A Strongly Quasiconvex PAC-Bayesian Bound

Niklas Thiemann, Christian Igel, Olivier Wintenberger, Yevgeny Seldin

Journal of Machine Learning Research 1–26, 2017 Algorithmic Learning Theory 2017

A Strongly Quasiconvex PAC-Bayesian Bound

Niklas Thiemann NIKLASTHIEMANN@GMAIL.COM
Department of Computer Science, University of Copenhagen

Christian Igel IGEL@DI.KU.DK
Department of Computer Science, University of Copenhagen

Olivier Wintenberger OLIVIER.WINTENBERGER@UPMC.FR
LSTA, Sorbonne Universités, UPMC Université Paris 06

Yevgeny Seldin SELDIN@DI.KU.DK
Department of Computer Science, University of Copenhagen

Editors: Steve Hanneke and Lev Reyzin

Abstract

We propose a new PAC-Bayesian bound and a way of constructing a hypothesis space, so that the bound is convex in the posterior distribution and also convex in a trade-off parameter between empirical performance of the posterior distribution and its complexity. The complexity is measured by the Kullback-Leibler divergence to a prior. We derive an alternating procedure for minimizing the bound. We show that the bound can be rewritten as a one-dimensional function of the trade-off parameter and provide sufficient conditions under which the function has a single global minimum. When the conditions are satisfied the alternating minimization is guaranteed to converge to the global minimum of the bound. We provide experimental results demonstrating that rigorous minimization of the bound is competitive with cross-validation in tuning the trade-off between complexity and empirical performance. In all our experiments the trade-off turned to be quasiconvex even when the sufficient conditions were violated.

1. Introduction

PAC-Bayesian analysis, where PAC stands for the Probably Approximately Correct frequentist learning model (Valiant, 1984), analyzes prediction accuracy of *randomized classifiers*. A randomized classifier is a classifier defined by a distribution ρ over a hypothesis class $\mathcal{H}$. A randomized classifier predicts by drawing a hypothesis from $\mathcal{H}$ according to ρ and applying it to make the prediction (McAllester, 1998). In many applications randomized prediction is replaced by a ρ-weighted majority vote (Germain et al., 2009).

PAC-Bayesian analysis provides some of the tightest generalization bounds in statistical learning theory (Germain et al., 2009). PAC-Bayesian bounds have a form of a trade-off between empirical performance of ρ and its complexity, measured by the Kullback-Leibler (KL) divergence (a.k.a. relative entropy) between ρ and a prior distribution π. Most of PAC-Bayesian literature relies on cross-validation to tune the trade-off. Cross-validation is an extremely powerful and practical heuristic for selecting model parameters, but it can potentially be misleading (Kearns et al., 1997; Kearns and Ron, 1999). It is also computationally expensive, especially for computationally demanding models, such as

kernel SVMs, since it requires training a large number of classifiers on almost the whole dataset. Derivation of theoretical results that would not require parameter cross-validation is a long-standing challenge for theoretical machine learning (Langford, 2005).

The need to rely on cross-validation stems from several reasons:

- Not all of the existing PAC-Bayesian bounds are convex in the posterior distribution ρ. For example, the most widely used PAC-Bayes-kl bound due to Seeger (2002) is non-convex. This makes it hard to minimize the bound with respect to the posterior distribution. In most papers the bound is replaced by a linear trade-off between empirical error and the KL divergence and the trade-off parameter is tuned by cross-validation.

  While it is possible to achieve convexity in the posterior distribution ρ by introducing an additional trade-off parameter (Catoni, 2007; Keshet et al., 2011), we are unaware of successful attempts to tune the additional trade-off parameter through rigorous bound minimization. In practice, the alternative bounds are replaced by the same linear trade-off mentioned above and tuned by cross-validation.

- The second obstacle is that, in order to keep the KL divergence between the posterior and the prior tractable, the set of posterior and prior distributions is often restricted. A popular example are Gaussian posteriors and Gaussian priors (Langford and Shawe-Taylor, 2002; McAllester, 2003; Langford, 2005). Even if the initial bound is convex in the posterior distribution, the convexity may be broken by such a restriction or reparametrization, as it happens in the Gaussian case (Germain et al., 2009).

- Even though PAC-Bayesian bounds are some of the tightest, we are unaware of examples, where their tightness is sufficient to compete with cross-validation in tuning the trade-off between complexity and empirical performance.

We propose a relaxation of Seeger's PAC-Bayes-kl inequality, which we name *PAC-Bayes-λ inequality* or *PAC-Bayes-λ bound* when referring to the right hand side of the inequality. The bound is convex in the posterior distribution ρ and has a convex trade-off between the empirical loss and KL divergence. The inequality is similar in spirit to the one proposed by Keshet et al. (2011), but it does not restrict the form of ρ and π. We provide an alternating procedure for minimizing the bound. We show that the bound can be rewritten as a continuous one-dimensional function of the trade-off parameter λ and that under certain sufficient conditions this function is strongly quasiconvex (it has a single global minimum and no other stationary points). This guarantees convergence of alternating minimization to the global optimum.

For infinite hypothesis spaces alternating minimization can be computationally intractable or require parametrization, which can break the convexity of the bound in the posterior distribution. We get around this difficulty by constructing a finite data-dependent hypothesis space. The hypothesis space is constructed by taking m subsamples of size r from the training data. Each subsample is used to train a weak classifier, which is then validated on the remaining $n-r$ points, where n is the sample size. We adapt our PAC-Bayesian bound and minimization procedure to this setting. Our analysis and minimization procedure work for any m, r, and any split of the data, including overlaps between training sets and overlaps

between validation sets. In particular, it can also be applied to aggregate models originating from a cross-validation split of the data. However, in cross-validation the training sets are typically large (of order n) and validation sets and the number of models are small. While the prediction accuracy is still competitive in this setting, the highest computational advantage from our approach is achieved when the relation is reversed and the training size r is taken to be small, roughly of order d, where d is the number of features, and the number of models m is taken to be large, roughly of order n. The construction of hypothesis space can be seen as sample compression (Laviolette and Marchand, 2007). However, unlike the common approach to sample compression, which considers all possible subsamples of a given size and thus computationally and statistically inefficient, we consider only a small subset of possible subsamples.

We provide experimental results on several UCI datasets showing that the prediction accuracy of our learning procedure (training m weak classifiers and weighting their predictions through minimization of the PAC-Bayes-λ bound) is comparable to prediction accuracy of kernel SVMs tuned by cross-validation. In addition, we show that when r is considerably smaller than n and m is of order n, the comparable prediction accuracy is achieved at a much lower computation cost. The computational speed-up is achieved because of the super-quadratic training time of kernel SVMs, which makes it much faster to train many weak SVMs on small training sets than one powerful SVM on a big training set.

In the following, we provide a brief review of PAC-Bayesian analysis, then present the PAC-Bayesian bound and its minimization procedure in Section 3, derive conditions for convergence of minimization procedure to the global minimum in Section 4, describe our construction of a hypothesis space and specialize our results to this construction in Section 5, and provide experimental validation in Section 6.

2. A Brief Review of PAC-Bayesian Analysis

To set the scene we start with a brief review of PAC-Bayesian analysis.

Notations

We consider a supervised learning setting with an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. We let $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ denote an independent identically distributed (i.i.d.) sample of size n drawn according to an unknown distribution $\mathcal{D}(X, Y)$. A hypothesis h is a function from the input to the output space $h : \mathcal{X} \to \mathcal{Y}$. We use $\mathcal{H}$ to denote a hypothesis class. We let $\ell : \mathcal{Y}^2 \to [0, 1]$ denote a bounded loss function. The loss of h on a sample (X, Y) is $\ell(h(X), Y)$ and the expected loss of h is denoted by $L(h) = \mathbb{E}\left[\ell(h(X), Y)\right]$. We use $\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$ to denote the empirical loss of h on S.

A randomized prediction rule parametrized by a distribution ρ over $\mathcal{H}$ is defined in the following way. For each prediction on a sample point X the rule draws a new hypothesis $h \in \mathcal{H}$ according to ρ and applies it to X. The expected loss of such prediction rule is $\mathbb{E}_{h\sim\rho}\left[L(h)\right]$ and the empirical loss is $\mathbb{E}_{h\sim\rho}\left[\hat{L}(h, S)\right]$. We use $\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{h\sim\rho}\left[\ln \frac{\rho(h)}{\pi(h)}\right]$ to denote the KL divergence between ρ and π. For Bernoulli distributions with biases p and q we use $\mathrm{kl}(p\|q)$ as a shorthand for $\mathrm{KL}([p, 1-p]\|[q, 1-q])$, the KL divergence between the

two distributions. Finally, we use $\mathbb{E}_\rho[\cdot]$ as a shorthand for $\mathbb{E}_{h\sim\rho}[\cdot]$ and $\mathbb{E}_S[\cdot]$ as a shorthand for $\mathbb{E}_{S\sim\mathcal{D}^n}[\cdot]$.

Change of Measure Inequality

The majority of PAC-Bayesian bounds are based on the following lemma.

Lemma 1 (Change of Measure Inequality) *For any function $f : \mathcal{H} \times (\mathcal{X}\times\mathcal{Y})^n \to \mathbb{R}$ and for any distribution π over $\mathcal{H}$, such that π is independent of S, with probability greater than $1-\delta$ over a random draw of S, for all distributions ρ over $\mathcal{H}$ simultaneously:*

$$\mathbb{E}_{h\sim\rho}[f(h,S)] \le \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln \mathbb{E}_{h\sim\pi}\left[\mathbb{E}_{S'}\left[e^{f(h,S')}\right]\right]. \tag{1}$$

The lemma is based on Donsker-Varadhan's variational definition of the KL divergence (Donsker and Varadhan, 1975), by which $\mathrm{KL}(\rho\|\pi) = \sup_f\left\{\mathbb{E}_\rho[f(h)] + \ln \mathbb{E}_\pi\left[e^{f(h)}\right]\right\}$, where the supremum is over all measurable functions $f : \mathcal{H} \to \mathbb{R}$. In the lemma, f is extended to be a function of h and S and then Markov's inequality is used to bound the expectation with respect to π by $\mathbb{E}_\pi\left[e^{f(h,S)}\right] \le \frac{1}{\delta}\mathbb{E}_{S'}\left[\mathbb{E}_\pi\left[e^{f(h,S')}\right]\right]$ with probability at least $1-\delta$. Independence of π and S allows to exchange the order of expectations, leading to the statement of the lemma. For a formal proof we refer to Tolstikhin and Seldin (2013).

PAC-Bayes-kl Inequality

Various choices of the function f in Lemma 1 lead to various forms of PAC-Bayesian bounds (Seldin et al., 2012). The classical choice is $f(h,S) = n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))$. The moment generating function of f can be bounded in this case by $\mathbb{E}_S\left[e^{f(h,S)}\right] \le 2\sqrt{n}$ (Maurer, 2004; Germain et al., 2015). This bound is used to control the last term in equation (1), leading to the PAC-Bayes-kl inequality (Seeger, 2002).

Theorem 2 (PAC-Bayes-kl Inequality) *For any probability distribution π over $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability greater than $1-\delta$ over a random draw of a sample S, for all distributions ρ over $\mathcal{H}$ simultaneously:*

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

3. PAC-Bayes-λ inequality and its Alternating Minimization

By inversion of the kl with respect to its second argument, inequality (2) provides a bound on $\mathbb{E}_\rho[L(h)]$. However, this bound is not convex in ρ and, therefore, inconvenient for minimization. We introduce a relaxed form of the inequality, which has an additional trade-off parameter λ. The inequality leads to a bound, which is convex in ρ for a fixed λ and convex in λ for a fixed ρ, making it amenable to alternating minimization. Theorem 3 is analogous to Keshet et al. (2011, Theorem 1) and a similar result can also be derived by using the techniques from Tolstikhin and Seldin (2013), as shown by Thiemann (2016).

Theorem 3 (PAC-Bayes-λ Inequality) *For any probability distribution π over $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability greater than $1-\delta$ over a random draw of a sample S, for all distributions ρ over $\mathcal{H}$ and $\lambda \in (0,2)$ simultaneously:*

$$\mathbb{E}_\rho\left[L(h)\right] \leq \frac{\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)n}. \tag{3}$$

We emphasize that the theorem holds for *all* values of $\lambda \in (0,2)$ simultaneously. This is in contrast to some other parametrized PAC-Bayesian bounds, for example, the one proposed by Catoni (2007), which hold for a *fixed* value of a trade-off parameter.

Proof We use the following analog of Pinsker's inequality (Marton, 1996, 1997; Samson, 2000; Boucheron et al., 2013, Lemma 8.4): for $p < q$

$$\mathrm{kl}(p\|q) \geq (q-p)^2/(2q). \tag{4}$$

By application of inequality (4), inequality (2) can be relaxed to

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

(McAllester, 2003). By using the inequality $\sqrt{xy} \leq \frac{1}{2}\left(\lambda x + \frac{y}{\lambda}\right)$ for all $\lambda > 0$, we have that with probability at least $1-\delta$ for all ρ and $\lambda > 0$

$$\mathbb{E}_\rho\left[L(h)\right] - \mathbb{E}_\rho\left[\hat{L}(h,S)\right] \leq \frac{\lambda}{2}\mathbb{E}_\rho\left[L(h)\right] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{\lambda n} \tag{6}$$

(Keshet et al., 2011). By changing sides

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

For $\lambda < 2$ we can divide both sides by $\left(1-\frac{\lambda}{2}\right)$ and obtain the theorem statement. ■

Since $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ is linear in ρ and $\mathrm{KL}(\rho\|\pi)$ is convex in ρ (Cover and Thomas, 2006), for a fixed λ the right hand side of inequality (3) is convex in ρ and the minimum is achieved by

$$\rho_\lambda(h) = \frac{\pi(h)e^{-\lambda n\hat{L}(h,S)}}{\mathbb{E}_\pi\left[e^{-\lambda n\hat{L}(h',S)}\right]}, \tag{7}$$

where $\mathbb{E}_\pi\left[e^{-\lambda n\hat{L}(h',S)}\right]$, a shorthand for $\mathbb{E}_{h'\sim\pi}\left[e^{-\lambda n\hat{L}(h',S)}\right]$, is a convenient way of writing the normalization factor, which covers continuous and discrete hypothesis spaces in a unified notation. Furthermore, for $t \in (0,1)$ and $c_1, c_2 \geq 0$ the function $\frac{c_1}{1-t} + \frac{c_2}{t(1-t)}$ is convex in t

(Tolstikhin and Seldin, 2013). Therefore, for a fixed ρ the right hand side of inequality (3) is convex in λ for $\lambda \in (0, 2)$. The minimum is achieved by

$$\lambda = \frac{2}{\sqrt{\frac{2n\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{\left(\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{n}}{\delta}\right)}+1}+1} \tag{8}$$

(Tolstikhin and Seldin, 2013). Note that the optimal value of λ is smaller than 1 and that for $n \geq 4$ it can be lower bounded as $\lambda \geq \frac{2}{\sqrt{2n+1}+1} \geq \frac{1}{\sqrt{n}}$. Alternating application of the update rules (7) and (8) monotonously decreases the bound, and thus converges to a local minimum.

Unfortunately, the bound is *not* jointly convex in ρ and λ (this can be verified by computing the Hessian of the first term and taking large n, so that the second term can be ignored). Joint convexity would have been a sufficient condition for convergence to the global minimum, but it is *not* a necessary condition. In the following section we show that under certain conditions alternating minimization is still guaranteed to converge to the global minimum of the bound despite absence of joint convexity.

4. Strong Quasiconvexity of the PAC-Bayes-λ Bound

We denote the right hand side of the bound in Theorem 3 by

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

By substituting the optimal value of ρ from equation (7) into $\mathcal{F}(\rho,\lambda)$ and applying the identity

$$\begin{aligned}\mathrm{KL}(\rho_\lambda\|\pi) &= \mathbb{E}_{\rho_\lambda}\left[\ln\frac{\rho_\lambda(h)}{\pi(h)}\right] = \mathbb{E}_{\rho_\lambda}\left[\ln\frac{e^{-n\lambda\hat{L}(h,S)}}{\mathbb{E}_\pi\left[e^{-n\lambda\hat{L}(h',S)}\right]}\right] \\ &= -n\lambda\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] - \ln\mathbb{E}_\pi\left[e^{-n\lambda\hat{L}(h,S)}\right] \end{aligned} \tag{9}$$

we can write $\mathcal{F}$ as a function of a single scalar parameter λ:

$$\begin{aligned}\mathcal{F}(\lambda) &= \frac{\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{1-\lambda/2} + \frac{\mathrm{KL}(\rho_\lambda\|\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda(1-\lambda/2)} \\ &= \frac{\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{1-\lambda/2} - \frac{\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{1-\lambda/2} + \frac{-\ln\mathbb{E}_\pi\left[e^{-n\lambda\hat{L}(h,S)}\right]+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda(1-\lambda/2)} \\ &= \frac{-\ln\mathbb{E}_\pi\left[e^{-n\lambda\hat{L}(h,S)}\right]+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda(1-\lambda/2)}.\end{aligned}$$

We note that $\mathcal{F}(\lambda)$ is not necessarily convex in λ. For example, taking $\mathcal{H} = \{h_1, h_2\}$, $\hat{L}(h_1,S) = 0$, $\hat{L}(h_2,S) = 0.5$, $\pi(h_1) = \pi(h_2) = \frac{1}{2}$, $n = 100$, and $\delta = 0.01$ produces a non-convex $\mathcal{F}$. However, we show that $\mathcal{F}(\lambda)$ is strongly quasiconvex under a certain condition

on the variance defined by

$$\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right] = \mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)^2\right] - \mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]^2.$$

We recall that a univariate function $f : \mathcal{I} \to \mathbb{R}$ defined on an interval $\mathcal{I} \subseteq \mathbb{R}$ is *strongly quasiconvex* if for any $x, y \in \mathcal{I}$ and $t \in (0,1)$ we have $f(tx + (1-t)y) < \max\{f(x), f(y)\}$.

Theorem 4 *$\mathcal{F}(\lambda)$ is continuous and if at least one of the two conditions*

$$2\,\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{4n}{\delta^2} > \lambda^2 n^2\, Var_{\rho_\lambda}\left[\hat{L}(h,S)\right] \tag{10}$$

or

$$\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] > (1-\lambda)n\, Var_{\rho_\lambda}\left[\hat{L}(h,S)\right] \tag{11}$$

is satisfied for all $\lambda \in \left[\sqrt{\frac{\ln\frac{2\sqrt{n}}{\delta}}{n}}, 1\right]$, then $\mathcal{F}(\lambda)$ is strongly quasiconvex for $\lambda \in (0,1]$ and alternating application of the update rules (7) and (8) converges to the global minimum of $\mathcal{F}$.

Proof The proof is based on inspection of the first and second derivative of $\mathcal{F}(\lambda)$. Calculation of the derivatives is provided in Appendix A. The existence of the first derivative ensures continuity of $\mathcal{F}(\lambda)$. By inspecting the first derivative we obtain that stationary points of $\mathcal{F}(\lambda)$ corresponding to $\mathcal{F}'(\lambda) = 0$ are characterized by the identity

$$2(1-\lambda)\left(\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{2\sqrt{n}}{\delta}\right) = \lambda^2 n \mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right].$$

The identity provides a lower bound on the value of λ at potential stationary points. Using the facts that $\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] \leq 1$ and for $\lambda \leq \frac{1}{2}$ the complement $(1-\lambda) \geq \frac{1}{2}$, for $n \geq 7$ we have

$$\lambda = \sqrt{\frac{2(1-\lambda)\left(\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}} \geq \min\left(\sqrt{\frac{\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}}\,,\ \frac{1}{2}\right) \geq \sqrt{\frac{\ln\frac{2\sqrt{n}}{\delta}}{n}}.$$

By expressing $\mathrm{KL}(\rho_\lambda\|\pi)$ via $\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]$ (or the other way around) and substituting it into the second derivative of $\mathcal{F}(\lambda)$ we obtain that if either of the two conditions of the theorem is satisfied at a stationary point then the second derivative of $\mathcal{F}(\lambda)$ is positive there. Thus, if (10) or (11) is satisfied for all $\lambda \in \left[\sqrt{\frac{\ln\frac{2\sqrt{n}}{\delta}}{n}}, 1\right]$ then all stationary points of $\mathcal{F}(\lambda)$ (if any exist) are local minima. Since $\mathcal{F}(\lambda)$ is a continuous one-dimensional function it means that $\mathcal{F}(\lambda)$ is strongly quasiconvex (it has a single global minimum and no other stationary points). Since alternating minimization monotonously decreases the value of $\mathcal{F}(\lambda)$ it is guaranteed to converge to the global minimum. ■

Next we show a sufficient condition for (10) to be satisfied for a finite $\mathcal{H}$ for all λ.

Theorem 5 *Let m be the number of hypotheses in $\mathcal{H}$ and assume that the prior $\pi(h)$ is uniform. Let $a = \frac{\sqrt{\ln \frac{4n}{\delta^2}}}{n\sqrt{3}}$, $b = \frac{\ln(3mn^2)}{\sqrt{n \ln \frac{2\sqrt{n}}{\delta}}}$, and $K = \frac{e^2}{12} \ln \frac{4n}{\delta^2}$. Let $x_h = \hat{L}(h,S) - \min_h \hat{L}(h,S)$. If the number of hypotheses for which $x_h \in (a,b)$ is at most K then $Var_{\rho_\lambda}\left[\hat{L}(h,S)\right] \leq \frac{\ln \frac{4n}{\delta^2}}{\lambda^2 n^2}$ for all $\lambda \in \left[\sqrt{\frac{\ln \frac{2\sqrt{n}}{\delta}}{n}}, 1\right]$ and $\mathcal{F}(\lambda)$ is strongly quasiconvex and its global minimum is guaranteed to be found by alternating application of the update rules (7) and (8).*

The theorem splits hypotheses in $\mathcal{H}$ into "good", "mediocre", and "bad". "Good" hypotheses are those for which $x_h \leq a$, meaning that the empirical loss $\hat{L}(h,S)$ is close to the best. "Mediocre" are those for which $x_h \in (a,b)$. "Bad" are those for which $x_h \geq b$. The theorem states that as long as the number of "mediocre" hypotheses is not too large, $\mathcal{F}(\lambda)$ is guaranteed to be quasiconvex.

Proof We have $\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right] = \mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S) - \min_h \hat{L}(h,S)\right]$. Under the assumption of a uniform prior $\rho_\lambda(h) = e^{-n\lambda x_h} / \sum_{h'} e^{-n\lambda x_{h'}}$. Since for $h^* = \arg\min_h \hat{L}(h,S)$ we have $x_{h^*} = 0$, the denominator satisfies $\sum_h e^{-n\lambda x_h} \geq 1$. Let $\mathcal{I}_{0a} = [0,a]$, $\mathcal{I}_{ab} = (a,b)$, and $\mathcal{I}_{b1} = [b,1]$ be the intervals corresponding to "good", "mediocre", and "bad" hypotheses. We have:

$$\begin{aligned}
\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right] &\leq \mathbb{E}_{\rho_\lambda}\left[x_h^2\right] \\
&= \frac{\sum_h x_h^2 e^{-n\lambda x_h}}{\sum_h e^{-n\lambda x_h}} \\
&= \frac{\sum_{x_h \in \mathcal{I}_{0a}} x_h^2 e^{-n\lambda x_h}}{\sum_h e^{-n\lambda x_h}} + \frac{\sum_{x_h \in \mathcal{I}_{ab}} x_h^2 e^{-n\lambda x_h}}{\sum_h e^{-n\lambda x_h}} + \frac{\sum_{x_h \in \mathcal{I}_{b1}} x_h^2 e^{-n\lambda x_h}}{\sum_h e^{-n\lambda x_h}} \\
&\leq a^2 + \sum_{x_h \in \mathcal{I}_{ab}} x_h^2 e^{-n\lambda x_h} + \sum_{x_h \in \mathcal{I}_{b1}} x_h^2 e^{-n\lambda x_h}.
\end{aligned}$$

We show a number of properties of the above expression. First, we recall that $\lambda \leq 1$. Therefore,

$$a^2 = \frac{\ln \frac{4n}{\delta^2}}{3n^2} \leq \frac{\ln \frac{4n}{\delta^2}}{3\lambda^2 n^2}.$$

For the second term, simple calculus gives $x^2 e^{-n\lambda x} \leq \frac{4}{e^2 n^2 \lambda^2}$. Since by the theorem assumption there are at most K hypotheses falling in $\mathcal{I}_{ab}$,

$$\sum_{h \in \mathcal{I}_{ab}} x_h^2 e^{-n\lambda x_h} \leq \frac{4K}{e^2 n^2 \lambda^2} \leq \frac{\ln \frac{4n}{\delta^2}}{3\lambda^2 n^2}.$$

For the last term we have

$$b > \frac{2}{\sqrt{n \ln \frac{2\sqrt{n}}{\delta}}} = \frac{2}{n}\sqrt{\frac{n}{\ln \frac{2\sqrt{n}}{\delta}}} \geq \frac{2}{\lambda n}$$

and we note that for $x \geq 2/\lambda n$ the function $x^2 e^{-n\lambda x}$ is monotonically decreasing in x. Since $\lambda \geq \sqrt{\frac{\ln \frac{2\sqrt{n}}{\delta}}{n}}$ we obtain

$$\sum_{x_h \in \mathcal{I}_{b1}} x_h^2 e^{-n\lambda x_h} \leq m b^2 e^{-n\lambda b} \leq m b^2 e^{-\sqrt{n \ln \frac{2\sqrt{n}}{\delta}} b} \leq m e^{-\sqrt{n \ln \frac{2\sqrt{n}}{\delta}} b} = m \frac{1}{3mn^2} = \frac{1}{3n^2} \leq \frac{\ln \frac{4n}{\delta^2}}{3\lambda^2 n^2}. \tag{12}$$

By taking all three terms together we arrive at

$$\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right] \leq \frac{\ln \frac{4n}{\delta^2}}{\lambda^2 n^2},$$

which implies condition (10) of Theorem 4 since $\mathrm{KL}(\rho_\lambda \| \pi) \geq 0$. ■

In Appendix B we provide a couple of relaxations of the conditions in Theorem 5. The first allows to trade the boundaries a and b of the intervals with the value of K and the second improves the value of b.

In our experiments presented in Section 6, $\mathcal{F}(\lambda)$ turned to be convex even when the sufficient conditions of Theorem 5 (including the relaxations detailed in Appendix B) were violated. This suggests that it may be possible to relax the conditions even further. At the same time, it is possible to construct artificial examples, where $\mathcal{F}(\lambda)$ is not quasiconvex. For example, taking $n = 200$, $\delta = 0.25$, and $m = \mathrm{round}\left(e^{0.74n\Delta}\right) + 1 \approx 2.7 \cdot 10^6$ hypotheses (where round is rounding to the nearest integer) with $\hat{L}(h_1, S) = 0$ and $\hat{L}(h_i, S) = \Delta = 0.1$ for all $i \in \{2, \ldots, m\}$ and a uniform prior leads to $\mathcal{F}(\lambda)$ with two local minima. The artificial example requires m to be of the order of $e^{n\lambda^\star \Delta}$, where $\lambda^\star$ is the value of λ at a stationary point of $\mathcal{F}(\lambda)$ and Δ is the loss of suboptimal hypotheses (in the example $\Delta = 0.1$). Thus, quasiconvexity is not always guaranteed, but it holds in a wide range of practical scenarios.

5. Construction of a Hypothesis Space

Computation of the partition function (the denominator in (7)) is not always tractable, however, it can be easily computed when $\mathcal{H}$ is finite. The crucial step is to construct a sufficiently powerful finite hypothesis space $\mathcal{H}$. Our proposal is to construct $\mathcal{H}$ by training m hypotheses, where each hypothesis is trained on r random points from S and validated on the remaining $n - r$ points. This construction resembles a cross-validation split of the data. However, in cross-validation r is typically large (close to n) and validation sets are non-overlapping. Our approach works for any r and has additional computational advantages when r is small. We do not require validation sets to be non-overlapping and overlaps between training sets are allowed. Below we describe the construction more formally.

Let $h \in \{1, \ldots, m\}$ index the hypotheses in $\mathcal{H}$. Let S_h denote the training set of h and $S \setminus S_h$ the validation set. S_h is a subset of r points from S, which are selected independently of their values (for example, subsampled randomly or picked according to a predefined partition of the data). We define the validation error of h by $\hat{L}^{\mathrm{val}}(h, S) = \frac{1}{n-r} \sum_{(X,Y) \in S \setminus S_h} \ell(h(X), Y)$. Note that the validation errors are $(n - r)$ i.i.d. random variables with bias $L(h)$ and, therefore, for $f(h, S) = (n - r)\,\mathrm{kl}(\hat{L}^{\mathrm{val}}(h, S) \| L(h))$ we have $\mathbb{E}_S\left[e^{f(h,S)}\right] \leq 2\sqrt{n - r}$. The following result is a straightforward adaptation of Theorem 3 to our setting. A proof sketch is provided in Appendix C.

	Mushrooms	Skin	Waveform	Adult	Ionosphere	AvsB	Haberman	Breast
|S|	2000	2000	2000	2000	200	1000	150	340
|T|	500	500	500	500	150	500	150	340
d	112	3	40	122	34	16	3	10

Table 1: **Datasets summary.** $|S| = n$ refers to the size of the training set and $|T|$ refers to the size of the test set, d refers to the number of features. "Breast" abbreviates "Breast cancer" dataset.

Theorem 6 *Let S be a sample of size n. Let $\mathcal{H}$ be a set of m hypotheses, where each $h \in \mathcal{H}$ is trained on r points from S selected independently of the composition of S. For any probability distribution π over $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability greater than $1-\delta$ over a random draw of a sample S, for all distributions ρ over $\mathcal{H}$ and $\lambda \in (0,2)$ simultaneously:*

$$\mathbb{E}_\rho\left[L(h)\right] \leq \frac{\mathbb{E}_\rho\left[\hat{L}^{\mathrm{val}}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n-r}}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)(n-r)}. \tag{13}$$

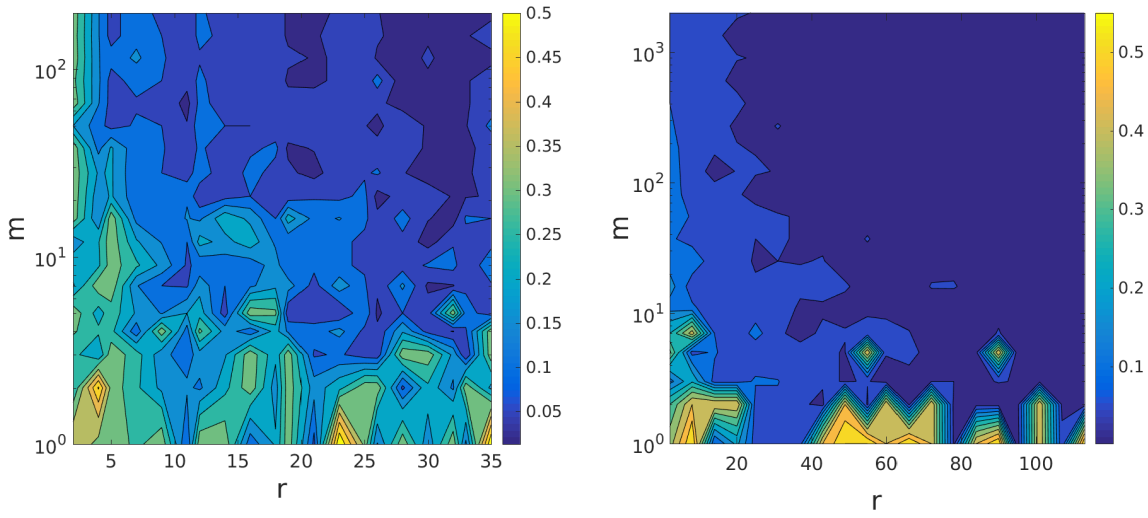
It is natural, but not mandatory to select a uniform prior $\pi(h) = 1/m$. The bound in equation (13) can be minimized by alternating application of the update rules in equations (7) and (8) with n being replaced by $n-r$ and $\hat{L}$ by $\hat{L}^{\mathrm{val}}$.

6. Experimental Results

In this section we study how PAC-Bayesian weighting of weak classifiers proposed in Section 5 compares with "strong" kernel SVMs tuned by cross-validation and trained on all training data. The experiments were performed on eight UCI datasets (Asuncion and Newman, 2007) summarized in Table 1. In our experiments we employed the SVM solver from LIBSVM (Chang and Lin, 2011).

We compared the prediction accuracy and run time of our prediction strategy with a baseline of RBF kernel SVMs tuned by cross-validation. For the baseline we used 5-fold cross-validation for selecting the soft-margin parameter, C, and the bandwidth parameter γ of the kernel $k(X_i, X_j) = \exp(-\gamma\|X_i - X_j\|^2)$. The value of C was selected from a grid, such that $\log_{10} C \in \{-3, -2, \ldots, 3\}$. The values for the grid of γ-s were selected using the heuristic proposed in Jaakkola et al. (1999). Specifically, for $i \in \{1, \ldots, n\}$ we defined $G(X_i) = \min_{(X_j,Y_j)\in S \wedge Y_i \neq Y_j} \|X_i - X_j\|$. We then defined a seed γ_J by $\gamma_J = \frac{1}{2\cdot\mathrm{median}(G)^2}$. Finally, we took a geometrically spaced grid around γ_J, so that $\gamma \in \{\gamma_J 10^{-4}, \gamma_J 10^{-2}, \ldots, \gamma_J 10^4\}$.

For our approach we selected m subsets of r points uniformly at random from the training set S. We then trained an RBF kernel SVM for each subset. The kernel bandwidth parameter γ was randomly selected for each subset from the same grid as used in the baseline. In all our experiments very small values of r, typically up to $d+1$ with d being the input dimension, were sufficient for successfully competing with the prediction accuracy of



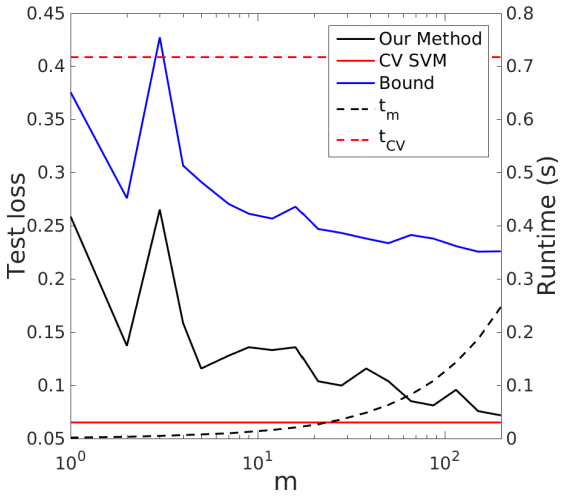
(a) Ionosphere dataset. $L_{\mathrm{CV}} = 0.06$. (b) Mushrooms dataset. $L_{\mathrm{CV}} = 0$

Figure 1: **Prediction accuracy of PAC-Bayesian aggregation vs. cross-validated SVM across different values of m and r.** The colors of the heatmap represent the difference between the zero-one loss of the ρ-weighted majority vote and the zero-one loss of the cross-validated SVM. The loss of the cross-validated SVM is given by L_{CV} in the caption.

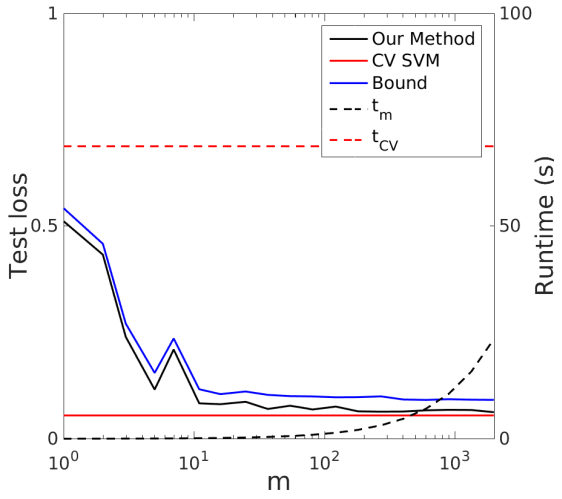
the baseline and provided the most significant computational improvement. For such small values of r it was easy to achieve perfect separation of the training points and, therefore, selection of C was unnecessary. The performance of each weak classifier was validated on $n - r$ points not used in its training. The weighting of classifiers ρ was then computed through alternating minimization of the bound in Theorem 6.

In most of PAC-Bayesian literature it is common to replace randomized prediction with ρ-weighted majority vote. From a theoretical point of view the error of ρ-weighted majority vote is bounded by at most twice the error of the corresponding randomized classifier, however, in practice it usually performs better than randomized prediction (Germain et al., 2009). In our main experiments we have followed the standard choice of using the ρ-weighted majority vote. In Appendix D.4 we provide additional experiments showing that in our case the improvement achieved by taking the majority vote was very minor compared to randomized prediction. We use the term *PAC-Bayesian aggregation* to refer to prediction with ρ-weighted majority vote.

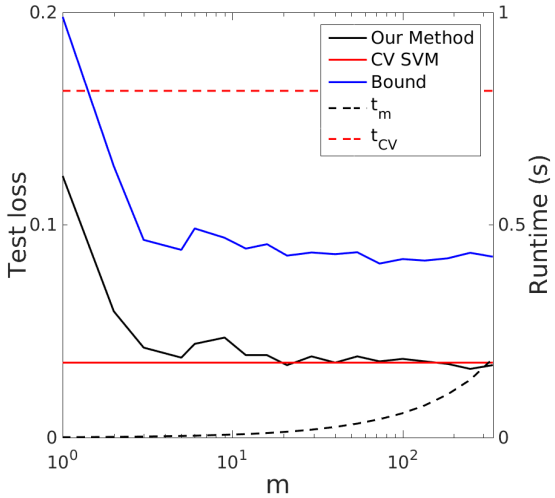
In the first two experiments we studied the influence of r and m on classification accuracy and run time. The complexity term in Theorem 6 (the second term on the right hand side of (13)) decreases with the decrease of the training set sizes r (because the size of the validation sets $n - r$ increases) and with the decrease of the number of hypotheses m (because $\pi(h) = 1/m$ increases). From the computational point of view it is also desirable to have small r and m, especially when working with expensive-to-train models, such as



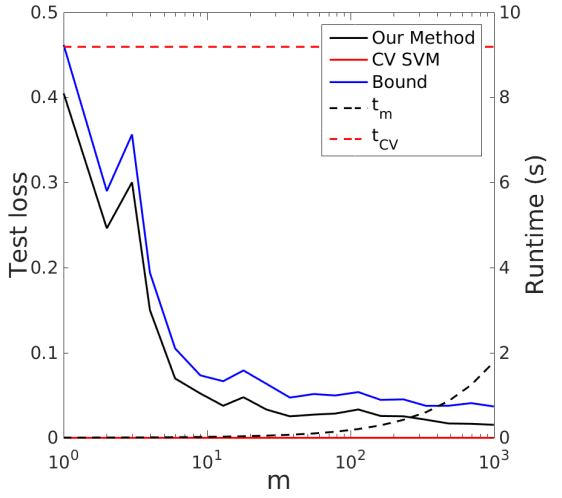
(*a*) Ionosphere dataset. $n = 200$, $r = d + 1 = 35$.



(*b*) Waveform dataset. $n = 2000$, $r = d + 1 = 41$.



(*c*) Breast cancer dataset. $n = 340$, $r = d + 1 = 11$.



(*d*) AvsB dataset. $n = 1000$, $r = d + 1 = 17$.

Figure 2: **Comparison of PAC-Bayesian aggregation with RBF kernel SVM tuned by cross-validation.** The solid red, black, and blue lines correspond, respectively, to the zero-one test loss of the cross-validated SVM, the loss of ρ-weighted majority vote, where ρ is a result of minimization of the PAC-Bayes-λ bound, and PAC-Bayes-kl bound on the loss of randomized classifier defined by ρ. The dashed black line represents the training time of PAC-Bayesian aggregation, while the red dashed line represents the training time of cross-validated SVM. The prediction accuracy and run time of PAC-Bayesian aggregation and PAC-Bayes-kl bound are given as functions of the hypothesis set size m.

	Mushrooms	Skin	Waveform	Adult	Ionosphere	AvsB	Haberman	Breast
m	130	27	140	28	24	160	23	50

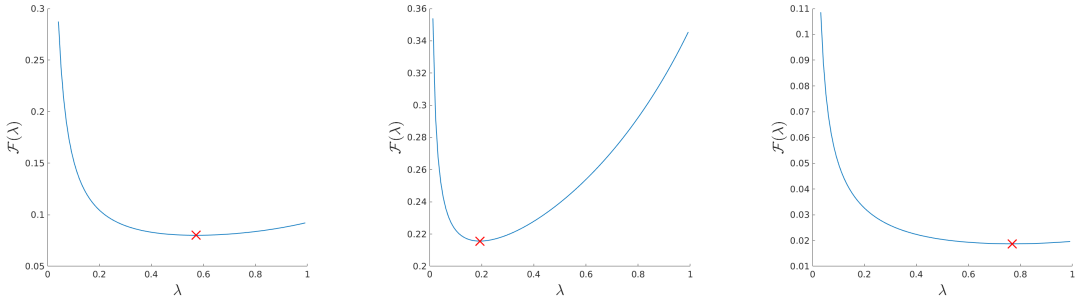
Table 2: **Average maximal values of m for which quasiconvexity was guaranteed by Theorem 5.**

kernel SVMs, which have super-quadratic training time. What pushes r and m up is the validation error, $\mathbb{E}_\rho\left[\hat{L}^{\text{val}}(h,S)\right]$.

In the first experiment we studied the influence of r and m on the prediction accuracy of PAC-Bayesian aggregation. We considered 20 values of m in $[1,n]$ and 20 values of r in $[2,d+1]$. For each pair of m and r the prediction accuracy of PAC-Bayesian aggregation was evaluated, resulting in a matrix of losses. To simplify the comparison we subtracted the prediction accuracy of the baseline, thus zero values correspond to matching the accuracy of the baseline. In Figure 1 we show a heatmap of this matrix for two UCI datasets and the results for the remaining datasets are provided in Appendix D.1. Overall, reasonably small values of m and r were sufficient for matching the accuracy of SVM tuned by cross-validation.

In the second experiment we provide a closer look at the effect of increasing the number m of weak SVMs when their training set sizes r are kept fixed. We picked $r=d+1$ and ran our training procedure with 20 values of m in $[1,n]$. In Figure 2 we present the prediction accuracy of the resulting weighted majority vote vs. prediction accuracy of the baseline for four datasets. The graphs for the remaining datasets are provided in Appendix D.1. We also show the running time of our procedure vs. the baseline. The running time of the baseline includes cross-validation and training of the final SVM on the whole training set, while the running time of PAC-Bayesian aggregation includes training of m weak SVMs, their validation, and the computation of ρ. In addition, we report the value of PAC-Bayes-kl bound from Theorem 2 on the expected loss of the randomized classifier defined by ρ. The kl divergence was inverted numerically to obtain a bound on the expected loss $\mathbb{E}_\rho[L(h)]$. The bound was adapted to our construction by replacing n with $n-r$ and $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ with $\mathbb{E}_\rho\left[\hat{L}^{\text{val}}(h,S)\right]$. We note that since the bound holds for any posterior distribution, it also holds for the distribution found by minimization of the bound in Theorem 6. However, since Theorem 6 is a relaxation of PAC-Bayes-kl bound, using PAC-Bayes-kl for the final error estimate is slightly tighter. The bound on the loss of ρ-weighted majority vote is at most a factor of 2 larger than than the bound for the randomized classifier. In calculation of the bound we used $\delta=0.05$. We conclude from the figure that relatively small values of m are sufficient for matching or almost matching the prediction accuracy of the baseline, while the run time is reduced dramatically. We also note that the bound is exceptionally tight.

In our last experiment we tested quasiconvexity of $\mathcal{F}(\lambda)$. Theorem 5 provided theoretical guarantee of quasiconvexity for small m and numerical evaluation has further shown that $\mathcal{F}(\lambda)$ was convex for all values of m used in our experiments. For testing the theoretical guarantees we increased m in steps of 10 until the sufficient condition for strong



(*a*) Breast cancer dataset with $|\mathrm{S}| = 340$.

(*b*) Adult dataset with $|\mathrm{S}| = 2000$.

(*c*) Mushrooms dataset with $|\mathrm{S}| = 2000$.

Figure 3: **Empirical evaluation of the shape of $\mathcal{F}(\lambda)$.** The blue curve represents the value of $\mathcal{F}(\lambda)$ and the red cross marks λ returned by alternating minimization.

quasiconvexity in Theorem 5 was violated. The condition included adjustment of interval boundaries a and b, as described in Appendix B.1, and improved value of b, as described in Appendix B.2. The experiments were repeated 10 times for each dataset, where in each experiment the training sets were redrawn and a new set of hypotheses was trained, leading to a new set of validation losses $\hat{L}^{\mathrm{val}}$. In Table 2 we report the average over the 10 repetitions of the maximal values of m with guaranteed quasiconvexity. Since $\mathcal{F}(\lambda)$ is always quasiconvex for $m \leq K(0,0)+1$, where $K(0,0) = \frac{e^2}{4} \ln \frac{4n}{\delta^2}$ (see equation (15) in Appendix B.1), we report the value of $K(0,0)+1$ whenever it was not possible to ensure quasiconvexity with larger m. When it was not possible to guarantee quasicovexity theoretically we tested the shape of $\mathcal{F}(\lambda)$ empirically. Figure 3 shows a few typical examples. The plots in Figure 3 were constructed in the following way: given a sample S, we trained m weak SVMs. We then computed the corresponding vector of validation losses, $\hat{L}^{\mathrm{val}}(h, S)$. For each value in a grid of λ-s, we computed the corresponding ρ according to equation (7). Finally, we substituted the value of ρ_λ and λ into equation (3) to get the value of the bound. In all our calculations we used a uniform prior and $\delta = 0.05$. The figure shows that $\mathcal{F}(\lambda)$ was convex in λ in all the cases.

7. Conclusion

We have presented a new PAC-Bayesian inequality, an alternating procedure for its minimization, and a way to construct a finite hypothesis space for which the bound and minimization procedure work particularly well. We have derived sufficient conditions for the minimization procedure to converge to the global optimum of the bound. We have shown that the procedure is competitive with cross-validation in tuning the trade-off between complexity and empirical performance of ρ. In addition, it provides tight high-probability generalization guarantees and achieves prediction accuracies on par with kernel SVMs tuned by cross-validation, but at a considerably lower computation cost.

In our experiments the bound turned to be convex even when the sufficient conditions of Theorem 5 were violated. It suggests that further relaxation of these conditions may be possible to achieve in future work.

Acknowledgments

We thank anonymous reviewers of this and earlier versions of the manuscript for valuable feedback. We also thank Oswin Krause for suggesting the use of the term "quasiconvexity" to describe the shape of $\mathcal{F}(\lambda)$. CI and YS acknowledge support by the Innovation Fund Denmark through the *Danish Center for Big Data Analytics Driven Innovation* (DABAI). OW would like to thank the mathematical department of the university of Copenhagen for his guest professorship in 2015-2016.

References

Arthur Asuncion and David J. Newman. UCI machine learning repository, 2007. URL www.ics.uci.edu/~mlearn/MLRepository.html.

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

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, 56, 2007.

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

Marc Claesen, Frank De Smet, Johan A.K. Suykens, and Bart De Moor. EnsembleSVM: A library for ensemble learning using support vector machines. *Journal of Machine Learning Research*, 15(1), 2014.

Ronan Collobert, Samy Bengio, and Yoshua Bengio. A parallel mixture of SVMs for very large scale problems. *Neural Computation*, 14(5), 2002.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2nd edition, 2006.

Monroe D. Donsker and S.R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

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.

Tommi Jaakkola, Mark Diekhans, and David Haussler. Using the fisher kernel method to detect remote protein homologies. In *In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 1999.

Michael Kearns, Yishay Mansour, Andrew Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27, 1997.

Michael J. Kearns and Dand Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11, 1999.

Joseph Keshet, David McAllester, and Tamir Hazan. Pac-bayesian approach for minimization of phoneme error rate. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6, 2005.

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

François Laviolette and Mario Marchand. PAC-Bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers. *Journal of Machine Learning Research*, 8, 2007.

Katalin Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6(3), 1996.

Katalin Marton. A measure concentration inequality for contracting Markov chains Erratum. *Geometric and Functional Analysis*, 7(3), 1997.

Andreas Maurer. A note on the PAC-Bayesian theorem. www.arxiv.org, 2004.

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

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51, 2003.

Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1), 2000.

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. PAC-Bayesian ensemble learning. Master's thesis, University of Copenhagen, 2016.

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

Giorgio Valentini and Thomas G. Dietterich. Low bias bagged support vector machines. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.

Leslie G. Valiant. A theory of the learnable. *Communications of the Association for Computing Machinery*, 27, 1984.

Appendix A. Calculation of the Derivatives of $\mathcal{F}(\lambda)$

We decompose $\mathcal{F}(\lambda)$ in the following way:

$$\mathcal{F}(\lambda) = \frac{-\ln \mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right] + \ln \frac{2\sqrt{n}}{\delta}}{n\lambda(1-\lambda/2)} = f(\lambda)g(\lambda),$$

where

$$f(\lambda) = -\frac{1}{n} \ln \mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right] + \frac{\ln \frac{2\sqrt{n}}{\delta}}{n},$$
$$g(\lambda) = \frac{1}{\lambda(1-\lambda/2)}.$$

For the derivatives of f and g we have:

$$f'(\lambda) = -\frac{\frac{d}{d\lambda}\mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right]}{n\mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right]} = \frac{\mathbb{E}_\pi \left[\hat{L}(h,S) e^{-n\lambda \hat{L}(h,S)} \right]}{\mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right]} = \mathbb{E}_{\rho_\lambda} \left[\hat{L}(h,S) \right] \geq 0.$$

$$\begin{aligned} f''(\lambda) &= \frac{\left(\frac{d}{d\lambda}\mathbb{E}_\pi \left[\hat{L}(h,S) e^{-n\lambda \hat{L}(h,S)} \right] \right) \mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right] - \left(\frac{d}{d\lambda}\mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right] \right) \mathbb{E}_\pi \left[\hat{L}(h,S) e^{-n\lambda \hat{L}(h,S)} \right]}{\mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right]^2} \\ &= \frac{-n\mathbb{E}_\pi \left[\hat{L}(h,S)^2 e^{-n\lambda \hat{L}(h,S)} \right]}{\mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right]} + n \left(\frac{\mathbb{E}_\pi \left[\hat{L}(h,S) e^{-n\lambda \hat{L}(h,S)} \right]}{\mathbb{E}_\pi \left[e^{-n\lambda \hat{L}(h,S)} \right]} \right)^2 \\ &= -n \left(\mathbb{E}_{\rho_\lambda} \left[\hat{L}(h,S)^2 \right] - \left(\mathbb{E}_{\rho_\lambda} \left[\hat{L}(h,S) \right] \right)^2 \right) \\ &= -n\mathrm{Var}_{\rho_\lambda} \left[\hat{L}(h,S) \right] \leq 0. \end{aligned}$$

$$g'(\lambda) = -\frac{(1-\lambda/2-\lambda/2)}{\lambda^2(1-\lambda/2)^2} = \frac{\lambda-1}{\lambda^2(1-\lambda/2)^2} \leq 0.$$

$$
\begin{aligned}
g''(\lambda) &= \frac{\lambda^2(1-\lambda/2)^2 - (\lambda-1)\left(2\lambda(1-\lambda/2)^2 - \lambda^2(1-\lambda/2)\right)}{\lambda^4(1-\lambda/2)^4} \\
&= \frac{\lambda(1-\lambda/2) - (\lambda-1)\left(2(1-\lambda/2) - \lambda\right)}{\lambda^3(1-\lambda/2)^3} \\
&= \frac{\lambda(1-\lambda/2) - 2(\lambda-1)(1-\lambda)}{\lambda^3(1-\lambda/2)^3} \\
&= \frac{\lambda(1-\lambda/2) + 2(\lambda-1)^2}{\lambda^3(1-\lambda/2)^3} \\
&= \frac{\lambda - \lambda^2/2 + 2\lambda^2 - 4\lambda + 2}{\lambda^3(1-\lambda/2)^3} \\
&= \frac{(3/2)\lambda^2 - 3\lambda + 2}{\lambda^3(1-\lambda/2)^3} \\
&= \frac{3\lambda^2 - 6\lambda + 4}{2\lambda^3(1-\lambda/2)^3} \\
&= \frac{3\left(\lambda-1\right)^2 + 1}{2\lambda^3(1-\lambda/2)^3} > 0.
\end{aligned}
$$

At a stationary point we have $\mathcal{F}'(\lambda) = f'(\lambda)g(\lambda) + g'(\lambda)f(\lambda) = 0$. By using the identity

$$
f(\lambda) = \lambda \mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] + \frac{\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}, \tag{14}
$$

which follows from (9), this gives

$$
\mathcal{F}'(\lambda) = \frac{\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{\lambda(1-\lambda/2)} + \frac{(\lambda-1)\left(\lambda\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]\right)}{\lambda^2(1-\lambda/2)^2} + \frac{(\lambda-1)\left(\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n\lambda^2(1-\lambda/2)^2} = 0.
$$

This can be rewritten as

$$
\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] + \frac{(\lambda-1)\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{1-\lambda/2} + \frac{(\lambda-1)\left(\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n\lambda(1-\lambda/2)} = 0,
$$

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

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

which characterizes the stationary points. By combining this with the identity (14) we obtain that at a stationary point

$$
f(\lambda) = \left(\lambda + \frac{\lambda^2}{2(1-\lambda)}\right)\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] = \frac{\lambda(1-\lambda/2)}{1-\lambda}\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right].
$$

For the second derivative we have $\mathcal{F}''(\lambda) = f''(\lambda)g(\lambda) + 2f'(\lambda)g'(\lambda) + g''(\lambda)f(\lambda)$. At a stationary point

$$\begin{aligned} g''(\lambda)f(\lambda) + 2f'(\lambda)g'(\lambda) &= \left(\frac{3\lambda^2 - 6\lambda + 4}{2\lambda^3(1-\lambda/2)^3}\frac{\lambda(1-\lambda/2)}{1-\lambda} - \frac{2(1-\lambda)}{\lambda^2(1-\lambda/2)^2}\right)\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] \\ &= \frac{\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{\lambda(1-\lambda/2)(1-\lambda)}. \end{aligned}$$

By plugging this into $\mathcal{F}''(\lambda)$ we obtain that at a stationary point (if such exists)

$$\begin{aligned} \mathcal{F}''(\lambda) &= \frac{-n\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{\lambda(1-\lambda/2)} + \frac{\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{\lambda(1-\lambda/2)(1-\lambda)} \\ &= \frac{1}{\lambda(1-\lambda/2)}\left(\frac{\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right]}{1-\lambda} - n\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right]\right). \end{aligned}$$

This expression is positive if

$$\mathbb{E}_{\rho_\lambda}\left[\hat{L}(h,S)\right] > (1-\lambda)n\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right]$$

or, equivalently (by characterization of a stationary point),

$$2\,\mathrm{KL}(\rho_\lambda\|\pi) + \ln\frac{4n}{\delta^2} > \lambda^2 n^2 \mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right].$$

Appendix B. Relaxation of the Sufficient Conditions in Theorem 5

In this section we propose a couple of relaxations of the conditions in Theorem 5. The first provides a possibility of tuning the intervals $\mathcal{I}_{0a}$, $\mathcal{I}_{ab}$, and $\mathcal{I}_{b1}$, and the second provides a slight improvement in the definition of b.

B.1. Tuning the intervals $\mathcal{I}_{0a}$, $\mathcal{I}_{ab}$, and $\mathcal{I}_{b1}$

We recall the definition $x_h = \hat{L}(h,S) - \min_h \hat{L}(h,S)$. In Theorem 5 we have tuned a and b so that the contribution to $\mathbb{E}_{\rho_\lambda}[x_h]$ from hypotheses falling into intervals $\mathcal{I}_{0a}$, $\mathcal{I}_{ab}$, and $\mathcal{I}_{b1}$ is equal. Obviously, this does not have to be the case. Take $\alpha \geq 0$ and $\beta \geq 0$, such that $\alpha + \beta \leq 1$ and define

$$a(\alpha) = \frac{\sqrt{\alpha \ln\frac{4n}{\delta^2}}}{n} \quad , \quad b(\beta) = \frac{\ln\left(\frac{1}{\beta}mn^2\right)}{\sqrt{n\ln\frac{2\sqrt{n}}{\delta}}} \quad , \quad K(\alpha,\beta) = \frac{e^2(1-\alpha-\beta)}{4}\ln\frac{4n}{\delta^2}. \tag{15}$$

If we can find any pair (α,β), such that the number of hypotheses for which $x_h \in (a(\alpha), b(\beta))$ is at most $K(\alpha,\beta)$ then $\mathrm{Var}_{\rho_\lambda}\left[\hat{L}(h,S)\right] \leq \frac{\ln\frac{4n}{\delta^2}}{\lambda^2 n^2}$ for all $\lambda \in \left[\sqrt{\frac{\ln\frac{2\sqrt{n}}{\delta}}{n}}, 1\right]$ and $\mathcal{F}(\lambda)$ is strongly quasiconvex. The proof is identical to the proof of Theorem 5 with α, $(1-\alpha-\beta)$, and β being the relative contributions to $\mathbb{E}_{\rho_\lambda}[x_h]$ from the intervals $\mathcal{I}_{0a}$, $\mathcal{I}_{ab}$, and $\mathcal{I}_{b1}$, respectively.

B.2. Refinement of the Boundary b

In the derivation in (12) we have dropped the factor b^2. If we would have kept it we could reduce the value of b. Let

$$b = \frac{\ln\left(3mn\frac{4(\ln(3mn))^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}\right)}{\sqrt{n\ln\frac{2\sqrt{n}}{\delta}}} = \frac{\ln\left(\frac{12mn(\ln(3mn))^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}\right)}{\sqrt{n\ln\frac{2\sqrt{n}}{\delta}}}.$$

Assuming that $\ln\left(\frac{12mn(\ln(3mn))^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}\right) \geq 2$ and $n \geq 5$ we have

$$\begin{aligned}
\sum_{x_h \in \mathcal{I}_{b1}} x_h^2 e^{-n\lambda x_h} &\leq mb^2 e^{-n\lambda b} \\
&= m\frac{\left(\ln(3mn) + \ln\left(\frac{4(\ln(3mn))^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}\right)\right)^2}{n\ln\frac{2\sqrt{n}}{\delta}} \frac{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}{3mn\times 4\left(\ln(3mn)\right)^2} \\
&= \frac{\ln\frac{4n}{\delta^2}}{3n^2}\frac{\left(\ln(3mn) + \ln\left(\frac{4(\ln(3mn))^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}\right)\right)^2}{4\left(\ln(3mn)\right)^2} \qquad (16) \\
&\leq \frac{\ln\frac{4n}{\delta^2}}{3n^2} \qquad (17) \\
&\leq \frac{\ln\frac{4n}{\delta^2}}{3\lambda^2 n^2}.
\end{aligned}$$

In step (17) we used the following auxiliary calculations. For $n \geq 5$ we have $\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2} \geq 4$. From here, for $n \geq 5$ we have $\frac{4(\ln(3mn))^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}} \leq (\ln(3mn))^2$. Furthermore, for $x \geq 0.5$ we have $x \geq (\ln x)^2$, leading to $\ln(3mn) + \ln\left((\ln(3mn))^2\right) \leq 2\ln(3mn)$, since $3mn \geq 0.5$. Thus, the second fraction in line (16) is bounded by 1.

B.3. Combining the two improvements

It is obviously possible to combine the two improvements by defining

$$b(\beta) = \frac{\ln\left(\frac{\frac{4mn}{\beta}\left(\ln\frac{mn}{\beta}\right)^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}\right)}{\sqrt{n\ln\frac{2\sqrt{n}}{\delta}}}$$

and $a(\alpha)$ and $K(\alpha,\beta)$ as before. We only have to check that the conditions $\ln\left(\frac{\frac{4mn}{\beta}\left(\ln\frac{mn}{\beta}\right)^2}{\ln\frac{2\sqrt{n}}{\delta}\ln\frac{4n}{\delta^2}}\right) \geq 2$ and $\frac{mn}{\beta} \geq 0.5$ are satisfied and, otherwise, tune further. Note that since $\hat{L}(h,S)$ is trivially upper bounded by 1, $b(\beta) > 1$ is vacuous.

Appendix C. A Proof Sketch of Theorem 6

In this section we provide a sketch of a proof of Theorem 6. The proof is a straightforward adaptation of the proof of Theorem 3.

Proof As we have already mentioned in the text, since the validation errors are $(n-r)$ i.i.d. random variables with bias $L(h)$, for $f(h,S) = (n-r)\,\mathrm{kl}(\hat{L}^{\mathrm{val}}(h,S)\|L(h))$ we have $\mathbb{E}_S\left[e^{f(h,S)}\right] \leq 2\sqrt{n-r}$ (Maurer, 2004). With this result replacing $f(h,S) = n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))$ and $\mathbb{E}_S\left[e^{f(h,S)}\right] \leq 2\sqrt{n}$ in the proof of Theorem 2 it is straightforward to obtain an analogue of Theorem 2. Namely, that for any probability distribution π over $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability greater than $1-\delta$ over a random draw of a sample S, for all distributions ρ over $\mathcal{H}$ simultaneously

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}^{\mathrm{val}}(h,S)\right]\middle\|\mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{n-r}}{\delta}}{n-r}. \tag{18}$$

And from here, directly following the steps in the proof of Theorem 3, we obtain Theorem 6. ■

Appendix D. Additional Experiments

In this section we present figures with experimental results for UCI datasets that could not be included in the body of the paper. We also present three additional experiments.

D.1. Additional Figures for the Main Experiments

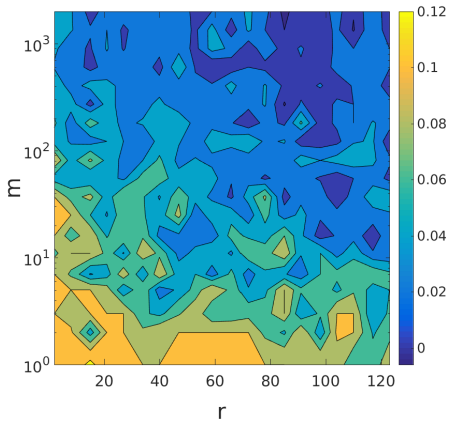
We present the outcomes of experiments in Section 6 for additional UCI datasets. Since the `skin` and `Haberman` datasets have low dimensionality ($d = 3$) we use $r = \sqrt{n}$ rather than $r = d+1$, to get a reasonable subsample size. Figure 4 continues the plots in Figure 1, and Figure 5 continues the plots in Figure 2.

D.2. Comparison with Uniform Weighting and Best Performing Classifier

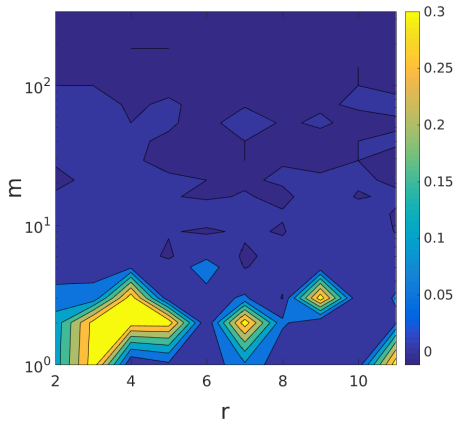
In Figures 6 and 7 we compare the prediction accuracy of ρ-weighted majority vote with uniformly weighted majority vote, which is popular in ensemble learning (Collobert et al., 2002; Valentini and Dietterich, 2003; Claesen et al., 2014). As a baseline the prediction accuracy of a cross-validated SVM is also shown. For the two datasets in Figure 7 we also include the prediction accuracy of SVM corresponding to the maximum value of ρ (which is the best performing SVM in the set). Due to significant overlap with the weighted majority vote, the latter graph is omitted for the datasets in Figure 6. Overall, in our setting the accuracy of ρ-weighted majority vote is comparable to the accuracy of the best classifier in the set and significantly better than uniform weighting.

D.3. Comparison of the Alternating Minimization with Grid Search for Selection of λ

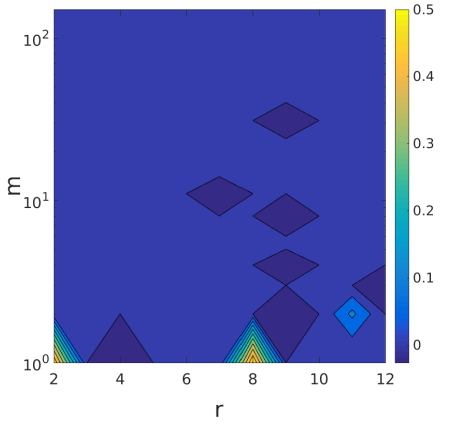
In this section we present a comparison between direct minimization of the PAC-Bayes-λ bound and selection of λ from a grid using a validation set. Table 3 shows how each dataset



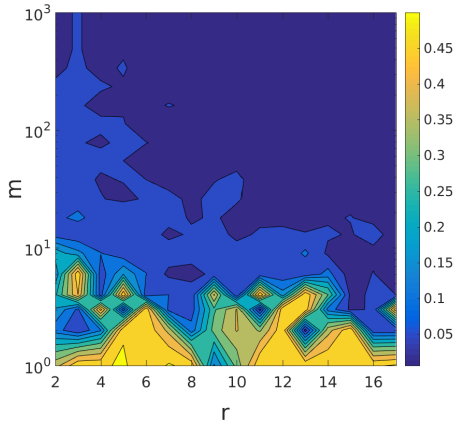
(*a*) Adult dataset. $L_{CV} = 0.15$



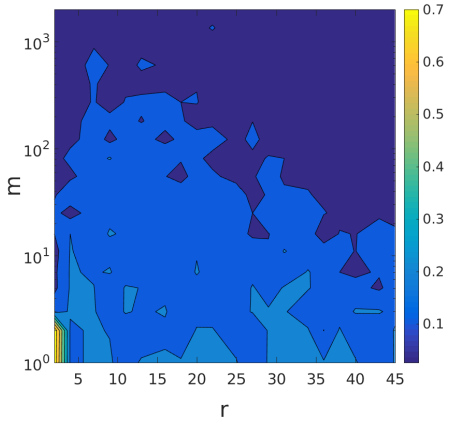
(*b*) Breast cancer dataset. $L_{CV} = 0.05$



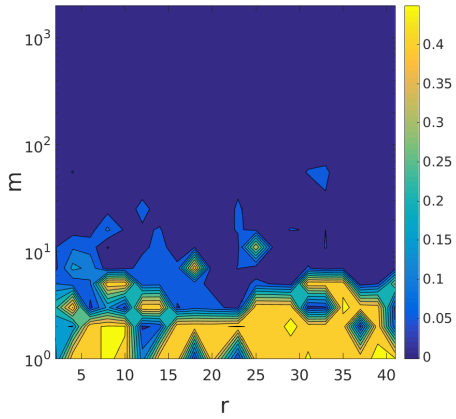
(*c*) Haberman dataset. $L_{CV} = 0.26$



(*d*) AvsB dataset. $L_{CV} = 0$

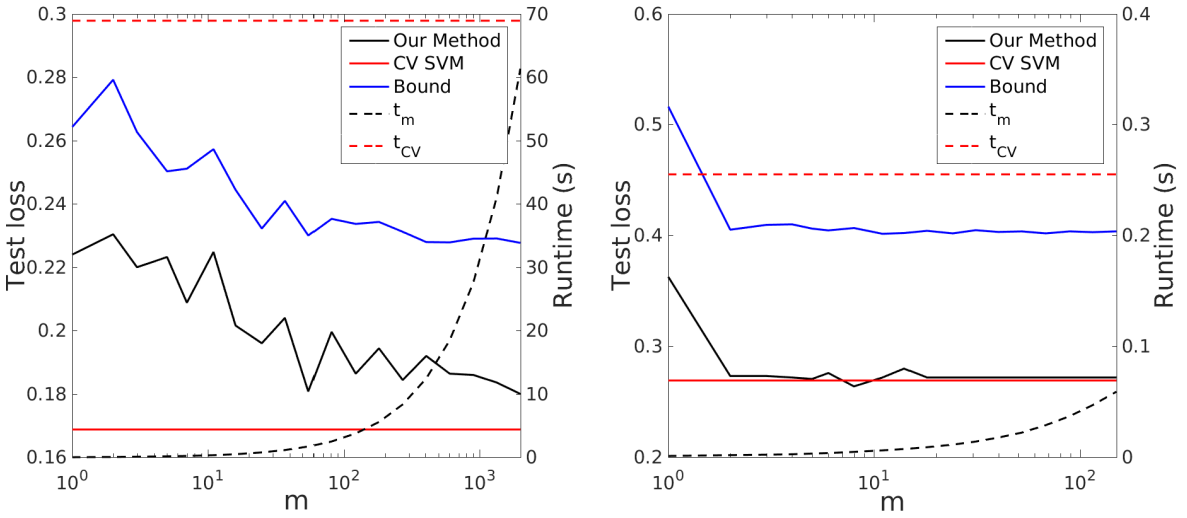


(*e*) Skin dataset. $L_{CV} = 0$

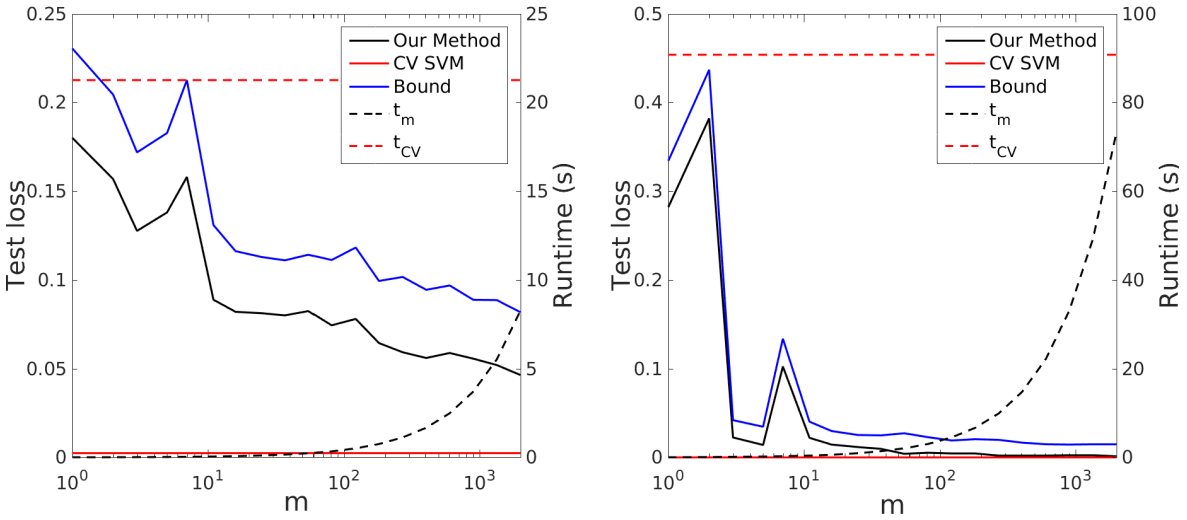


(*f*) Waveform dataset. $L_{CV} = 0.06$

Figure 4: **Prediction accuracy of PAC-Bayesian aggregation vs. cross-validated SVM across different values of m and r.** The colors of the heatmap represent the difference between the zero-one loss of the ρ-weighted majority vote and the zero-one loss of the cross-validated SVM. The loss of the cross-validated SVM is given by L_{CV} in the caption.

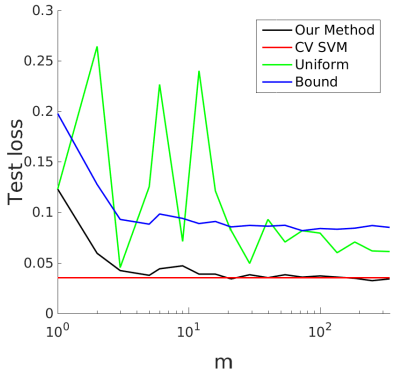


(*a*) Adult dataset. $n = 2000$, $r = d + 1 = 123$. (*b*) Haberman dataset. $n = 150$, $r = \sqrt{n} = 12$.

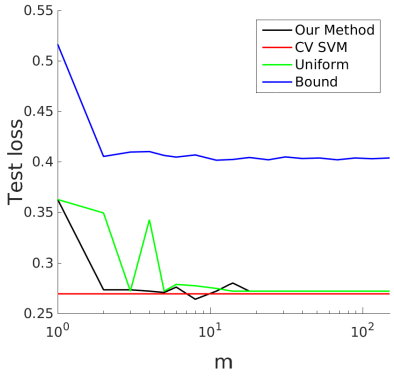


(*c*) Skin dataset. $n = 2000$, $r = \sqrt{n} = 45$. (*d*) Mushrooms dataset. $n = 2000$, $r = d + 1 = 113$.

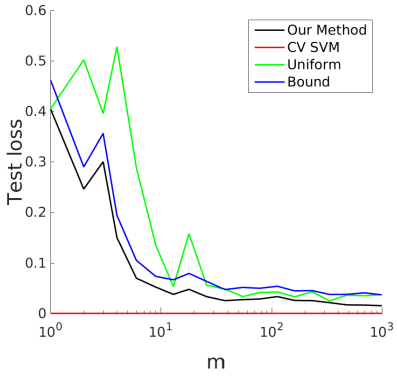
Figure 5: **Comparison of PAC-Bayesian aggregation with RBF kernel SVM tuned by cross-validation.** The solid red, black, and blue lines correspond, respectively, to the zero-one test loss of the cross-validated SVM, the loss of ρ-weighted majority vote, where ρ is a result of minimization of the PAC-Bayes-λ bound, and PAC-Bayes-kl bound on the loss of randomized classifier defined by ρ. The dashed black line represents the training time of PAC-Bayesian aggregation, while the red dashed line represents the training time of cross-validated SVM. The prediction accuracy and run time of PAC-Bayesian aggregation are given as functions of the hypothesis set size m.



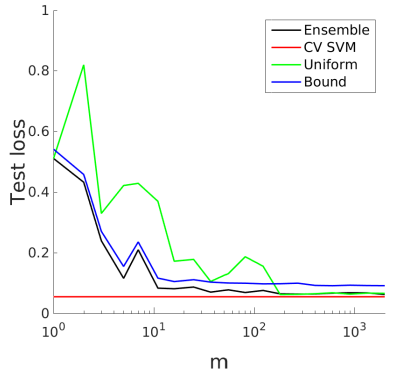
(*a*) Breast cancer dataset. $n = 340$, $r = d + 1 = 11$.



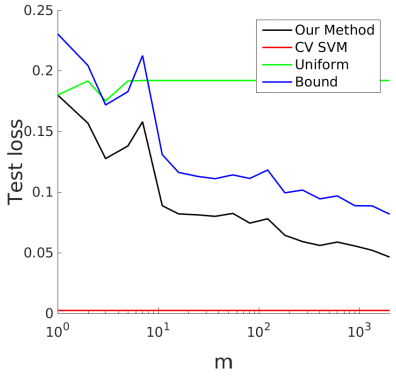
(*b*) Haberman dataset. $n = 150$, $r = \sqrt{n} = 12$.



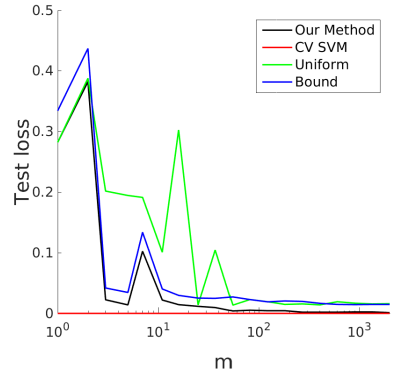
(*c*) AvsB dataset. $n = 1000$, $r = d + 1 = 17$.



(*d*) Waveform dataset. $n = 2000$, $r = d + 1 = 41$.

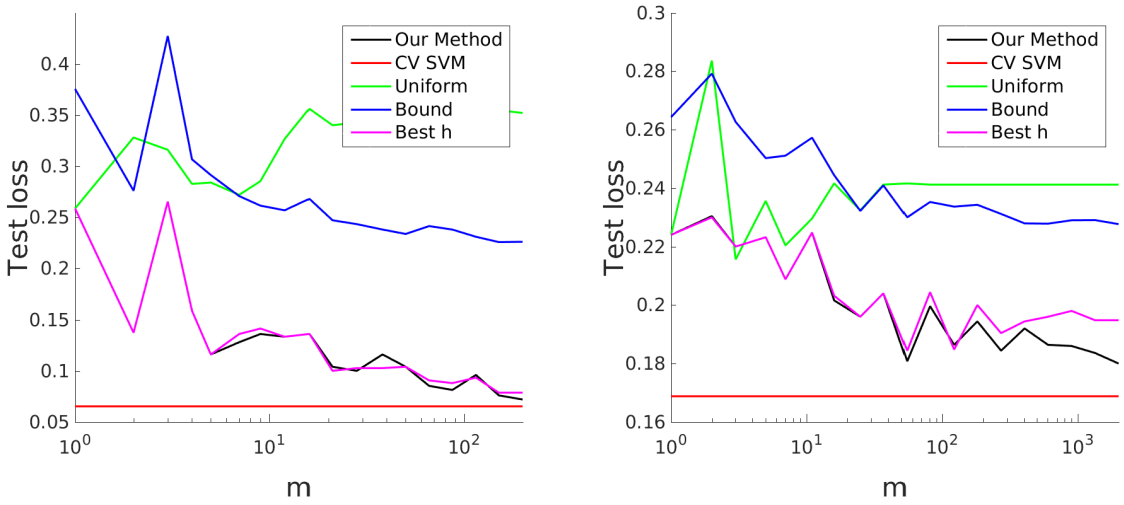


(*e*) Skin dataset. $n = 2000$, $r = \sqrt{n} = 45$.



(*f*) Mushrooms dataset. $n = 2000$, $r = d + 1 = 113$.

Figure 6: **Prediction performance of ρ-weighted majority vote ("Our Method"), uniform majority vote ("Uniform"), and cross-validated SVM ("CV SVM") together with the PAC-Bayes kl bound ("Bound").**



(*a*) Ionosphere dataset. $n = 200$, $r = d + 1 = 35$. (*b*) Adult dataset. $n = 2000$, $r = d + 1 = 123$.

Figure 7: **Prediction performance of weighted majority vote, uniform majority vote, SVM corresponding to the maximum of ρ (Best h), and cross-validated SVM together with the PAC-Bayes-kl bound**.

Name	|S|	|V|	|T|
Mushrooms	2000	500	1000
Skin	2000	500	1000
Waveform	2000	600	708
Adult	2000	500	685
Ionosphere	150	75	126
AvsB	700	500	355
Haberman	150	50	106
Breast cancer	300	100	283

Table 3: **Sizes of dataset partitions used in Figure 8**. |S| refers to the size of the training set, |V| refers to the size of the validation set, and |T| refers to the size of the test set.

is partitioned into training, validation, and test sets. The grid of λ-s was constructed by taking nine evenly spaced values in $[0.05; 1.9]$. For each λ we evaluated on the validation set the performance of the majority vote weighted by the distribution $\rho(\lambda)$ defined in equation (7), and picked the one with the lowest validation error. Note that the grid search had access to additional validation data that was unavailable to the alternating minimization procedure. Figure 8 presents the results. We conclude that the bound minimization performed comparably to validation in our experiments.

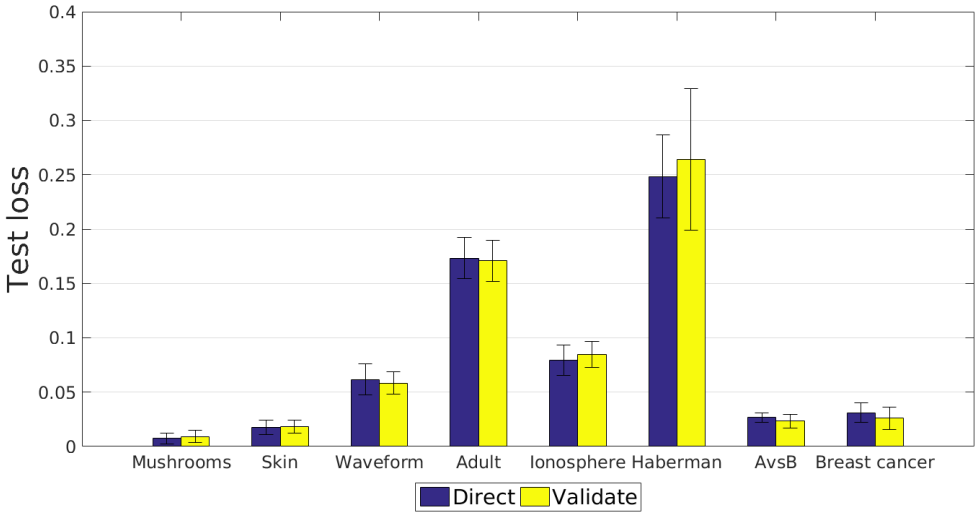


Figure 8: **Comparison of the Alternating Minimization with Grid Search for Selection of λ.** We show the loss on the test set obtained by direct minimization of λ ("Direct") and grid search ("Validate"). Error bars correspond to one standard deviation over 5 splits of the data into training, validation, and test set.

D.4. Comparison of ρ-weighted Majority Vote with Randomized Classifier and Empirically Best Classifier

In this section we compare the performance of ρ-weighted majority vote with the performance of randomized classifier defined by ρ and the performance of the best out of m weak classifiers (measured by the validation loss). The comparison is provided in Figure 9. Furthermore, in Table 4 we provide the number of hypotheses that took up 50% of the posterior mass ρ. While the performance of the randomized classifier is close to the performance of the best weak classifier, the distribution of posterior mass ρ over several classifiers improves the generalization bound and reduces the risk of overfitting when m is large. In other words, randomized classifier makes learning with large m safer. In our experiments the majority vote provided slight, but not significant improvement over the randomized classifier.

Name	$\#(h)$ that make 50% of ρ-mass
Mushrooms	2
Skin	1
Waveform	3
Adult	4
Ionosphere	2
Haberman	26
AvsB	2
Breast cancer	12

Table 4: **The number of hypotheses that make up 50% of the posterior mass ρ.**

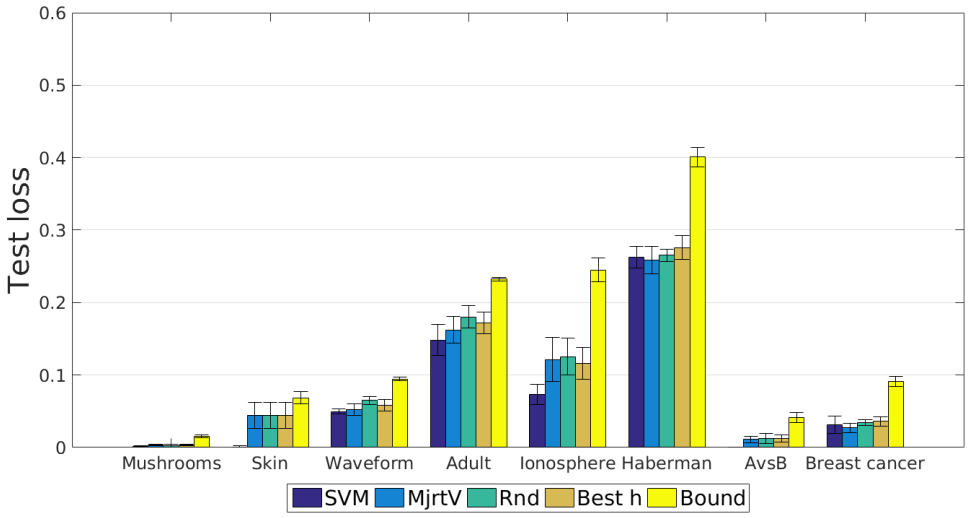


Figure 9: **Comparison of cross-validated SVM (SVM), ρ-weighted majority vote (MjrtV), randomized classifier ρ (Rnd), the best (empirically) out of m weak classifiers (Best h), and the PAC-Bayesian bound (Bound).** The comparison is for the maximal value of m ($m = n$) and the same values of r as given in the main experiments in Figures 2 and 5. Error bars correspond to one standard deviation over 5 splits of the data into training, validation, and test set.