Chebyshev-Cantelli PAC-Bayes-Bennett Inequality for the Weighted Majority Vote

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a new second-order oracle bound for the expected risk of a weighted majority vote. The bound is based on a novel parametric form of the Chebyshev-Cantelli inequality (a.k.a. one-sided Chebyshev's), which is amenable to efficient minimization. The new form resolves the optimization challenge faced by prior oracle bounds based on the Chebyshev-Cantelli inequality, the C-bounds [Germain et al., 2015], and, at the same time, it improves on the oracle bound based on second order Markov's inequality introduced by Masegosa et al. [2020]. We also derive the PAC-Bayes-Bennett inequality, which we use for empirical estimation of the oracle bound. The PAC-Bayes-Bennett inequality improves on the PAC-Bayes-Bernstein inequality by Seldin et al. [2012]. We provide an empirical evaluation demonstrating that the new bounds can improve on the work by Masegosa et al. [2020]. Both the parametric form of the Chebyshev-Cantelli inequality and the PAC-Bayes-Bennett inequality may be of independent interest for the study of concentration of measure in other domains.

1 Introduction

Weighted majority vote is a central technique for combining predictions of multiple classifiers. It is an integral part of random forests [Breiman, 1996, 2001], boosting [Freund and Schapire, 1996], gradient boosting [Friedman, 1999, 2001, Mason et al., 1999, Chen and Guestrin, 2016], and it is also used to combine predictions of heterogeneous classifiers. It is part of the winning strategies in many machine learning competitions. The power of the majority vote is in the cancellation of errors effect: when the errors of individual classifiers are independent or anticorrelated and the error probability of individual classifiers is smaller than 0.5, then the errors average out and the majority vote tends to outperform the individual classifiers.

Generalization bounds for weighted majority vote and theoretically-grounded approaches to weight tuning are decades-old research topics. Berend and Kontorovich [2016] derived an optimal solution under the assumption of known error rates and independence of errors of individual classifiers, but neither of the two assumptions is typically satisfied in practice.

In absence of the independence assumption, the most basic result is the first order oracle bound, which is based on Markov's inequality and bounds the expected loss of ρ-weighted majority vote by twice the ρ-weighted average of expected losses of the individual classifiers. This finding is so old and basic that Langford and Shawe-Taylor [2002] call it "the folk theorem". The ρ-weighted average of the expected losses is then bounded using PAC-Bayesian bounds, turning the oracle bound into an empirical bound [McAllester, 1998, Seeger, 2002, Langford and Shawe-Taylor, 2002]. While the translation from oracle to empirical bound is quite tight [Germain et al., 2009, Thiemann et al., 2017], the first order oracle bound ignores the correlation of errors, which is the main power of the majority vote. As a result, its minimization overconcentrates the weights on the best-performing classifiers,

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

effectively reducing the majority vote to a very few or even just a single best classifier, which leads to a significant deterioration of the test error on hold-out data [Lorenzen et al., 2019, Masegosa et al., 2020].

In order to take correlation of errors into account, Lacasse et al. [2007] derived second order oracle bounds, the C-bounds, which are based on the Chebyshev-Cantelli inequality. The ideas were further developed by Laviolette et al. [2011], Germain et al. [2015], and Laviolette et al. [2017]. However, they were only able to optimize the bounds in highly restrictive settings and, additionally, second order oracle quantities in the denominator of the oracle bounds led to looseness in translation to empirical bounds.

Masegosa et al. [2020] proposed an alternative second-order oracle bound, the tandem bound, based on second order Markov's inequality. While the second order Markov's inequality is weaker than the Chebyshev-Cantelli inequality, the resulting bound has no oracle quantities in the denominator, which allows tight translation to an empirical bound. Additionally, Masegosa et al. proposed an efficient procedure for minimization of their empirical bound. They have shown that minimization of the empirical bound does not lead to deterioration of the test error.

Our work can be seen as a bridge between the tandem bound and the C-bounds, and as an improvement of both. The key novelty is a new parametric form of Chebyshev-Cantelli inequality, which preserves the tightness of Chebyshev-Cantelli, but avoids oracle quantities in the denominator. This allows both efficient translation to empirical bounds and efficient minimization. We derive two new second order oracle bounds based on the new inequality, one using the tandem loss and the other using the tandem loss with an offset. For empirical estimation of the latter we derive a PAC-Bayes-Bennett inequality. The overall contributions can be summarized as follows:

1. We propose a new parametric form of the Chebyshev-Cantelli inequality, which has no variance in the denominator and preserves tightness of the original bound. The new form allows efficient minimization and empirical estimation.
2. We propose two new second order oracle bounds for the weighted majority vote based on the new form of the Chebyshev-Cantelli inequality. The bounds have two advantages: (1) they are amenable to tight translation to empirical bounds; and (2) the resulting empirical bounds are amenable to efficient minimization.
3. We derive a PAC-Bayes-Bennett inequality, which improves on the PAC-Bayes-Bernstein inequality by Seldin et al. [2012]. The inequality is used for bounding the tandem loss with an offset.

2 Problem setup

The problem setup and notations are borrowed from Masegosa et al. [2020].

Multiclass classification. We let $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an independent identically distributed sample from $\mathcal{X} \times \mathcal{Y}$, drawn according to an unknown distribution D, where $\mathcal{Y}$ is finite and $\mathcal{X}$ is arbitrary. A hypothesis is a function $h : \mathcal{X} \to \mathcal{Y}$, and $\mathcal{H}$ denotes a space of hypotheses. We evaluate the quality of h using the zero-one loss $\ell(h(X), Y) = \mathbb{1}(h(X) \neq Y)$, where $\mathbb{1}(\cdot)$ is the indicator function. The expected loss of h is denoted by $L(h) = \mathbb{E}_{(X,Y)\sim D}[\ell(h(X), Y)]$ and the empirical loss of h on a sample S of size n is denoted by $\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$.

Randomized classifiers. A *randomized classifier* (a.k.a. Gibbs classifier) associated with a distribution ρ on $\mathcal{H}$, for each input X randomly draws a hypothesis $h \in \mathcal{H}$ according to ρ and predicts $h(X)$. The expected loss of a randomized classifier is given by $\mathbb{E}_{h\sim\rho}[L(h)]$ and the empirical loss by $\mathbb{E}_{h\sim\rho}[\hat{L}(h, S)]$. To simplify the notation we use $\mathbb{E}_D[\cdot]$ as a shorthand for $\mathbb{E}_{(X,Y)\sim D}[\cdot]$ and $\mathbb{E}_\rho[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho}[\cdot]$.

Ensemble classifiers and majority vote. Ensemble classifiers predict by taking a weighted aggregation of predictions by hypotheses from $\mathcal{H}$. The ρ-weighted majority vote MV_ρ predicts $\mathrm{MV}_\rho(X) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}_\rho[\mathbb{1}(h(X) = y)]$, where ties can be resolved arbitrarily.

3 A review of prior first and second order oracle bounds

If majority voting makes an error, we know that at least a ρ-weighted half of the classifiers have made an error and, therefore, $\ell(\mathrm{MV}_\rho(X), Y) \leq \mathbb{1}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$. This observation leads to the well-known first order oracle bound for the loss of weighted majority vote.

Theorem 1 (First Order Oracle Bound)**.**

$$L(\mathrm{MV}_\rho) \leq 2\mathbb{E}_\rho[L(h)].$$

Proof. We have $L(\mathrm{MV}_\rho) = \mathbb{E}_D[\ell(\mathrm{MV}_\rho(X), Y)] \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$. By applying Markov's inequality to random variable $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ we have:

$$L(\mathrm{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5) \leq 2\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] = 2\mathbb{E}_\rho[L(h)]. \qquad \square$$

PAC-Bayesian analysis can be used to bound $\mathbb{E}_\rho[L(h)]$ in Theorem 1 in terms of $\mathbb{E}_\rho[\hat{L}(h, S)]$, thus turning the oracle bound into an empirical one. The disadvantage of the first order approach is that $\mathbb{E}_\rho[L(h)]$ ignores correlations of predictions, which is the main power of the majority vote.

Masegosa et al. [2020] have used second order Markov's inequality, by which for a non-negative random variable Z and $\varepsilon > 0$

$$\mathbb{P}(Z \geq \varepsilon) = \mathbb{P}\big(Z^2 \geq \varepsilon^2\big) \leq \frac{\mathbb{E}\left[Z^2\right]}{\varepsilon^2}.$$

For pairs of hypotheses h and h' they have defined *tandem loss* $\ell(h(X), h'(X), Y) = \mathbb{1}(h(X) \neq Y \wedge h'(X) \neq Y) = \mathbb{1}(h(X) \neq Y)\mathbb{1}(h'(X) \neq Y)$, also termed joint error by Lacasse et al. [2007], which counts an error only if both h and h' err on a sample (X, Y). The corresponding *expected tandem loss* is defined by

$$L(h, h') = \mathbb{E}_D[\mathbb{1}(h(X) \neq Y)\mathbb{1}(h'(X) \neq Y)].$$

Lacasse et al. [2007] and Masegosa et al. [2020] have shown that expectation of the second moment of the weighted loss equals expectation of the tandem loss. Using ρ^2 as a shorthand for the product distribution $\rho \times \rho$ over $\mathcal{H} \times \mathcal{H}$ and $\mathbb{E}_{\rho^2}[L(h, h')]$ as a shorthand for $\mathbb{E}_{h\sim\rho, h'\sim\rho}[L(h, h')]$, the result is the following.

Lemma 2 (Masegosa et al., 2020)**.** *In multiclass classification*

$$\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] = \mathbb{E}_{\rho^2}[L(h, h')].$$

By combining second order Markov's inequality with Lemma 2, Masegosa et al. have shown the following result.

Theorem 3 (Masegosa et al., 2020)**.** *In multiclass classification*

$$L(\mathrm{MV}_\rho) \leq 4\mathbb{E}_{\rho^2}[L(h, h')].$$

Lacasse et al. [2007] have used the Chebyshev-Cantelli inequality to derive a different form of a second order oracle bound. We use $\mathbb{V}[Z]$ to denote the variance of Z in the statement of Chebyshev-Cantelli inequality.

Theorem 4 (Chebyshev-Cantelli inequality)**.** *For* $\varepsilon > 0$

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq \varepsilon) \leq \frac{\mathbb{V}[Z]}{\varepsilon^2 + \mathbb{V}[Z]}.$$

Theorem 4 together with Lemma 2 leads to the following result, known as the *C-bound*.

Theorem 5 (Lacasse et al., 2007, Masegosa et al., 2020)**.** *If* $\mathbb{E}_\rho[L(h)] \leq \frac{1}{2}$*, then*

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L(h, h')] - \mathbb{E}_\rho[L(h)]^2}{\frac{1}{4} + \mathbb{E}_{\rho^2}[L(h, h')] - \mathbb{E}_\rho[L(h)]}.$$

Masegosa et al. have shown that the Chebyshev-Cantelli inequality is always at least as tight as second order Markov's inequality (below we provide an alternative and more intuitive proof of this fact) and, therefore, the oracle bound in Theorem 5 is always at least as tight as the oracle bound in Theorem 3. However, the presence of $\mathbb{E}_{\rho^2}[L(h, h')]$ and $\mathbb{E}_\rho[L(h)]$ in the denominator make empirical estimation and optimization of the bound in Theorem 5 impractical, and Theorem 3 was the only practically applicable second order bound so far.

4 Main Contributions

We present three main contributions: (1) a new form of the Chebyshev-Cantelli inequality, which is convenient for optimization; (2) an application of the inequality to the analysis of weighted majority vote; and (3) a PAC-Bayes-Bennett inequality, which is used to bound the risk with an offset in the bound for weighted majority vote. We start with the new form of Chebyshev-Cantelli inequality, which can be seen as a refinement of second order Markov's inequality or as an intermediate step in the proof of the Chebyshev-Cantelli inequality.

Theorem 6. *For any $\varepsilon > 0$ and $\mu < \varepsilon$*

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2}.$$

Proof.

$$\mathbb{P}(Z \geq \varepsilon) = \mathbb{P}(Z-\mu \geq \varepsilon-\mu) \leq \mathbb{P}\big((Z-\mu)^2 \geq (\varepsilon-\mu)^2\big) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2}. \qquad \square$$

The inequality can also be written as

$$\mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}\left[Z\right] + \mu^2}{(\varepsilon-\mu)^2}. \tag{1}$$

The bound is minimized by $\mu^* = \mathbb{E}\left[Z\right] - \frac{\mathbb{V}[Z]}{\varepsilon - \mathbb{E}[Z]}$, which can be verified by taking a derivative of the bound with respect to μ. Note that μ^* can take negative values. Substitution of μ^* into the bound and simple algebraic manipulations recover the Chebyshev-Cantelli inequality, whereas $\mu = 0$ recovers second order Markov's inequality. The main advantage of Theorem 6 over the Chebyshev-Cantelli inequality is ease of estimation and optimization due to absence of the variance term in the denominator.

Equation (1) leads to two new second order oracle bounds for the weighted majority vote, given in Theorems 7 and 8.

Theorem 7. *In multiclass classification*

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L(h,h')] - 2\mu\mathbb{E}_\rho[L(h)] + \mu^2}{(0.5-\mu)^2}.$$

Proof. As in the previous section, we take $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$, so that $L(\mathrm{MV}_\rho) \leq \mathbb{P}(Z \geq 0.5)$. The result follows by (1) and the calculations of $\mathbb{E}\left[Z^2\right]$ and $\mathbb{E}\left[Z\right]$ from the previous section. $\square$

For $\mu = 0$, Theorem 7 recovers Theorem 3, but if $\mu^* = \mathbb{E}_\rho[L(h)] - \frac{\mathbb{E}_{\rho^2}[L(h,h')] - \mathbb{E}[L(h)]^2}{0.5 - \mathbb{E}_\rho[L(h)]} \neq 0$, then substitution of μ^* into the theorem yields a tighter oracle bound. At the same time, substitution of μ^* recovers Theorem 5, but the great advantage of Theorem 7 is that the bound allows easy empirical estimation and optimization with respect to ρ, due to absence of $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ in the denominator. Thus, Theorem 7 has the oracle tightness of Theorem 5 and the ease of estimation and optimization of Theorem 3. The oracle quantities $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ can be bounded using PAC-Bayes-kl or PAC-Bayes-λ inequalities, as discussed in the next section.

In order to present the second oracle bound we introduce a new quantity. For a pair of hypotheses h and h' and a constant μ, we define *tandem loss with μ-offset*, for brevity *μ-tandem loss*, as

$$\ell_\mu(h(X), h'(X), Y) = (\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}(h'(X) \neq Y) - \mu). \tag{2}$$

Note that it can take negative values. We denote its expectation by

$$L_\mu(h,h') = \mathbb{E}_D[\ell_\mu(h(X), h'(X), Y)] = \mathbb{E}_D[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}(h'(X) \neq Y) - \mu)].$$

With $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ as before, we have

$$\begin{aligned}\mathbb{E}\left[(Z-\mu)^2\right] &= \mathbb{E}_D[(\mathbb{E}_\rho[(\mathbb{1}(h(X) \neq Y) - \mu)])^2] \\ &= \mathbb{E}_{\rho^2}[\mathbb{E}_D[(\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}(h'(X) \neq Y) - \mu)]] = \mathbb{E}_{\rho^2}[L_\mu(h,h')].\end{aligned}$$

Now we present our second oracle bound.

Theorem 8. *In multiclass classification*

$$L(\mathrm{MV}_\rho) \leq \frac{\mathbb{E}_{\rho^2}[L_\mu(h,h')]}{(0.5-\mu)^2}.$$

Proof. The result follows by Theorem 6 and the calculation above. □

In order to discuss the advantage of Theorem 8, we define the variance of the μ-tandem loss

$$\mathbb{V}_\mu(h,h') = \mathbb{E}_D[((\mathbb{1}(h(X)\neq Y)-\mu)(\mathbb{1}(h'(X)\neq Y)-\mu) - L_\mu(h,h'))^2].$$

If the variance of the μ-tandem loss is small, we can use Bernstein-type inequalities to obtain tighter estimates compared to kl-type inequalities.

We bound the μ-tandem loss using our next contribution, the PAC-Bayes-Bennett inequality, which improves on the PAC-Bayes-Bernstein inequality derived by Seldin et al. [2012] and may be of independent interest. The inequality holds for any loss function bounded from above, we use $\tilde{\ell}$ and matching tilde-marked quantities to distinguish it from the zero-one loss ℓ. We let $\tilde{L}(h) = \mathbb{E}_D[\tilde{\ell}(h(X),Y)]$ and $\tilde{\mathbb{V}}(h) = \mathbb{E}_D[(\tilde{\ell}(h(X),Y)-\tilde{L}(h))^2]$ be the expected tilde-loss of h and its variance and let $\hat{\tilde{L}}(h,S) = \frac{1}{n}\sum_{i=1}^n \tilde{\ell}(h(X_i),Y_i)$ be the empirical tilde-loss of h on a sample S.

Theorem 9 (PAC-Bayes-Bennett inequality)**.** *Let $\tilde{\ell}(\cdot,\cdot)$ be an arbitrary loss function bounded from above by b and assume that $\tilde{\mathbb{V}}(h)$ is finite for all h. Let $\phi(x) = e^x - x - 1$. Then for any distribution π on $\mathcal{H}$ that is independent of S and any $\gamma > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over a random draw of S, for all distributions ρ on $\mathcal{H}$ simultaneously:*

$$\mathbb{E}_\rho[\tilde{L}(h)] \leq \mathbb{E}_\rho[\hat{\tilde{L}}(h,S)] + \frac{\phi(\gamma b)}{\gamma b^2}\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\gamma n}.$$

The proof is based on a change of measure argument combined with Bennett's inequality, the details are provided in Appendix A. Note that the result holds for a fixed (but arbitrary) γ. In case of optimization with respect to γ a union bound has to be applied. For a fixed ρ the bound is convex in γ and for a fixed γ it is convex in ρ, although it is not necessarily jointly convex in ρ and γ. See Appendix D for optimization details. The PAC-Bayes-Bennett inequality is identical to the PAC-Bayes-Bernstein inequality by Seldin et al. [2012, Theorem 7], except that in the latter the coefficient in front of $\mathbb{E}_\rho[\tilde{\mathbb{V}}[h]]$ is $(e-2)\gamma$ instead of $\frac{\phi(\gamma b)}{\gamma b^2}$. The result improves on the result by Seldin et al. in three ways. First, it requires only one-sided instead of two-sided boundedness of the loss. Second, in the result by Seldin et al. γ is restricted to the $(0,1/b]$ interval, whereas in our result γ is unrestricted from above. And finally, we can rewrite the coefficient in front of the variance as $\frac{\phi(\gamma b)}{\gamma b^2} = \frac{\phi(\gamma b)}{\gamma^2 b^2}\gamma$, and $\frac{\phi(\gamma b)}{\gamma^2 b^2}$ is a monotonically increasing function of γ, which in the interval $\gamma \in (0,1/b]$ satisfies $\lim_{\gamma\to 0}\frac{\phi(\gamma b)}{\gamma^2 b^2} = \frac{1}{2}$ and for $\gamma = 1/b$ gives $\frac{\phi(\gamma b)}{\gamma^2 b^2} = (e-2)$. Thus, PAC-Bayes-Bennett is always at least as tight as PAC-Bayes-Bernstein and, at the same time, for $\gamma < 1/b$ it improves the constant coefficient in front of the variance from $(e-2) \approx 0.72$ down to 0.5 for $\gamma \to 0$.

5 From oracle to empirical bounds

We obtain empirical bounds on the oracle quantities $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ in Theorem 7 and $\mathbb{E}_{\rho^2}[L_\mu(h,h')]$ in Theorem 8 by using PAC-Bayesian inequalities. The empirical counterpart of the expected tandem loss is the empirical tandem loss

$$\hat{L}(h,h',S) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(h(X_i)\neq Y_i)\mathbb{1}(h'(X_i)\neq Y_i).$$

For bounding $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ we use either PAC-Bayes-kl or PAC-Bayes-λ inequalities, both cited below. We use $\mathrm{KL}(\rho\|\pi)$ to denote the Kullback-Leibler divergence between distributions ρ and π on $\mathcal{H}$ and $\mathrm{kl}(p\|q)$ to denote the Kullback-Leibler divergence between two Bernoulli distributions with biases p and q.

Theorem 10 (PAC-Bayes-kl Inequality, Seeger, 2002)**.** *For any probability distribution π on $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability at least $1-\delta$ over a random draw of a sample S, for all distributions ρ on $\mathcal{H}$ simultaneously:*

$$\mathrm{kl}\left(\mathbb{E}_\rho[\hat{L}(h,S)]\middle\|\mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\mathrm{KL}(\rho\|\pi)+\ln(2\sqrt{n}/\delta)}{n}. \tag{3}$$

Theorem 11 (PAC-Bayes-λ Inequality, Thiemann et al., 2017)**.** *For any probability distribution π on $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability at least $1-\delta$ over a random draw of a sample S, for all distributions ρ on $\mathcal{H}$ and all $\lambda \in (0,2)$ and $\gamma > 0$ simultaneously:*

$$\mathbb{E}_\rho\left[L(h)\right] \leq \frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi)+\ln(2\sqrt{n}/\delta)}{\lambda\left(1-\frac{\lambda}{2}\right)n}, \tag{4}$$

$$\mathbb{E}_\rho\left[L(h)\right] \geq \left(1-\frac{\gamma}{2}\right)\mathbb{E}_\rho[\hat{L}(h,S)] - \frac{\mathrm{KL}(\rho\|\pi)+\ln(2\sqrt{n}/\delta)}{\gamma n}. \tag{5}$$

The PAC-Bayes-λ inequality is an optimization-friendly relaxation of the PAC-Bayes-kl inequality. Therefore, for optimization of ρ we use the PAC-Bayes-λ inequality, the upper bound for $\mathbb{E}_{\rho^2}[L(h,h')]$ and the lower bound for $\mathbb{E}_\rho[L(h)]$, but once we have converged to a solution we use PAC-Bayes-kl to compute the final bound. The kl form provides both an upper and a lower bound through the upper and lower inverse of the kl. We note that $\mathrm{KL}(\rho^2\|\pi^2) = 2\,\mathrm{KL}(\rho\|\pi)$ [Germain et al., 2015, Page 814]. Taking the oracle bound from Theorem 7 and bounding the oracle quantities using Theorem 11 we obtain the following result.

Theorem 12. *For any parameter grid $\{\mu_1,\ldots,\mu_k\}$, any distribution π on $\mathcal{H}$ that is independent of S, and any $\delta \in (0,1)$, with probability at least $1-\delta$ over a random draw of S, for all distributions ρ on $\mathcal{H}$, all $\lambda \in (0,2)$, all $\gamma > 0$, and all values of μ in the grid simultaneously:*

$$L(\mathrm{MV}_\rho) \leq \frac{1}{(0.5-\mu)^2}\Bigg(\frac{\mathbb{E}_{\rho^2}[\hat{L}(h,h',S)]}{1-\frac{\lambda}{2}} + \frac{2\,\mathrm{KL}(\rho\|\pi)+\ln(4k\sqrt{n}/\delta)}{\lambda\left(1-\frac{\lambda}{2}\right)n} - 2\mu\left(\left(1-\frac{\gamma}{2}\right)\mathbb{E}_\rho[\hat{L}(h,S)] - \frac{\mathrm{KL}(\rho\|\pi)+\ln(4k\sqrt{n}/\delta)}{\gamma n}\right) + \mu^2\Bigg).$$

Proof. The result follows by substitution of the bounds on $\mathbb{E}_{\rho^2}[L(h,h')]$ and $\mathbb{E}_\rho[L(h)]$ by Theorem 11 into Theorem 7. The $4k$ factor comes from a union bound over the grid and the two empirical bounds. □

We note that both the loss and the tandem loss are Bernoulli random variables and that Bernoulli random variables have the maximal variance given the range and expectation. Therefore, PAC-Bayes-kl is tighter than PAC-Bayes-Bennett in this case. However, the empirical counterpart of the expected μ-tandem loss is the empirical μ-tandem loss

$$\hat{L}_\mu(h,h',S) = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}(h(X_i)\neq Y_i)-\mu)(\mathbb{1}(h'(X_i)\neq Y_i)-\mu),$$

and the μ-tandem losses are not Bernoulli. Therefore, we use the PAC-Bayes-Bennett inequality, which provides an advantage if the variance of the μ-tandem losses happens to be small. The expected and empirical variance of the μ-tandem losses of a pair of hypotheses h and h' are, respectively, defined by

$$\mathbb{V}_\mu(h,h') = \mathbb{E}_D[((\mathbb{1}(h(X)\neq Y)-\mu)(\mathbb{1}(h'(X)\neq Y)-\mu) - L_\mu(h,h'))^2],$$

$$\hat{\mathbb{V}}_\mu(h,h',S) = \frac{1}{n-1}\sum_{i=1}^{n}\left((\mathbb{1}(h(X_i)\neq Y_i)-\mu)(\mathbb{1}(h'(X_i)\neq Y_i)-\mu) - \hat{L}_\mu(h,h',S)\right)^2.$$

The empirical variance $\hat{\mathbb{V}}_\mu(h,h',S)$ is an unbiased estimate of $\mathbb{V}_\mu(h,h')$.

Since the PAC-Bayes-Bennett inequality is stated in terms of the oracle variance $\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)]$, we use the result by Tolstikhin and Seldin [2013, Equation (15)] to bound it in terms of the empirical variance. For a general loss function $\tilde{\ell}(\cdot,\cdot)$ (not necessarily within $[0,1]$), we define the empirical variance of the loss of h by $\hat{\tilde{\mathbb{V}}}(h,S) = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{\ell}(h(X_i),Y_i) - \tilde{L}(h))^2$. We recall that $\tilde{L}$, $\tilde{\mathbb{V}}$, and $\hat{\tilde{L}}$ were defined above Theorem 9. We note that the result of Tolstikhin and Seldin assumes that the losses are bounded in the $[0,1]$ interval. Rescaling to a general range introduces the squared range factor c^2 in front of the last term in the inequality, since scaling a random variable by c scales the variance by c^2.

Theorem 13 (Tolstikhin and Seldin, 2013)**.** *Let $\tilde{\ell}(\cdot,\cdot)$ be an arbitrary loss function with range c. Then for any distribution π on $\mathcal{H}$ that is independent of S, any $\lambda \in \left(0, \frac{2(n-1)}{n}\right)$, and any $\delta \in (0,1)$, with probability at least $1-\delta$ over a random draw of the sample S, for all distributions ρ on $\mathcal{H}$ simultaneously:*

$$\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] \le \frac{\mathbb{E}_\rho[\hat{\tilde{\mathbb{V}}}(h)]}{1-\frac{\lambda n}{2(n-1)}} + \frac{c^2\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}\right)}{n\lambda\left(1-\frac{\lambda n}{2(n-1)}\right)}.$$

We note that, similar to the PAC-Bayes-Bennett inequality, but in contrast to the PAC-Bayes-kl and PAC-Bayes-λ inequalities, the inequality above holds for a fixed value of λ and in case of optimization over λ a union bound has to be applied.

The last thing that is left is to bound the value and range of μ-tandem losses defined in equation (2).

Lemma 14. *For $\mu < 0.5$ the range of the μ-tandem loss $\ell_\mu(\cdot,\cdot,\cdot)$ is $K_\mu = \max\{1-\mu, 1-2\mu\}$ and it is upper bounded $\ell_\mu(\cdot,\cdot,\cdot) \le (1-\mu)^2$.*

A proof is provided in Appendix B. Taking together the results of Theorems 8, 9, 13, and Lemma 14 we obtain the following result.

Theorem 15. *For any parameter grid $\{\mu_1,\ldots,\mu_{k_\mu}\}$, $\{\gamma_1,\ldots,\gamma_{k_\gamma}\}$, and $\{\lambda_1,\ldots,\lambda_{k_\lambda}\}$ any distribution π on $\mathcal{H}$ that is independent of S, and any $\delta \in (0,1)$, with probability at least $1-\delta$ over a random draw of S, for all distributions ρ on $\mathcal{H}$, and all values of μ, γ, and λ in the corresponding grids simultaneously:*

$$L(\mathrm{MV}_\rho) \le \frac{1}{(0.5-\mu)^2}\Bigg(\mathbb{E}_{\rho^2}[\hat{L}_\mu(h,h',S)] + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2k_\mu k_\gamma k_\lambda}{\delta}}{\gamma n} + \frac{\phi(\gamma(1-\mu)^2)}{\gamma(1-\mu)^4}\left(\frac{\mathbb{E}_{\rho^2}[\hat{\mathbb{V}}_\mu(h,h',S)]}{1-\frac{\lambda n}{2(n-1)}} + \frac{K_\mu^2\left(2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2k_\mu k_\gamma k_\lambda}{\delta}\right)}{n\lambda\left(1-\frac{\lambda n}{2(n-1)}\right)}\right)\Bigg).$$

Proof. The result follows by reverse substitution of the result of Lemma 14 into Theorem 13, then into Theorem 9, and finally into Theorem 8. The factor $2k_\mu k_\gamma k_\lambda$ comes from a union bound over the parameter grid and the bounds in Theorems 9 and 13. □

6 Experiments

We start with a simulated comparison of the oracle bounds and then present an empirical evaluation on real data.

Comparison of the oracle bounds

Figure 1 depicts a comparison of the second order oracle bound based on the Chebyshev-Cantelli inequality (Theorems 5, 7 and 8, which, as oracle bounds, are equivalent) and the second order oracle bound based on the second order Markov's inequality (Theorem 3). We plot the ratio of the right hand side of the bound in Theorem 7 for the optimal value $\mu^* = \mathbb{E}_\rho[L(h)] - \frac{\mathbb{E}_{\rho^2}[L(h,h')] - \mathbb{E}_\rho[L(h)]^2}{0.5-\mathbb{E}_\rho[L(h)]}$ to the value of the right hand side of the bound in Theorem 3. A simple calculation shows that if $\mathbb{E}_{\rho^2}[L(h,h')] = 0.5\mathbb{E}_\rho[L(h)]$, then $\mu^* = 0$, which

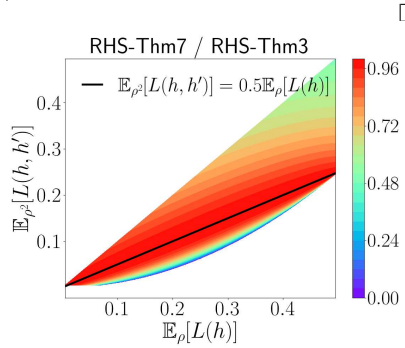

Figure 1: Theorem 7 vs. Theorem 3

recovers the bound in Theorem 3. The line $\mathbb{E}_{\rho^2}[L(h,h')] = 0.5\mathbb{E}_\rho[L(h)]$ is shown in black in Figure 1. We also note that $\mathbb{E}_\rho[L(h)]^2 \leq \mathbb{E}_{\rho^2}[L(h,h')] \leq \mathbb{E}_\rho[L(h)]$, which defines the feasible region in Figure 1. Whenever $\mathbb{E}_{\rho^2}[L(h,h')] \neq 0.5\mathbb{E}_\rho[L(h)]$ the Chebyshev-Cantelli inequality provides an improvement over second order Markov's inequality. The region above the black line, where $\mathbb{E}_{\rho^2}[L(h,h')] > 0.5\mathbb{E}_\rho[L(h)]$, is the region of high correlation of errors and in this case majority vote yields little improvement over individual classifiers. In this region the first order oracle bound is tighter than the second order oracle bounds (see Appendix C). The region below the black line, where $\mathbb{E}_{\rho^2}[L(h,h')] < 0.5\mathbb{E}_\rho[L(h)]$, is the region of low correlation of errors. In this region the second order oracle bounds are tighter than the first order oracle bound. Note that the potential for improvement below the black line is much higher than above it.

Empirical evaluation on real datasets

We studied the empirical performance of the bounds using standard random forest [Breiman, 2001] and a combination of heterogeneous classifiers on a subset of data sets from UCI and LibSVM repositories [Dua and Graff, 2019, Chang and Lin, 2011]. An overview of the data sets is given in Appendix E.1. The number of points varied from 3000 to 70000 with dimensions $d < 1000$. For each data set, we set aside 20% of the data for the test set $\mathrm{S_{test}}$ and used the remaining data S for ensemble construction, weight optimization and bound evaluation. We evaluate the classifiers and bounds obtained by minimizing the tandem bound TND [Masegosa et al., 2020, Theorem 9], which is the empirical bound on the oracle tandem bound in Theorem 3, the μ-tandem bound $\mathrm{C}\mu\mathrm{TND}$ (Theorem 12), and the tandem bound with an offset COTND (Theorem 15). We made 5 repetitions of each experiment.

Ensemble construction and minimization of the bounds. We follow the construction used by Masegosa et al. [2020]. The idea is to generate multiple random splits of the data set S into pairs of subsets $S = T_h \cup S_h$, such that $T_h \cap S_h = \emptyset$. Each hypothesis is trained on T_h and the empirical loss on S_h provides an unbiased estimate of its expected loss. Note that the splits cannot depend on the data. For our experiments, we generate these splits by bagging, where out-of-bag (OOB) samples S_h provide unbiased estimates of expected losses of individual hypotheses h. The resulting set of hypotheses produces an ensemble. As in the work of Masegosa et al., two modifications are required to apply the bounds: the empirical losses $\hat{L}(h,S)$ in the bounds are replaced by the validation losses $\hat{L}(h,S_h)$, and the sample size n is replaced by the minimal validation size $\min_h |S_h|$. For pairs of hypotheses (h,h'), we take the overlaps of their validation sets $S_h \cap S_{h'}$ to calculate an unbiased estimate of their tandem loss $\hat{L}(h,h',S_h \cap S_{h'})$, μ-tandem loss $\hat{L}_\mu(h,h',S_h \cap S_{h'})$, and the variance of the μ-tandem loss $\hat{\mathbb{V}}_\mu(h,h',S_h \cap S_{h'})$, which replaces the corresponding empirical losses in the bounds. The sample size is then replaced by $\min_{h,h'} |S_h \cap S_{h'}|$. The details on bound minimization are provided in Appendix D.

Optimizing weighted random forest. In the first experiment we compare TND, $\mathrm{C}\mu\mathrm{TND}$, and COTND bounds in the setting studied by Masegosa et al. [2020]. We take 100 fully grown trees, use the Gini criterion for splitting, and consider $\sqrt{d}$ features in each split. Figure 2a compares the loss of the random forest on $\mathrm{S_{test}}$ using either uniform weighting ρ_u or optimized weighting ρ^* found by minimization of the three bounds (we exclude the first order bound from the comparison, since it was shown by Masegosa et al. that it significantly deteriorates the test error of the ensemble). While $\mathrm{C}\mu\mathrm{TND}$ often performs similar or slightly worse than TND, we find that optimizing using COTND rarely degrades, but often improves accuracy. Figure 2b compares the tightness of the optimized $\mathrm{C}\mu\mathrm{TND}$ and COTND bounds to the optimized TND bound. The $\mathrm{C}\mu\mathrm{TND}$ is generally comparable to TND, while COTND is consistently looser than TND, mainly due to the union bounds. Numerical values for loss and bounds can be found in Tables E.2 and E.3 in Appendix E.2.

Ensembles with heterogeneous classifiers. In the second experiment, we consider ensembles of heterogeneous classifiers (Linear Discriminant Analysis, k-Nearest Neighbors, Decision Tree, Logistic Regression, and Gaussian Naive Bayes). A detailed description is provided in Appendix E.3. Compared to random forests, the variation in performance of ensemble members is larger here.

Figure 2c compares the ratio of the loss of the majority vote with optimized weighting to the loss of majority vote with uniform weighting on $\mathrm{S_{test}}$ for ρ^* found by minimization of the first order bound (FO), TND, $\mathrm{C}\mu\mathrm{TND}$, and COTND. The numerical values are given in Table E.4 in the

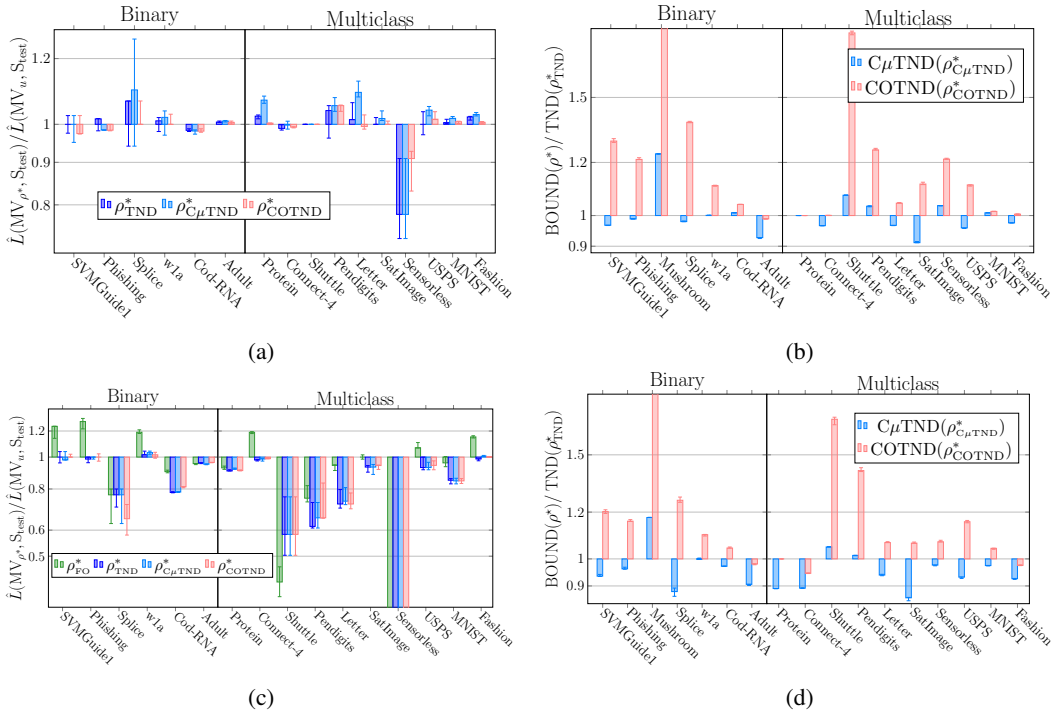

Figure 2: **(a,b) Optimized weighted random forest. (c,d) Ensembles with heterogeneous classifiers.** The median, 25%, and 75% quantiles of: (a,c) the ratio $\hat{L}(\mathrm{MV}_{\rho^*}, \mathrm{S}_{\mathrm{test}})/\hat{L}(\mathrm{MV}_u, \mathrm{S}_{\mathrm{test}})$ of the test loss of the majority vote with optimized weighting ρ^* generated by TND, CμTND and COTND to the test loss of majority vote with uniform weighting, and (b,d) the ratio $\mathrm{BOUND}(\rho^\star)/\,\mathrm{TND}(\rho^\star_{\mathrm{TND}})$ of the CμTND and COTND bounds to the TND bound with the corresponding optimized weighting . The plots are on a logarithmic scale. Values above 1 represent degradation and values below 1 represent improvement. Data sets with $L(\mathrm{MV}_u, \mathrm{S}_{\mathrm{test}}) = 0$ are left out in (a,c).

supplementary material. We observed that optimizing the FO tends to improve the ensemble accuracy in some cases but degrade in others. However, TND, CμTND, and COTND almost always improve the performance w.r.t. the uniform weighting. Table E.4 also shows that choosing the best single hypothesis gives almost identical results as optimizing FO. Figure 2d compares the tightness of the CμTND and COTND bounds relative to the TND bound. The numerical values are given in Table E.5 in the supplementary material. In this case, we have that CμTND is usually tighter than TND, while COTND is usually looser than TND.

7 Discussion

We derived an optimization-friendly form of the Chebyshev-Cantelli inequality and applied it to derive two new forms of second order oracle bounds for the weighted majority vote. The new oracle bounds bridge between the C-bounds [Germain et al., 2015] and the tandem bound [Masegosa et al., 2020] and take the best of both: the tightness of the Chebyshev-Cantelli inequality and the optimization and estimation convenience of the tandem bound. We also derived the PAC-Bayes-Bennett inequality, improving on the PAC-Bayes-Bernstein inequality by Seldin et al. [2012].

Our paper opens several directions for future research. One of them is a better treatment of parameter search in parametric bounds that would be tighter than a union bound over a grid. It would also be interesting to find other applications for the new form of Chebyshev-Cantelli inequality and the PAC-Bayes-Bennett inequality.

References

Daniel Berend and Aryeh Kontorovich. A finite sample analysis of the naive Bayes classifier. *Journal of Machine Learning Research*, 2016.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In O. Bousquet, U.v. Luxburg, and G. Rätsch, editors, *Advanced Lectures in Machine Learning*. Springer, 2004.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2), 1996.

Leo Breiman. Random forests. *Machine learning*, 45(1), 2001.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pages 785–794, 2016.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambertw function. *Advances in Computational mathematics*, 5(1), 1996.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

Ciprian Florescu and Christian Igel. Resilient backpropagation (rprop) for batch-learning in tensorflow. In *International Conference on Learning Representations (ICLR) Workshop*, 2018.

Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1996.

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38, 1999.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

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. *Journal of Machine Learning Research*, 16, 2015.

Christian Igel and Michael Hüsken. Empirical evaluation of the improved rprop learning algorithms. *Neurocomputing*, 50, 2003.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2002.

François Laviolette, Mario Marchand, and Jean-Francis Roy. From PAC-Bayes bounds to quadratic programs for majority votes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general ensemble methods with an application to multiclass classification. *Neurocomputing*, 219, 2017.

Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. On PAC-Bayesian bounds for random forests. *Machine Learning*, 2019.

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

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1999.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

David McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Conference on Learning Theory (COLT)*, 1998.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3, 2002.

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

Niklas Thiemann, Christian Igel, and Yevgeny Seldin. PAC-Bayesian aggregation without cross-validation. `http://arxiv.org/abs/1608.05610`, 2016.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2017.

Ilya Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

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] All assumptions are stated in the statements of the theorems.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Majority vote is a meta-algorithm and, as such, inherits the societal impact of the base algorithms.
   (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]
   (b) Did you include complete proofs of all theoretical results? [Yes]
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] Section 6, Appendices E, D.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 6, Appendices E, D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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] Section E.
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 the code and in the paper.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] New code in the supplementary.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We use publicly available datasets from the UCI and LibSVM repositories.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We use publicly available datasets from the UCI and LibSVM repositories. The data does not contain personally identifiable information or offensive content.
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]