On Margins and Generalisation for Voting Classifiers

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the generalisation properties of majority voting on finite ensembles of classifiers, proving margin-based generalisation bounds via the PAC-Bayes theory. These provide state-of-the-art guarantees on a number of classification tasks. Our central results leverage the Dirichlet posteriors studied recently by Zantedeschi et al. [1] for training voting classifiers; in contrast to that work our bounds apply to non-randomised votes via the use of margins. Our contributions add perspective to the debate on the "margins theory" proposed by Schapire et al. [2] for the generalisation of ensemble classifiers.

1 Introduction

Weighted ensemble methods are among the most widely-used and effective algorithms known in machine learning. Variants of boosting [3, 4] are state-of-the-art in a wide variety of tasks [5, 6] and methods such as random forest [7] are among the most commonly-used in machine learning competitions [see, *e.g.*, 8, 9], valued both for their excellent results and interpretability. Even when these algorithms do not directly produce the best learners for a task, the best performance in competitions is often obtained by an ensemble of "strong learners"—the output of a collection of different algorithms trained on the data—contrasted to the weak learners usually considered in the ensemble learning literature.

Among the oldest ideas to explain the performance of ensemble classifiers, and machine learning methods in general, is the concept of margins. First introduced to analyse the Perceptron algorithm [10], margins relate closely to the idea of confidence in predictions in ensemble learning, with a large margin implying that a considerable weighted fraction of voters chose the same answer. This was first leveraged to obtain early margin-based generalisation bounds for ensembles by Schapire et al. [2], in an attempt to understand the excellent generalisation of boosting, a surprising result given classical Vapnik-Chervonenkis theory. This "margins theory" was explored further in a number of works [11–13] and is among the leading explanations for the success of such methods and boosting in particular.

The same thread of margin bounds for ensemble methods has also been taken up in parallel in PAC-Bayes theory by Langford and Seeger [14], Biggs and Guedj [15]. PAC-Bayes provides a natural framework both for deriving margin bounds, and for considering ensemble methods in general, particularly majority votes where the largest-weighted ensemble prediction is taken. Within the framework, the weightings are typically considered as the parameter of a categorical distribution over individual voters. PAC-Bayes theorems [see the comprehensive surveys of 16, 17] then directly provide generalisation bounds for the performance of this "randomised" proxy for the majority vote, *a.k.a.* Gibbs classifier. These can then be de-randomised by such margin-based techniques, or through a variety of oracle bounds [18–21], motivating new learning algorithms [22–28].

Uniquely among PAC-Bayesian approaches, Zantedeschi et al. [1] instead consider Dirichlet distributions over the voters. Any sample from this distribution already implies a vector of voting weights, and it is on the performance and optimisation of these "stochastic majority votes" they primarily

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

focus. As an aside, they provide an oracle result which allows their bounds to be de-randomised, but this introduces an irreducible factor such that the bound on the true fixed vote can never be less than double that of the stochastic version. It also neglects to leverage the generally high confidence of predictions obtained by their algorithm.

Our contribution. By combining tools from margin bounds and the use of Dirichlet majority votes, we provide a new margin bound for non-randomised majority votes. This is in contrast to Zantedeschi et al. [1] which primarily considers stochastic majority votes. Our bound empirically compares very favourably to existing margin bounds and in contrast to them are applicable to multi-class classification. Remarkably, our empirical results are also sharper than existing PAC-Bayesian ones, even when the algorithm optimising those bounds is used.

Our primary tool is a new result relating the margin loss of these stochastic votes to the misclassification loss of the non-randomised ones in a surprisingly sharp way. This tool can additionally be utilised alongside a further idea from Zantedeschi et al. [1] to obtain an alternative form of the bound which is more amenable to optimisation. Through this work we provide further support to the margins theory for ensembles, showing that near-sharp bounds based on margins alone can be obtained on a variety of real-world tasks.

Outline. The rest of this section introduces the problem setup, notation and summarises main results. Section 2 provides background on PAC-Bayes, Dirichlet majority votes and margin bounds, relating them to our new results. Section 3 states and summarises our new theoretical results, giving the most relevant proofs (all remaining proofs are deferred to appendices). Section 4 empirically evaluates these new results before we conclude with an overall discussion in Section 5.

1.1 Notation and setting

Majority voting algorithms combine the predictions of a finite set of "base" classifiers, $\mathcal{H}$, from $\mathcal{X}$ to $\mathcal{Y} = [c] := \{1, \ldots, c\}$. The classifiers $h_i \in \mathcal{H}$ take the form $h_i : \mathcal{X} \to \mathcal{Y}$ for $i \in [d]$ so that $|\mathcal{H}| = d$. Majority votes consider as set of weightings $\boldsymbol{\theta}$ in Δ^d, the simplex, and return the highest-weighted overall prediction. Using the indicator function $\boldsymbol{I}_{\mathcal{A}}$ of a set $\mathcal{A}$, this is expressed as

$$f_{\boldsymbol{\theta}}(x) = \operatorname*{argmax}_{k \in \mathcal{Y}} \sum_{i \in [d]} \theta_i \boldsymbol{I}_{h_i(x)=k}.$$

We are primarily interested in learning a weighting $\boldsymbol{\theta}$ with small misclassification risk (and guarantees of this) based on a sample $S \sim \mathcal{D}^m$, where $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$ is the data-generating distribution and $m \in \mathbb{N}_+$ the sample size. We let $\mathcal{M}_1^+(\mathcal{A})$ denote the set of probability measures on a set $\mathcal{A}$. For $h \in \mathcal{H}$ the misclassification loss is $\ell_0(h, x, y) := \boldsymbol{I}_{h(x) \neq y}$, the misclassification out-of-sample risk is $L_0(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell_0(h, x, y)$ and a hat denotes the in-sample estimate of this quantity, $\hat{L}_0(h) := \mathbb{E}_{(x,y)\sim\mathrm{Uniform}(S)} \ell_0(h, x, y)$. In a slight abuse of notation we will also often write the risk of the majority vote $L_0(\boldsymbol{\theta}) = L_0(f_{\boldsymbol{\theta}})$ and similarly for its empirical counterpart.

The margin of majority vote $f_{\boldsymbol{\theta}}$ on example (x, y) is derived from the minimal gap between the total weight assigned to the true class y and to any other predicted class:

$$M(\boldsymbol{\theta}, x, y) := \frac{1}{2} \sum_{i : h_i(x) = y} \theta_i - \frac{1}{2} \max_{k \neq y} \sum_{i : h_i(x) = k} \theta_i.$$

The corresponding margin loss is $\ell_\gamma(\boldsymbol{\theta}, x, y) := \boldsymbol{I}_{M(\boldsymbol{\theta}, x, y) \leq \gamma}$ for margin $\gamma \geq 0$, with the corresponding in-sample and out-of-sample risks notated as $\hat{L}_\gamma(\boldsymbol{\theta})$ and $L_\gamma(\boldsymbol{\theta})$ respectively.

1.2 Overview of results

Our main result is a margin bound of the following form: with high probability $\geq 1 - \delta$ over the sample, simultaneously for any $\boldsymbol{\theta} \in \Delta^d$ and $K > 0$,

$$L(\boldsymbol{\theta}) \leq O\left(\hat{L}_\gamma(\boldsymbol{\theta}) \; + \; e^{-K\gamma^2} \; + \; \frac{\mathbb{D}_{\mathrm{Dir}}(K\boldsymbol{\theta}, \mathbf{1}) + \log \frac{m}{\delta}}{m} \right) \tag{1}$$

where $\mathbb{D}_{\mathrm{Dir}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the KL divergence between Dirichlet random vectors with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, with $\mathbf{1}$ a vector of ones implying a uniform Dirichlet prior distribution on the simplex. The term

$e^{-K\gamma^2}$ is a de-randomisation penalty. The parameter K is chosen freely in an arbitrary data-dependent way to balance the requirements of the different terms: it must be large enough to decrease this exponential term, while too-large a parameter increases the KL divergence from the uniform prior. This result is surprisingly strong; in particular there is no dependence on the dimensionality (*i.e.*, number of voters d) in the exponential term, an advantage discussed further in Section 3.2.

In Equation (1), $\hat{L}_\gamma(\boldsymbol{\theta})$ is the 0-1 valued γ-margin loss which enables comparison with existing margin bounds for trained weighted ensembles. We further consider a second scenario, where the generalization bound is also used to train the model itself. We note that the γ-margin loss $\hat{L}_\gamma(\boldsymbol{\theta})$ appearing in Equation (1) has null gradients, so the bound cannot be directly optimised by gradient descent. To rectify this we also prove a variation of the bound, replacing the above loss by its expectation under a Dirichlet stochastic vote, $\mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(K\boldsymbol{\theta})}\hat{L}_\gamma(\boldsymbol{\xi})$, which is bounded in differentiable closed form to give an alternative, optimisation-friendly bound.

In our evaluations we focus on these two complementary scenarios, obtaining state-of-the-art empirical results. Across different scenarios and tasks our results outperform both existing margin bounds (including a sharpened version of the result from Biggs and Guedj [15] which may be of independent interest), and PAC-Bayes bounds, even when it is not used as the objective. Further, in contrast to existing margin bounds our results also hold for multi-class majority votes.

2 Background

2.1 PAC-Bayes bounds

PAC-Bayes bounds are among the tightest known generalisation bounds, as for example the only framework in which non-vacuous generalisation bounds for neural networks have been obtained [see *e.g.* 29–36]. However, unlike many other such bounds they usually apply to randomised Gibbs(-like) prediction functions rather than deterministic ones. These are typically re-drawn for every new test evaluation. Thus a high-probability bound is obtained on the expectation of the risk w.r.t. the PAC-Bayes posterior Q, with the complexity of $Q \in \mathcal{M}_1^+(\mathcal{H})$ appearing in the bound in terms of a Kullback-Leibler (KL) divergence from a pre-chosen PAC-Bayes prior $P \in \mathcal{M}_1^+(\mathcal{H})$ [which is not required to be a true prior in the Bayesian sense – see the discussion in 16]. A particularly sharp [as discussed in 37] and widely-used result is given in Theorem 1, valid for any bounded loss function ℓ with values in $[0, 1]$.

Theorem 1 (Seeger et al. [38], Maurer [39])**.** *For any $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}_+$, prior $P \in \mathcal{M}_1^+(\mathcal{H})$ and $\delta \in (0, 1)$, with probability $\geq 1-\delta$ over $S \sim D^m$, simultaneously for all $Q \in \mathcal{M}_1^+(\mathcal{H})$*

$$\mathbb{E}_{h\sim Q} L(h) \leq \mathrm{kl}^{-1}\left(\mathbb{E}_{h\sim Q}\hat{L}(h),\ \frac{1}{m}\left(\mathrm{KL}(Q, P) + \log\frac{2\sqrt{m}}{\delta}\right)\right)$$

where the generalised inverse $\mathrm{kl}^{-1}(u, c) := \sup\{v \in [0, 1] : \mathrm{kl}(u, v) \leq c\}$ and $\mathrm{kl}(u, v) := u \log\frac{u}{v} + (1-u)\log\frac{1-u}{1-v}$ is a KL divergence between Bernoulli random variables.

The above bound uses the inverse small-kl function which will be seen in our later results and a number of pre-existing ones. To lend intuition we note that $\mathrm{kl}^{-1}(u, c) \in O(u + c)$, giving Equation (1) from Theorem 2 when using a uniform prior. The following upper bounds are also useful: $\mathrm{kl}^{-1}(u, c) \leq u + \sqrt{c/2}$ giving "slow-rates" and $\mathrm{kl}^{-1}(u, c) \leq u + \sqrt{2cu} + 2c$. From this we can see that when the loss $\hat{L} \to 0$ then the overall rate improves to $O(1/m)$, so the small-kl formulation interpolates between the traditional fast and slow rate regimes of learning theory.

2.2 Margin bounds

In the learning theory literature there exists a rich tradition of using the concept of a margin, which quantifies the confidence of predictions, to explain generalisation. This is particularly evident in the case of voting algorithms such as boosting, where traditional Vapnik-Chervonenkis based techniques predict classical overfitting which is not ultimately observed. The "margins theory" was developed by Schapire et al. [2] to explain this discrepancy. By considering the weightings $\boldsymbol{\theta}$ as the parameter of a categorical distribution, they proved a bound of the form (holding with probability greater than $1-\delta$

over the sample, as for all bounds in this section) $L_0(\boldsymbol{\theta}) \leq L_\gamma(\boldsymbol{\theta}) + \tilde{O}\left(\frac{1}{\gamma\sqrt{m}}\right)$. Although there was initially some debate about the validity of the theory [40], eventually Gao and Zhou [12, Theorem 4] provided the following improved bound which further supported that a large-margin voting classifier could generalise: simultaneously for any $\gamma > \sqrt{2/d}$ and $\boldsymbol{\theta} \in \Delta^d$,

$$L_0(\boldsymbol{\theta}) \leq \mathrm{kl}^{-1}\left(\hat{L}_\gamma(\boldsymbol{\theta}),\ \frac{1}{m}\left(\frac{2\log(2d)}{\gamma^2}\log\frac{2m^2}{\log d} + \log\frac{dm}{\delta}\right)\right) + \frac{\log d}{m}. \tag{2}$$

More recently, a similar bound (proved through a PAC-Bayesian method based on Seeger et al. [38]) was proved in Biggs and Guedj [15, Theorem 8]. Here we give a bound provided as an intermediate step in their proof that is strictly (and empirically considerably) sharper than their final result: for any fixed margin $\gamma > 0$, simultaneously for any $\boldsymbol{\theta} \in \Delta^d$

$$L_0(\boldsymbol{\theta}) \leq \mathrm{kl}^{-1}\left(\hat{L}_\gamma(\boldsymbol{\theta}) + \frac{1}{m},\ \frac{1}{m}\left(\lceil 2\gamma^{-2}\log m\rceil \log d + \log\frac{2\sqrt{m}}{\delta}\right)\right) + \frac{1}{m}. \tag{3}$$

Since $\gamma \in (0, \frac{1}{2})$ for non-vacuous results, a union bound argument can be used to extend the above to fixed-precision γ, and this result has the advantage of being valid for small γ as are often observed empirically.

Our contributions. Firstly we mention the smaller contribution of the improved form of the bound from Biggs and Guedj [15] given in Equation (3); a proof is given in Appendix B alongside further refinements and evaluation. However we show that in many cases even this improved version and Equation (2) give weak or vacuous results. As a result of this weakness (and thus perhaps null result for the margins theory applied to voting classifiers) we present a new margin bound in Theorem 2 based on Dirichlet distributions as a theoretical intermediate step. This is also valid in the multi-class case, unlike the above results which are only for binary classification. Empirically the bound is observed to give an enormous improvement in tightness than the existing margin bounds and in some cases is near-sharp.

2.3 Dirichlet stochastic majority votes

In most results from the PAC-Bayes framework, and in the proof of the existing results given in Section 2.2, the majority vote weightings $\boldsymbol{\theta}$ are considered the parameters of a categorical distribution over voters. Zantedeschi et al. [1] instead consider PAC-Bayesian bounds (specifically, Theorem 1) applied to a hypothesis class of majority votes of the form $f_{\boldsymbol{\xi}}$, where $\boldsymbol{\xi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ is drawn from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$. This distribution has mean $\mathbb{E}\boldsymbol{\xi} = \boldsymbol{\alpha}/\sum_{i=1}^d \alpha_i$ with a larger sum $\sum_{i=1}^d \alpha_i$ giving a more concentrated or peaked distribution (see Appendix A for more details).

Since $\boldsymbol{\xi}$ is randomised, the bounds from Zantedeschi et al. [1] apply to "stochastic majority votes" rather than the more typical deterministic ones we consider here. However, the use of such Dirichlet distributions over voters in the PAC-Bayes bounds rather than the more usual categorical ones is a major step forward as it allows the correlation between voters to be more carefully considered. We will utilise and de-randomise these as a stepping stone to ones for deterministic predictors $f_{\boldsymbol{\theta}}$ directly.

As is common in the PAC-Bayes literature, Zantedeschi et al. [1] use their new bound as an optimisation objective to obtain a new algorithm, here using stochastic gradient descent. The bound with Dirichlet posterior obtained directly from Theorem 1 includes the expected misclassification loss with respect to the Dirichlet parameters, $\mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\ell(f_{\boldsymbol{\xi}}, x, y)$, which has null gradient for any sampled $\boldsymbol{\xi}$. They therefore additionally upper bound this term by the differentiable closed form

$$\mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\ell(f_{\boldsymbol{\xi}}, x, y) \leq I_{\frac{1}{2}}\left(\sum_{i:h_i(x)=y}\alpha_i,\ \sum_{i:h_i(x)\neq y}\alpha_i\right), \tag{4}$$

where $I_z(a, b)$ is the regularised incomplete beta function, which has a sigmoidal shape. The inequality is sharp in the binary classification case, and is used in the training objective and final evaluation of their method. As an aside, Zantedeschi et al. [1] also proved an oracle bound which allows their result to be de-randomised, but this introduces a irreducible factor of two.

Our contributions. Firstly, we provide a new margin bound for majority vote algorithms utilising Dirichlet posteriors as a theoretical stepping stone. We show that this bound gives sharper bounds on

the misclassification loss than the bound from Zantedeschi et al. [1], doing better than the irreducible factor, even when applied to the output of their algorithm. We show further that the bound is also tighter when applied to the outputs of other PAC-Bayes algorithms derived from "categorical"-type posteriors. Finally, we give an altered form of the bound involving the expectation of the margin loss $\mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\ell_\gamma(f_{\boldsymbol{\xi}}, x, y)$ and a result analogous to Equation (4) for this case. Through this we are able to obtain a new PAC-Bayes objective which is compared to existing PAC-Bayes optimisation methods.

3 Main results

Our main results use the idea of Dirichlet stochastic majority votes from Zantedeschi et al. [1] as an intermediate step to prove new margin bounds for deterministic majority votes. In this section, first we give our main result in Theorem 2 and discuss further. In Section 3.1 we give an alternative bound obtained by a very similar method which is more amenable to optimisation, and we provide proofs for these results in Section 3.2.

The central step in these proofs is in constructing a proxy Dirichlet distribution $\boldsymbol{\xi} \sim \mathrm{Dir}(K\boldsymbol{\theta})$ over voters, the loss of which is bounded à la PAC-Bayes, and de-randomised using margins to obtain bounds directly for $f_{\boldsymbol{\theta}}$. The primary complexity term appearing in our bounds is therefore $\mathbb{D}_{\mathrm{Dir}}(K\boldsymbol{\theta}, \boldsymbol{\beta})$, the KL divergence between Dirichlet distributions with parameters $K\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. As with PAC-Bayes priors, $\boldsymbol{\beta}$ can be chosen in arbitrary sample-independent fashion, but we typically choose it as a vector of ones, giving a uniform distribution on the simplex as prior as in Equation (1). The bounds also involve a de-randomisation penalty of $O(e^{-K\gamma^2})$ where γ is the margin appearing in the loss; this term upper bounds the difference between our randomised proxy $\boldsymbol{\xi}$ and its mean $\boldsymbol{\theta}$ and gets smaller with K as the distribution concentrates tightly around its mean. This parameter K can be optimised in any data-dependent way to obtain the tightest final bound.

Theorem 2. *For any $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}_+$, margin $\gamma > 0$, $\delta \in (0, 1)$, and prior $\boldsymbol{\beta} \in \mathbb{R}_+^d$, with probability at least $1 - \delta$ over the sample $S \sim \mathcal{D}^m$ simultaneously for every $\boldsymbol{\theta} \in \Delta^d$ and $K > 0$,*

$$L_0(\boldsymbol{\theta}) \leq \mathrm{kl}^{-1}\left(\hat{L}_\gamma(\boldsymbol{\theta}) + e^{-(K+1)\gamma^2},\ \frac{\mathbb{D}_{\mathrm{Dir}}(K\boldsymbol{\theta}, \boldsymbol{\beta}) + \log\frac{2\sqrt{m}}{\delta}}{m}\right) + e^{-(K+1)\gamma^2}.$$

Theorem 2 differs from the existing margin bounds of Equations (2) and (3), and Schapire et al. [2] in a specific and significant way, with $\boldsymbol{\theta}$ appearing not only in the loss function $\hat{L}_\gamma(\boldsymbol{\theta})$, but *also* in the KL complexity term. Empirically we find our bound to be an improvement but it is possible to generate scenarios where the pre-existing bounds are non-vacuous while ours is not, since the KL divergence is unbounded for certain choices of $\boldsymbol{\theta}$, for example when one of the components is exactly zero. This difference arises because the existing bounds all use the idea of a categorical distribution with parameter $\boldsymbol{\theta}$ in their proofs (which has KL divergence from a uniform prior upper bounded by $\log d$), while we use a Dirichlet. This gains us the surprisingly tight de-randomisation result (Theorem 4) used in all proofs.

3.1 PAC-Bayes bound as objective

We note here that it is non-trivial to directly obtain a training objective for optimisation from Theorem 2, due to the non-differentiability of the margin loss $\hat{L}_\gamma(\boldsymbol{\theta})$. Therefore, in order to compare results with a wide variety of methods that optimise PAC-Bayes bounds [including those used by 1, as baselines], we obtain a relaxed and differentiable formulation in Theorem 3 for direct optimisation.

Theorem 3. *Under the conditions of Theorem 2 the following bound also holds*

$$L_0(\boldsymbol{\theta}) \leq \mathrm{kl}^{-1}\left(\mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(K\boldsymbol{\theta})}\hat{L}_\gamma(\boldsymbol{\xi}),\ \frac{\mathbb{D}_{\mathrm{Dir}}(K\boldsymbol{\theta}, \boldsymbol{\beta}) + \log\frac{2\sqrt{m}}{\delta}}{m}\right) + e^{-4(K+1)\gamma^2}.$$

Using the incomplete Beta function $I_z(a, b)$ we also have the following result, which is sharp in the binary classification case,

$$\mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\ell_\gamma(\boldsymbol{\xi}, x, y) \leq I_{\frac{1}{2}+\gamma}\left(\sum_{i:h_i(x)=y}\alpha_i,\ \sum_{i:h_i(x)\neq y}\alpha_i\right).$$

Theorem 3 has a stronger PAC-Bayesian flavour than Theorem 2, with an expected loss under some distribution appearing (complicating the final optimisation of K), while Theorem 2 takes a form much closer to that of a classical margin bound. The second part of the result is analogous to Equation (4) used by Zantedeschi et al. [1]. We combine both parts to calculate the overall bound in closed form and obtain gradients for optimisation.

3.2 Proof of main results

The proof of Theorems 2 and 3 essentially follow from applying a simple PAC-Bayesian bound in combination with the key Theorem 4 below. In some sense this is our most important and novel result. Our whole approach is largely motivated by its surprising tightness; in particular there is no dependence on the dimension, which is avoided by careful use of the aggregation property of the Dirichlet distribution. This surprise arises because to obtain a tightly concentrated Dirichlet distribution on $\boldsymbol{\xi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, the concentration parameter $K = \sum_{i=1}^d \alpha_i$ must grow linearly with the dimension. In fact, even a uniform distribution (which will be less peaked than our final posterior) has $\sum_{i=1}^d \alpha_i = d$, so the de-randomisation step is effectively very cheap in higher dimensions.

Theorem 4. *Let $\boldsymbol{\theta} \in \Delta^d$ and $K > 0$. Then for any $\gamma > 0$ and (x, y),*

$$\ell_0(\boldsymbol{\theta}, x, y) \leq \mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(K\boldsymbol{\theta})}\ell_\gamma(\boldsymbol{\xi}, x, y) + e^{-4(K+1)\gamma^2},$$
$$\mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(K\boldsymbol{\theta})}\ell_\gamma(\boldsymbol{\xi}, x, y) \leq \ell_{2\gamma}(\boldsymbol{\theta}, x, y) + e^{-4(K+1)\gamma^2}.$$

For our proofs we first recall the aggregation property of the Dirichlet distribution: if $(\xi_1, \ldots, \xi_d) \sim \mathrm{Dir}((\alpha_1, \ldots, \alpha_d))$, then $(\xi_1, \ldots, \xi_{d-1} + \xi_d) \sim \mathrm{Dir}((\alpha_1, \ldots, \alpha_{d-1} + \alpha_d))$. We further note the following crucial concentration-of-measure result. The aforementioned lack of dimensionality in Theorem 4 is possible because Theorem 5 depends only on $\sum_{i=1}^d \alpha_i$, and this value is unchanged by aggregation, which avoids the dimension dependence that would otherwise be introduced by the requirement $\|\boldsymbol{u}\|_2 = 1$ below.

Theorem 5 (41)**.** *Let $\boldsymbol{X} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, $t > 0$, and $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\|_2 = 1$. Then*

$$\mathbb{P}_{\boldsymbol{X}}\left\{\boldsymbol{u}\cdot(\boldsymbol{X} - \mathbb{E}\boldsymbol{X}) > t\right\} \leq \exp\left(-2\left(\sum_{i=1}^d \alpha_i + 1\right)t^2\right).$$

Proof of Theorem 2 and Theorem 3. The proof of our main results is completed by applying the PAC-Bayes bound Theorem 1 with the γ-margin loss to a Dirichlet prior and posterior with parameters $\boldsymbol{\beta}$ and $K\boldsymbol{\theta}$ respectively. Substituting the first part of Theorem 4 gives the first part of Theorem 3, and additionally substituting the second part and re-scaling $\gamma \to \gamma/2$ gives Theorem 2.

For the second part of Theorem 3, define $w = \{i : h_i(x) = y\}$ for fixed (x, y) so $W := \sum_{i\in w} \xi_i \sim \mathrm{Beta}\left(\sum_{i\in w}\alpha_i, \sum_{i\notin w}\alpha_i\right)$ by the aggregation property of the Dirichlet distribution. Then

$$\mathbb{E}_{\boldsymbol{\xi}}\ell_\gamma(\boldsymbol{\xi}, x, y) \leq \mathbb{E}_{\boldsymbol{\xi}}\left\{W \geq \frac{1}{2} - \gamma\right\} = 1 - I_{\frac{1}{2}-\gamma}\left(\sum_{j\in w}\alpha_j, \sum_{i\notin w}\alpha_j\right) = I_{\frac{1}{2}+\gamma}\left(\sum_{j\notin w}\alpha_j, \sum_{i\in w}\alpha_j\right)$$

using $\boldsymbol{I}_{a_i - \max_{j\neq i} a_j \leq 2\gamma} \leq \boldsymbol{I}_{a_i - \sum_{j\neq i} a_j \leq 2\gamma} = \boldsymbol{I}_{\sum_{j\neq i} a_j \geq \frac{1}{2}-\gamma}$ for $\boldsymbol{a} \in \Delta^c$ (with equality for $c = 2$ classes), and that $I_z(a, b)$ is the CDF of a Beta distribution with parameters (a, b). □

Proof of Theorem 4. Define $\gamma_2 > \gamma_1$ such that $\gamma := \gamma_2 - \gamma_1$, and $\boldsymbol{\alpha} = K\boldsymbol{\theta}$. From the trivial inequality $\boldsymbol{I}_{x\in A} - \boldsymbol{I}_{x\in B} \leq \boldsymbol{I}_{x\in A}\boldsymbol{I}_{x\notin B}$ we derive

$$\begin{aligned}
\Delta := \ell_{\gamma_1}(\boldsymbol{\theta}, x, y) - \mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\ell_{\gamma_2}(\boldsymbol{\xi}, x, y) &= \mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\boldsymbol{I}_{M(\boldsymbol{\theta},x,y)\leq 0} - \boldsymbol{I}_{M(\boldsymbol{\xi},x,y)\leq\gamma}] \\
&\leq \mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\boldsymbol{I}_{M(\boldsymbol{\theta},x,y)\leq 0}\boldsymbol{I}_{M(\boldsymbol{\xi},x,y)>\gamma}] \leq \mathbb{E}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}[\boldsymbol{I}_{M(\boldsymbol{\xi},x,y)-M(\boldsymbol{\theta},x,y)>\gamma}] \\
&= \mathbb{P}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left\{\sum_{i:h_i(x)=y}\xi_i - \max_{j'\neq y}\sum_{i:h_i(x)=k'}\xi_i - \sum_{i:h_i(x)=y}\theta_i + \max_{j\neq y}\sum_{i:h_i(x)=k}\theta_i > 2\gamma\right\}
\end{aligned}$$

$$\leq \mathbb{P}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left\{\sum_{i:h_i(x)=y}\xi_i - \sum_{i:h_i(x)=k}\xi_i - \sum_{i:h_i(x)=y}\theta_i + \sum_{i:h_i(x)=k}\theta_i > 2\gamma\right\}$$

where in the last inequality we set $k = \mathrm{argmax}_{k\neq y}\sum_{i:h_i(x)=k}\theta_i$, and use that $\max_j \sum_{i:h_i(x)=j}\theta_i - \max_j \sum_{i:h_i(x)=j}\xi_i \leq \max_j \sum_{i:h_i(x)=j}\theta_i - \sum_{i:h_i(x)=k}\xi_i$ for any k. We rewrite the above in vector form (with inner product denoted $\boldsymbol{u}\cdot\boldsymbol{v}$) as

$$\Delta \leq \mathbb{P}_{\boldsymbol{\xi}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left\{\underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\0\end{bmatrix}}_{\boldsymbol{u}}\cdot\left(\underbrace{\begin{bmatrix}\sum_{i:h_i(x)=y}\xi_i\\ \sum_{i:h_i(x)=k}\xi_i\\ \sum_{i:h_i(x)\notin\{k,y\}}\xi_i\end{bmatrix}}_{\tilde{\boldsymbol{\xi}}} - \underbrace{\begin{bmatrix}\sum_{i:h_i(x)=y}\theta_i\\ \sum_{i:h_i(x)=k}\theta_i\\ \sum_{i:h_i(x)\notin\{k,y\}}\theta_i\end{bmatrix}}_{\mathbb{E}\tilde{\boldsymbol{\xi}}}\right) > \sqrt{2}\gamma\right\}$$

$$= \mathbb{P}_{\tilde{\boldsymbol{\xi}}\sim\mathrm{Dir}(\tilde{\boldsymbol{\alpha}})}\left\{\boldsymbol{u}\cdot(\tilde{\boldsymbol{\xi}} - \mathbb{E}\tilde{\boldsymbol{\xi}}) > \sqrt{2}\gamma\right\}$$

where by the aggregation property of the Dirichlet distribution $\tilde{\boldsymbol{\xi}} \sim \mathrm{Dir}(\tilde{\boldsymbol{\alpha}})$ with

$$\tilde{\boldsymbol{\alpha}} := \left[\sum_{i:h_i(x)=y}\alpha_i, \sum_{i:h_i(x)=k}\alpha_i, \sum_{i:h_i(x)\notin\{k,y\}}\alpha_i\right]^T.$$

Applying Theorem 5 we obtain $\Delta \leq e^{-4(\sum_i \tilde{\alpha}_i+1)\gamma^2} = e^{-4(\sum_{i=1}^d \alpha_i+1)\gamma^2}$. This gives the first inequality by setting $\gamma_1 = 0, \gamma_2 = \gamma$. Setting $\gamma_1 = \gamma, \gamma_2 = 2\gamma$ and swapping $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ gives an almost identical proof (with some signs reversed) of the second inequality. □

4 Empirical evaluation

In this section we empirically validate our results against existing PAC-Bayesian and margin bounds on several classification datasets from UCI [42], LIBSVM[1] and Zalando [43]. Since our main result in Theorem 2 is not associated with any particular algorithm, we use $\boldsymbol{\theta}$ outputted from PAC-Bayes-derived algorithms to evaluate this result against other margin bounds (Figure 1) and PAC-Bayes bounds (Figure 2). We then compare optimisation of our secondary result Theorem 3 with optimising those PAC-Bayes bounds directly (Figure 3). All generalisation bounds given are evaluated with a probability $1-\delta=0.95$. Further details not provided here including tabulated results, description of datasets, training mechanisms and compute are provided in Appendix C.

Strong and weak voters. Similarly to Zantedeschi et al. [1] we consider both using data-independent and data-dependent voters. This brings our experimental setup in line with a common workflow for machine learning practicioners: the training set is sub-divided into a set for training several different strong algorithms, and a second set on which the weightings of these are optimised. More specifically, the weak voter setting, used only for binary classification, uses axis-aligned decision stumps (denoted *stumps*), with thresholds evenly spread over the input space (6 per feature and per class). The stronger voters (denoted *rf*) are learned from half of the training data, while the other half is used for evaluating and optimising the different generalisation bounds (note this reduces m). These take the form of random forests [7] of $M=10$ trees optimising Gini impurity score on $\frac{n}{2}$ bagged samples and $\sqrt{d}$ drawn features for each tree, with unbounded maximal depth.

Optimising γ and K in bounds. In reporting margin bounds we optimise over a grid of margin γ values in $(0, \frac{1}{2})$, and additionally over K for Theorem 2. Since Theorem 2 and Equation (3) as stated require a fixed margin, we apply a union bound over the values in the grid, replacing δ in these bounds with δ/N where N is the number of grid points.

Existing PAC-Bayes bounds. We compare to state-of-the-art PAC-Bayesian bounds (and derived algorithms) for weighted majority vote classifiers: the First Order [18], the Second Order [21], Binomial [20] (with the number of voters set to 100) and the two Chebyshev-Cantelli-based [28] empirical bounds from categorical-type Gibbs classifiers with parameter $\boldsymbol{\theta}$, and we refer to these as

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

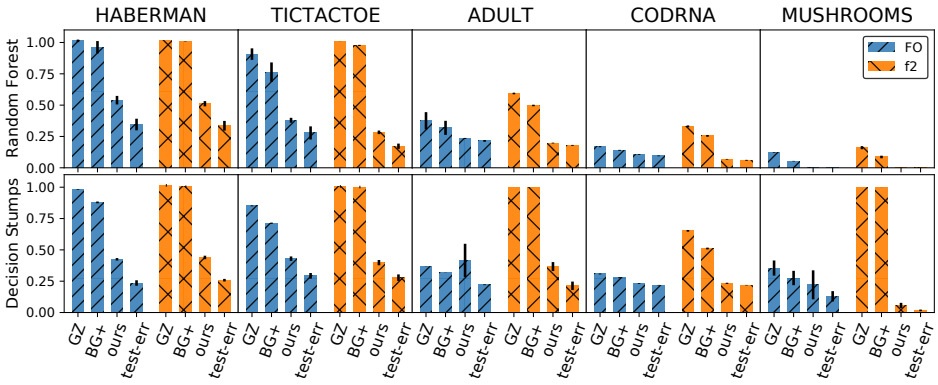

Figure 1: Theorem 2 (**ours**) compared with the margin bounds of Equation (3) (**BG+**), Equation (2) (**GZ**), and the test error. Settings are *rf* (first row) and *stumps* (second row) on the given datasets, with $\boldsymbol{\theta}$ output by optimising either *FO* or *f2* (first and second column groupings respectively).

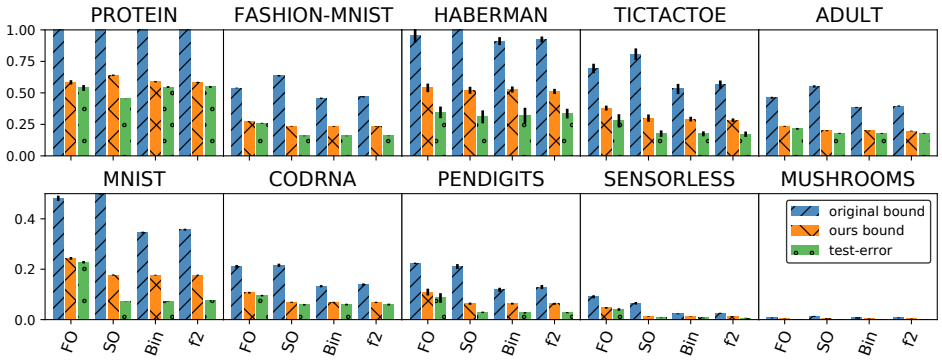

Figure 2: Theorem 2 (*our bound*) compared with the bounds of *FO*, *SO*, *Bin* or *f2* (*original bound*), and test errors. For each column grouping, $\boldsymbol{\theta}$ is the output from optimising the corresponding PAC-Bayes bound for *rf* on the given dataset.

FO, *SO*, *Bin*, *CCPBB* and *CCTND* respectively (more details are given in Appendix C). We denote by *f2* the factor two bound derived in Zantedeschi et al. [1, Annex A.4] from Dirichlet majority votes. All prior distributions for PAC-Bayes bounds, including ours, are set to uniform. We also refer by the same names to the outputs of optimising these bounds with stochastic gradient descent; details on training and initialisation are given in Appendix C.

Description of figures. In Figure 1 we compare Theorem 2 with the existing margin bound of Equation (2) and the improved Biggs and Guedj [15] bound given in Equation (3). Since Equation (3) is strictly better than the original result and the latter was vacuous in almost all cases considered (see Appendix B), we do not include it. All datasets are for binary classification as the existing results only cover this case, and the $\boldsymbol{\theta}$ values considered are the outputs of either the FO- or f2-optimisation using either the weak or the strong voters described above. Figure 2 extends this evaluation of Theorem 2 to improve generalisation results, by applying it to the models optimised with the PAC-Bayes bounds *FO*, *SO*, *Bin* and *f2* as objective. In this case, we consider both binary and multiclass datasets. In Figure 3 we directly compare the outputs of optimising state-of-the-art PAC-Bayesian bounds with our optimisation-ready variant result Theorem 3. These experiments were carried out on strong voters, as standard in the literature [*e.g.* 21, 26, 28].

5 Discussion and conclusion

We observe overall that in many cases the existing margin and PAC-Bayes bounds are insufficient to explain the generalisation observed, while our new bound is consistently tight, and sometimes sharp (*i.e.* it approaches the true test error).

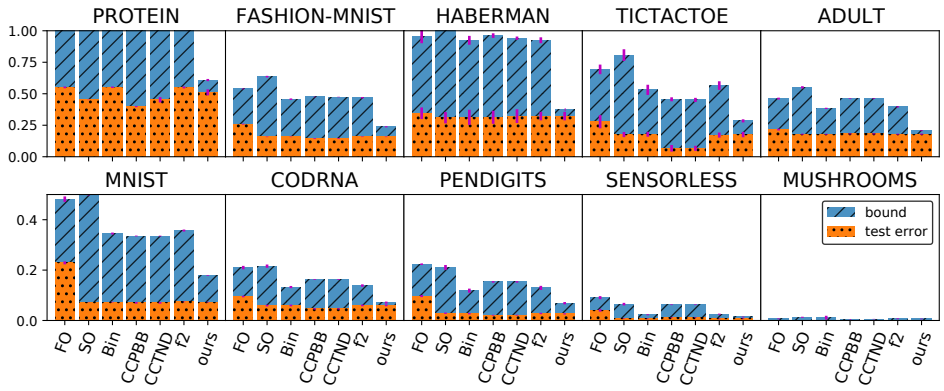

Figure 3: Theorem 3 (**ours**) as optimisation objective compared to other PAC-Bayes results (*FO*, *SO*, *Bin*, *CCPBB* and *CCTND*) as objectives in the *rf* setting. For each objective the test error and bound associated with the objective is shown.

Figure 1 demonstrates that existing margin bounds can be insufficient to explain the generalisation observed, which could be construed as a null result for the "margins theory". However, our new bound obtains empirically very sharp results in almost all cases, reaffirming to the theory. Note that due to the non-convexity of our bound, the reported values are local minima and can potentially be improved by applying a thorougher search for the optimal K, still giving a similarly valid bound. For instance, simply by enlarging the search space for K our bound drops to 0.36 ± 0.10 on *ADULT* with decision stumps as voters, beating existing bounds also in this setting. Unlike the existing results, $\boldsymbol{\theta}$ also arises in the complexity (KL divergence) term and so the bound is not equally tight for every $\boldsymbol{\theta}$ at fixed margin loss. Further examination of this property could add additional nuance and perspective to the theory.

When comparing to existing PAC-Bayes bounds in Figure 2, remarkably Theorem 2 is *always* tighter than just using the bound which is being optimised. We speculate that this arises partially due to the irreducible factors appearing in those bounds; for example the *FO* or *f2* bounds can never be tighter than twice the train loss of the associated Gibbs classifier, while ours has no such limitation. This result is quite valuable as it demonstrates that Theorem 2 can be readily used in an algorithm-free manner: the choice of learning algorithm is up to the practitioner, but the bound will then often provide an excellent guarantee on the obtained weights $\boldsymbol{\theta}$.

Finally, in Figure 3, our optimisation-friendly variant bound Theorem 3 is seen to be competitive in terms of test error while giving an improved-or-equal final bound on all datasets. When considering the less-common setting of binary stumps (see Appendix C) we found that sometimes this objective converged to a sub-optimal local minimum. We speculate that this arises due to the highly non-convex nature of the objective combined with a strong K-inflating gradient signal from the $O(e^{-K\gamma^2})$ term. Thus future work to improve these results even further could start with the use of the quasi-convex small-kl relaxation from Thiemann et al. [44]. We note however that this is overall less important than our main results, as both our bounds are still extremely tight when used in an algorithm-free way and applied to the output of another algorithm as discussed above.

Overall, we note that in many cases (a majority in Figure 2) our main bound of Theorem 2 is very close to the test set bound and thus cannot actually be improved any further, with the problem of providing sharp guarantees based on the training data alone effectively solved in many cases.

Conclusion. We obtain empirically very strong generalisation bounds for voting classifiers using margins. We believe these are highly relevant to the community, since voting-based classifiers and margin-maximising algorithms are among the most popular and influential in machine learning. Dirichlet majority votes have already obtained excellent results in the stochastic setting [1], but our new result in Theorem 4 showing they are well-approximated by their mean should open new directions in the more conventional deterministic setting.

Our results also have practical relevance: for example, in the strong voter machine learning workflow described above, instead of setting data aside as a test set, this data can be freed up to learn even

stronger voters, since a strong out-of-sample ensemble guarantee can still be provided even *without* a test set.

In future work we hope to expand these results further to other (non-majority) voting schemes like those with score-output voters [as in *e.g.* 2], and ensembles of voters with finite VC dimension.

References

[1] Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning stochastic majority votes by minimizing a pac-bayes generalization bound. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 455–467, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0415740eaa4d9decbc8da001d3fd805f-Abstract.html.

[2] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998. doi: 10.1214/aos/1024691352.

[3] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

[5] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Inf. Fusion*, 81: 84–90, 2022. doi: 10.1016/j.inffus.2021.11.011. URL https://doi.org/10.1016/j.inffus.2021.11.011.

[6] Didrik Nielsen. Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU, 2016.

[7] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[8] Robert M. Bell and Yehuda Koren. Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl.*, 9 (2):75–79, dec 2007. ISSN 1931-0145. doi: 10.1145/1345448.1345465. URL https://doi.org/10.1145/1345448.1345465.

[9] Thomas Uriot, Dario Izzo, Luís F Simões, Rasit Abay, Nils Einecke, Sven Rebhan, Jose Martinez-Heras, Francesca Letizia, Jan Siminski, and Klaus Merz. Spacecraft collision avoidance challenge: design and results of a machine learning competition. *Astrodynamics*, pages 1–20, 2021.

[10] A. B. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622, New York, NY, USA, 1962. Polytechnic Institute of Brooklyn.

[11] Liwei Wang, Masashi Sugiyama, Cheng Yang, Zhi-Hua Zhou, and Jufu Feng. On the margin explanation of boosting algorithms. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 479–490. Omnipress, 2008. URL http://colt2008.cs.helsinki.fi/papers/08-Wang.pdf.

[12] Wei Gao and Zhi-Hua Zhou. On the doubt about margin explanation of boosting. *Artif. Intell.*, 203:1–18, 2013. doi: 10.1016/j.artint.2013.07.002. URL https://doi.org/10.1016/j.artint.2013.07.002.

[13] Allan Grønlund, Lior Kamma, and Kasper Green Larsen. Margins are insufficient for explaining gradient boosting. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/146f7dd4c91bc9d80cf4458ad6d6cd1b-Abstract.html.

[14] John Langford and Matthias Seeger. Bounds for averaging classifiers, 2001. URL https://www.cs.cmu.edu/~jcl/papers/averaging/averaging_tech.pdf.

[15] Felix Biggs and Benjamin Guedj. On margins and derandomisation in PAC-Bayes. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3709–3731. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/biggs22a.html.

[16] Benjamin Guedj. A primer on pac-bayesian learning. *CoRR*, abs/1901.05353, 2019. URL http://arxiv.org/abs/1901.05353.

[17] Pierre Alquier. User-friendly introduction to pac-bayes bounds. *CoRR*, abs/2110.11216, 2021. URL https://arxiv.org/abs/2110.11216.

[18] John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

[19] John Shawe-Taylor and David R. Hardoon. PAC-Bayes Analysis Of Maximum Entropy Classification. In *AISTATS*, 2009.

[20] Alexandre Lacasse, François Laviolette, Mario Marchand, and Francis Turgeon-Boutin. Learning with randomized majority votes. In José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part II*, volume 6322 of *Lecture Notes in Computer Science*, pages 162–177. Springer, 2010. doi: 10.1007/978-3-642-15883-4_11. URL https://doi.org/10.1007/978-3-642-15883-4_11.

[21] Andrés R. Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/386854131f58a556343e056f03626e00-Abstract.html.

[22] Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 769–776. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/hash/779efbd24d5a7e37ce8dc93e7c04d572-Abstract.html.

[23] Jean-Francis Roy, François Laviolette, and Mario Marchand. From pac-bayes bounds to quadratic programs for majority votes. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 649–656. Omnipress, 2011. URL https://icml.cc/2011/papers/379_icmlpaper.pdf.

[24] Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: from a pac-bayesian analysis to a learning algorithm. *J. Mach. Learn. Res.*, 16:787–860, 2015. doi: 10.5555/2789272.2831140. URL https://dl.acm.org/doi/10.5555/2789272.2831140.

[25] François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general ensemble methods with an application to multiclass classification. *Neurocomputing*, 2017.

[26] Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. On pac-bayesian bounds for random forests. *Mach. Learn.*, 108(8-9):1503–1522, 2019. doi: 10.1007/s10994-019-05803-4. URL https://doi.org/10.1007/s10994-019-05803-4.

[27] Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. Self-bounding Majority Vote Learning Algorithms by the Direct Minimization of a Tight PAC-Bayesian C-Bound. In *ECML-PKDD 2021*, pages 167–183, 2021.

[28] Yi-Shan Wu, Andrés R. Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Chebyshev-Cantelli PAC-Bayes-Bennett inequality for the weighted majority vote. *CoRR*, abs/2106.13624, 2021. URL https://arxiv.org/abs/2106.13624.

[29] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Conference on Uncertainty in Artificial Intelligence 33.*, 2017.

[30] Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems 31*, pages 8430–8441. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8063-data-dependent-pac-bayes-priors-via-differential-privacy.pdf.

[31] 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 *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJgqqsAct7.

[32] Gaël Letarte, Pascal Germain, Benjamin Guedj, and Francois Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6872–6882. Curran Associates, Inc., 2019.

[33] Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, and Daniel M. Roy. On the role of data in PAC-Bayes bounds. *CoRR*, abs/2006.10929, 2020. URL https://arxiv.org/abs/2006.10929.

[34] Maria Perez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvari. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021. URL http://jmlr.org/papers/v22/20-879.html.

[35] Felix Biggs and Benjamin Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10):1280, 2021. doi: 10.3390/e23101280. URL https://doi.org/10.3390/e23101280.

[36] Felix Biggs and Benjamin Guedj. Non-vacuous generalisation bounds for shallow neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1963–1981. PMLR, 2022. URL https://proceedings.mlr.press/v162/biggs22a.html.

[37] Andrew Y. K. Foong, Wessel P. Bruinsma, David R. Burt, and Richard E. Turner. How tight can PAC-Bayes be in the small data regime? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4093–4105, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/214cfbe603b7f9f9bc005d5f53f7a1d3-Abstract.html.

[38] Matthias Seeger, John Langford, and Nimrod Megiddo. An improved predictive accuracy bound for averaging classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, number CONF, pages 290–297, 2001.

[39] Andreas Maurer. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004. URL https://arxiv.org/abs/cs.LG/0411099.

[40] Leo Breiman. Prediction Games and Arcing Algorithms. *Neural Computation*, 11(7):1493–1517, 10 1999. ISSN 0899-7667. doi: 10.1162/089976699300016106. URL https://doi.org/10.1162/089976699300016106.

[41] Olivier Marchal and Julyan Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017.

[42] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[43] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, cs.LG/1708.07747, 2017.

[44] Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex pac-bayesian bound. In Steve Hanneke and Lev Reyzin, editors, *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, volume 76 of *Proceedings of Machine Learning Research*, pages 466–492. PMLR, 2017. URL http://proceedings.mlr.press/v76/thiemann17a.html.

[45] Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics lecture notes-monograph series. Institute of Mathematical Statistics, 2007. ISBN 9780940600720. URL https://books.google.fr/books?id=acnaAAAAMAAJ.

[46] 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 - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553419.

[47] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

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, e.g. discussion in Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is of a primarily theoretical nature and therefore no negative societal impact is foreseen. All data is from publicly-available datasets.
   (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] Included in the theorem statements and setting introduction in Section 1.1
   (b) Did you include complete proofs of all theoretical results? [Yes] Included primarily in the main paper with some extra result in the appendices.
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]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Appendix C
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In Appendix C
   (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] In Appendix C
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? [Yes] In Appendix C
   (b) Did you mention the license of the assets? [Yes] In Appendix C
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include links in Appendix C
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All data was publicly available.
   (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] We did not conduct any such research.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not conduct any such research.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not conduct any such research.

A Properties of the Dirichlet distribution

The Dirichlet measure has probability density function w.r.t. Lebesgue measure given by:

$$f(x_1, \ldots, x_d; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{d} x_i^{\alpha_i - 1}$$

where $B(\boldsymbol{\alpha})$ is the multivariate Beta function,

$$B(\boldsymbol{\alpha}) := \frac{\prod_{i=1}^{d} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{d} \alpha_d)}.$$

The mean of a Dirichlet is $\mathbb{E}_{\xi \sim \mathrm{Dir}(\boldsymbol{\alpha})} \boldsymbol{\xi} = \boldsymbol{\alpha} / \sum_{i=1}^{d} \alpha_i$.

The KL divergence between two Dirichlet distributions is the following, given in *e.g.* Zantedeschi et al. [1]:

$$\mathbb{D}_{\mathrm{Dir}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log \frac{B(\boldsymbol{\beta})}{B(\boldsymbol{\alpha})} + \sum_{i=1}^{d} (\alpha_i - \beta_i)(\psi(\alpha_i) - \psi(\alpha_0)) = \log B(\boldsymbol{\beta}) - \mathbb{H}_{\mathrm{Dir}}(\boldsymbol{\alpha}).$$

B Additional details on margin bounds

Here we first note the original result from Biggs and Guedj [15] that is adapted in Equation (3); since this is obtained by applying an upper bound to the inverse small-kl and an additional step, it is strictly looser than the result we give in Equation (3). Biggs and Guedj [15] also uses a dimension doubling trick to allow negative weights (as they consider only the binary case), which we remove here to replace the factor $\log(2d)$ by $\log d$.

Theorem 6. *For any margin $\gamma > 0$, $\delta \in (0, 1)$, sample size $m \in \mathbb{N}$, each of the following results holds with probability at least $1 - \delta$ over the sample $S \sim \mathcal{D}^m$ simultaneously for any $\boldsymbol{\theta} \in \Delta^d$,*

$$L_0(\boldsymbol{\theta}) \leq \hat{L}_\gamma(\boldsymbol{\theta}) + \sqrt{\frac{C}{m} \cdot \hat{L}_\gamma(\boldsymbol{\theta})} + \frac{C + \sqrt{C} + 2}{m}, \tag{5}$$

where $C := 2\log(2/\delta) + \frac{19}{4}\gamma^{-2} \log d \log m$.

B.1 Definition of the margin

We here note that the definition of the margin given in Gao and Zhou [12] and Biggs and Guedj [15] is slightly different from our own, leading to a scaling of the margin definition by a factor of one-half. We show this below.

Both the above papers consider prediction functions like $F_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{d} \theta_i h_i(x)$ with output set $\mathcal{Y} = \{+1, -1\}$. The functions $h_i(x)$ can be positive or negative. The margin is defined as as $yF_{\boldsymbol{theta}}(x)$. We translate this into our equivalent but scaled version as follows:

$$yF_{\boldsymbol{\theta}}(x) = y \left(\sum_{i:h_i(x)=1} \theta_i - \sum_{i:h_i(x)=-1} \theta_i \right) = \sum_{i:h_i(x)=y} \theta_i - \sum_{i:h_i(x)=-y} \theta_i \tag{6}$$

which is double the margin as we define it. Thus $\ell_\gamma(\boldsymbol{\theta}, x, y) = \boldsymbol{I}_{M(\boldsymbol{\theta}, x, y) \leq \gamma} = \boldsymbol{I}_{yF_{\boldsymbol{\theta}} \geq 2\gamma}$ and the condition on the margin $yF(x) \geq \sqrt{8/d}$ given in Gao and Zhou [12] translates to $M(\boldsymbol{\theta}, x, y) > \sqrt{2/d}$ as we give.

B.2 Proof of Theorem 6 and Equation (3)

For completeness we provide here short proofs of Equation (3) and Theorem 6. The central proposition used in Biggs and Guedj [15] to prove their margin bound and these results for voting algorithms is the following, proved implicitly there and here adapted to our setting.

Theorem 7 (Biggs and Guedj [15]). *Let $\boldsymbol{\theta} \in \Delta^d$ and define $\rho = \mathrm{Categ}(\boldsymbol{\theta})$ and $\boldsymbol{i} \sim \rho^T$ as T i.i.d. samples from ρ indexed by $j \in [T]$. Then for any $\gamma > 0, T \in \mathbb{N}_+$ and $(x, y \in \{+1, -1\})$,*

$$\ell_0(\boldsymbol{\theta}, x, y) \leq \mathbb{E}_{\boldsymbol{i}\sim\rho^T} \ell_\gamma^C(\boldsymbol{i}, x, y) + e^{\frac{1}{2}T\gamma^2}$$
$$\mathbb{E}_{\boldsymbol{i}\sim\rho^T} \ell_\gamma^C(\boldsymbol{i}, x, y) \leq \ell_\gamma(\boldsymbol{\theta}, x, y) + e^{\frac{1}{2}T\gamma^2}$$

where we have defined the margin loss for a sum of Categoricals as $\ell_\gamma^{\mathrm{C}}(\boldsymbol{i}, x, y) = \boldsymbol{I}_{yT^{-1}\sum_{t=1}^T h_{i_t}(x) \leq \gamma}$.

Proof of Equation (3). We apply the PAC-Bayes bound Theorem 1 to ℓ_γ^C with ρ^T as defined in Theorem 7 and a uniform prior of the same form, π^T. We then substitute the results from Theorem 7 to show that

$$L_0(\boldsymbol{\theta}) \leq \mathrm{kl}^{-1}\left(\hat{L}_\gamma(\boldsymbol{\theta}) + e^{-\frac{1}{2}T\gamma^2},\ \frac{\mathrm{KL}(\rho^T, \pi^T) + \log\frac{2\sqrt{m}}{\delta}}{m}\right) + e^{-\frac{1}{2}T\gamma^2}.$$

With a uniform prior, $\mathrm{KL}(\rho^T, \pi^T) = T\mathbb{D}_{\mathrm{Cat}}(\theta, d^{-1}\mathbf{1}) \leq T\log d$. Substitution of this upper bound and $T = \lceil 2\gamma^{-2}\log m \rceil$ gives the result. □

Proof of Theorem 6. Beginning with Equation (3), we relax the ceiling using $\gamma \leq \frac{1}{2}$ and $m \geq 2$ for a non-vacuous bound to obtain

$$L_0(\boldsymbol{\theta}) \leq \mathrm{kl}^{-1}\left(\hat{L}_\gamma(\boldsymbol{\theta}) + \frac{1}{m},\ \frac{C}{2m}\right) + \frac{1}{m}$$

with $C := 2\log(2/\delta) + \frac{19}{4}\gamma^{-2}\log d \log m$. Then using the small-kl upper bound $\mathrm{kl}^{-1}(u, c) \leq u + \sqrt{2cu} + 2c$ we have

$$\begin{aligned} L_0(\boldsymbol{\theta}) &\leq \hat{L}_\gamma(\boldsymbol{\theta}) + \frac{2}{m} + \sqrt{\left(\hat{L}_\gamma(\boldsymbol{\theta}) + \frac{1}{m}\right)\frac{C}{m}} + \frac{C}{m} \\ &\leq \hat{L}_\gamma(\boldsymbol{\theta}) + \frac{2}{m} + \sqrt{\frac{C}{m}\cdot\hat{L}_\gamma(\boldsymbol{\theta})} + \frac{C + \sqrt{C}}{m} \end{aligned}$$

which is the result given. □

Proof of Theorem 7. Using the same method as the beginning of the proof of Equation (6),

$$\begin{aligned} &\ell_0(\boldsymbol{\theta}, x, y) - \mathbb{E}_{\boldsymbol{i}\sim\rho^T}\ell_\gamma^C(\boldsymbol{i}, x, y) \\ &\quad = \mathbb{E}_{\boldsymbol{i}\sim\rho^T}[\boldsymbol{I}_{yF(x)\leq 0} - \boldsymbol{I}_{yT^{-1}\sum_{t=1}^T h_{i_t}(x)\leq\gamma}] \\ &\quad \leq \mathbb{P}_{\boldsymbol{i}\sim\rho^T}\left(\frac{1}{2}y(F(x) - T^{-1}\sum_{t=1}^T h_{i_t}(x)) > \frac{1}{2}\gamma\right) \\ &\quad \leq \exp\left(-\frac{1}{2}T\gamma^2\right). \end{aligned}$$

In the last line we used Hoeffding's inequality for a sum of T random variables bounded in $[-1, 1]$. The other side follows using an identical method with the margin losses reversed. □

B.3 Further improvement to the bound

A question which naturally arises from looking at the proof of Equation (3) and Theorem 6 is whether we can do better by choosing T in a more optimal way, rather than just setting it to $\lceil 2\gamma^{-2}\log m \rceil$. We thus prove a bound here which is valid for the optimal choice of T; in practice this is seen to be slightly tighter than Equation (3), although the improvement from Theorem 6 to that result is far greater.

For any $\boldsymbol{\theta} \in \Delta^d$ with probability at least $1-\delta$ over the sample,

$$.L_0(\boldsymbol{\theta}) \leq \inf_{T \in \mathbb{N}_+} \left[\mathrm{kl}^{-1}\left(\hat{L}_\gamma(\boldsymbol{\theta}) + e^{-\frac{1}{2}T\gamma^2}, \frac{T(\log d - \mathbb{H}_{\mathrm{Categ}}[\boldsymbol{\theta}]) + \log \frac{m}{\delta}}{m} \right) + e^{-\frac{1}{2}T\gamma^2} \right] \tag{7}$$

A slightly weaker version of this result, with an extra $m^{-1}\log(2\sqrt{m})$ term, can be proved from

$$L_0(\boldsymbol{\theta}) \leq \mathrm{kl}^{-1}\left(\hat{L}_\gamma(\boldsymbol{\theta}) + e^{-\frac{1}{2}T\gamma^2}, \frac{T\log d + \log \frac{2\sqrt{m}}{\delta}}{m} \right) + e^{-\frac{1}{2}T\gamma^2},$$

which is shown in the proof of Equation (3). We note that the optimal T depends on the data only through $\hat{L}_\gamma(\boldsymbol{\theta}) \in \{0, m^{-1}, 2m^{-1}, \ldots, 1\}$. The last possibility gives a trivial bound. A union bound over the m non-vacuous possibilities gives Equation (7) with the extra logarithmic factor.

In order to remove this term, we use a slightly more sophisticated argument applied to a different PAC-Bayes bound (Theorem 8) given below. This result uses the function (defined for $C > 0, p \in [0,1]$)

$$\Phi_C(p) = -\frac{1}{C}\log(1 - p + pe^{-C})$$

which relates to the small KL (Theorem 9).

Theorem 8 (Catoni [45])**.** *Given data distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, prior $P \in \mathcal{M}^1(\mathcal{H})$, $C > 0$ and $\delta \in (0,1)$, the following hold each with probability $\geq 1-\delta$ over $S \sim D^m$, for all $Q \in \mathcal{M}^1(\mathcal{H})$*

$$\mathbb{E}_{h\sim Q} L(h) \leq \Phi_C^{-1}\left(\mathbb{E}_{h\sim Q}\hat{L}(h) + \frac{\mathrm{KL}(Q,P) + \log\frac{1}{\delta}}{Cm} \right)$$

Theorem 9 (Germain et al. [46], Proposition 2.1)**.** *For any $0 \leq q \leq p < 1$,*

$$\sup_{C>0} \left[C\Phi_C(p) - Cq \right] = \mathrm{kl}(q,p).$$

Proof of Equation (7). We substitute Theorem 7 into Theorem 8 with the categorical loss and a uniform prior, π^T. and KL upper bound as in the above proof. as we obtain for any data-independent $C > 0, T \in \mathbb{N}_+, \gamma > 0$ that

$$L_0(\boldsymbol{\theta}) - e^{-\frac{1}{2}T\gamma^2} \leq \Phi_C^{-1}\left(\frac{k}{m} + e^{-\frac{1}{2}T\gamma^2} + \frac{T\log d + \log\frac{1}{\delta}}{Cm} \right).$$

where $k := m\hat{L}_\gamma(\boldsymbol{\theta})$ is the number of margin errors.

Since the only quantity on the left hand side in this bound unknown before we see data is the value of k, there exists a C_k dependent on the value of k that optimises the bound, and a T_k that depends on this pair. Since there are only m such values giving non-vacuous bounds ($k = m$ is trivially vacuous), we can apply a union bound over all these bounds with $\delta = \delta/m$ to give the following with probability $\geq 1-\delta$:

$$L_0(\boldsymbol{\theta}) \leq \min_{T\in\mathbb{N}_+} \min_{C>0} \left[e^{-\frac{1}{2}T\gamma^2} + \Phi_C^{-1}\left(\frac{k}{m} + e^{-\frac{1}{2}T\gamma^2} + \frac{T\log d + \log\frac{m}{\delta}}{Cm} \right) \right].$$

Applying the inversion of Theorem 9 gives the second result. □

B.4 Comparison of margin bounds

In Figure 4 we compare the various bounds given above in a non-experimental way, fixing the margin loss $\hat{L}_\gamma$ to a particular value and seeing how the bounds change if that value of the loss is achieved for different values of the margin $\gamma \in (0, 0.5)$. Since (uniquely among the bounds), the value of $\boldsymbol{\theta}$ appears in our bound Theorem 2, we show three different sampled possible values, drawn uniformly from the simplex.

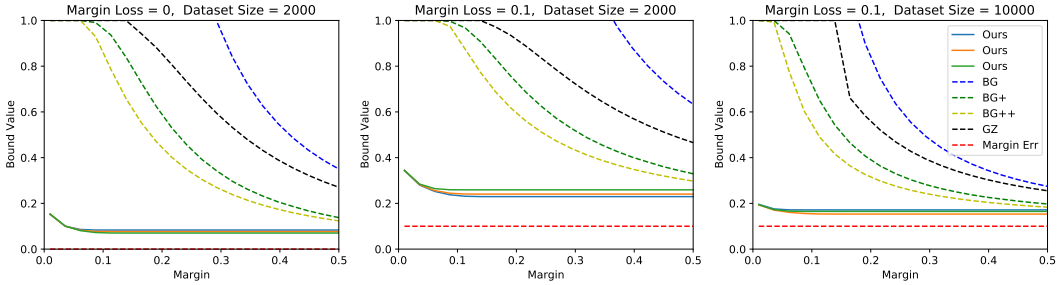

Figure 4: Values of different bounds versus margin at margin error $\hat{L}_\gamma$ (0 or 0.1). Dimension $d = 100$, probability $\delta = 0.5$ and dataset size m (2000 or 10000) are also fixed. The bounds are Theorem 2 (**ours**) with three different samples $\boldsymbol{\theta} \sim \mathrm{Uniform}(\Delta^d)$, compared with the margin bounds of Theorem 6 (**BG**), Equation (3) (**BG+**), Equation (7) (**BG++**), Equation (2) (**GZ**), and the margin error $\hat{L}_\gamma$.

The results for "categorical"-based bounds demonstrate that the refined bounds Equations (3) and (7) are much tighter that the result as given in Theorem 6 by Biggs and Guedj [15]. Both these refinements are also tighter than Equation (2) from Gao and Zhou [12]. We used Equation (3) in the main paper because it is closer to an exiting result (as it appears in the proof from 15), and is not much worse than the refinement Equation (7), particularly when compared to our far stronger new result Theorem 2.

This figure also shows that, at least for some values of $\boldsymbol{\theta}$, this new bound can be far tighter than all the existing bounds. One interesting facet of this is that the bound is improved very little for γ above a certain point, quite a different behaviour to the other bounds. Empirically this was seen too in our other experiments, with the optimised γ often being quite small. Of course, for some values of $\boldsymbol{\theta}$ this bound will be weaker, but we observe the same kind of results in our main experimental results, where this is a learned value.

C Additional experimental details and evaluations

Dataset descriptions.

We provide the description of the classification datasets considered in our empirical evaluation.

***Haberman* (UCI)** prediction of survival of $n = 306$ patients who had undergone surgery from $d = 3$ anonymized features.

***TicTacToe* (UCI)** determination of a win for player x at TicTacToe game of any of the $n = 958$ board configurations ($d = 9$ categorical states).

***Mushrooms* (UCI)** prediction of edibility of $n = 8,124$ mushroom sample, given their $d = 22$ categorical features describing their aspect.

***Adult* (LIBSVM a1a)** determining whether a person earns more than 50K a year ($n = 32,561$ people and $d = 123$ binary features).

***CodRNA* (LIBSVM)** detection of non-coding RNAs among $n = 59,535$ instances and from $d =$ features.

***Pendigits* (UCI)** recognition of hand-written digits (10 classes, $d = 9$ features and $n = 12,992$).

***Protein* (LIBSVM)** $d = 357$ features, $n = 24,387$ instances and 3 classes.

***Sensorless* (LIBSVM)** prediction of motor condition ($n = 58,509$ instances and 11 classes), with intact and defective components, from $d = 48$ features extracted from electric current drive signals.

***MNIST* (LIBSVM)** prediction of hand-written digits ($n = 70,000$ instances and 10 classes) from $d = 28 \times 28$ gray-scale images.

***Fashion-MNIST* (Zalando)** prediction of cloth articles ($n = 70,000$ instances and 10 classes) from $d = 28 \times 28$ gray-scale images.

In all experiments, we convert all categorical features to numerical using an ordinal encoder and we standardize all features using the statistics of the training set.

Baseline descriptions. We report the generalization bounds of the literature used for training weighted majority vote classifiers in our comparison. We additionally note: $\mathrm{Cat}(\boldsymbol{\theta})$ the categorical distribution over the base classifiers (with θ_i the weight associated to voter $h_i \in \mathcal{H}$), and $\mathbb{D}_{\mathrm{Cat}}(\boldsymbol{\theta}), \boldsymbol{\pi})$ the KL-divergence between two categorical distribution with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$; $\ell_{\mathrm{TND}}(h, h', x, y) := \boldsymbol{I}_{h(x)\neq y \wedge h'(x)\neq y}$ the tandem loss proposed in Masegosa et al. [21] and $\hat{L}_{\mathrm{TND}}(h, h') := \mathbb{E}_{(x,y)\sim \mathrm{Uniform}(S)} \ell_{\mathrm{TND}}(h, h', x, y)$ its in-sample estimate; $\ell_{\mathrm{Bin}}(\boldsymbol{\theta}, N, x, y) := \sum_{k=\frac{N}{2}}^{N} \binom{N}{k} M(\boldsymbol{\theta}, x, y)^k (1 - M(\boldsymbol{\theta}, x, y))^{(N-k)}$ the probability that among N voters randomly drawn from $\mathrm{Cat}(\boldsymbol{\theta})$ at least $\frac{N}{2}$ of them are incorrect, as defined in Lacasse et al. [20].

- First Order [FO, 18]:
  For any $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}_+$, $\delta \in (0, 1)$, and prior $\boldsymbol{\pi} \in \Delta^d$, with probability at least $1 - \delta$ over the sample $S \sim \mathcal{D}^m$ simultaneously for every $\boldsymbol{\theta} \in \Delta^d$,

$$L_0(\boldsymbol{\theta}) \leq 2\,\mathrm{kl}^{-1}\left(\mathbb{E}_{h\sim \mathrm{Cat}(\boldsymbol{\theta})} \hat{L}_0(h),\ \frac{\mathbb{D}_{\mathrm{Cat}}(\boldsymbol{\theta}, \boldsymbol{\pi}) + \log \frac{2\sqrt{m}}{\delta}}{m}\right).$$

- Second Order [SO, 21]:
  For any $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}_+$, $\delta \in (0, 1)$, and prior $\boldsymbol{\pi} \in \Delta^d$, with probability at least $1 - \delta$ over the sample $S \sim \mathcal{D}^m$ simultaneously for every $\boldsymbol{\theta} \in \Delta^d$,

$$L_0(\boldsymbol{\theta}) \leq 4\,\mathrm{kl}^{-1}\left(\mathbb{E}_{h\sim \mathrm{Cat}(\boldsymbol{\theta}), h'\sim \mathrm{Cat}(\boldsymbol{\theta})} \hat{L}_{\mathrm{TND}}(h, h'),\ \frac{2\mathbb{D}_{\mathrm{Cat}}(\boldsymbol{\theta}, \boldsymbol{\pi}) + \log \frac{2\sqrt{m}}{\delta}}{m}\right).$$

- Binomial [Bin, 20]:
  For any $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}_+$, $N \in \mathbb{N}_+$, $\delta \in (0, 1)$, and prior $\boldsymbol{\pi} \in \Delta^d$, with probability at least $1 - \delta$ over the sample $S \sim \mathcal{D}^m$ simultaneously for every $\boldsymbol{\theta} \in \Delta^d$,

$$L_0(\boldsymbol{\theta}) \leq 2\,\mathrm{kl}^{-1}\left(\mathbb{E}_{(x,y)\sim \mathrm{Uniform}(S)} \ell_{\mathrm{Bin}}(\boldsymbol{\theta}, N, x, y),\ \frac{N\mathbb{D}_{\mathrm{Cat}}(\boldsymbol{\theta}, \boldsymbol{\pi}) + \log \frac{2\sqrt{m}}{\delta}}{m}\right).$$

- Chebyshev-Cantelli tandem loss bound [CCTND, 28, Theorem 12];
- Chebyshev-Cantelli tandem loss bound with an offset [CCPBB, 28, Theorem 15];
- Dirichlet Factor-Two [f2, 1]:
  For any $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}_+$, $\delta \in (0, 1)$, and prior $\boldsymbol{\beta} \in \mathbb{R}_+^d$, with probability at least $1 - \delta$ over the sample $S \sim \mathcal{D}^m$ simultaneously for every $\boldsymbol{\theta} \in \Delta^d$ and $K > 0$,

$$L_0(\boldsymbol{\theta}) \leq 2\,\mathrm{kl}^{-1}\left(\mathbb{E}_{\boldsymbol{\xi}\sim \mathrm{Dir}(K\boldsymbol{\theta})} \hat{L}(\boldsymbol{\xi}),\ \frac{\mathbb{D}_{\mathrm{Dir}}(K\boldsymbol{\theta}, \boldsymbol{\beta}) + \log \frac{2\sqrt{m}}{\delta}}{m}\right).$$

Optimisation of PAC-Bayesian bounds. To optimize the baselines *CCPBB* and *CCTND*, we rely on the code released by its authors [2], with the Gradient Descent option and building random forests as described in our main text. When optimising the PAC-Bayesian bounds *FO*, *SO*, *Bin*, *f2* and ours, we initialize $\boldsymbol{\theta}$'s to be uniform, i.e. $\theta_i = 1/d$, and $K = 2$. We then optimise the posterior parameters of the method ($\alpha = K\theta$ for Dirichlet, and θ for Categorical distributions) with the Adam optimiser [47] with running average coefficients $(0.9, 0.999)$, batch size equal to 100 and learning rate set to 0.1. All methods are run for a maximum of 100 epochs with patience of 25 epochs for early stopping and a learning rate scheduler reducing it by a factor of 10 with 2 epochs patience.

check Vale?

At each run of an algorithm, we randomly split a dataset into training and test sets of sizes $80\% - 20\%$ respectively, and optimise/evaluate the bounds only with the half of the training set that was not

[2] https://github.com/StephanLorenzen/MajorityVoteBounds/tree/44cec987865ddce01cd27076019394538cee85ca/NeurIPS2021

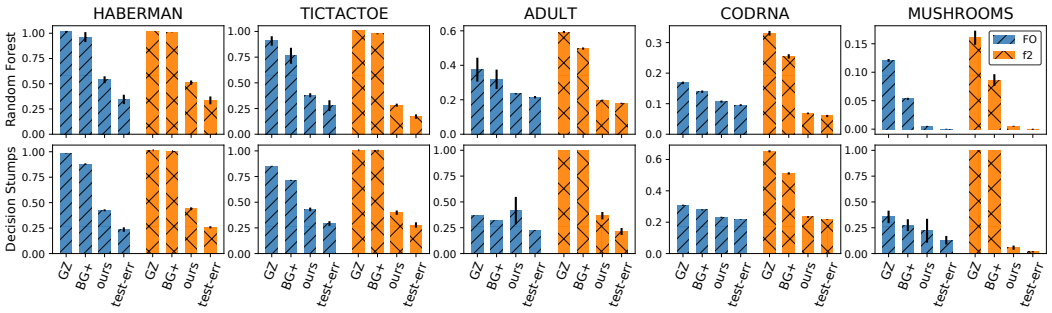

Figure 5: Theorem 2 (**ours**) compared with the margin bounds of Equation (3) (**BG+**), Equation (2) (**GZ**), and the test error. Settings are *rf* (first row) and *stumps* (second row) on the given datasets, with $\boldsymbol{\theta}$ output by optimising either *FO* or *f2* (first and second column groupings respectively).

used for learning the voters (in the case of data-dependent ones). Note that we do not make use of a validation set, as we use the risk certificates as estimate of the test error for model selection. Finally, we report the value of Seeger's "small-kl" bound of Theorem 1, even when a different type of bound has been optimised (*e.g.* for the *CCPBB* and *CCTND* baselines), and we average all results over 5 different trials.

Margin bound comparison. Given a pre-trained model, hence fixed θ and initial K_{init} (which is different from 1. only for the models trained via Dirichlet bounds), we search for its optimal risk certificate by evaluating a given bound at $1,000$ values of γ, spaced evenly on a log scale with base 10 and in the interval $[10^{-4}, 0.5)$. For our margin bound, for each of these γ values we also optimise $K \in [K_{init}, K_{init}\ 2^{16}]$ using the golden-section search technique to obtain the tightest upper bound. Notice that this does not add significant computational overhead to the search. Also for these experiments, the bounds are evaluated with the portion of training data that was not used for learning the voters.

Compute. All experiments were run on a virtual machine with 16 vCPUs and $128Gb$ of RAM.

C.1 Additional results

In Figure 5, Figure 6 and Figure 7 we report the results from Figure 1, Figure 2 and Figure 3 in the main text. Here we deploy a different scale per dataset so that they can be easily read, also when the bounds and test errors are very small. Additionally, in Figure 8 we provide the test errors and risk certificates obtained by optimising the generalization bounds with decision stumps as voters. Although our certificates are always the tightest, we found that in some cases our method converges to sub-optimal solutions. We speculate that this arises due to the highly non-convex nature of the objective combined with a strong K-inflating gradient signal from the $O(e^{-K\gamma^2})$ term. Thus future work to improve these results even further could start with the use of the quasi-convex small-kl relaxation from Thiemann et al. [44]. We note however that this is overall less important than our main results, as both our bounds are still extremely tight when used in an algorithm-free way and applied to the output of another algorithm.

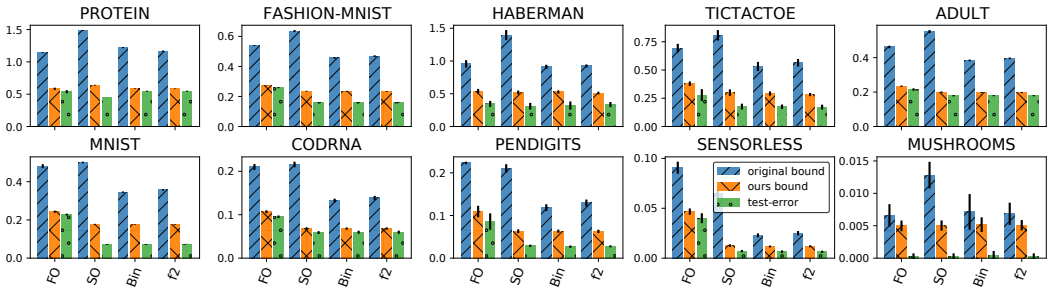

Figure 6: Theorem 2 (*our bound*) compared with the bounds of *FO*, *SO*, *Bin* or *f2* (*original bound*), and test errors. For each column grouping, $\boldsymbol{\theta}$ is the output from optimising the corresponding PAC-Bayes bound for *rf* on the given dataset.

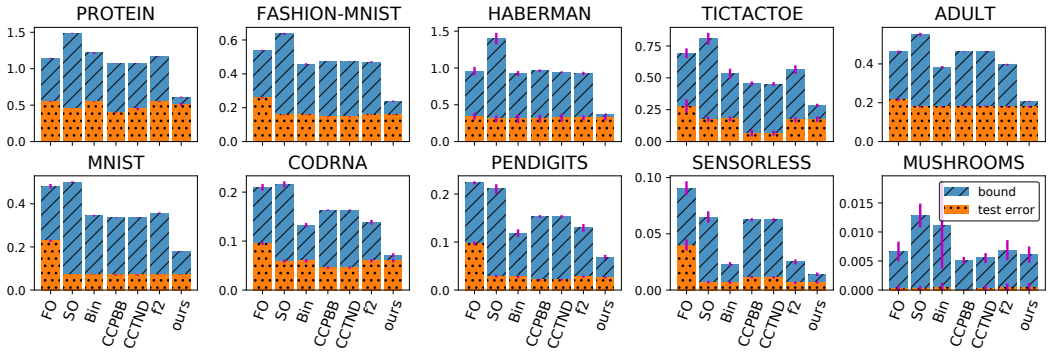

Figure 7: Theorem 3 (**ours**) as optimisation objective compared to other PAC-Bayes results (*FO*, *SO*, *Bin*, *CCPBB*, *CCTND* and *f2*) as objectives, with a Random Forest as set of voters.

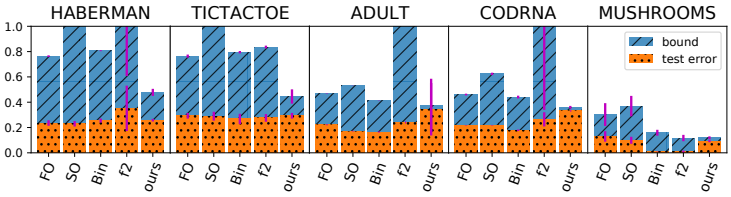

Figure 8: Theorem 3 (**ours**) as optimisation objective compared to other PAC-Bayes results (*FO*, *SO*, *Bin*, *f2*) as objectives, with decision stumps as voters.