Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound

Valentina Zantedeschi∗1,2,3, Paul Viallard†4, Emilie Morvant†4, Remi Emonet†4, Amaury Habrard†4, Pascal Germain‡5, and Benjamin Guedj§1,2,3

1Inria, Lille - Nord Europe
2Inria-London - The Inria London Programme
3University College London, Centre for Artificial Intelligence
4Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d’Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France
5Université Laval, Québec, Canada

Abstract

We investigate a stochastic counterpart of majority votes over finite ensembles of classifiers, and study its generalization properties. While our approach holds for arbitrary distributions, we instantiate it with Dirichlet distributions: this allows for a closed-form and differentiable expression for the expected risk, which then turns the generalization bound into a tractable training objective. The resulting stochastic majority vote learning algorithm achieves state-of-the-art accuracy and benefits from (non-vacuous) tight generalization bounds, in a series of numerical experiments when compared to competing algorithms which also minimize PAC-Bayes objectives – both with uninformed (data-independent) and informed (data-dependent) priors.

1 Introduction

By combining the outcomes of several predictors, ensemble methods [Dietterich 2000] have been shown to provide models that are more accurate and more robust than each predictor taken singularly. The key to their success lies in harnessing the diversity of the set of predictors [Kuncheva 2004]. Among ensemble methods, weighted Majority Votes (MV) classifiers assign a score to each base classifier (a.k.a. voter) and output the most common prediction, given by the weighted majority. When voters have known probabilities of making an error and make independent predictions, the optimal weighting is given by the so-called Naive Bayes rule [Berend and Kontorovich 2015]. However, in most situations these assumptions are not satisfied, giving rise to the need for techniques that estimate the optimal combination of voter predictions from the data.

Among them, PAC-Bayesian based methods are well-grounded approaches for optimizing the voter weighting. Indeed, PAC-Bayes theory (introduced by Shawe-Taylor and Williamson 1997, McAllester 1999 – we refer to [Guedj 2019] for a recent survey and references therein) provides not only bounds on the true error of a MV through Probably Approximately Correct (PAC) generalization bounds (see e.g. Catoni 2007, Seeger 2002, Maurer 2004, Langford and Shawe-Taylor 2002, Germain et al. 2015), but is also suited to derive theoretically grounded learning algorithms (see e.g. Germain et al. 2009, Roy et al. 2011).
Contrary to the most classical PAC bounds [Valiant, 1984], as VC-dimension [Vapnik, 2000] or Rademacher-based bounds [Mohri et al., 2012], PAC-Bayesian guarantees do not stand for all hypotheses (i.e. are not expressed as a worst-case analysis) but stand in expectation over the hypothesis set. They involve a hypothesis space (formed by the base predictors), a prior distribution on it (i.e. an a priori weighting) and a posterior distribution (i.e. an a posteriori weighting) evaluated on a learning sample. The prior brings some prior knowledge on the combination of predictors, and the posterior distribution is learned (adjusted) to lead to good generalization guarantees; the deviation between the prior and the posterior distributions plays a role in generalization guarantee and is usually captured by the Kullback-Leibler (KL) divergence. In their essence, PAC-Bayesian results do not bound directly the risk of the deterministic MV, but bound the expected risk of one (or several) base voters randomly drawn according to the weight distribution of the MV [Langford and Shawe-Taylor, 2002] [Lacasse et al., 2006, 2010] [Germain et al., 2015] [Masegosa et al., 2020]. This randomization scheme leads to upper bounds on the true risk of the MV that are then used as a proxy to derive PAC-Bayesian generalization bounds. However, the obtained risk certificates are generally not tight, as they depend on irreducible constant factors, and when optimized they can lead to sub-optimal weightings. Indeed, by considering a random subset of base predictors, state-of-the-art methods do not fully leverage the diversity of the whole set of voters. This is especially a problem when the voters are weak, and learning to combine their predictions is critical for good performance.

Our contributions. In this paper, we propose a new randomization scheme. We consider the voter weighting associated to a MV as a realization of a distribution of voter weightings. More precisely, we analyze with the PAC-Bayesian framework the expected risk of a MV drawn from the posterior distribution of MVs. The main difference with the literature is that we propose a stochastic MV, while previous works aim at studying randomized evaluations of the true risk of the deterministic MV. Doing so, we are able to derive tight empirical PAC-Bayesian bounds for our model directly on its expected risk, in Section 4. We further propose, in Section 3 two approaches for optimizing the generalization bounds, hence learning the optimal posterior: the first optimizes an analytical and differentiable form of the empirical risk that can be derived when considering Dirichlet distributions; the second optimizes a Monte Carlo approximation of the expected risk and can be employed with any form of posterior. In our experiments, reported in Section 5 we first compare these two approaches, highlighting in which regimes one is preferable to the other. Finally, we assess our method’s performance on real benchmarks w.r.t. the performance of PAC-Bayesian approaches also learning MV classifiers. These results indicate that our models enjoy generalization bounds that are consistently tight and non-vacuous both when studying ensembles of data-independent predictors and when studying ensembles of data-dependent ones.

Societal impact. Our work abides by the ethical guidelines enforced in contemporary research in machine learning. Given the theoretical nature of our contributions we do not foresee immediate potentially negative societal impact.

2 Notation and background

In this section, we formally define weighted Majority Vote (MV) classifiers and review the principal PAC-Bayesian approaches for learning them.

2.1 Weighted majority vote classifiers

Consider the data random variable \((X, Y)\), taking values in \(\mathcal{X} \times \mathcal{Y}\) with \(\mathcal{X} \subseteq \mathbb{R}^d\) a \(d\)-dimensional representation space and \(\mathcal{Y}\) the set of labels. We denote \(\mathcal{P}\) the (unknown) data distribution of \((X, Y)\). We define a set (dictionary) of base classifiers \(D=\{h_j : \mathcal{X} \rightarrow \mathcal{Y}\}_{j=1}^M\). The weighted majority vote classifier is a convex combination of the base classifiers from \(D\). Formally, a MV is parameterized by a weight vector \(\theta \in [0, 1]^M\),
The current paper stands in the context of PAC-Bayesian MV learning. The PAC-Bayesian framework has been employed to study the generalization guarantees of randomized classifiers. It is known to provide tight risk certificates that can be used to derive self-bounding learning algorithms [e.g. Germain et al., 2015; Lorenzen et al., 2019; Masegosa et al., 2020]. In this sense, $f_\theta$ corresponds to the MV predictor $f_\theta(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{h \sim \mathcal{C}(\theta)} [1(h(x) = y)]$ and $W_\theta$ corresponds to the expected ratio of wrong predictors $W_\theta(X, Y) = \mathbb{E}_{h \sim \mathcal{C}(\theta)} [1(h(X) \neq Y)]$. The PAC-Bayesian analysis provides a sensible way to find such a categorical distribution, called posterior distribution, that leads to a model with low true risk $R(f_\theta)$. However, an important caveat is that the PAC-Bayesian generalization bounds cannot be derived directly on the true risk $R(f_\theta)$, without making assumptions on the distribution of $W_\theta$ and raising fidelity problems.
Thus, a common approach is to consider oracle upper bounds on the true risk in terms of quantities from which PAC-Bayesian bounds can be derived, that are typically related to the statistical moments of $W_{\theta}$. By doing so, oracle bounds act as a proxy for estimating the cumulative density function for $W_{\theta} \geq 0.5$ [Langford and Shawe-Taylor, 2002; Germain et al., 2006; Lacasse et al., 2010; Masegosa et al., 2020]. Generalization guarantees for $R(f_{\theta})$ are hence derived, involving the empirical counterpart of the oracle bound, the KL-divergence between the posterior categorical distribution $C(\theta)$ and a prior one. An overview of the existing oracle bounds described below is represented in Figure 1.

**First order bound.** The most classical “factor two” oracle bound [Langford and Shawe-Taylor, 2002] is derived considering the relation between MV’s risk and the first moment of $W_{\theta}$, a.k.a. the expected risk of the randomized classifier: $R_1(\theta) = \mathbb{E}_{h \sim C(\theta)} R(h) = \mathbb{E}_P W_{\theta}$. By applying Markov’s inequality, we have $R(f_{\theta}) \leq 2R_1(\theta)$. This “factor two” bound is close to zero only if the expectation of the risk of a single base classifiers drawn according to $C(\theta)$ is itself close to zero. Therefore, it does not take into account the correlation between base predictors, which is key to characterize how a MV classifier can achieve strong predictions even when its base classifiers are individually weak (as observed when performing e.g. Boosting [Schapire and Singer, 1999]). This explains why $R_1(\theta)$ can be a very loose estimate of $R(f_{\theta})$ when the base classifiers are adequately decorrelated.

**Randomized MV.** A generalization of the first order approach was proposed in [Lacasse et al., 2010], where the true risk of the MV is estimated by drawing multiple ($N$) base hypotheses and computing the probability that at least $\frac{N}{2}$ make an error (which is given by the binomial with parameter $W_{\theta}$: $W_{\theta}^N(X, Y) = \sum_{k=\frac{N}{2}}^{N} \binom{N}{k} W_{\theta}^k (1 - W_{\theta})^{N-k}$). The higher $N$, the better $W_{\theta}^N(X, Y)$ approximates the true risk, but the looser the bound, as the KL term worsens by a factor of $N$. Moreover, with this approach it is not possible to derive generalization bounds directly on the true risk, and the corresponding oracle bound still presents a factor two: $R(f_{\theta}) \leq 2 \mathbb{E}_P W_{\theta}^N$.

**Second order bound.** A parallel line of works focuses on improving the bounds by accounting for voter correlations, i.e. considering the agreement and/or disagreement of two random voters. Masegosa et al., 2020 recently proposed a new upper bound for the true risk depending on the second moment of $W_{\theta}$, a.k.a. tandem loss or joint error, the risk that two predictors make a mistake on the same point: $R_2(\theta) = \mathbb{E}_{h \sim C(\theta), h' \sim C(\theta)} \mathbb{1}(h(x) \neq Y \wedge h'(x) \neq Y) = \mathbb{E}_P W_{\theta}^2$. By applying the second order Markov’s inequality, we have $R(f_{\theta}) \leq 4R_2(\theta)$. Masegosa et al., 2020 show that in the worst case (i.e. when the base classifiers are identical) the second order bound could be twice worse than the first order bound ($R_2(\theta) \approx 2R_1(\theta)$), while in the best case (i.e. the $M$ base classifiers are perfectly decorrelated), the second order bound could be an order of magnitude tighter ($R_2(\theta) \approx \frac{1}{4}R_1(\theta)$).

**C-bound.** Originally proposed by Breiman, 2001 in the context of Random Forest, a bound known as the C-Bound in the PAC-Bayesian literature [Lacasse et al., 2006; Roy et al., 2011; Germain et al., 2015; Laviolette et al., 2017] is derived by considering explicitly the joint error and disagreement between two base predictors. Using Chebyshev-Cantelli’s inequality, the C-Bound can be written in terms of the first and second moments of $W_{\theta}$: $R(f_{\theta}) \leq \frac{R_2(\theta) - \mathbb{E}_{h \sim C(\theta)}(\mathbb{E}_P \mathbb{1}(h(x) \neq Y))^2}{R_2(\theta) - R_1(\theta) + \frac{1}{4}}$. This bound is tighter than the second order one if the disagreement $\mathbb{E}_{h, h'} \mathbb{E}_P (\mathbb{1}(h(x) \neq h'(x)))$ is greater than $R_2(\theta)$ [Masegosa et al., 2020].

From a practical point of view, one could minimize the generalization bounds of one of the above methods to learn a weight distribution over an ensemble of predictors. However, this could lead to sub-optimal MV classifiers. To illustrate this behavior we plot in Figure 3 the decision surfaces learnt by the minimization of a PAC-Bayesian bound on each of the aforementioned oracle bounds. These plots provide evidence that, when the base classifiers are weak, state-of-the-art PAC-Bayesian methods do not necessarily build powerful ensembles (failing to improve upon a Naive Bayes approach [Berend and Kontorovich, 2015]). First Order

---

1 Oracle bounds are expressed in terms of the unknown data distribution; their exact value cannot be computed.
concentrates on few base classifiers, as previously observed by [Lorenzen et al. 2019] and [Masegosa et al. 2020], while Second Order and C-Bound fail to leverage the diversity in the whole set of classifiers. On the contrary, Randomized is able to fit the problem by drawing more than 2 voters, but it provides generalization bounds that are loose even when the learned model exhibits good generalization capabilities, as in this case.

3 Our approach: stochastic weighted majority vote

In our framework for deriving generalization bound for Majority Vote classifiers, we consider \( f_\theta \) as a realization of a distribution of majority vote classifiers, with parameter \( \theta \in \Theta \subseteq \mathbb{R}^M \) and probability measure \( \rho \), as represented in Figure 2. The main advantage of considering a stochastic majority vote is that it allows to derive and optimize PAC-Bayesian generalization bounds directly on its true risk and to fully leverage the set of base classifiers. The true risk of the proposed stochastic weighted majority vote takes into account the whole distribution of MVs \( \rho(\theta) \), as follows:

\[
\int_\Theta R(f_\theta)\rho(d\theta) = \int_\Theta \mathbb{E}_P 1(W_\theta(X, Y) \geq 0.5)\rho(d\theta) = \mathbb{E}_P \int_\Theta 1(W_\theta(X, Y) \geq 0.5)\rho(d\theta). \tag{4}
\]

Given its stochastic nature, in order to evaluate and/or minimize Equation (4) we can either (i) compute its closed form or (ii) approximate it (e.g. through Monte Carlo methods). In both cases, assumptions have to be made on the form of \( \rho \). As the components of \( \theta \) are constrained to sum to one, \( \theta \) lies in the \((M-1)\)-simplex: \( \Theta = \Delta^{(M-1)} \), hence natural choices for its probability measure are e.g. the Dirichlet or the Logit Normal distributions. In the following, we show that under Dirichlet assumptions, we can derive an analytical and differentiable form of the risk of a stochastic MV.

3.1 Exact risk under Dirichlet assumptions

First of all, we recall that the probability density function of the Dirichlet distribution is defined by:

\[
\theta \sim \mathcal{D}(\alpha_1, \ldots, \alpha_M), \quad \rho(\theta) = \frac{1}{B(\alpha)} \prod_{j=1}^M (\theta_j)^{\alpha_j-1}, \tag{5}
\]

with \( \alpha = [\alpha_j \in \mathbb{R}^+]_{j=1}^M \) the vector of concentration parameters and \( B(\alpha) \) a normalization factor (see App A for its definition). Notice that by taking \( \alpha \) as the unit vector, the distribution corresponds to a uniform distribution over the \((M-1)\)-simplex.
Under these assumptions for the MV distribution, an closed form can be derived for the expected risk:

**Lemma 1.** For a given \((x, y) \sim \mathcal{P}\), let \(w = \{j | h_j(x) \neq y\}\) be the set of indices of the base classifiers that misclassify \((x, y)\) and \(c = \{j | h_j(x) = y\}\) be the set of indices of the base classifiers that correctly classify \((x, y)\). The expected error (or 0-1-loss) for \((x, y)\) of the stochastic majority vote under \(\theta \sim \mathcal{D}(\alpha_1, \ldots, \alpha_M)\) is equal to

\[
\int_{\Theta} \mathbb{1}(W_{\theta}(x, y) \geq 0.5) \rho(d\theta) = I_{0.5} \left( \sum_{j \in c} \alpha_j, \sum_{j \in w} \alpha_j \right),
\]

with \(I_{0.5}(\cdot)\) the regularized incomplete beta function evaluated at 0.5.

**Proof.** We rewrite \(W_{\theta}\) as \(W_{\theta}(x, y) = \sum_{j=1}^M \theta_j \mathbb{1}(h_j(x) \neq y) = \sum_{j \in w} \theta_j\), and use the aggregation property of Dirichlet distributions (Lemma 2 in Appendix) to show that \(W_{\theta}\) follows a bivariate Dirichlet distribution (a.k.a. Beta distribution) over the two sets of wrong and correct base classifiers:

\[
P[W_{\theta}(x, y) = \omega] = \mathbb{P} \left[ \sum_{j \in w} \theta_j = \omega \right] \mathbb{P} \left[ \sum_{j \in c} \theta_j = 1 - \omega \right]
= \mathcal{D} \left( \sum_{j \in w} \alpha_j, \sum_{j \in c} \alpha_j \right)
\quad \text{by aggregation.}
\]

Finally, notice that the expected error is related to the cumulative probability function of \(W_{\theta}\), the incomplete beta function \(I_p : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow [0, 1]:\)

\[
\int_{\Theta} \mathbb{1}(W_{\theta}(x, y) \geq 0.5) \rho(d\theta) = \int_{0.5}^{1} \mathbb{P}[dW_{\theta}(x, y)]
= 1 - I_{0.5} \left( \sum_{j \in w} \alpha_j, \sum_{j \in c} \alpha_j \right) = I_{0.5} \left( \sum_{j \in c} \alpha_j, \sum_{j \in w} \alpha_j \right).
\]

Equation \((9)\) follows by symmetry of the incomplete beta function: \(I_p(a, b) = 1 - I_{1-p}(b, a)\).

The expected risk can be then expressed as follows:

\[
R(\rho) = \mathbb{E}_\mathcal{P} \int_{\Theta} \mathbb{1}(W_{\theta}(x, y) \geq 0.5) \rho(d\theta) = \mathbb{E}_\mathcal{P} I_{0.5} \left( \sum_{j \in c} \alpha_j, \sum_{j \in w} \alpha_j \right).
\]

Importantly, this exact form of the risk is differentiable, hence can be directly optimized by gradient-based methods.

### 3.2 Monte Carlo approximated risk

We now propose a relaxed Monte Carlo (MC) optimization scheme for those distributions that do not admit an analytical form of the expected risk, unlike the Dirichlet one. This second strategy is also suited to speed up training, in some cases, as the derivatives of the exact risk depend on the hyper-geometric function and can be slow to evaluate (see App. A.4). With the approximated scheme, in order to update \(\alpha\) by gradient descent we need to relax the true risk as the gradients of the 0-1-loss are always null for discrete \(W_{\theta}\). In practice, we make use of a tempered sigmoid loss \(\sigma_c(x) = \frac{1}{1 + \exp(-c x)}\) with slope parameter \(c \in \mathbb{R}^+\). De facto this corresponds to solving a relaxation of the problem and not its exact form [Nesterov, 2005]. At each iteration of the MC optimization algorithm we perform:

1. Draw a sample \(\{\theta_t \sim \rho(\alpha)\}_{t=1}^T\) using the implicit reparameterization trick [Figurnov et al., 2018, Jankowiak and Obermeyer, 2018].
2. Compute the relaxed empirical risk $\sum_{t=1}^{T} \hat{R}_{\sigma_c}(\theta) = \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_c(W_\theta(x_i, y_i) - 0.5)$;

3. Update $\alpha$ by gradient descent.

Notice that when considering Dirichlet distributions for the posterior and the prior, at inference time the empirical PAC-Bayesian bounds can still be evaluated using the exact form of Lemma 1.

A drawback of the approximated scheme is that it has a complexity linear in the number of MC draws $T$, but also linear in the number of predictors $M$, as sampling over the simplex requires sampling from $O(M)$ distributions, one per base classifier. As an example, sampling from a Dirichlet over the $(M-1)$-simplex is usually implemented as sampling from $M$ Gamma distributions and normalizing the samples so that they lie on the simplex. In contrast, the exact formulation’s complexity is constant in $M$ as it depends only on the sets of wrong and of correct predictors, hence on a constant number of variables (2) no matter the number of predictors $M$. In Section 5.1 we empirically study the trade-off between training time and accuracy, showing in which regimes it is more convenient to optimize the relaxed MC risk than optimizing the exact one, and vice versa.

4 PAC-Bayesian generalization guarantees

We now derive PAC-Bayesian generalization upper bounds for the proposed stochastic MV. In our context, upper bounds can be derived for studying the gap between true and empirical risk considering a prior distribution $\pi$ over the hypothesis space $\Theta$. In this paper, we make use of one of the tightest classical PAC-Bayesian bound [Seeger 2002, Maurer 2004].

**Theorem 1** (Seeger’s bound). For any $\pi$ over $\Theta$ and $\delta \in (0, 1)$ with probability at least $1-\delta$ over samples $S = \{(x_i, y_i) \sim P\}_{i=1}^{n}$ of size $n$ we have simultaneously for any posterior $\rho$ over $\Theta$:

$$
\int_\Theta R(f_\rho)\rho(d\theta) \leq \text{kl}^{-1}\left(\int_\Theta \hat{R}(f_\rho)\rho(d\theta) , \frac{\text{KL}(\rho, \pi) + \text{ln}\left(\frac{2\sqrt{n}}{3}\right)}{n}\right),
$$

(11)

with $\hat{R}(f_\rho) = \frac{1}{n} \sum_{i=1}^{n} 1(W_\rho(x_i, y_i) \geq 0.5)$ the empirical risk on sample $S$, $\text{KL}(\rho, \pi) = \int_\Theta \rho(\theta) \log \frac{\rho(\theta)}{\pi(\theta)} d\theta$ the KL divergence and $\text{kl}^{-1}(q, \epsilon)$ the inverse of the binary KL divergence defined as $\text{kl}^{-1}(q, \epsilon) = \max\{p \in [0, 1] \mid \text{kl}(q, p) \leq \epsilon\}$.

A proof of Theorem 1 can be found in Seeger [2002]. The $\text{kl}^{-1}$ function can be evaluated via the bisection method and optimized by gradient descent, as proposed in [Reeb et al.] 2018. Note that our contributions do not restrict the choice of generalization bound.

Importantly, Theorem 1 is valid when the prior $\pi$ is independent from the data. Thus it cannot be evaluated with base classifiers learned from the training sample. However, it is known that considering a data-dependent prior can lead to tighter PAC-Bayes bounds [Dziugaite et al.] 2021. Following recent works on PAC-Bayesian bounds with data-dependent priors [Thiemann et al.] 2017, [Mhammedi et al.] 2019, we derive a cross-bounding certificate that allows us to learn and evaluate the set of base classifiers without held-out data. More precisely, we split the training data $S$ into two subsets $(S_{\leq m} = \{(x_i, y_i) \in S\}_{i=1}^{m}$ and $S_{> m} = \{(x_i, y_i) \in S\}_{i=m+1}^{n}$ and we learn a set of base classifiers on each data split independently (determining the hypothesis spaces $\Theta_{\leq m}$ and $\Theta_{> m}$). We refer to the prior distribution over $\mathcal{D}_{\leq m}$ as $\rho_{\leq m}$ and to the prior distribution over $\mathcal{D}_{> m}$ as $\rho_{> m}$. In the same way, we can then define a posterior distribution per hypothesis space: $\rho_{\leq m}$ and $\rho_{> m}$. The following theorem shows that we can bound the expected risk of any convex combination of the two posteriors, as long as their empirical risks are evaluated on the data split that was not used for learning their respective priors.
Theorem 2 (Seeger’s bound with informed priors). Let \( \pi_{\leq m} \) and \( \rho_{\leq m} \) be the prior and posterior distributions on \( \Theta_{\leq m} \), and \( \pi_{> m} \) and \( \rho_{> m} \) the prior and posterior distributions on \( \Theta_{> m} \). For any \( p \in (0, 1) \) and \( \delta \in (0, 1) \) with probability at least \( 1 - \delta \) over samples \( S = \{ (x_i, y_i) \sim \mathcal{P} \}_{i=1}^n \) we have

\[
\begin{align*}
\mathrm{kl} \left( pR(\rho_{> m}) + (1 - p)R(\rho_{\leq m}) \middle| \tilde{p}R_{\leq m}(\rho_{> m}) + (1 - p)\tilde{R}_{> m}(\rho_{\leq m}) \right) \\
\leq \frac{p}{m} KL(\rho_{> m}, \pi_{> m}) + \frac{(1 - p)}{n - m} KL(\rho_{\leq m}, \pi_{\leq m}) + \frac{4\sqrt{m(n-m)}}{n},
\end{align*}
\]

with \( R(\rho_{> m}) = \int_{\Theta_{> m}} R(f_\theta) \rho(d\theta) \), and \( R(\rho_{\leq m}) = \int_{\Theta_{\leq m}} R(f_\theta) \rho(d\theta) \),

and \( \tilde{R}_{\leq m}(\rho_{> m}) = \int_{\Theta_{\leq m}} \tilde{R}(f_\theta) \rho(d\theta) \), and \( \tilde{R}_{> m}(\rho_{> m}) = \int_{\Theta_{> m}} \tilde{R}(f_\theta) \rho(d\theta) \).

The result follows by \( \mathrm{kl} \)'s convexity. The complete proof is reported in App. \[3\]. In practice, following [Mhammedi et al. 2019] we set \( m = \frac{n}{2} \) and \( p = \frac{m}{n} \), and we learn the base classifiers by empirical risk minimization.

Comparison with existing bounds Until now, we considered bounds for the expected risk over the space of MVs but not for a single realization \( f_\theta \), unlike state-of-the-art methods. Under Dirichlet assumptions, we can bound the risk of the expected MV \( f_\theta \) by twice the expected risk (see App. \[3.5\]): \( R(f_\theta) = R(\mathbb{E}_{\rho(\theta)} f_\theta) \leq 2E_{\rho(\theta)} R(f_\theta) \). The obtained oracle bound would comprise an irreducible factor, but its KL term would not degrade, unlike state-of-the-art bounds that account for voter correlation [Lacasse et al. 2006, 2010, Masegosa et al. 2020]. The empirical bound on the deterministic MV could be also refined leveraging works on the disintegration of PAC-Bayesian guarantees [Blanchard and Fleuret 2007, Catoni 2007, Rivasplata et al. 2020, Viallard et al. 2021].

A downside of our method is that the complexity of the KL term grows with the number of base classifiers \( M \), unlike the KL on categoricals that tends to 0 in its limit. This results in making the generalization bounds increasingly looser and conservative with growing \( M \), even for low empirical risks. From a generalization perspective, our guarantees are hence able to reflect the complexity of the model, expressed as the size of the hypothesis space \( M \). However, our risk certificates do not account for redundancy in the voter set. For instance, they are not able to distinguish scenarios where base classifiers are highly correlated (hence less complex hypothesis space) from scenarios where base classifiers are independent (more complex). An expedient for ensuring that generalization bounds are tight consists in learning the hypothesis space: the number of base predictors can then be limited without degrading the performance of the majority vote.

5 Experiments

In this section, we empirically evaluate StocMV, and we compare its generalization bounds and test errors to those obtained with PAC-Bayesian methods learning majority votes. We show that our method allows to derive generalization bounds that are consistently tight (i.e. close to the test errors) and non-vacuous (i.e. smaller than 1) both when studying ensembles of weak predictors and when studying ensembles of strong ones.

In the following, we consider Dirichlet distributions for the prior and the posterior of our method, and refer to the model obtained by optimizing the exact risk as \textit{exact} and the one obtained by optimizing the approximated one as \textit{MC}. We consider as baselines the PAC-Bayesian methods described in Section \[2.1\]. We refer to the methods optimizing the First Order, Second Order and Randomized empirical bound as \textit{FO}, \textit{SO} and \textit{Rnd} respectively. We do not compare with the C-Bound, as it is hard to optimize on large-scale datasets and existing algorithms are suited only for binary classification. All generalization bounds are evaluated with a probability \( 1 - \delta = 0.95 \) and all prior distributions are set to the uniform (we provide a study for different priors in App. \[C.3\]). The posterior parameters (\( c \) for our method, \( \theta \) for the others) are initialized uniformly in \([0.01, 2]\) (and normalized to sum to 1 for \textit{SO}, \textit{FO} and \textit{Rnd}). Finally, for \textit{MC} the sigmoid’s slope parameter \( c \) is set to 100 and for \textit{Rnd} the number of voters drawn at each iteration is set to \( N = 100 \). Code
was implemented in pytorch [Paszke et al., 2019] and all experiments were run on a virtual machine with 8 vCPUs and 128Gb of RAM.

5.1 Comparison of exact and MC variants

For this set of experiments, we optimize Seeger's Bound (Equation 1) by (batch) Gradient Descent, for 1,000 iterations and with learning rate equal to 0.1. We study the performance of our method on the binary classification two-moons dataset, with 2 features, 2 classes and \( N(0, 0.05) \) Gaussian noise. We draw 1,000 points for training, and 1,000 points for testing. Figure 4 reports a comparison of exact and MC variants in terms of error, generalization bound and training time (in seconds). Increasing the number of MC draws \( T \) unsurprisingly allows to recover exact’s performance, but at a higher computational cost for \( M > 10^3 \) in this case. At the same time, when \( M > 10^2 \), MC achieves exact’s errors even with \( T = 1 \), probably because of increased redundancy in the base predictor set. We also observe that the error rates and bounds gradually degrade for higher values of \( M \) for both methods. This is due to the KL term increasing with \( M \), as highlighted in Section 4, becoming a too strong regularization during training and making the bound looser.

5.2 Experiments on real benchmarks

We now compare the considered methods on real datasets and on two different scenarios, depending on the type of PAC-Bayesian bounds that are evaluated: When making use of data-independent priors, we chose as voters axis-aligned decision stumps, with thresholds evenly spread over the input space (10 per feature); When making use of data-dependent priors, we build Random Forests [Breiman, 2001] as set of voters, with \( M = 100 \) trees learned bagging \( \frac{n}{2} \) points and sampling \( \sqrt{d} \) random features to ensure voter diversity, optimizing Gini impurity score and, unless stated otherwise, without bounding their maximal depth.

We consider several classification datasets from UCI [Dua and Graff, 2017], LIBSVM[^2] and Zalando [Xiao et al., 2017], of different number of features and of instances. Their descriptions and details on any pre-processing are provided in App. C.3. We train the models by Stochastic Gradient Descent (SGD) using Adam [Kingma and Ba, 2015] with (0.9, 0.999) running average coefficients, batch size equal to 1024 and learning rate equal to 0.1 with a scheduler reducing this parameter of a factor of 10 with 2 epochs patience. We fix the maximal number of epochs to 100 and patience equal to 25 for early stopping, and for MC we fix \( T = 10 \) to increase randomness.

We report the test errors and generalization bounds in Figure 5 (additional results are reported in the appendix, in Tables 1 and 2 and Figure 15): We compare the different methods on binary datasets and with data-independent priors in Figure 5a and on multi-class datasets and with data-dependent priors in Figure 5b.

[^2]: https://www.csie.ntu.edu.tw/~cjlin/libsvm/
First we notice that the bounds obtained by our method are consistently non vacuous and tighter than those obtained by the baselines on all datasets. Regarding the error rates, our method’s performance is generally aligned with the baselines, while it achieved error rates significantly lower than FO and SO on the perfectly separable two-moons dataset. Sensitivity to noise could explain why our method does not outperform the baselines on the studied real problems, as these usually present label and input noise. Indeed our learning algorithm optimizes the $01$-loss, which does not distinguish points with margins close or far from 0.5 because of its discontinuity in $W_\theta^0$. Preliminary results reported in App C.2 seem to confirm this supposition.

Finally, to gain a better understanding of the relation between base classifier strength and performance of the models obtained with the different methods, we further study their performance and generalization guarantees with varying voter strength. As hypothesis set, we learn a Random Forest with 100 decision trees, as before, but for this experiment we bound their maximal depth between 1 and 10. Constraining the tree depth allows to indirectly control how well the voters fit the training set (as shown in Appendix, Figure 14). Following Lorenzen et al. [2019], we assess the voter strength by computing the expected accuracy of a random voter. All methods’ results improve overall when voters get stronger, even though FO at a slower pace. Notice that on the considered datasets SO is the most sensitive method, particularly suffering from weak base predictors. Our method generally provides test errors comparable with the best baselines and consistently tighter bounds.

6 Future work

We propose a stochastic version of the classical majority vote classifier, and we directly analyze and optimize its expected risk through the PAC-Bayesian framework. The benefits on the model accuracy of this direct optimization are however reduced in presence of input noise, and fostering robustness in noisy contexts is the subject of future work. Another potential improvement would consist in tackling the discussed looseness of our generalization bounds with increasing number of base predictors, by accounting for redundancy in the hypothesis space.
Figure 6: Test error rates and PAC-Bayesian bound values as a function of the maximal depth of the voters (decision trees here). We report the means and standard deviations over 4 different runs, and mark with a red horizontal line the threshold above which Seeger’s bounds are vacuous. Additional results are available in the appendix.

Acknowledgement

This work was supported by the French Project APRIORI ANR-18-CE23-0015. Experiments presented in this paper were carried out using the Grid’5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). P. Germain is supported by the Canada CIFAR AI Chair Program, and the NSERC Discovery grant RGPIN-2020-07223.

References


Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. A PAC-Bayes risk bound for

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of

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. JMLR, 2015.


Martin Jankowiak and Fritz Obermeyer. Pathwise Derivatives Beyond the Reparameterization Trick. In
ICML, 2018.


Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes

Alexandre Lacasse, François Laviolette, Mario Marchand, and Francis Turgeon-Boutin. Learning with


François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general

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

Andrés R. Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian


Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected bernstein inequality. In


Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen,
Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary
DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and
Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In NeurIPS,
2019.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for


A Derivation details under Dirichlet assumptions

A.1 Dirichlet distribution

The probability density function of the Dirichlet distribution is defined as follows:

$$\theta \sim \mathcal{D}(\alpha_1, \ldots, \alpha_M), \quad \rho(\theta) = \frac{1}{B(\alpha)} \prod_{j=1}^{M} (\theta_j)^{\alpha_j - 1},$$  (12)

with $\alpha = [\alpha_j \in \mathbb{R}^+]_{j=1}^{M}$ the vector of the distribution parameters and $B(\alpha)$ the normalized multi-variate beta function, defined using the gamma function $\Gamma$:

$$B(\alpha) = \frac{\prod_{j=1}^{M} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{M} \alpha_j)}, \quad \Gamma(\alpha) = \int_{0}^{\infty} t^{\alpha-1} e^{-t} dt. \quad (13)$$

A.2 Aggregation property of Dirichlet distributions

Lemma 2. If $\theta \sim \mathcal{D}(\alpha_1, \ldots, \alpha_M)$, then for any $j \in [1, M]$ and $j' \in [1, M] \setminus \{j\}$ the variable $\theta'$ formed by dropping $\theta_j$ and $\theta_{j'}$ and adding their sum also follows a Dirichlet distribution

$$(\theta_1, \ldots, \theta_{M'}, \theta_j + \theta_{j'}) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_M, \alpha_j + \alpha_{j'}).$$

A.3 KL divergence between Dirichlet distributions

Let $\rho(\theta) = \mathcal{D}(\alpha)$ and $\pi(\theta) = \mathcal{D}(\beta)$, with $\alpha_0 = \sum_{j=1}^{M} \alpha_j$ and $\beta_0 = \sum_{j=1}^{M} \beta_j$. The KL divergence between $\rho$ and $\pi$ is equal to:

$$KL(\rho, \pi) = \ln(\Gamma(\alpha_0)) - \sum_{j=1}^{M} \ln(\Gamma(\alpha_j)) - \ln(\Gamma(\beta_0)) + \sum_{j=1}^{M} \ln(\Gamma(\beta_j)) + \sum_{j=1}^{M} (\alpha_j - \beta_j)(\psi(\alpha_j) - \psi(\alpha_0)) \quad (14)$$

with $\psi(\alpha)$ the digamma function: the first derivative of $\ln(\Gamma(\alpha))$ (see Figure 7).

Proof. We have

$$KL(\rho, \pi) = \int_{\Theta} \rho(\theta) \ln \left( \frac{\rho(\theta)}{\pi(\theta)} \right) d\theta \quad (15)$$

$$= \int_{\Theta} \ln \left( \frac{B(\beta)}{B(\alpha)} \prod_{j=1}^{M} \frac{\theta_j^{\alpha_j - 1}}{\theta_j^{\beta_j - 1}} \right) \rho(\theta) d\theta \quad (16)$$

$$= \int_{\Theta} \left( \ln B(\beta) - \ln B(\alpha) + \ln \left( \prod_{j=1}^{M} \theta_j^{\alpha_j - \beta_j} \right) \right) \rho(\theta) d\theta \quad (17)$$

$$= \int_{\Theta} \ln B(\beta) - \ln B(\alpha) + \sum_{j=1}^{M} (\alpha_j - \beta_j) \ln \theta_j \rho(\theta) d\theta \quad (18)$$

$$= \ln B(\beta) - \ln B(\alpha) + \sum_{j=1}^{M} (\alpha_j - \beta_j) \int_{\Theta} \ln \theta_j \rho(\theta) d\theta$$

$$= \sum_{j=1}^{M} \ln(\Gamma(\beta_j)) - \ln(\Gamma(\beta_0)) + \ln(\Gamma(\alpha_0)) - \sum_{j=1}^{M} \ln(\Gamma(\alpha_j))$$

$$+ \sum_{j=1}^{M} (\alpha_j - \beta_j)(\psi(\alpha_j) - \psi(\alpha_0)). \quad (20)$$
Figure 8: Oracle bound on the expected MV.

Figure 7: Functions of the gamma family.

Equation (20) follows by definition of Dirichlet’s geometric mean:
\[ \int \ln \theta_j \rho(\theta) d\theta = \psi(\alpha_j) - \psi(\alpha_0). \]

\[ \square \]

A.4 Partial derivatives for gradient-based optimization

When optimizing our objective functions by gradient descent, we make use of the following partial derivatives w.r.t. the Dirichlet parameters \( \{\alpha_j\}_{j=1}^M \).

The partial derivatives of the KL divergence are
\[ \nabla_{\alpha_j} KL(\rho, \pi) = \psi(\alpha_0) - \psi(\alpha_j) + \psi(\alpha_j) - \psi(\alpha_0) + (\alpha_j - \beta_j)(\psi'(\alpha_j) - \psi'(\alpha_0)) \]  
\[ = (\alpha_j - \beta_j)(\psi'(\alpha_j) - \psi'(\alpha_0)). \]  

Equation (20) follows by definition of Dirichlet’s geometric mean:
\[ \int \ln \theta_j \rho(\theta) d\theta = \psi(\alpha_j) - \psi(\alpha_0). \]

For a given data point \((x, y) \sim P\), recall that \( w = \{j \mid h_j(x) \neq y\} \) is the set of indices of the base classifiers that misclassify \((x, y)\) and \( c = \{j \mid h_j(x) = y\} \) is the set of indices of the base classifiers that correctly classify \((x, y)\). Let us define \( w(\alpha) = \sum_{j \in w} \alpha_j \) and \( c(\alpha) = \sum_{j \in c} \alpha_j = \alpha_0 - w(\alpha) \). For any \( j \in c \) we have
\[ \nabla_{\alpha_j} I_{0.5}(c(\alpha), w(\alpha)) = \]  
\[ \left( \ln 0.5 - \psi(c(\alpha)) + \psi(\alpha_0) \right) I_{0.5}(c(\alpha), w(\alpha)) \]  
\[ - 0.5^c(\alpha) \frac{\Gamma(c(\alpha))\Gamma(\alpha_0)}{\Gamma(w(\alpha))} \cdot 3F_2(c(\alpha), c(\alpha), 1 - w(\alpha); c(\alpha) + 1, c(\alpha) + 1; 0.5) \]  
\[ ) \]  
\[ (23) \]

and for any \( j \in w \) we have
\[ \nabla_{\alpha_j} I_{0.5}(c(\alpha), w(\alpha)) = \]  
\[ \left( - \ln 0.5 + \psi(w(\alpha)) - \psi(\alpha_0) \right) I_{0.5}(c(\alpha), w(\alpha)) \]  
\[ + 0.5^w(\alpha) \frac{\Gamma(w(\alpha))\Gamma(\alpha_0)}{\Gamma(c(\alpha))} \cdot 3F_2(w(\alpha), w(\alpha), 1 - c(\alpha); w(\alpha) + 1, w(\alpha) + 1; 0.5) \]  
\[ ) \]  
\[ (24) \]

where \( 3F_2(a, b, c; d, e; z) \) is the generalized hyper-geometric function:
\[ 3F_2(a, b, c; d, e; z) = \sum_{t=1}^{T} \frac{(a)^t (b)^t (c)^t z^t}{(d)^t (e)^t t!} \]
We now derive an oracle bound for the risk of the expected majority vote, given a distribution $\alpha$ with vector $p$. Equation (26) follows by linearity of $p$.

For a given $(x, y) \sim P$, let us define $w(x, y) = \{j|\hat{h}_j(x) \neq y\}$ the set of indices of the base classifiers that misclassify $(x, y)$ and $c(x, y) = \{j|\hat{h}_j(x) = y\}$ be the set of indices of the base classifiers that correctly classify $(x, y)$. The risk of the expected MV can then be measured as follows:

$$R(f_{\hat{\theta}}) = \mathbb{P}(W_{\hat{\theta}} \geq 0.5) = \mathbb{P} \left[ \sum_{j} \frac{\alpha_j}{\alpha_0} \geq 0.5 \right]$$

Equation (26) follows by linearity of $W_{\hat{\theta}}$.

We recall that the expected risk is given by

$$\mathbb{E}_{(x, y) \sim P} I_{0.5} \left( \sum_{j \in w(x, y)} \alpha_j, \sum_{j \in c(x, y)} \alpha_j \right) = \mathbb{E}_{(x, y) \sim P} I_{0.5} \left( \alpha_0 - \alpha_{w(x, y)}, \alpha_{w(x, y)} \right)$$

with $\alpha_{w(x, y)} = \sum_{j \in w(x, y)} \alpha_j$. As $I_{0.5}(\cdot)$ is monotonically increasing in its second argument and $I_{0.5}(\frac{\alpha_0}{2}, \frac{\alpha_0}{2}) = 0.5$, we can then relate the two risks:

$$R(f_{\hat{\theta}}) \leq 2\mathbb{E}_{\rho} R(f_{\theta}).$$

See Figure 8 for an illustration.

### A.5 Oracle bound on expected Majority Vote

We now derive an oracle bound for the risk of the expected majority vote, given a distribution $\rho(\theta)$ having a Dirichlet form and concentration vector $\alpha = \{\alpha_j\}_{j=1}^M$. The oracle bound can then be used to derive empirical PAC-Bayesian bounds on the true risk of the expected MV. The expected MV is parameterized by the mean weighting vector $\hat{\theta}$, which for a Dirichlet distribution is given by

$$\hat{\theta} = \mathbb{E}_{\rho, \theta} = \frac{\alpha}{\alpha_0} \quad \text{with} \quad \alpha_0 = \sum_{j=1}^{M} \alpha_j.$$
Theorem 2 (Seeger’s bound with informed priors). Let $\pi_{\leq m}$ and $\rho_{\leq m}$ be the prior and posterior distributions on $\Theta_{\leq m}$ and $\theta_{> m}$ and $\rho_{> m}$ the prior and posterior distributions on $\Theta_{> m}$. For any $p \in (0, 1)$ and $\delta \in (0, 1)$ with probability at least $1 - \delta$ over samples $S = \{(x_i, y_i) \sim P\}_{i=1}^n$ of size $n$ we have simultaneously:

$$
\text{kl}
\left(
\frac{p(R(\rho_{> m}) + (1 - p)R(\rho_{\leq m}))}{(1 - p)R(\rho_{\leq m})} \right)
\leq
\frac{p KL(\rho_{> m}, \pi_{> m})}{n - m} + \frac{\ln 4^{\sqrt{n(m - m)}}}{n}
$$

with $R(\rho_{> m}) = \int_{\Theta_{> m}} R(f_\theta)p(d\theta)$, and $R(\rho_{\leq m}) = \int_{\Theta_{\leq m}} R(f_\theta)p(d\theta)$, and $\hat{R}_{\leq m}(\rho_{> m}) = \int_{\Theta_{\leq m}} \hat{R}(f_\theta)p(d\theta)$.

Proof. From the joint convexity of the binary kl divergence, we have for any $p \in [0, 1]$

$$
\text{kl}
\left(
\frac{pR(\rho_{> m}) + (1 - p)R(\rho_{\leq m})}{p\hat{R}_{\leq m}(\rho_{> m}) + (1 - p)\hat{R}_{> m}(\rho_{\leq m})}
\right)
\leq
\frac{p KL(\rho_{> m}, \pi_{> m})}{n - m} + \frac{\ln 4^{\sqrt{n(m - m)}}}{n}
$$

For $\forall \rho_{> m}$ defined on $\Theta_{> m}$:

$$
\mathbb{P}_{S_{\leq m} \sim P_{m}} \left[
\frac{p \text{kl}
\left(
\frac{R(\rho_{> m})}{\hat{R}_{\leq m}(\rho_{> m})}
\right)
\leq
\frac{p \frac{KL(\rho_{> m}, \pi_{> m})}{m} + \frac{\ln 4^{\sqrt{n(m - m)}}}{m}}{1 - \frac{\delta}{2}}
\right]
\geq
1 - \frac{\delta}{2};
\right)
\right)
$$

and for $\forall \rho_{\leq m}$ defined on $\Theta_{\leq m}$:

$$
\mathbb{P}_{S_{> m} \sim P_{n-m}} \left[
\frac{(1 - p) \text{kl}
\left(
\frac{R(\rho_{\leq m})}{\hat{R}_{> m}(\rho_{\leq m})}
\right)
\leq
\frac{(1 - p) \frac{KL(\rho_{\leq m}, \pi_{\leq m})}{n - m} + \frac{\ln 4^{\sqrt{n(m - m)}}}{n - m}}{1 - \frac{\delta}{2}}
\right]
\geq
1 - \frac{\delta}{2};
\right)
$$

Combining Equation (30) and Equation (31) using the union bound, we obtain the desired result with $1 - \delta$ probability.

C Additional experimental results

C.1 Analysis of the role of bound regularization on the posterior

Synthetic dataset We study the behavior of the method during training on a simple toyset, built from two normal distributions $\mathcal{N}(1([-1, 0], \text{diag}([0.1, 1]))$ and $\mathcal{N}_2([1, 0], \text{diag}([0.1, 1]))$, one per class, so that the two classes are almost perfectly separable on the first dimension. From each class distribution for training, $n$ points are drawn, and 1,000 points for testing.

Base predictors and prior We fix the set of base voters to $M=4$ axis-aligned decision stumps: 2 per class and centered in 0 on each dimension, so that the problem is well specified as the optimal classifier is in the predictor set. Then, we set the parameters $\beta$ of the prior distribution all to 0.1 to encourage sparse solutions for the posterior, and we initialize the prior’s parameters by drawing $\alpha_j \sim \mathcal{U}(0.01, 2)$.
Study of the impact of optimizing the PAC-Bayesian bound  In Figure 9, we compare the evolution of $\|\alpha\|_2$ (left), its gradient (center) and $KL(\alpha, \beta)$ (right) during training using the exact method. In each plot we compare the values obtained when optimizing Seeger’s Bound (Equation (1)) with those obtained when optimizing only the empirical risk. 

Figure 10: Comparison of exact and MC variants as a function of the number of training points $n$, for $M = 16$ and different number of MC draws $T$.

C.2 Results on two-moons dataset

Comparison of exact and MC variants  In Figure 10 we report a second comparison of exact and MC variants in terms of error, generalization bound and training time (in seconds). This time we fix the number of base classifiers $M$ and vary the size of the training set $n$. We first remark that $T = 100$ samples for $MC$ are consistently enough to recover exact’s performance, while allowing for significantly faster optimization. In general, as $MC$ is easily parallelizable, its training times have a better dependence on $n$.

Full comparison with baselines  We report the full comparison of FO, SO, Rnd and our method on the binary classification two-moons dataset. We report in Figure 11 the results obtained after optimizing the respective PAC-Bayesian bounds. Notice that our bounds are tighter than the baselines, for any value of $M$: this is principally due to the fact that our method consistently obtains lower error rates, being able to
better fit the training set (as shown in Figure 3), but also to the fact that our bound is not based on an oracle upper bound on the true risk, unlike the baselines.

Sensitivity to noise We observe that on two-moons our method outperforms the baselines also in terms of error rates. However, this does not seem to be the case on real benchmarks where its test errors are generally aligned with those of the baselines. We conduct an additional experiment to study whether this phenomenon is due to sensitivity to noise, such as input noise. Indeed our learning algorithm optimizes the 01-loss, which does not distinguish points with margins close or far from 0.5 because of its discontinuity in $W_{\theta}$. In Figure 12, we assess training and test errors, and Seeger’s bound values with increasing input noise. For this experiment, we generated training and test sets as before, and with Gaussian noise $\mathcal{N}(0, \sigma^2)$ added to the inputs. As expected, all methods degrade with increasing noise. In particular, the test errors of exact and Rnd worsen the fastest and the benefits of using them vanish from $\sigma^2 > 0.35$.

C.3 Additional results on real benchmarks

Dataset descriptions We consider several classification datasets from UCI [Dua and Graff 2017], LIB-SVM [https://www.csie.ntu.edu.tw/~cjlin/libsvm/], and Zalando [Xiao et al. 2017], of different number of features and of points:

- 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);
At each run of an algorithm, we randomly split a dataset in training and test sets of sizes 80% – 20% respectively, apart from Svmguide1, Pendigits, Protein, Shuttle, MNIST and Fashion-MNIST for which we keep the original train-test splits. Note that we do not make use of a validation set, as we use the PAC-Bayesian bounds as estimate of the test error for model selection. Finally, we convert all categorical features to numerical using an ordinal encoder and z-score all features using the statistics of the training set.

**Choice of prior** In all the other experiments, we fixed the prior distribution (parameterized by $[\beta_j]_{j=1}^{M}$) to the uniform, i.e. $\beta_j = 1$, $\forall j$. This choice was to make the comparison with the baselines as fair as possible, as their prior was also fixed to the uniform (categorical). However, we can bias the sparsity of the posterior, or conversely its concentration, by choosing a different value for the prior distribution parameters. In some cases, tuning the prior parameters allows to obtain better performance, as reported in Figure 13. In particular, on Protein encouraging sparser solutions generally provides better results, confirmed by the fact that the best baseline on this dataset, FO, is known to output sparse solutions. On the contrary, on datasets where methods accounting for voter correlation outperform FO, such as on MNIST, encouraging solutions to be concentrated and close to the simplex mean yields better performance. In general, these results suggest that the choice of prior distribution has a high impact on the learned model’s performance and tuning its concentration parameters would be a viable option for improving results.

**C.4 Impact of voter strength**

We report the complete study on the impact of voter strength on the learned models. More precisely we provide results for additional datasets as well as the study of the expected strength of a voter as a function of the tree maximal depth. Recall that as hypothesis set, we learn a Random Forest with 100 decision trees for which we bound the maximal depth between 1 and 10. In Figure 14 we can see that limiting the maximal depth is an effective way for controlling the strength of the voters, measured as the expected accuracy of a random voter. Apart from Protein, where decision trees do not seem to be a good choice of base predictor, increasing the strength of the voters generally yields more powerful ensembles for all methods. Our method
has error rates comparable with the best baselines and enjoys tight and non-vacuous generalization guarantees for any tree depth. Finally, by comparing SO’s training and test errors we notice that this method tends to overfit the dataset especially when the base classifiers are weaker (tree depth close to 1).
Figure 13: Study of impact of prior on posterior’s performance. We fix all $M$ prior’s parameters to $\beta$, on the x axis: the smaller $\beta$, the sparser the posterior is encouraged to be. We plot average and standard deviations over 4 trials.

C.5 Summary of main results

We finally report the detailed comparison on real benchmarks in Figure [15] and Tables [1] and [2].
Figure 14: Comparison voter strength (1st column), training error (2nd column), test error (3rd column) and Seeger’s bound (4th column) as a function of the tree maximal depth. We mark with a red horizontal line the threshold above which the bounds are vacuous. Results are averaged over 4 trials.
Figure 15: Comparison of First Order (FO), Second Order (SO) and Randomized MV (Rnd) and our methods (exact, MC) in terms of training and test error rates and PAC-Bayesian bound values. Each row of subfigures corresponds to a dataset, where we marked its number of features $d$ and number of training instances $n$. The dashed horizontal line in the rightmost column plots marks the threshold above which the bounds are vacuous. We report the means (bars) and standard deviations (vertical, black lines) over 10 different runs.
Table 1: Summary of the main results on binary datasets with data agnostic prior. For the PAC-Bayesian methods, we report the performance of the model obtained after optimizing Seeger’s Bound (Theorem 1). Bayesian Naive Bayes (BayesianNB) corresponds to the weighting strategy proposed in Berend and Kontorovich [2015]. For each method we report the average (± standard deviations) of the test errors and generalization bounds (for the PAC-Bayesian methods) over 10 runs. We bold the results (test error and bound value) that are significantly better (smaller) than the other results for a dataset. Notice that the PAC-Bayesian methods generally improve upon BayesianNB and that our method consistently provides the tightest and non-vacuous generalization bounds over all the datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>HABERMAN</th>
<th>TICTACTOE</th>
<th>SVMGUIDE</th>
<th>MUSHROOMS</th>
<th>PHISHING</th>
<th>CODRNA</th>
<th>ADULT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BayesianNB</td>
<td>25.65 ± 2.10</td>
<td>30.21 ± 3.06</td>
<td>34.51 ± 0.00</td>
<td>11.65 ± 0.59</td>
<td>44.49 ± 0.05</td>
<td>66.67 ± 0.00</td>
<td>24.07 ± 0.00</td>
</tr>
<tr>
<td>Fist Order</td>
<td>25.48 ± 3.37</td>
<td>29.22 ± 2.77</td>
<td>6.83 ± 0.00</td>
<td>4.60 ± 0.36</td>
<td>11.26 ± 0.58</td>
<td>23.52 ± 0.35</td>
<td>22.14 ± 0.36</td>
</tr>
<tr>
<td>Second Order</td>
<td>26.45 ± 2.99</td>
<td>30.52 ± 3.55</td>
<td>6.83 ± 0.00</td>
<td>4.60 ± 0.36</td>
<td>10.10 ± 0.73</td>
<td>22.93 ± 1.43</td>
<td><strong>17.02 ± 0.29</strong></td>
</tr>
<tr>
<td>RandomizedMV</td>
<td>26.45 ± 2.19</td>
<td>27.76 ± 3.71</td>
<td><strong>6.04 ± 0.32</strong></td>
<td><strong>1.08 ± 0.22</strong></td>
<td><strong>6.84 ± 0.41</strong></td>
<td>12.41 ± 0.23</td>
<td><strong>16.24 ± 0.37</strong></td>
</tr>
<tr>
<td>ours-exact</td>
<td>29.84 ± 7.34</td>
<td>30.91 ± 2.10</td>
<td><strong>6.11 ± 0.25</strong></td>
<td><strong>1.30 ± 0.35</strong></td>
<td>8.12 ± 0.35</td>
<td><strong>11.77 ± 0.16</strong></td>
<td>23.49 ± 1.99</td>
</tr>
<tr>
<td>ours-MC</td>
<td>30.10 ± 7.21</td>
<td>30.84 ± 2.38</td>
<td><strong>6.27 ± 0.40</strong></td>
<td><strong>1.23 ± 0.23</strong></td>
<td>7.74 ± 0.42</td>
<td>13.13 ± 3.43</td>
<td>24.07 ± 0.00</td>
</tr>
<tr>
<td></td>
<td>Seeger’s b.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>75.24 ± 1.78</td>
<td>76.48 ± 1.48</td>
<td>18.62 ± 0.00</td>
<td>12.31 ± 0.22</td>
<td>25.75 ± 0.32</td>
<td>49.45 ± 0.47</td>
<td>46.93 ± 0.19</td>
</tr>
<tr>
<td>Second Order</td>
<td>110.60 ± 1.65</td>
<td>100.25 ± 0.99</td>
<td>31.06 ± 0.36</td>
<td>21.27 ± 0.31</td>
<td>35.67 ± 0.21</td>
<td>60.59 ± 0.20</td>
<td>53.33 ± 0.09</td>
</tr>
<tr>
<td>RandomizedMV</td>
<td>89.19 ± 10.59</td>
<td>79.55 ± 0.90</td>
<td>26.37 ± 1.34</td>
<td>11.49 ± 0.27</td>
<td>24.64 ± 0.20</td>
<td>36.24 ± 0.13</td>
<td>41.06 ± 0.13</td>
</tr>
<tr>
<td>ours-exact</td>
<td><strong>46.97 ± 17.69</strong></td>
<td><strong>42.49 ± 0.53</strong></td>
<td><strong>11.38 ± 0.61</strong></td>
<td><strong>4.77 ± 0.26</strong></td>
<td><strong>13.50 ± 0.19</strong></td>
<td><strong>15.16 ± 0.17</strong></td>
<td><strong>25.42 ± 1.07</strong></td>
</tr>
<tr>
<td>ours-MC</td>
<td><strong>47.34 ± 17.57</strong></td>
<td><strong>43.07 ± 0.72</strong></td>
<td><strong>11.79 ± 0.68</strong></td>
<td><strong>5.42 ± 0.20</strong></td>
<td>14.62 ± 0.28</td>
<td><strong>16.48 ± 2.86</strong></td>
<td><strong>26.46 ± 0.09</strong></td>
</tr>
</tbody>
</table>
Table 2: Summary of the main results on multi-class datasets with informed priors. For the PAC-Bayesian methods, we report the performance of the model obtained after optimizing Seeger’s Bound with informed priors (Theorem 2). Bayesian Naive Bayes (BayesianNB) corresponds to the weighting strategy proposed in Berend and Kontorovich [2015]. For each method we report the average (± standard deviations) of the test errors and generalization bounds (for the PAC-Bayesian methods) over 10 runs. We bold the results (test error and bound value) that are significantly better (smaller) than the other results for a dataset. Notice that the PAC-Bayesian methods generally improve upon BayesianNB and that our method consistently provides the tightest and non-vacuous generalization bounds over all the datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>PENDIGITS</th>
<th>PROTEIN</th>
<th>SHUTTLE</th>
<th>SENSORLESS</th>
<th>MNIST</th>
<th>FASHION-MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BayesianNB</td>
<td>82.41 ± 0.11</td>
<td>82.09 ± 0.00</td>
<td>20.90 ± 0.03</td>
<td>86.27 ± 0.21</td>
<td>86.09 ± 0.09</td>
<td>89.14 ± 0.05</td>
</tr>
<tr>
<td>Fist Order</td>
<td>13.01 ± 1.12</td>
<td>54.51 ± 0.69</td>
<td>0.08 ± 0.01</td>
<td>2.94 ± 0.12</td>
<td>21.45 ± 0.17</td>
<td>25.94 ± 0.22</td>
</tr>
<tr>
<td>Second Order</td>
<td>7.59 ± 0.28</td>
<td>64.15 ± 0.39</td>
<td><strong>0.05 ± 0.01</strong></td>
<td>0.41 ± 0.06</td>
<td><strong>13.50 ± 0.08</strong></td>
<td>21.67 ± 0.14</td>
</tr>
<tr>
<td>RandomizedMV</td>
<td>7.86 ± 0.19</td>
<td>55.24 ± 0.47</td>
<td>0.07 ± 0.00</td>
<td>0.43 ± 0.05</td>
<td><strong>13.62 ± 0.11</strong></td>
<td>21.80 ± 0.07</td>
</tr>
<tr>
<td>ours-exact</td>
<td>8.04 ± 0.23</td>
<td>55.00 ± 0.78</td>
<td>0.07 ± 0.01</td>
<td>0.46 ± 0.05</td>
<td>13.84 ± 0.08</td>
<td>21.90 ± 0.05</td>
</tr>
<tr>
<td>ours-MC</td>
<td>8.05 ± 0.23</td>
<td>54.74 ± 0.80</td>
<td>0.07 ± 0.01</td>
<td>0.46 ± 0.05</td>
<td>13.84 ± 0.09</td>
<td>21.90 ± 0.05</td>
</tr>
<tr>
<td>Seeger’s b.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fist Order</td>
<td>20.65 ± 0.45</td>
<td>112.29 ± 0.46</td>
<td>0.32 ± 0.02</td>
<td>6.80 ± 0.26</td>
<td>46.18 ± 0.37</td>
<td>52.12 ± 0.16</td>
</tr>
<tr>
<td>Second Order</td>
<td>16.38 ± 0.20</td>
<td>139.81 ± 0.23</td>
<td>0.42 ± 0.01</td>
<td>3.82 ± 0.06</td>
<td>44.24 ± 0.12</td>
<td>57.09 ± 0.08</td>
</tr>
<tr>
<td>RandomizedMV</td>
<td>8.32 ± 0.16</td>
<td>122.92 ± 0.41</td>
<td>0.28 ± 0.01</td>
<td>1.29 ± 0.05</td>
<td>30.46 ± 0.14</td>
<td>42.93 ± 0.12</td>
</tr>
<tr>
<td>ours-exact</td>
<td><strong>4.34 ± 0.11</strong></td>
<td><strong>59.28 ± 0.60</strong></td>
<td><strong>0.15 ± 0.01</strong></td>
<td><strong>0.67 ± 0.03</strong></td>
<td><strong>15.57 ± 0.07</strong></td>
<td><strong>21.83 ± 0.06</strong></td>
</tr>
<tr>
<td>ours-MC</td>
<td><strong>4.35 ± 0.11</strong></td>
<td><strong>59.25 ± 0.63</strong></td>
<td><strong>0.15 ± 0.01</strong></td>
<td><strong>0.67 ± 0.03</strong></td>
<td><strong>15.49 ± 0.17</strong></td>
<td><strong>21.83 ± 0.06</strong></td>
</tr>
</tbody>
</table>