\Delta \in \mathcal{C}$ and $\delta \in (0,1)$. Choose some fixed $\overline{R}_S(Q) =: q \in [0,1]$ and $\mathrm{KL}(Q\|P) =: \mathrm{KL} \in [0,\infty)$. If $q > 0$, then there exists a $\beta \in (0,\infty)$ such that $\overline{p}_\Delta \geq \overline{p}_{C_\beta}$. Moreover if $q = 0$, then $\overline{p}_\Delta \geq \lim_{\beta\to\infty} \overline{p}_{C_\beta}$.*

Remark 2. *By Thm 4, for all $\Delta \in \mathcal{C}$, we have $\overline{p}_\Delta \geq \inf_{\beta>0} \overline{p}_{C_\beta}$, and, by Prop 2.1 of Germain et al. [2009], $\inf_{\beta>0} \overline{p}_{C_\beta} = B[\mathrm{kl}(q, \cdot\,), \frac{1}{N}(\mathrm{KL} + \log\frac{1}{\delta})]$. Hence, for all $\Delta \in \mathcal{C}$, it holds that $\overline{p}_\Delta \geq B[\mathrm{kl}(q, \cdot\,), \frac{1}{N}(\mathrm{KL} + \log\frac{1}{\delta})]$, which is also shown directly in the proof of Thm 4 (Eq* (17)*). Note that optimising β in this way is illegal, and would typically require a union bound to be valid.*

We defer the proof of Thm 4 to the end of this section. We numerically verify Thm 4 by optimising $\overline{p}_\Delta$ with respect to an arbitrary convex Δ for various settings of fixed q and KL. To parametrise a general convex Δ, we use a one-hidden-layer neural network with positive weights at the output layer and softplus nonlinearities. The inversion performed by B is approximated numerically by discretising the inputs to Δ and detecting an upcrossing. Gradients are then approximated using the inverse function theorem: $\frac{\mathrm{d}}{\mathrm{d}\theta} B[f_\theta, c(\theta)] = (\partial_\theta c(\theta) - \partial_\theta f_\theta(x))/\partial_x f_\theta(x)$. See App F for details. Figs 2a and 2b show the difference between the numerically optimised Δ and the best Catoni bound for two settings of fixed q and KL. In both cases, $\overline{p}_\Delta - \inf_{\beta>0} \overline{p}_{C_\beta}$ appears to converge to zero from above, as expected from Thm 4. Interestingly, App F shows that the learned Δ can deviate substantially from C_β, suggesting there are choices for Δ besides Catoni's which achieve $\inf_{\Delta\in\mathcal{C}} \overline{p}_\Delta$.

For a fixed dataset, Thm 4 states that the tightest bound is one of the Catoni bounds; precisely: $\inf_{\Delta\in\mathcal{C}} \overline{p}_\Delta = \inf_{\beta>0} \overline{p}_{C_\beta}$. The more interesting question is whether, when S is sampled randomly,



(a) (b) (c) (d)

Figure 2: **The tightest Catoni bound is the optimal generic PAC-Bayes bound for a fixed dataset, but not the optimal expected bound for a random dataset.** We optimise a convex function with H hidden units to minimise $\overline{p}_\Delta$ with $N = 30$. **(a) and (b) consider fixed** q **and** KL (precise values below) and show the difference with the best Catoni bound (Thm 4); **(c) and (d) consider random** q **and** KL and show the *expected* difference with the optimistic MLS bound (Cor 3). Shaded regions show the minimum and maximum over ten initialisations. All plots show the MLS bound (dotted red) and (c) and (d) show the optimal Catoni bound with parameter β^* (dashed orange). All runs quickly converged to non-vacuous values. (a): $(q, \mathrm{KL}) = (2\%, 1)$, $\beta^* \approx 2.24$, $H = 256$. (b): $(q, \mathrm{KL}) = (5\%, 2)$, $\beta^* \approx 1.84$, $H = 256$. (c): $(q, \mathrm{KL}) \in \{(2\%, 1), (5\%, 2)\}$ uniformly, $\beta^* \approx 1.99$, $H = 512$. (d): $(q, \mathrm{KL}) \in \{(0.3, 1), (0.4, 50)\}$ uniformly, $\beta^* \approx 2.32$, $H = 1024$.

one of the Catoni bounds can still achieve the tightest bound (in expectation). The answer is no: Fig 2d gives a numerical counterexample where $\inf_{\Delta \in \mathcal{C}} \mathbb{E}[\overline{p}_\Delta] < \mathbb{E}[\overline{p}_{\mathrm{kl}}] < \inf_{\beta > 0} \mathbb{E}[\overline{p}_{C_\beta}]$. Since the Catoni family of bounds cannot generally achieve the tightest bound in expectation, which Δ do? And how tight is $\inf_{\Delta \in \mathcal{C}} \mathbb{E}[\overline{p}_\Delta]$? Whilst we do not have a full answer, we establish a simple lower bound on $\inf_{\Delta \in \mathcal{C}} \mathbb{E}[\overline{p}_\Delta]$. Define the *optimistic MLS bound* $\underline{p}$ as the quantity from Rem 2, which equals the MLS bound without the $\frac{1}{N} \log \mathcal{I}_{\mathrm{kl}}(N)$ term:

$$\underline{p} \coloneqq B[\mathrm{kl}(q, \cdot), \tfrac{1}{N}(\mathrm{KL}(Q \| P) + \log \tfrac{1}{\delta})]. \tag{7}$$

The optimistic MLS bound is *not* proven to be a valid generalisation bound. Rem 2 then yields:

Corollary 3. *Consider the setting from Thm 3. Then the expected optimistic MLS bound* $\mathbb{E}[\underline{p}]$ *gives a lower bound on all expected generalisation bounds obtained through the generic PAC-Bayes theorem (Thm 3). That is, for any distribution over datasets, any prior, and any learning algorithm,*

$$\inf_{\Delta \in \mathcal{C}} \mathbb{E}[\overline{p}_\Delta] \geq \mathbb{E}[\underline{p}] \tag{8}$$

Moreover, there exists a distribution over datasets, a prior, and a posterior such that equality holds. For example, let (x, y) *be constant almost surely, which reduces to the setting of Thm 4.*

Fig 1 shows how Cor 3 fits into the picture so far. The optimistic MLS bound is at least as tight as the bound achieved by any Δ, but Cor 3 does not establish the existence of a Δ which achieves it. Cor 3 has practical utility: for any choice of Δ and corresponding generalisation bound, the optimistic MLS bound can quantify potential slack in the bound due to a suboptimal choice of Δ. App H considers an example of this application of Cor 3 in the simplified scenario where $\overline{R}_S(Q) = \frac{1}{2}$ almost surely. The optimistic MLS bound recovers the Chernoff test set bound (Thm 2) when setting $Q = P$. Since the binomial tail bound (Thm 1) is strictly tighter than the Chernoff bound, this shows there does not exist a Δ such that the generic PAC-Bayes bound (Thm 3) recovers the Binomial tail bound when $Q = P$; this is illustrated in Fig 1. What is unclear, however, is whether there always exists a Δ such that Thm 3 recovers the Chernoff test set bound; or, alternatively, such that the optimistic MLS bound is attained. A positive answer to the latter would establish that the optimistic MLS bound is a valid generalisation bound.[3] As a first piece of evidence, the traces from Figs 2c and 2d suggest that a convex function could actually achieve $\mathbb{E}[\underline{p}]$; see App G for more traces. We leave a full resolution of this question as an open problem; see Sec 5. Interestingly, Fig 2c shows that a Catoni bound is sometimes *nearly* optimal even in the stochastic case; we will see another example of this in Fig 3.

We end this section with the proof of Thm 4. Recall that the Catoni family of bounds follows from Thm 3 by considering $\Delta(q, p) = C_\beta(q, p) \coloneqq \mathcal{F}_\beta(p) - \beta q$ with $\mathcal{F}_\beta(p) \coloneqq -\log(p(e^{-\beta} - 1) + 1)$ and $\beta > 0$. To simplify the notation, we denote $\alpha = \frac{1}{N}(\mathrm{KL} + \log \frac{1}{\delta}) \in (0, \infty)$.

Proof of Thm 4. The proof proceeds in three steps. In the first two steps, we lower bound $\frac{1}{N} \log \mathcal{I}_\Delta(N)$ and upper bound Δ. In the third step, we use these bounds to lower bound $B[\Delta(q, \cdot), \alpha + \frac{1}{N} \log \mathcal{I}_\Delta(N)]$ and identify the result with a particular Catoni bound.

[3] By Rem 2, $\overline{p}_\Delta \geq \underline{p}$, so $\mathbb{E}[\overline{p}_\Delta] = \mathbb{E}[\underline{p}]$ implies that $\overline{p}_\Delta = \underline{p}$ a.s., meaning that $\underline{p}$ is a valid gen. bound.

Lower bound on $\frac{1}{N}\log \mathcal{I}_\Delta(N)$: Since $\Delta \in \mathcal{C}$, it is equal to its own double convex conjugate: $\Delta(q,p) = \Delta^{**}(q,p) = \sup_{c_q,c_p\in\mathbb{R}} (c_q q + c_p p - \Delta^*(c_q,c_p))$, where * denotes convex conjugation. Let $X \sim \mathrm{Bin}(r,N)$. Then

$$\mathcal{I}_\Delta(N) = \sup_{r\in[0,1]} \mathbb{E}[e^{N\Delta(X/N,r)}] = \sup_{r\in[0,1]} \mathbb{E}[e^{\sup_{c_q,c_p\in\mathbb{R}}(c_q X + N c_p r - N\Delta^*(c_q,c_p))}] \tag{9}$$

$$\geq \sup_{r\in[0,1]} \sup_{c_q,c_p\in\mathbb{R}} e^{N c_p r - N\Delta^*(c_q,c_p)} \mathbb{E}[e^{c_q X}] \tag{10}$$

where $\mathbb{E}[e^{c_q X}] = (r(e^{c_q}-1)+1)^N$ is the moment-generating function of X. Consequently, taking log, dividing by N, and noting that $\frac{1}{N}\log \mathbb{E}[e^{c_q X}] = -\mathcal{F}_{-c_q}(r)$,

$$\tfrac{1}{N}\log \mathcal{I}_\Delta(N) \geq A \quad \text{where} \quad A \coloneqq \sup_{c_q,c_p\in\mathbb{R}}[-\Delta^*(c_q,c_p) + \sup_{r\in[0,1]}(c_p r - \mathcal{F}_{-c_q}(r))]. \tag{11}$$

Upper bound on Δ: We upper bound Δ by making Δ^* as small as possible without exceeding the supremum from (11). Note that A is finite, because Δ^* is proper. Define $\tilde{\Delta}^*$ as follows: $\tilde{\Delta}^*(c_q,c_p) = -A + \sup_{r\in[0,1]}(c_p r - \mathcal{F}_{-c_q}(r))$. Note that $\tilde{\Delta}^*$ is proper, convex as a pointwise supremum of convex functions, and l.s.c. as a supremum of l.s.c. functions. In fact, $\tilde{\Delta}^*$ is finite for all inputs. As the notation suggests, define $\tilde{\Delta} \coloneqq (\tilde{\Delta}^*)^*$. Then $\tilde{\Delta}^*$ is indeed the convex conjugate of $\tilde{\Delta}$, because $\tilde{\Delta}^* \in \mathcal{C}$, so it is equal to its own double convex conjugate. Moreover,

$$\tilde{\Delta}(q,p) = A + \sup_{c_q,c_p\in\mathbb{R}}[c_q q + c_p p - \sup_{r\in[0,1]}(c_p r - \mathcal{F}_{-c_q}(r))] \tag{12}$$

$$= A + \sup_{c_q\in\mathbb{R}}[c_q q + \sup_{c_p\in\mathbb{R}}[c_p p - \mathcal{F}^*_{-c_q}(c_p)]] \tag{13}$$

$$= A + \sup_{c_q\in\mathbb{R}}[c_q q + \mathcal{F}_{-c_q}(p)], \tag{14}$$

by observing that $p \mapsto \mathcal{F}_{-c_q}(p) \in \mathcal{C}$, so it is equal to its own double convex conjugate. Therefore,

$$\tilde{\Delta}(q,p) = A + \sup_{c_q\in\mathbb{R}} C_{-c_q}(q,p) \overset{\text{(i)}}{=} A + \mathrm{kl}(q,p) \tag{15}$$

where (i) follows from a direct computation; see Lem E.1 (App E). *Claim:* For all $q,p \in [0,1]$, $\tilde{\Delta}(q,p) \geq \Delta(q,p)$. This follows from the definitions and finiteness of $\tilde{\Delta}^*$ and A: for all $c_q, c_p \in \mathbb{R}$,

$$-\tilde{\Delta}^*(c_q,c_p) + \sup_{r\in[0,1]}(c_p r - \mathcal{F}_{-c_q}(r)) = A \geq -\Delta^*(c_q,c_p) + \sup_{r\in[0,1]}(c_p r - \mathcal{F}_{-c_q}(r)), \tag{16}$$

which means that $\tilde{\Delta}^* \leq \Delta^*$, so $\tilde{\Delta} \geq \Delta$ by the order-reversing property of the convex conjugate.

Conclusion: Assume that $\overline{p}_\Delta < 1$; otherwise, any $\beta > 0$ works. To begin with, use the previous steps:

$$\overline{p}_\Delta = B[\Delta(q,\,\cdot\,), \alpha + \tfrac{1}{N}\log \mathcal{I}_\Delta(N)] \overset{(11),\,\text{claim}}{\geq} B[\tilde{\Delta}(q,\,\cdot\,), \alpha + A] \overset{(15)}{=} B[\mathrm{kl}(q,\,\cdot\,), \alpha] = \underline{p}. \tag{17}$$

Since $\alpha > 0$, clearly $\underline{p} > q$, so $0 \leq q < \underline{p} < 1$. Hence, if $q > 0$, then there exists a $\beta > 0$ such that $\mathrm{kl}(q,\underline{p}) = C_\beta(q,\underline{p})$ (Lem E.2; App E). Using that $p \mapsto C_\beta(q,p)$ is continuous and strictly increasing for all $\beta > 0$, we have that $\underline{p} = B[C_\beta(q,\,\cdot\,), \alpha]$, so

$$\overline{p}_\Delta \geq B[C_\beta(q,\,\cdot\,), \alpha] \overset{\text{(i)}}{=} B[C_\beta(q,\,\cdot\,), \alpha + \tfrac{1}{N}\log \mathcal{I}_{C_\beta}(N)] = \overline{p}_{C_\beta}, \tag{18}$$

where (i) uses that $\frac{1}{N}\log \mathcal{I}_{C_\beta}(N) = 0$ (Lem E.3; App E). If $q = 0$, then $\mathrm{kl}(0,\underline{p}) = \lim_{\beta\to\infty} C_\beta(0,\underline{p})$ (Lem E.2; App E), so $\overline{p}_\Delta \geq B[\lim_{\beta\to\infty} C_\beta(0,\cdot), \alpha]$ and conclude like in (18) using Lem E.4 (App E). □

4 Meta-Learning the Tightest Bounds for Synthetic Classification

We now consider, for a particular distribution over tasks, how tight each bound can be made in expectation. Two questions naturally arise: *Which PAC-Bayes bounds are tightest?* and *Can PAC-Bayes bounds be tighter than test set bounds?* Our goal is *not* to compare these bounds when using standard practice, but to see how tight they can be *in principle* if we use every tool in our toolbox to minimise the expected bounds. While these optimisations will be impractical for large models and datasets, they can provide some statistical insight. Code to reproduce all experiments can be found at `github.com/account/repo`.

Learning Algorithm. Certain learning algorithms may work better with test set bounds, and others with PAC-Bayes bounds. Instead of choosing a fixed algorithm, we *meta-learn* [Schmidhuber, 1987,

Thrun and Pratt, 2012] separate algorithms to optimise each bound in expectation: we parametrise a hypothesis space $\mathcal{H}_\theta$ and a *posterior map* $Q_\theta \colon Z^N \to \mathcal{M}_1(\mathcal{H}_\theta)$ by a finite dimensional vector θ, which is trained to optimise the expected bound. This is explained in more detail below. This way, we obtain algorithms that are optimised for each bound. After meta-learning, we can further refine each PAC-Bayes posterior by minimising the PAC-Bayes bound, see App I.4.

Task Distribution. In meta-learning, we refer to a data-generating distribution D and dataset $S \sim D^N$ as a *task*. We consider a *distribution* over tasks, $D \sim \mathcal{T}$, and aim to find the best expected bounds for this distribution achievable by an optimised algorithm.[4] We choose especially simple learning tasks — synthetic 1-dimensional binary classification problems, generated by thresholding Gaussian process (GP) samples — which allows us to fully control the task distribution and easily inspect predictive distributions visually to diagnose learning. App I.1 contains full details.

Priors. The choice of prior is crucial in PAC-Bayes, and the role of *data-dependent priors* (DDPs) [Ambroladze et al., 2007, Parrado-Hernández et al., 2012] has been gaining increased attention. This involves splitting the dataset into $N = N_{\text{prior}} + N_{\text{risk}}$ datapoints. The DDP is allowed to depend on the *prior set* of size N_{prior} (standard priors use $N_{\text{prior}} = 0$), and the empirical risk is computed on the *risk set* of size N_{risk}. Crucially, *the bound is valid when the posterior depends on all N datapoints*. Recently, Dziugaite et al. [2021] showed that DDPs can lead to tighter expected bounds than the optimal non-data-dependent prior, and are sometimes even *required* to obtain non-vacuous bounds. In our experiments we meta-learn a DDP as a map from the prior set to the prior, $P_\theta \colon Z^{N_{\text{prior}}} \to \mathcal{M}_1(\mathcal{H})$ (P_θ is not the same map as Q_θ, but we amalgamate all meta-learned parameters into the single vector θ to ease notation). To compare PAC-Bayes DDPs against test set bounds, we sweep the prior/train set proportion from 0 to 0.8 and see what the tightest value obtained is. Strictly this would require a union bound over the proportions, but here we are primarily interested in comparing the various bounds against each other on an even footing, and vary the proportion for illustrative purposes.

The Meta-Learning Objective. We now discuss meta-learning in more detail. During meta-training, θ is trained to optimise the expected PAC-Bayes generalisation bound over the task distribution:

$$\mathbb{E}_{D\sim\mathcal{T}}\mathbb{E}_{S\sim D^N}\, B\big[\Delta_\theta(\overline{R}_{S_{\text{risk}}}(Q_\theta(S)),\,\cdot\,),\, \tfrac{1}{N_{\text{risk}}}\big(\mathrm{KL}(Q_\theta(S)\|P_\theta(S_{\text{prior}})) + \log\tfrac{\mathcal{I}_{\Delta_\theta}(N_{\text{risk}})}{\delta}\big)\big], \quad (19)$$

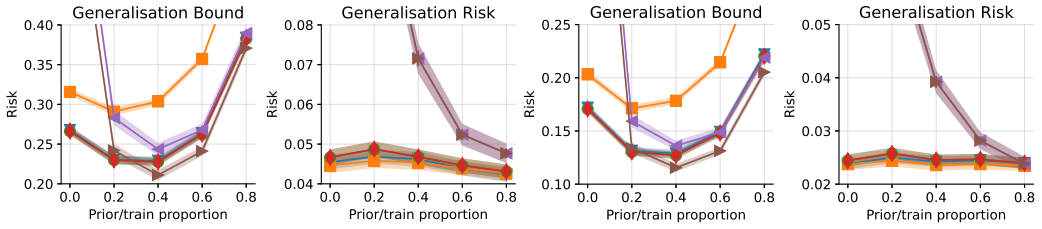
where the θ in Δ_θ denotes that some bounds (Catoni and learned convex) have meta-learnable parameters. Alternatively, for a meta-learner that minimises a test set bound, the objective is simply $\mathbb{E}_{D\sim\mathcal{T}}\mathbb{E}_{S\sim D^N}\,\overline{R}_{S_{\text{test}}}(Q_\theta(S_{\text{train}}))$, since all test set bounds are monotonic in the test set risk. We use the $0/1$ loss. As the classifiers are stochastic, the empirical risk is still differentiable with respect to θ. In contrast to PAC-Bayes, the predictor that minimises the test set bound can be made deterministic after θ is learned, since it tends to eventually learn essentially deterministic classifiers; see App I.2. We sample $T = 80\,000$ tasks $D_t \sim \mathcal{T}$, with associated datasets $S_t \sim D_t^N$. These form the *meta-trainset*. Additionally, we sample 1024 tasks that form a *meta-testset* used to estimate the average bounds over $\mathcal{T}$ after meta-training. For the PAC-Bayes bounds, we then Monte Carlo estimate (19). Hence, the final objective for a PAC-Bayes meta-learner is (a minibatched version of):

$$\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\, B\big[\Delta_\theta(\overline{R}_{S_{t,\text{risk}}}(Q_\theta(S_t)),\,\cdot\,),\, \tfrac{1}{N_{\text{risk}}}\big(\mathrm{KL}(Q_\theta(S_t)\|P_\theta(S_{t,\text{prior}})) + \log\tfrac{\mathcal{I}_{\Delta_\theta}(N_{\text{risk}})}{\delta}\big)\big]. \quad (20)$$

Similarly, the objective for the test set bound meta-learner is $\frac{1}{T}\sum_{t=1}^{T}\overline{R}_{S_{t,\text{test}}}(Q_\theta(S_{t,\text{train}}))$. The bounds we compute on datasets in the meta-testset, after meta-training is complete and θ is frozen, are valid even though θ was optimised on the meta-trainset. This highlights a contrast between our procedure and the PAC-Bayes meta-learning in Amit and Meir [2018], Rothfuss et al. [2020], Liu et al. [2021] and Farid and Majumdar [2021]. While those works use PAC-Bayes to analyse generalisation of a meta-learner on new tasks, we use PAC-Bayes to analyse generalisation *within* individual tasks.

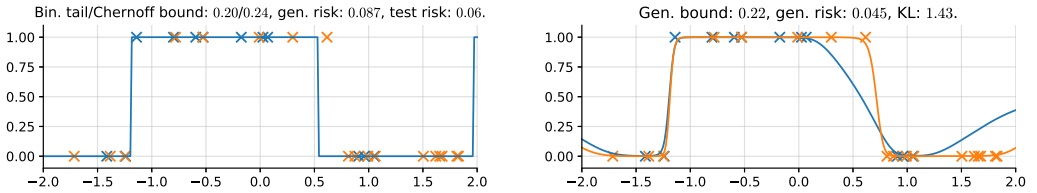
Parametrising the Meta-Learner and Hypothesis Space. We now describe how to parametrise the hypothesis space $\mathcal{H}_\theta$ and the maps Q_θ, P_θ. We meta-learn a feature map $\phi_\theta \colon \mathbb{R} \to \mathbb{R}^K$ and choose $\mathcal{H}_\theta = \{h_w : h_w(x) = \mathrm{sign}\langle w, \phi_\theta(x)\rangle,\, w \in \mathbb{R}^K\}$. For Q_θ and P_θ Gaussian, this hypothesis space allows us to compute the empirical Gibbs risk without Monte Carlo integration; see App I.3 for details. For the form of Q_θ, we take inspiration from Neural Processes (NPs) [Garnelo et al., 2018a,b, Kim et al., 2019]. NPs use neural networks to flexibly parametrise a map from datasets to predictive distributions that respects the permutation invariance of datasets [Zaheer et al., 2017].

[4] We could also consider drawing all datasets from a *single* task D, which would more directly match Sec 3. We regard this case as less interesting, since we would often want a bound to perform well on a variety of tasks.



(a) $N = 30$ datapoints. (b) $N = 60$ datapoints.

Figure 3: **Average generalisation bound and actual generalisation risk** ($\pm$ two standard errors) for CNN-NP meta-learners trained to optimise Catoni (▼), MLS (■), optimistic MLS (●), learned convex (♦), Chernoff test set (◀), and binomial tail test set (▶) bounds. Catoni, optimistic MLS, and learned convex overlap. The generalisation risks for Chernoff and binomial tail test set bounds are identical as they share the same meta-learner; only the bound computation differs. The bounds are valid except for optimistic MLS, which should be a lower bound on the best bound achievable with Thm 3. Corresponding plots for the MLP-NP are in App J.2.



(a) **Binomial tail/Chernoff test set** bounds, showing the learned hypothesis (—), the train set (×) of size 12 and the test set (×) of size 18.

(b) **Learned convex** bound with data-dependent prior, showing the prior (—) and posterior (—) predictive, prior set (×) of size 12 and risk set (×) of size 18.

Figure 4: **Predictions on one of the 1D datasets in the meta-test set** with $N = 30$ and prior/train proportion 0.4. For each method, we report the generalisation bound and actual generalisation risk. For the test set model, we also show the risk on the test set, and for the PAC-Bayes model we show the KL-divergence. The learned convex bound meta-learner has learned a DDP that provides a "first guess" given the prior set, which is then refined by the posterior. Figures for other PAC-Bayes bounds and datasets are provided in App J.1.

They are regularly benchmarked on 1D meta-learning tasks, making them ideally suited. We make a straightforward modification to NPs to make them output Gaussian measures over weight vectors $w \in \mathbb{R}^K$. Hence, they act as parametrisable maps from Z^N to the set of Gaussian measures on $\mathbb{R}^K$.

We considered two kinds of NP, one based on multilayer perceptrons (MLP-NP) and another based on convolutional neural networks (CNN-NP) (detailed in Apps I.5 and I.6) Although the MLP-NP is very flexible, the state-of-the-art in NPs on 1D tasks is given by CNN-based NPs [Gordon et al., 2020, Foong et al., 2020, Bruinsma et al., 2021]. We use an architecture closely based on the *Gaussian Neural Process* [Bruinsma et al., 2021], which outputs full-covariance Gaussians. As expected, we found the CNN-NP to produce tighter (or comparable) average bounds to the MLP-NP, while using far fewer parameters, and training much more reliably and quickly. Hence, we focus on the CNN-NP, but report some results for the MLP-NP in App J.2. Hyperparameter details are given in App I.7.

Results. We show example classification tasks and average bounds on the meta-test set in Figs 3 and 4. Note that the test set classifier became deterministic and makes hard predictions whereas the PAC-Bayes classifier shows uncertainty; see App I.2 for a discussion. The MLS bound is loosest, which is unsurprising as it has no optimisable parameters to adapt to $\mathcal{T}$.[5] Surprisingly, the results for Catoni, optimistic MLS, and learned convex are nearly identical. As long as optimisation has succeeded reasonably, Cor 3 then implies that one of the Catoni bounds is very nearly optimal among all convex functions for this task distribution — there is not much "slack" from choosing suboptimal Δ here. We also see that the Catoni and learned convex bounds with prior proportion 0.4 are tighter than any Chernoff test set bound. Hence, *PAC-Bayes can provide slightly tighter (or comparable) generalisation bounds to a Chernoff test set bound*. However, we see that *the binomial tail test set*

[5]This is in contrast with usual applications of PAC-Bayes, where one does not have a meta-dataset with which to optimise parameters of the bound. In that setting, it can be an advantage to not have tunable parameters.

bound with train set proportion 0.4 *leads to the tightest generalisation bounds overall*. Cor 3 sheds light on this behaviour: the optimal generic PAC-Bayes bound reduces, at best, to the Chernoff test set bound when the posterior equals the prior. However, the Chernoff bound is itself looser than the binomial tail bound. Of course, the posterior does not equal the prior here, but Cor 3 indicates there is an extra source of looseness that PAC-Bayes has to overcome relative to the binomial tail bound. Finally, although the test set meta-learner leads to the tightest generalisation bounds, its generalisation risk is roughly double that of the PAC-Bayes meta-learner when the prior/train set proportion is 0.4.

5 Conclusions, Open Problems, and Limitations

PAC-Bayes presents a potentially attractive framework for obtaining tight generalisation bounds in the small-data regime. We have investigated the tightness of PAC-Bayes and test set bounds in this regime both theoretically and experimentally. Theoretically, we showed that the generic PAC-Bayes theorem of Germain et al. [2009] and Bégin et al. [2016] which encompasses a wide range of PAC-Bayes bounds, cannot produce tighter bounds in expectation than the expression obtained by discarding the $\log(2\sqrt{N})/N$ term in the Langford and Seeger [2001] bound (*i.e.*, the *optimistic MLS bound*; Cor 3). Although we did not prove that the optimistic MLS bound is a valid generalisation bound, numerical evidence suggests (Figs 2c and 2d) that there may exist a convex function Δ which achieves it, at least for the distributions over empirical risk and KL-divergence we considered. This suggests the following open problem:

Open Problem 1. *For an arbitrary distribution over datasets, does there exist a choice of* Δ *such that the expected optimistic MLS bound is attained (Cor 3)? If not, how close can one get to the expected optimistic MLS bound?*

If such a Δ exists, then that would imply the optimistic MLS bound is a valid generalisation bound (see Sec 3) and resolve Problem 6.1.2 of Langford [2002] in the affirmative.

We then considered, in a controlled experimental setting where meta-learning all parameters of the bounds and learning algorithms was feasible, whether PAC-Bayes bounds could be tighter than test set bounds. Although we found PAC-Bayes competitive with Chernoff bounds, both were outperformed by the binomial tail test set bound. This motivates a second open problem:

Open Problem 2. *Can a PAC-Bayes bound be found that relaxes gracefully to the binomial tail test set bound (Thm 1) when the posterior is equal to the prior?*

Resolving these problems could have a significant impact on the tightness of PAC-Bayes applied to small-data, and clarify our understanding of the relationship between PAC-Bayes and test set bounds.

Limitations. In this paper, we concern ourselves with understanding the tightness of bounds in what might be called the *standard PAC-Bayes setting* of supervised learning, bounded losses, i.i.d. data and Gibbs risk. We also focus on bounds that are 'first order' in the sense that they rely only on the empirical Gibbs risk, though extending the analysis to consider other PAC-Bayes theorems (*e.g.* Tolstikhin and Seldin [2013]) would be of interest, especially with regards to Open Problems 1 and 2. For many practical applications in which performance guarantees are needed (*e.g.* health care), the i.i.d. assumption should be considered carefully, as it is likely an unrealistic simplification. Furthermore, Gibbs classifiers are not often used in practice. To address these and other concerns, PAC-Bayes has been generalised in many directions beyond the scope of the standard setting we consider. Examples include bounds for non-i.i.d. data [Seldin et al., 2012, Alquier and Guedj, 2018, Rivasplata et al., 2020], unbounded losses [Germain et al., 2016], derandomised classifiers [Blanchard and Fleuret, 2007, Viallard et al., 2021], and Bayes risk [Germain et al., 2015, Masegosa et al., 2020]. Bounds based on other divergences besides the KL have also been proposed [Bégin et al., 2016, Alquier and Guedj, 2018]. As our proof relies primarily on tools from convex analysis, and Jensen's inequality is ubiquitous in PAC-Bayes bounds, it would be interesting to see if our arguments can be extended beyond the limited setting we focus on. Finally, our meta-learning experiments only considered 1D classification, and the results might not necessarily be representative of more realistic datasets. Scaling these up is an important, but potentially challenging, avenue for future work.

References

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

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. *Advances in neural information processing systems*, 2007.

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, 2018.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2016.

Gilles Blanchard and François Fleuret. Occam's hammer. In *International Conference on Computational Learning Theory*, pages 112–126. Springer, 2007.

Wessel P Bruinsma, James Requeima, Andrew YK Foong, Jonathan Gordon, and Richard E Turner. The Gaussian neural process. In *Proceedings of the 3nd Symposium on Advances in Approximate Bayesian Inference*, Jan 2021.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. volume 56 of *Lectures Notes and Monographs*. Institute of Mathematical Statistics, Beachwood Ohio, USA, Dec 2007.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, pages 1078–1086. PMLR, 2018.

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

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

Alec Farid and Anirudha Majumdar. PAC-bus: Meta-learning bounds via PAC-Bayes and uniform stability. *arXiv preprint arXiv:2102.06589*, 2021.

Andrew Y. K. Foong, Wessel P. Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard E. Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. In *Advances in Neural Information Processing Systems*, Jul 2020.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

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

Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario March, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015.

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

Jonathan Gordon, Wessel P Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. In *International Conference on Learning Representations (ICLR), 8th*, 2020.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations, ICLR*, 2014.

John Langford. *Quantitatively tight sample complexity bounds*. PhD thesis, Carnegie Mellon University, 2002.

John Langford. Tutorial on practical prediction theory for classification. *Journal of machine learning research*, 6(3), 2005.

John Langford and Rich Caruana. (Not) bounding the true error. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

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

Tianyu Liu, Jie Lu, Zheng Yan, and Guangquan Zhang. PAC-Bayes bounds for meta-learning with data-dependent prior. *arXiv preprint arXiv:2102.03748*, 2021.

Andres Masegosa, Stephan Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. In *Advances in Neural Information Processing Systems*, 2020.

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

David McAllester. A PAC-Bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

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

David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

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

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *arXiv preprint arXiv:2007.12911*, 2020.

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

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. *arXiv preprint arXiv:2006.13057*, 2020.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. *arXiv preprint arXiv:2002.05551*, 2020.

Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning how to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 1987.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12): 7086–7093, 2012.

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

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

Ilya O Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems*, 2013.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the derandomization of PAC-Bayesian bounds. *arXiv preprint arXiv:2102.08649*, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems 30*, 2017.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. In *Proceedings of the 7th International Conference on Learning Representations*, May 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the limitations section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We provide a brief discussion in the limitations section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See the theory section.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the theory section and lemmas section in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code is included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the appendix on hyperparameters.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, we report error bars over random seed in Fig 2 and over datasets in Fig 3. We do not rerun the meta-learning experiments with different initialisations as this is too computationally expensive.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We used roughly 500-1000 GPU-hours on NVIDIA Tesla V100 and GeForce RTX 2080 graphics cards using both an internal cluster and Amazon Web Services. See the appendix on hyperparameters.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]