Optimal PAC-Bayesian Posteriors for Stochastic Classifiers and their use for Choice of SVM Regularization Parameter

Puja Sahu PUJA.SAHU@IITB.AC.IN and **Nandyala Hemachandra** NH@IITB.AC.IN
Indian Institute of Technology Bombay, Mumbai, India

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

PAC-Bayesian set up involves a stochastic classifier characterized by a posterior distribution on a classifier set, offers a high probability bound on its averaged true risk and is robust to the training sample used. For a given posterior, this bound captures the trade off between averaged empirical risk and KL-divergence based model complexity term. Our goal is to identify an optimal posterior with the least PAC-Bayesian bound. We consider a finite classifier set and 5 distance functions: KL-divergence, its Pinsker's and a sixth degree polynomial approximations; linear and squared distances. Linear distance based model results in a convex optimization problem and we obtain a closed form expression for its optimal posterior. For uniform prior, this posterior has full support with weights negative-exponentially proportional to number of misclassifications. Squared distance and Pinsker's approximation bounds are possibly quasi-convex and are observed to have single local minimum. We derive fixed point equations (FPEs) using partial KKT system with strict positivity constraints. This obviates the combinatorial search for subset support of the optimal posterior. For uniform prior, exponential search on a full-dimensional simplex can be limited to an ordered subset of classifiers with increasing empirical risk values. These FPEs converge rapidly to a stationary point, even for a large classifier set when a solver fails. We apply these approaches to SVMs generated using a finite set of SVM regularization parameter values on 9 UCI datasets. The resulting optimal posteriors (on the set of regularization parameters) yield stochastic SVM classifiers with tight bounds. KL-divergence based bound is the tightest, but is computationally expensive due to its non-convex nature and multiple calls to a root finding algorithm. Optimal posteriors for all 5 distance functions have lowest 10% test error values on most datasets, with that of linear distance being the easiest to obtain.

Keywords: KL divergence, generalized Pinsker's inequality, convex optimization, constrained non-convex optimization, Fixed Point Equations, averaged true risk, Bayesian posterior, high probability bounds on true risk

1. Introduction and Motivation

Often we are faced with the issue of choosing a parameter of the learning algorithm, since this parameter has a significant role in determining the performance of the resulting classifier. For example, consider the Support Vector Machine (SVM) algorithm for classification with the regularization parameter, $\lambda > 0$. This parameter is a user input which trades off between model complexity and training error. The optimal classifier that we get, depends heavily on the sample S that is used for training and the value of the parameter, λ. We can control only this parameter value for obtaining a classifier with low (training) error, *but not* the given data. For a given training sample, we can choose the best value of the parameter from a prefixed set of values, which yields a classifier with the lowest error. However, this is a long drawn process. Additionally, there is no guarantee that the chosen value will yield a classifier having low(est) error on another sample from the same distribution. This implies that the best parameter value is sample dependent and that there is no unique value which is best for almost all the samples. However, if we determine the set of λ values with lowest $\rho\%$ error rates on each sample, we observe a recurring subset of λ values across the training samples. (*See Appendix A in the Suppl. file for an illustration.*) Thus, we have an ensemble of values to pick from. The PAC-Bayesian approach does such a stochastic selection.

PAC-Bayesian Bounds and Optimal Posteriors PAC-Bayesian approach assumes an arbitrary but fixed prior distribution on the space of classifiers and outputs a posterior distribution on this space, corresponding to a stochastic classifier. This approach provides a probabilistic bound on the difference between the posterior averaged true and empirical risk of a stochastic classifier as measured by a convex distance function. For a given posterior, these bounds offer a trade-off between averaged empirical risk and a term which encompasses model complexity of the stochastic classifier. The bound is computed based on a single sample but with a high probability guarantee over different samples (from the same distribution). We are interested in the 'optimal PAC-Bayesian posterior'. For a chosen distance function, the optimal posterior is defined as the one which minimizes the corresponding PAC-Bayesian bound. By design, these bounds and the resulting optimal posterior are robust to the choice of training sample, addressing the above sample bias.

Relevant Work PAC-Bayesian bounds were proposed by McAllester (2003); Seeger (2002) and refined further by Maurer (2004); Langford (2005); McAllester (2013) using Bayesian priors and posteriors on the classifier space to provide better performance guarantees. Several authors improvised the bounds for the choice of distance function they considered. While Maurer (2004) provided a bound for the KL-divergence as the distance function, ϕ, by tightening up the threshold with a factor of $\sqrt{m}$ instead of m, Germain et al. (2009) generalized the framework of PAC-Bayesian bounds for a broader class of convex ϕ functions and relaxed the constraints on tail bounds of empirical risks of the classifiers. Catoni (2007) made an important contribution by considering bounds which are independent of distance function ϕ, and instead require a parameter $C > 0$. Choice of C can influence the bound on the performance of stochastic classifier just as the choice of ϕ. Ambroladze et al. (2006) specialized PAC-Bayesian bounds using spherical Gaussian distributions on the space of linear classifiers. Bégin et al. (2016) introduced bounds based on Rényi divergence between posterior and prior distributions. We limit ourselves to KL-divergence based bounds.

All of the above consider a continuous (SVM) classifier space (n-dimensional Euclidean space) and continuous prior as well as posterior distributions on it (spherical Gaussian distributions) whereas we consider a finite set of classifiers such as those generated by a finite set of regularization parameter values for the SVM. Our PAC-Bayesian bounds are derived for the set up with a discrete prior distribution, and five different distance functions between posterior averaged empirical risk and posterior averaged true risk.

Contributions We consider *optimal PAC-Bayesian posterior* which minimizes the PAC-Bayesian bound for a given distance function. We consider a finite classifier set and five distance functions: KL-divergence and its two approximations based on Pinsker's inequality and its improvised version (a sixth degree polynomial), linear distance and squared distance. The linear distance based optimal posterior is obtained via a convex program; is shown to have full support, with weights proportional to negative-exponential number of misclassifications when prior is uniform. Bounds based on KL-divergence as distance function and its sixth degree approximation are non-convex. Squared distance and Pinsker's approximation are possibly quasi-convex because they are observed to have single local minimum. We simplify the search for optimal posteriors via Fixed Point equations deduced from the partial KKT system with strict positivity constraints. We use these approaches on the set of SVMs generated by a finite set of regularization parameter values. This leads us to the notion of a *stochastic SVM* characterized by an optimal posterior on the regularization parameter set. KL-distance yields the tightest bound, but is non-convex and has computational overhead of determining the root. All five distance functions have good generalization performance (lowest 10% test error values) on most datasets considered, except for Bupa dataset and two almost linearly separable datasets, Banknote and Mushroom. Table 1 describes theoretical and computational aspects of these optimal posteriors.

Outline In Section 2, we consider PAC-Bayesian optimal posterior as the one minimizing the bound, and propose a Fixed Point (FP) scheme based on the partial KKT system. We analyze optimal posteriors for five distance functions: KL-distance (Section 4), its approximations (Section 5), linear and squared distances (Sections 6 and 7). These approaches are applied to a set of SVMs (Section 8) with summary in Section 9.

2. PAC-Bayesian Bound Minimization, Optimal Posteriors and the Fixed Point Approach

We recall the general version of the PAC-Bayesian theorem Germain et al. (2009); Bégin et al. (2016) for a given distance function and describe the notion of a PAC-Bayesian optimal posterior which minimizes the bound derived from the PAC-Bayesian theorem.

Theorem 1 (PAC-Bayesian Theorem Germain et al. (2009); Bégin et al. (2016)) *For any data distribution $\mathcal{D}$ over input space $\mathcal{X} \times \mathcal{Y}$, the following bound holds for any prior P over the set of classifiers $\mathcal{H}$ and any $\delta \in (0,1)$, where the probability is over random i.i.d. samples $S_m = \{(x_i, y_i) | i = 1, \ldots, m\}$ of size m drawn from $\mathcal{D}$, for any convex function $\phi : [0,1] \times [0,1] \to \mathbb{R}$:*

$$\mathbb{P}_{S_m}\left\{\forall Q \ on \ \mathcal{H}: \ \phi\left(\mathbb{E}_Q[\hat{l}], \mathbb{E}_Q[l]\right) \le \frac{KL[Q||P] + \ln\left(\frac{\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{h\sim P}e^{m\phi(\hat{l},l)}}{\delta}\right)}{m}\right\} \ge 1-\delta. \quad (1)$$

Here, Q is an arbitrary posterior distribution on $\mathcal{H}$, which may depend on the sample S_m and on the prior P. $\mathbb{E}_Q[\hat{l}] := \mathbb{E}_{h\sim Q}\sum_{i=1}^m \frac{1}{m}[l(h, \mathbf{x}_i, y_i)]$ denotes the averaged empirical risk and $\mathbb{E}_Q[l] := \mathbb{E}_{h\sim Q}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[l]$ denotes averaged true risk of a classifier $h \in \mathcal{H}$ computed using a loss function, $l(h, \mathbf{x}, y) : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [a, b)$ (here, $0 \le a < b$).

For a choice of distance function, ϕ, the upper bound on $\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{h\sim P}e^{m\phi(\hat{l}(h),l(h))}$ determines the tightness of PAC-Bayesian bound. Bégin et al. (2016) give $\mathcal{I}_\phi^K(m) := \sup_{l\in[0,1]}\left[\sum_{k=0}^m \binom{m}{k} l^k (1-l)^{m-k} e^{m\phi(\frac{k}{m},l)}\right]$ as an upper bound on $\mathbb{E}_{S_m\sim\mathcal{D}^m}\mathbb{E}_{h\sim P}e^{m\phi(\hat{l}(h),l(h))}$.

Distance fn ϕ	**Theoretical Aspects**			
	$\mathcal{I}_{\phi}^{\mathbf{K}}(m)$	**Convexity**	**Global min**	**Fixed Point (FP)**
ϕ_{lin}	Not required	Convex	$q_i^* \propto p_i e^{-m\hat{l}_i}$	Not required
ϕ_{sq}	$\sum\limits_{k=0}^{m} \binom{m}{k} 0.5^m e^{m\left(\frac{k}{m}-0.5\right)^2}$ approximated by $2\sqrt{m}$	possibly quasi-convex	closed form may not exist	$q_{i,\text{sq, KL}}^{FP} \propto p_i e^{\left(-2\sqrt{m}\hat{l}_i \sqrt{\sum\limits_{i=1}^{H} q_{i,\text{sq, KL}}^{\text{FP}} \ln \frac{q_{i,\text{sq, KL}}^{\text{FP}}}{p_i} + \ln \frac{\mathcal{I}_{\text{sq}}^{K}(m)}{\delta}}\right)}$
kl	$2\sqrt{m}$ (due to Maurer (2004))	Non-convex; Difference of Convex (DC) functions	closed form may not exist	$q_{i,\text{kl, KL}}^{FP}$ satisfies: $q_i = p_i \exp\left\{ \sum\limits_{i=1}^{H} q_i \ln \frac{q_i}{p_i} - m\left(\sum\limits_{i=1}^{H} \hat{l}_i q_i - \hat{l}_i\right) \left[\ln\left(\frac{(1-r)\sum\limits_{i=1}^{H} \hat{l}_i q_i}{r(1-\sum\limits_{i=1}^{H} \hat{l}_i q_i)} \right) \right] \right\}$
ϕ_{P}	approximated by $2\sqrt{m}$	possibly quasi-convex	closed form may not exist	$q_{i,\text{P, KL}}^{FP} \propto p_i e^{\left(-2\sqrt{2m}\hat{l}_i \sqrt{\sum_{i=1}^{H} q_{i,\text{P, KL}}^{FP} \ln \frac{q_{i,\text{P, KL}}^{FP}}{p_i} + \ln \frac{2\sqrt{m}}{\delta}}\right)}$
ϕ_{CH}	$0.9334m$ (due to Sahu and Hemachandra (2018))	shown to be non-convex	closed form may not exist	$q_{i,\text{CH, KL}}^{FP} \propto p_i \exp\left\{-(2m-1)\hat{l}_i \frac{2\sqrt{r_{\text{CH}}(R(Q_{\text{CH, KL}}^{FP}))}}{\frac{\partial r_{\text{CH}}}{\partial R}}\right\}$ ($r_{\text{CH}}(R(Q))$ is the root of ϕ_{CH} for a given $\mathbb{E}_Q[\hat{l}]$ in (16))

Distance fn ϕ	**Computations**		
	Solver (Ipopt) output	**Global minima**	**Fixed Point (FP)**
ϕ_{lin}	identifies global minima	identified analytically	Not required
ϕ_{sq} ϕ_{P} ϕ_{CH}	identifies a unique (local) minima even with different initializations	closed form may not exist	matches solver output
kl	identifies multiple local minima with different initializations; throws up error for large H; especially for almost separable data	closed form may not exist	identifies same stationary point even with different initializations

Table 1: An outline of theoretical aspects and computational results for optimal posteriors $Q_{\phi,\text{KL}}^* = \{q_{i,\phi,\text{KL}}^*\}_{i=1}^{H}$ for minimization of the PAC-Bayesian bound, $B_{\phi,\text{KL}}(Q)$, based on KL-divergence $KL[Q||P] = \sum_{i=1}^{H} q_i \ln \frac{q_i}{p_i}$ between a posterior Q and a prior P on the classifier space $\mathcal{H}$. We consider five different distance functions, ϕ: KL-divergence $kl(\hat{l}, l) = \hat{l} \ln \frac{\hat{l}}{l} + (1-\hat{l}) \ln \left(\frac{1-\hat{l}}{1-l}\right)$, its Pinsker's approximation $\phi_{\text{P}}(\hat{l}, l) = 2(l-\hat{l})^2$ and a tighter approximation (a sixth degree polynomial) $\phi_{\text{CH}} = (l-\hat{l})^2 + \frac{2}{9}(l-\hat{l})^4 + \frac{16}{135}(l-\hat{l})^6$; linear $\phi_{\text{lin}}(\hat{l}, l) = l - \hat{l}$ and squared distances $\phi_{\text{sq}}(\hat{l}, l) = (l-\hat{l})^2$ for $l, \hat{l} \in (0,1)$. H denotes the classifier set size and H^* denotes the size of the support set of the optimal posterior $Q_{\phi,\text{KL}}^*$. $\hat{l}_i$ denotes empirical risk value of classifier $h_i \in \mathcal{H}$ computed on a sample of size m. $\mathcal{I}_{\phi}^{K}(m)$ is a sample size based constant for a distance function ϕ. It is a component of the bound function $B_{\phi,\text{KL}}(Q)$.

Thus, with the above upper bound on the right hand side threshold, (1) becomes:

$$\mathbb{P}_{S_m}\left\{\forall Q \text{ on } \mathcal{H}:\ \phi\left(\mathbb{E}_Q[\hat{l}], \mathbb{E}_Q[l]\right) \leq \frac{KL[Q||P] + \ln\left(\frac{\mathcal{I}_\phi^K(m)}{\delta}\right)}{m}\right\} \geq 1-\delta. \tag{2}$$

For illustrating the role of this upper bound, $Q^*_{\text{sq, KL}}$ is computed with two values: $\mathcal{I}_{sq}^K(m)$ defined by Bégin et al. (2016) and $2\sqrt{m}$ by Maurer (2004). Bounds with $\mathcal{I}_{sq}^K(m)$ are tighter than those with $2\sqrt{m}$, and test error rates increase only marginally (Please see Table 2).

2.1. Optimal posteriors via PAC-Bayesian bound minimization

The PAC-Bayesian theorem (2) gives the following high probability upper bound on averaged true risk, $\mathbb{E}_Q[l]$, assuming distance function $\phi(\mathbb{E}_Q[\hat{l}],\cdot)$ is invertible for given $\mathbb{E}_Q[\hat{l}]$:

$$B_{\phi,\text{KL}}(Q) \equiv B_{\phi,\text{KL}}(\mathbb{E}_Q[\hat{l}], S_m, \delta, P) = f_\phi\left(\mathbb{E}_Q[\hat{l}], \phi^{-1}_{\mathbb{E}_Q[\hat{l}]}\left(\frac{KL[Q||P] + \ln\left(\frac{\mathcal{I}_\phi^K(m)}{\delta}\right)}{m}\right)\right), \tag{3}$$

where $\phi^{-1}_{\mathbb{E}_Q[\hat{l}]}(K) = b$ implies $\phi(\mathbb{E}_Q[\hat{l}], b) = K$ for some $b \in (0,1)$ and a given $K > 0$. Generally $f_\phi(\cdot,\cdot)$ is the sum of its arguments except when ϕ is KL-distance function. That is, bound function $B_{\phi,\text{KL}}(Q)$ is the sum of averaged empirical risk, $\mathbb{E}_Q[\hat{l}]$, and a model complexity term which depends on system parameters, S_m, δ, P. We are interested in determining an optimal posterior distribution $Q^*_{\phi,\text{KL}}$ which minimizes $B_{\phi,\text{KL}}(Q)$ for a given ϕ.

2.2. The fixed point approach to determine PAC-Bayesian optimal posterior

To characterize the minimum of $B_{\phi,\text{KL}}(Q)$, we make use of the first order KKT conditions which are necessary for a stationary point of a non-convex problem. These KKT conditions require the objective function and the active constraints to be differentiable at the local minimum. We derive fixed point (FP) equations for the optimal posterior for various distance functions in (8), (12), (17) and (25) (with derivations in supplemenatry file). These FP equations use KKT system with strict positivity constraints due to which complementary slackness conditions are automatically satisfied; hence called '*partial*' KKT system. We consider strict positivity constraints on posterior weights to avoid the combinatorial problem of choosing the subset of classifiers which form the support set of the optimal posterior. Computations illustrate that these FP equations always converge to a stationary point at a very fast rate, even for a large classifier set when a non-convex solver fails to identify a local solution. (Please see Table 3 for an illustration of such cases.)

We work with a finite set of classifiers: $\mathcal{H} = \{h_i\}_{i=1}^H$ of size H. The prior, $P = \{p_i\}_{i=1}^H$ and posterior, $Q = \{q_i\}_{i=1}^H$ are discrete distributions on $\mathcal{H}$, where $p_i, q_i \geq 0\ \forall i = 1,\ldots,H$ with $\sum_{i=1}^H p_i = 1$ and $\sum_{i=1}^H q_i = 1$. For differentiability required by KKT conditions, our objective function should have open domain, that is, the interior of the H-dimensional probability simplex: $int(\Delta^H) = \{(q_1,\ldots,q_H)|q_i > 0\ \forall i = 1,\ldots,H; \sum_{i=1}^H q_i = 1\}$. In computations, we consider $q_i \geq \epsilon\ \forall i = 1,\ldots,H$ for $\epsilon > 0$ to ensure existence of a minimizer in $int(\Delta^H)$. Our FP equations are derived using partial KKT system on $int(\Delta^H)$.

3. Optimal posterior, $Q^*_{\phi,\mathbf{KL}}$, for uniform prior

We consider the special case of uniform prior on entire $\mathcal{H}$. We want to identify the optimal posterior $Q^*_{\phi,\mathrm{KL}}$ with the H-dimensional probability simplex as the feasible region. We show below that it is enough to restrict the search space to certain subsets of this simplex. This reduces the computational complexity of the search from exponential scale to linear scale.

Theorem 2 *Consider a uniform prior distribution on the set $\mathcal{H}$ of classifiers, and a given set of posterior weights $Q = \{q_j\}_{j=1}^{H'}$. We have three choices of distance function $\phi = \{\phi_{lin}, \phi_{sq}, kl\}$. Then among all subsets $\mathcal{H}' \subset \mathcal{H}$ of size H', the smallest bound value $B_{\phi,KL}(Q, \mathcal{H}')$ corresponding to the given posterior weights Q is achieved when $\mathcal{H}'$ is the subset formed by the first H' elements of the ordered set of classifiers ranked by non-decreasing empirical risk values, $\hat{l}_1 \leq \hat{l}_2 \leq \ldots \leq \hat{l}_H$.*

Proof (*Please see Appendix C in suppl. file for other distance functions*) We consider linear distance based bound, $B_{\text{lin, KL}}(Q, \mathcal{H}')$ under the given set up, defined as follows:

$$B_{\text{lin, KL}}(Q, \mathcal{H}') := \sum_{i\in\mathcal{H}'} \hat{l}_i q_i + \frac{\sum\limits_{i\in\mathcal{H}'} q_i \ln q_i + \ln H + \ln\left(\frac{\mathcal{I}^K_{\text{lin}}(m)}{\delta}\right)}{m} \tag{4}$$

For a given set of posterior weights $\{q_j\}_{j=1}^{H'}$, the term $\sum_{i\in\mathcal{H}'} q_i \ln q_i$ of the bound $B_{\text{lin, KL}}(Q, \mathcal{H}')$ is invariant of the support set $\mathcal{H}'$ as long as its cardinality is H'. Thus $B_{\text{lin, KL}}(Q, \mathcal{H}')$ is the smallest when the sum $\sum_{i\in\mathcal{H}'} \hat{l}_i q_i$ is minimized. This will happen when $\mathcal{H}'$ consists of classifiers with smallest H' values in the set $\{\hat{l}_i\}_{i=1}^{H}$. Furthermore, if the elements of $\mathcal{H}'$ are ordered by non-decreasing empirical risk values, $\hat{l}_1 \leq \hat{l}_2 \leq \ldots \leq \hat{l}_{H'}$, the weights $\{q_j\}_{j=1}^{H'}$ should be ordered non-increasingly. So, the theorem holds for linear distance function. ■

Corollary 1 *As a consequence of the above Theorem 2, for determining the (globally) optimal posterior $Q^*_{\phi,KL}$, it is sufficient to compare the bound values corresponding to the best posteriors on ordered subsets of $\mathcal{H}$, ranked by non-decreasing $\hat{l}_i$ values. These ordered subsets can be uniquely identified by their size.*

4. Optimal PAC-Bayesian Posterior using KL-distance

The most commonly referenced version of the PAC-Bayesian theorem was given by Seeger (2002) and improved by Maurer (2004), as given below:

Theorem 3 (PAC-Bayesian Theorem for KL-distance Maurer (2004)) *For any data distribution $\mathcal{D}$ over input space $\mathcal{X} \times \mathcal{Y}$, the following bound holds for any prior P over the set of classifiers $\mathcal{H}$ and any $\delta \in (0, 1)$, where the probability is over random i.i.d. samples $S_m = \{(x_i, y_i) | i = 1, ..., m\}$ of size m drawn from $\mathcal{D}$:*

$$\mathbb{P}_{S_m}\left\{\forall Q \text{ on } \mathcal{H}: \, kl\left(\mathbb{E}_Q[\hat{l}], \mathbb{E}_Q[l]\right) \leq \frac{KL[Q||P] + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}{m}\right\} \geq 1 - \delta. \tag{5}$$

Here, Q is an arbitrary posterior distribution on $\mathcal{H}$, which may depend on the sample S_m and on the prior P, and where $kl(p,q) = p\ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right)$ for any $p, q \in (0,1)$.

The upper bound on the averaged true risk $\mathbb{E}_Q[l]$ corresponding to the above PAC-Bayesian theorem is obtained as:

$$B_{\text{kl, KL}}(Q) = \sup_{r\in(0,1)} \left\{ r : kl\left(\mathbb{E}_Q[\hat{l}], r\right) \leq \frac{KL[Q||P] + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}{m} \right\} \tag{6}$$

An inverse $kl(\cdot,\cdot)$ function does not exist since it is not a monotone function, and so the bound $B_{\text{kl, KL}}(Q)$ does not have an explicit form. However, we can employ a numerical root finding algorithm such as that described in Sahu and Hemachandra (2018) (Algo. (KLroots)) to obtain $B_{\text{kl, KL}}(Q)$ for a given instance of system parameters.

4.1. The KL-distance bound minimization problem

For a finite classifier space $\mathcal{H} = \{h_i\}_{i=1}^H$, this optimization problem can be described as:

$$\min_{q_1,\ldots,q_H, r} r \tag{7a}$$

$$\text{s.t.} \quad \left(\sum_{i=1}^H \hat{l}_i q_i\right) \ln\left(\frac{\sum\limits_{i=1}^H \hat{l}_i q_i}{r}\right) + \left(1 - \sum_{i=1}^H \hat{l}_i q_i\right) \ln\left(\frac{1 - \sum\limits_{i=1}^H \hat{l}_i q_i}{1-r}\right) = \frac{\sum\limits_{i=1}^H q_i \ln\frac{q_i}{p_i} + \ln\frac{2\sqrt{m}}{\delta}}{m} \tag{7b}$$

$$r \geq \sum_{i=1}^H \hat{l}_i q_i \tag{7c}$$

$$\sum_{i=1}^H q_i = 1,\ q_i \geq 0,\ \forall i = 1,\ldots,H \tag{7d}$$

Here, r is the right root of $kl\left(\mathbb{E}_Q[\hat{l}], r\right) = \frac{KL[Q||P]+\ln\left(\frac{2\sqrt{m}}{\delta}\right)}{m}$ for a given $\mathbb{E}_Q[\hat{l}]$. The above is known to be a non-convex problem with a difference of convex (DC) equality constraint (7b). The constraint (7c) is a strict inequality which is relaxed for modelling purpose to have a feasible region with a closed domain.

4.2. The posterior based on fixed point scheme, $Q^{\mathbf{FP}}_{\mathbf{kl,KL}}$

We derive FP equation for KL-distance based bound optimization problem below:

Theorem 4 *The bound minimization problem (7) for the bound $B_{kl,\ KL}(Q)$ has a stationary point $Q^{FP}_{kl,KL}$ which can be obtained as the solution to the following fixed point equation:*

$$q_i = p_i \exp\left\{\sum_{i=1}^H q_i \ln\frac{q_i}{p_i} - m\left(\sum_{i=1}^H \hat{l}_i q_i - \hat{l}_i\right)\left[\ln\left(\frac{(1-r)\sum_{i=1}^H \hat{l}_i q_i}{r(1-\sum_{i=1}^H \hat{l}_i q_i)}\right)\right]\right\} \ \forall i = 1,\ldots,H \tag{8}$$

where r is the solution to (7b) and (7c) for a given $Q = (q_1,\ldots,q_H)$.

Proof The Lagrangian function for (7) can be written as follows:

$$\mathcal{L}_{\text{kl, KL}} = r - \beta_0 \left[\left(\sum_{i=1}^{H} \hat{l}_i q_i \right) \ln \left(\frac{\sum_{i=1}^{H} \hat{l}_i q_i}{r} \right) + \left(1 - \sum_{i=1}^{H} \hat{l}_i q_i \right) \ln \left(\frac{1 - \sum_{i=1}^{H} \hat{l}_i q_i}{1 - r} \right) \right. \\ \left. - \frac{\left(\sum_{i=1}^{H} q_i \ln \frac{q_i}{p_i} + \ln \frac{2\sqrt{m}}{\delta} \right)}{m} \right] - \beta_1 \left(r - \sum_{i=1}^{H} \hat{l}_i q_i \right) - \mu_0 \left(\sum_{i=1}^{H} q_i - 1 \right) - \sum_{i=1}^{H} \mu_i q_i \quad (9)$$

Due to the strict inequality constraint (7c), complementary slackness conditions for a stationary point imply that the Lagrange multiplier β_1 should vanish at optimality ($\beta_1 = 0$).

We assume that $q_i > 0 \forall i = 1, \ldots, H$, since otherwise $\ln q_i = \ln(0)$ is undefined. Even if we use fact that $\lim_{x \to 0^+} \ln x = -\infty$ to define $\frac{\partial \mathcal{L}_{\text{kl, KL}}}{\partial q_j}$ for some $j \in [H]$, the KKT condition will mean that μ_j is infeasible. Therefore, for a stationary point, we have $q_i > 0$. And the complementary slackness conditions imply that $\mu_i = 0$ for all $i = 1, \ldots, H$.

At an optimal solution, derivatives of $\mathcal{L}_{\text{kl, KL}}$ with respect to primal variables r and q_is, should be set to zero. By solving for these derivatives, we get the FP equation (8) which identifies a stationary point of (7). (*Please see Appendix D.1 in suppl. file for details.*) ■

Note: The requirement that $q_i > 0 \ \forall \ i = 1, \ldots, H$ holds true for the KKT system of a generic PAC-Bayesian bound minimization because of KL-divergence measure between posterior and prior distributions; so, we assume this condition for the other four ϕs also.

KL-distance based bound minimization is non-convex with multiple stationary points which makes it difficult to identify the global minimum even by FP scheme. The iterative root finding algorithm adds to the computational complexity of the bound minimization algorithm. Therefore, in the next section, we look for simpler and easily invertible approximations to KL-distance function in the PAC-Bayesian bound minimization.

5. Optimal Posterior for PAC-Bayesian Bound Minimization based on approximations to KL-distance function

We explore two approximations to the KL-distance function: a known Pinsker's approximation and another tighter approximation based on improvised Pinsker's inequality.

5.1. Optimal PAC-Bayesian Posterior based on Pinsker's approximation

Based on Pinsker's inequality Fedotov et al. (2003), we get the following second order polynomial approximation to $kl(l, l')$: $\phi_{\text{P}}(l, l') = 2(l - l')^2 \quad \forall l, l' \in [0, 1] \times [0, 1]$ which serves as a distance function in the PAC-Bayesian theorem:

$$\mathbb{P}_{S_m} \left\{ \forall Q \text{ on } \mathcal{H} : 2 \left(\mathbb{E}_Q \left[\hat{l} \right] - \mathbb{E}_Q[l] \right)^2 \leq \frac{KL[Q||P] + \ln \left(\frac{2\sqrt{m}}{\delta} \right)}{m} \right\} \geq 1 - \delta. \quad (10)$$

The associated PAC-Bayesian bound function is:

$$B_{\text{P, KL}}(Q) := \sum_{i=1}^{H} \hat{l}_i q_i + \sqrt{\frac{\sum_{i=1}^{H} q_i \ln \frac{q_i}{p_i} + \ln \left(\frac{2\sqrt{m}}{\delta} \right)}{2m}}. \quad (11)$$

Algorithm 1: FP klKL: Fixed point solution for PAC-Bayesian bound with KL-distance

```
Input: $\delta \in (0,1), m, H, \{\hat{l}_i\}_{i=1}^{H}, \{p_i\}_{i=1}^{H}, \texttt{tol} > 0$
Output: Fixed point solution: $\{q_{i,\text{kl, KL}}^{FP}\}_{i=1}^{H}$
/* Intialize $Q^0 = \{q_i^0\}_{i=1}^{H}$ with a random distribution from $\Delta^H$ simplex   */
$q_i^0 \sim \exp(1),\ \forall i = 1, \ldots, H$
$q_i^0 \leftarrow \frac{q_i^0}{\sum_{j=1}^{H} q_j^0}\ \forall i = 1, \ldots, H$
$RHS \leftarrow \frac{\sum_{i=1}^{H} q_i^0 \ln \frac{q_i^0}{p_i} + \ln \frac{2\sqrt{m}}{\delta}}{m}$
$r \leftarrow \text{KL}\textsc{roots}(\sum_{i=1}^{H} \hat{l}_i q_i^0, RHS)_2$
$q_i^1 \leftarrow p_i \exp\left\{\sum_{i=1}^{H} q_i^0 \ln \frac{q_i^0}{p_i} - m\left(\sum_{i=1}^{H} \hat{l}_i q_i^0 - \hat{l}_i\right)\left[\ln\left(\frac{(1-r)\sum_{i=1}^{H} \hat{l}_i q_i^0}{r(1-\sum_{i=1}^{H} \hat{l}_i q_i^0)}\right)\right]\right\}\ \forall i = 1, \ldots, H$
do
    for $i = 1$ to $H$ do
        $q_i^0 \leftarrow q_i^1$
    end
    $RHS \leftarrow \frac{\sum_{i=1}^{H} q_i^0 \ln \frac{q_i^0}{p_i} + \ln \frac{2\sqrt{m}}{\delta}}{m}$
    $r \leftarrow \text{KL}\textsc{roots}(\sum_{i=1}^{H} \hat{l}_i q_i^0, RHS)_2$
    $q_i^1 \leftarrow p_i \exp\left\{\sum_{i=1}^{H} q_i^0 \ln \frac{q_i^0}{p_i} - m\left(\sum_{i=1}^{H} \hat{l}_i q_i^0 - \hat{l}_i\right)\left[\ln\left(\frac{(1-r)\sum_{i=1}^{H} \hat{l}_i q_i^0}{r(1-\sum_{i=1}^{H} \hat{l}_i q_i^0)}\right)\right]\right\}\ \forall i = 1, \ldots, H$
while $\|q^1 - q^0\| > \texttt{tol}$
return $\{q_i^1\}_{i=1}^{H}$
```

We wish to determine the optimal posterior $Q^*_{\text{P, KL}}$ which minimizes $B_{\text{P, KL}}(Q)$ subject to the constraints given in (7d). The convexity of this bound function could not be established, but computationally this bound minimization problem is observed to have single local minimum. We propose that (11) is possibly quasi-convex. Based on the proof for Theorem 4 for KL-distance function, we identify the following FP equation for stationary point of (11):

$$q_{i,\text{P, KL}}^{FP} = \frac{p_i e^{\left(-2\sqrt{2m}\hat{l}_i \sqrt{\sum_{i=1}^{H} q_{i,\text{P, KL}}^{FP} \ln \frac{q_{i,\text{P, KL}}^{FP}}{p_i} + \ln \frac{2\sqrt{m}}{\delta}}\right)}}{\sum_{i=1}^{H} p_i e^{\left(-2\sqrt{2m}\hat{l}_i \sqrt{\sum_{i=1}^{H} q_{i,\text{P, KL}}^{FP} \ln \frac{q_{i,\text{P, KL}}^{FP}}{p_i} + \ln\left(\frac{2\sqrt{m}}{\delta}\right)}\right)}} \quad \forall i = 1, \ldots, H. \tag{12}$$

5.2. Optimal PAC-Bayesian Posterior based on improvised Pinsker's approximation, $\phi_{\mathbf{CH}}$

A lower bound for KL-divergence $kl(l, l')$ given by an improvised version of Pinsker's inequality Fedotov et al. (2003) is the following tighter sixth degree polynomial approximation:

$$\phi_{\text{CH}}(l, l') = (l - l')^2 + \frac{2}{9}(l - l')^4 + \frac{16}{135}(l - l')^6 \quad \forall l, l' \in [0,1] \times [0,1] \tag{13}$$

ϕ_{CH} is a valid distance function since it satisfies the Seeger's assumptions Seeger (2002).

Theorem 5 (Sahu and Hemachandra (2018)) *PAC-Bayesian theorem with ϕ_{CH} is:*

$$\mathbb{P}_{S_m}\left\{\forall Q \text{ on } \mathcal{H} : \phi_{CH}\left(\mathbb{E}_Q\left[\hat{l}\right], \mathbb{E}_Q[l]\right) \leq \frac{KL[Q||P] + \ln\left(\frac{K_{\phi_{CH}}}{\delta}\right)}{2m-1}\right\} \geq 1-\delta, \tag{14}$$

$$\textit{where } K_{\phi_{CH}} := 4m \times \left[1 - e^{-\phi_{CH}\left(\frac{1}{2}\right)}\right] \approx 0.9334m. \tag{15}$$

Due to its structure, $\phi_{\text{CH}}(\hat{l}, \cdot)$ has a single positive real root and has a PAC-Bayesian bound:

$$B_{\text{CH, KL}}(Q) := \mathbb{E}_Q[\hat{l}] + \sqrt{r_{\text{CH}}(R(Q))} \tag{16a}$$

$$\text{where, } r_{\text{CH}}(R(Q)) = -\frac{5}{8} + \sqrt[3]{\left(\frac{1225}{512} + \frac{135}{32}R(Q)\right) + \frac{5}{32}\sqrt{729R^2(Q) + \frac{6615}{8}R(Q) + \frac{208980}{256}}}$$
$$+ \sqrt[3]{\left(\frac{1225}{512} + \frac{135}{32}R(Q)\right) - \frac{5}{32}\sqrt{729R^2(Q) + \frac{6615}{8}R(Q) + \frac{208980}{256}}}, \tag{16b}$$

$$R(Q) = \frac{KL[Q||P] + \ln\left(\frac{K_{\phi_{\text{CH}}}}{\delta}\right)}{2m-1} = \frac{\sum_{i=1}^{H} q_i \ln \frac{q_i}{p_i} + \ln\left(\frac{K_{\phi_{\text{CH}}}}{\delta}\right)}{2m-1}. \tag{16c}$$

The optimal posterior distribution $Q^*_{\text{CH, KL}}$ is the one which minimizes $B_{\text{CH, KL}}(Q)$ in (16).

Lemma 1 *The bound function $B_{CH,\ KL}(Q)$ defined in (16) is a non-convex function and hence the associated bound minimization problem is non-convex program.*

We identify the following FP equation for a stationary point for minimizing (16), based on the partial KKT system:

$$q^{FP}_{i,\text{CH, KL}} = \frac{p_i \exp\left\{-(2m-1)\hat{l}_i \frac{2\sqrt{r_{\text{CH}}(R(Q^{FP}_{\text{CH, KL}}))}}{\frac{\partial r_{\text{CH}}}{\partial R}}\right\}}{\sum_{i=1}^{H} p_i \exp\left\{-(2m-1)\hat{l}_i \frac{2\sqrt{r_{\text{CH}}(R(Q^{FP}_{\text{CH, KL}}))}}{\frac{\partial r_{\text{CH}}}{\partial R}}\right\}} \quad \forall i = 1, \ldots, H. \tag{17}$$

6. Optimal PAC-Bayesian Posterior using Linear Distance Function

One of the simplest distance functions is the linear distance function, $\phi_{\text{lin}}(\hat{l}, l) = l - \hat{l}$ for $\hat{l}, l \in [0, 1]$. The PAC-Bayesian bound in this case takes the following simplified form:

$$\mathbb{P}_{S_m}\left\{\forall Q \text{ on } \mathcal{H} : \mathbb{E}_Q[l] - \mathbb{E}_Q[\hat{l}] \leq \frac{KL[Q||P] + \ln\left(\frac{\mathcal{I}^K_{\text{lin}}(m)}{\delta}\right)}{m}\right\} \geq 1-\delta \tag{18}$$

where $\mathcal{I}^K_{\text{lin}}(m) := \sup\limits_{l\in[0,1]} \left[\sum_{k=0}^{m} \binom{m}{k} l^k (1-l)^{m-k} e^{m\left(l-\frac{k}{m}\right)}\right]$ is a function of the sample size, m.

Thus, the corresponding PAC-Bayesian bound is:

$$B_{\text{lin, KL}}(Q) := \mathbb{E}_Q[\hat{l}] + \frac{KL[Q||P] + \ln\left(\frac{\mathcal{I}^K_{\text{lin}}(m)}{\delta}\right)}{m}. \tag{19}$$

We want to find the optimal distribution $Q^*_{\text{lin, KL}}$ which minimizes the bound $B_{\text{lin, KL}}(Q)$.

Remark 1 *For $m \geq 1028$, computing $\mathcal{I}^K_{lin}(m)$ is difficult due to storage limitations in the range of floating point numbers – gives $\mathcal{I}^K_{lin}(m)$ as NaN. As it is just an additive term in the bound, it does not influence the optimal solution. Hence we can determine $Q^*_{lin,\ KL}$ even for large m as shown in Table 2, but is needed for computing $B_{lin,\ KL}(Q^*_{lin,\ KL})$.*

6.1. The linear distance bound minimization problem

For a finite classifier space $\mathcal{H} = \{h_i\}_{i=1}^H$, this optimization problem can be described as:

$$\begin{aligned} \min_{q_1,\ldots,q_H} \; & \sum_{i=1}^H \hat{l}_i q_i + \frac{\sum_{i=1}^H q_i \ln \frac{q_i}{p_i}}{m} \\ \text{s. t.} \; & \sum_{i=1}^H q_i = 1, \; q_i \geq 0 \quad \forall i = 1, \ldots, H. \end{aligned} \tag{20}$$

6.2. Convexity of the bound function, $B_{\text{lin, KL}}(Q)$

The bound function $B_{\text{lin, KL}}(Q)$ is convex in Q since it is a positive affine transformation of $KL[Q\|P]$, which in turn is convex in Q. Also, the feasible region is the H-dimensional probability simplex which is a closed convex set. Hence (20) is a convex optimization problem. Thus, KKT conditions are both necessary and sufficient for (20).

6.3. The optimal posterior, $Q^*_{\text{lin, KL}}$

Theorem 6 *The distribution $Q^*_{lin,\ KL} = (q^*_{1,lin,\ KL}, \ldots, q^*_{H,lin,\ KL})$ where*

$$q^*_{i,lin,\ KL} = \frac{p_i e^{-m\hat{l}_i}}{\sum_{i=1}^H p_i e^{-m\hat{l}_i}} \;\; \forall i = 1, \ldots, H \tag{21}$$

is the optimal PAC-Bayesian posterior which minimizes the bound $B_{lin,\ KL}(Q)$ in (19).

Proof Since this is a differentiable convex OP, we identify the global minimizer (21) using the associated KKT system. (*Please refer to details in Appendix E.2 in suppl. file*) ■

Remark 2 *$Q^*_{lin,\ KL}$ in (21) is a Boltzmann distribution for a given P. In case of uniform prior, the optimal posterior weight ($q^*_{i,lin,\ KL}$) on a classifier is negative-exponentially proportional to the number of misclassifications ($m\hat{l}_i$) it makes on the (validation) sample.*

Theorem 7 *When the prior is a uniform distribution on the set $\mathcal{H}$ of classifiers, the optimal posterior $Q^*_{lin,\ KL}$ for the bound minimization problem (20) has full support. That is, all the classifiers in $\mathcal{H}$ will have strictly positive posterior weight at optimality.*

Proof Using the result of Theorem 2, it is sufficient to compare the bound values corresponding to the best posteriors for all ordered subsets of $\mathcal{H}$, ranked by non-decreasing $\hat{l}_i$ values, to determine the optimal posterior for (20). Using Theorem 6, the optimal posterior $Q^*_{\text{lin, KL}}(H')$ on an ordered subset of classifiers of size $H' \in [H]$ is given as:

$$q^*_{\text{i, lin, KL}}(H') = \begin{cases} \frac{e^{-m\hat{l}_i}}{\sum_{i=1}^{H'} e^{-m\hat{l}_i}} & \forall i = 1, \ldots, H' \\ 0 & \forall i = H'+1, \ldots, H, \end{cases}$$

and the optimal objective value is:

$$\begin{aligned} B_{\text{lin, KL}}(Q^*_{\text{lin, KL}}(H')) &= \sum_{i=1}^{H} \hat{l}_i q^*_{i,lin,KL} + \frac{\sum_{i=1}^{H} q^*_{i,lin,KL} \ln(q^*_{i,lin,KL} H)}{m} \\ &= \frac{\ln H - \ln\left(\sum_{i=1}^{H'} e^{-m\hat{l}_i}\right)}{m} \end{aligned}$$

The bound, $B_{\text{lin, KL}}(Q^*_{\text{lin, KL}}(H'))$ is a decreasing function of $H' = 1, \ldots, H$. Therefore the least bound value is achieved when all classifiers are assigned strictly positive weights, that is, the optimal posterior has full support. (*Details are in Appendix E.2 in suppl. file*) ■

Remark 3 *We believe that this full support for the optimal posterior, $Q^*_{lin,\ KL}$, is due to the KL-divergence measure on the right hand side threshold of the PAC-Bayesian bound, (18). As an implication, $Q^*_{lin,\ KL}$ considers even the worst performing classifier but with infinitesimally positive (negative-exponential) posterior weight.*

7. Optimal PAC-Bayesian Posterior using Squared Distance Function

We now consider a widely used squared distance function McAllester (2003); Seeger (2002) between the averaged empirical risk and the averaged true risk : $\phi_{\text{sq}}\left(\hat{l}, l\right) = \left(\hat{l} - l\right)^2$ for $\hat{l}, l \in [0,1]$. With ϕ_{sq}, the PAC-Bayesian theorem takes the following form:

$$\mathbb{P}_{S_m}\left\{\forall Q \text{ on } \mathcal{H} : \left(\mathbb{E}_Q[\hat{l}], \mathbb{E}_Q[l]\right)^2 \leq \frac{KL[Q||P] + \ln\left(\frac{\mathcal{I}^K_{\text{sq}}(m)}{\delta}\right)}{m}\right\} \geq 1 - \delta, \tag{22}$$

where $\mathcal{I}^K_{\text{sq}}(m) := \sup_{l\in[0,1]} \left[\sum_{k=0}^m \binom{m}{k} l^k (1-l)^{m-k} e^{m\left(\frac{k}{m}-l\right)^2}\right]$ is a function of the sample size, m.

The above PAC-Bayesian statement gives the following high probability upper bound:

$$B_{\text{sq, KL}}(Q) := \mathbb{E}_Q[\hat{l}] + \sqrt{\frac{KL[Q||P] + \ln\left(\frac{\mathcal{I}^K_{\text{sq}}(m)}{\delta}\right)}{m}}. \tag{23}$$

We identify the constant term $\mathcal{I}^K_{\text{sq}}(m)$ in (23) based on Bégin et al. (2016)'s result.

Lemma 2 *For a given sample size, m, $\mathcal{I}^K_{sq}(m) := \sum_{k=0}^{m} \binom{m}{k} 0.5^m e^{2m\left(\frac{k}{m}-0.5\right)^2}$.*

Remark 4 *On a machine equipped with 4 Intel Xeon 2.13 GHz cores and 64 GB RAM, we couldn't compute $\mathcal{I}^K_{sq}(m)$ for $m \geq 1028$ due to storage limitations for floating point numbers. Therefore, we upper bound it by $2\sqrt{m}$ for $m \geq 8$ Bégin et al. (2016).*

7.1. The squared distance bound minimization problem

We want to determine the optimal posterior $Q^*_{\text{sq, KL}}$ which minimizes $B_{\text{sq, KL}}(Q)$. For a finite classifier space $\mathcal{H} = \{h_i\}_{i=1}^{H}$, this optimization problem can be described as:

$$\begin{aligned} &\min_{q_1,\ldots,q_H} \sum_{i=1}^{H} \hat{l}_i q_i + \sqrt{\frac{\sum_{i=1}^{H} q_i \ln \frac{q_i}{p_i} + \ln\left(\frac{\mathcal{I}^K_{\text{sq}}(m)}{\delta}\right)}{m}} \\ &\text{s. t.} \sum_{i=1}^{H} q_i = 1,\ q_i \geq 0 \quad \forall i = 1, \ldots, H. \end{aligned} \tag{24}$$

The convexity of this bound function could not be established, but computationally this bound minimization problem is observed to have a single local minimum, hinting at quasi-convexity of $B_{\text{sq, KL}}(Q)$. (*Please see Appendices F.1 and F.2 in Suppl. file for proof.*)

7.2. The posterior based on fixed point scheme, $Q^{FP}_{\textbf{sq,KL}}$

We can identify a FP solution for (24) based on the partial KKT system by setting the derivatives of the Lagrange function for (24) to zero, and using the complementary slackness conditions, we get the FP equation (25). (*Proof details are in Appendix F.3 in Suppl.file.*)

Theorem 8 *The bound minimization problem (24) has a stationary point which can be obtained as the solution to the following fixed point equation:*

$$q^{FP}_{i,sq,\ KL} = \frac{p_i e^{\left(-2\sqrt{m}\hat{l}_i \sqrt{\sum_{i=1}^{H} q^{FP}_{i,sq,\ KL} \ln \frac{q^{FP}_{i,sq,\ KL}}{p_i} + \ln \frac{\mathcal{I}^K_{sq}(m)}{\delta}}\right)}}{\sum_{i=1}^{H} p_i e^{\left(-2\sqrt{m}\hat{l}_i \sqrt{\sum_{i=1}^{H} q^{FP}_{i,sq,\ KL} \ln \frac{q^{FP}_{i,sq,\ KL}}{p_i} + \ln\left(\frac{\mathcal{I}^K_{sq}(m)}{\delta}\right)}\right)}}, \quad \forall i = 1, \ldots, H \tag{25}$$

8. Choice of Regularization Parameter for SVMs

For computations, we included nine datasets from UCI repository Dheeru and Karra Taniskidou (2017) with small to moderate number of examples (306 examples to 5463 examples) and small to moderate number of features (3 features to 57 features). These datasets span a variety ranging from almost linearly separable (Banknote, Mushroom and Wave datasets) to moderately inseparable (Wdbc, Mammographic and Ionosphere datasets) to inseparable data (Spambase, Bupa and Haberman datasets). SVMs on these datasets have varying ranges and degrees of variation in their empirical risk values. We consider a finite set of SVM regularization parameter values $\Lambda = \{\lambda_i\}_{i=1}^{H}$, say, between 0 and an upper bound $\lambda_0 > 0$, since small values of λ_i's are preferable. We took $\Lambda = \{0.1, 0.11, \ldots, 20\}$ at a granularity of 0.01. SVM QP (with RBF kernels) was implemented using `ksvm` function in `kernlab` package Karatzoglou et al. (2004) in *R (version 3.1.3 (2015-03-09))*. The Gaussian width parameter is estimated by `kernlab` using `sigest` function which estimates 0.1 and 0.9 quantiles of squared distance between the data points.

Each of these datasets was partitioned such that 80% of the examples formed a composition of training set and validation set (in equal proportion) used for constructing the set $\mathcal{H} = \{h(\lambda_i)|\lambda_i \in \Lambda\}_{i=1}^{H}$ of SVM classifiers and remaining 20% used for computing their test error rates. The training set size (m), validation set size (v) and test set size (t) are in the ratio $m : v : t = 0.4 : 0.4 : 0.2$. The role of the validation set is to compute the empirical risk $\hat{l}_i$ of the SVM $h(\lambda_i) \in \mathcal{H}$ which will be used for deriving the PAC-Bayesian bound. We follow the scheme provided in Bégin et al. (2016); Thiemann et al. (2017) to generate the set $\mathcal{H}$. Each classifier $h(\lambda_i) \in \mathcal{H}$ is trained on m training examples subsampled from this composite set and validated on the remaining v examples. Overlaps between training sets of different classifiers are allowed. Same is true for their validation sets.

The PAC-Bayesian bound minimization problem for finding the optimal posterior was implemented in AMPL Interface and solved using `Ipopt` software package *(version 3.12 (2016-05-01))* Wächter and Biegler (2006), a library for large-scale nonlinear optimization (`http://projects.coin-or.org/Ipopt`). All computations were done on a machine equipped with 12 Intel Xeon 2.20 GHz cores and 64 GB RAM. We summarize comparisons among optimal posteriors for different distance functions in Table 2.

Fixed point solutions can be more reliable than solver output In case of KL-distance based bound, we observe that the FP scheme is able to converge to a stationary point even when solver fails to identify a local solution, as seen in Table 3. More such cases are illustrated in Table 5 in supplementary file with 7 other datasets.

9. Conclusion and Future Directions

We considered the PAC-Bayesian bound minimization problem for a finite classifier set with 5 distance functions. The optimal posterior weights are negative-exponentially decreasing with empirical risk values. For linear distance and uniform prior, weights are negative-exponentially proportional to number of misclassifications. Since some of these minimization problems are non-convex, we proposed fixed point (FP) iterates to identify posteriors with good test error rates. We apply these ideas for choosing SVM regularization parameter via an optimal posterior on the regularization parameter set, yielding a stochastic SVM.

As a part of the future work, we wish to investigate the convergence of FP iterates, and the reason for uniqueness of local minimum for some non-convex cases. For a comparative study, we can consider the PAC-Bayesian counterpart based on Rényi divergence between posterior and prior (proposed by Bégin et al. (2016)) for the distance functions considered.

Acknowledgments

We thank the Reviewers for their comments which helped us in improvising this paper.

Dataset	**PAC-Bayesian Bound**, $B^*_{\phi,\text{KL}}$					**Average Test Error**, $T_{\phi,\text{KL}}$				
	$B^*_{\text{lin, KL}}$	$B^*_{\text{sq, KL}}$	$B^*_{\text{P, KL}}$	$B^*_{\text{CH, KL}}$	$B^*_{\text{kl, KL}}$	$T_{\text{lin, KL}}$	$T_{\text{sq, KL}}$	$T_{\text{P, KL}}$	$T_{\text{CH, KL}}$	$T_{\text{kl, KL}}$
Spambase	NaN	0.20046	0.17361	0.17958	**0.15332**⋆	**0.15684**	**0.15392**	**0.15423**	**0.15434**	**0.15487**⋆
Bupa	0.27005	0.38167 *0.34547*	0.29265	0.30537	**0.23851**⋆	**0.13207**	0.145801 *0.14873*	0.13631	0.13382	**0.11998**⋆
Mammographic	0.29518	0.34187 *0.31290*	0.28790	0.29659	**0.26063**⋆	**0.20462**	**0.21120** *0.21386*	**0.20716**	**0.20628**	**0.20519**⋆
Wdbc	0.20706	0.26000 *0.22122*	0.20236	0.21646	**0.14759**⋆	**0.06489**	**0.06901** *0.07052*	**0.06650**	**0.06584**	**0.06541**⋆
Banknote	0.13647	0.13225 *0.10343*	0.09538	0.10672	**0.02051**	0.00161	0.00561 *0.00592*	0.00500	0.00469	**0.00037**
Mushroom	NaN	0.06584	0.04702	0.05399	**0.00489**	**8.92e-05**	0.00066	0.00057	0.00053	**1.39e-05**
Ionosphere	0.20816	0.30151 *0.25884*	0.22508	0.24011	**0.14707**⋆	**0.04494**	**0.04781** *0.04899*	**0.04393**	**0.04553**	**0.04359**⋆
Waveform	NaN	0.12875	0.10335	0.11103	**0.06338**	0.05847	**0.05175**	**0.05276**	**0.05345**	0.05792
Haberman	**0.37277**	0.48385 *0.43977*	**0.39769**	0.41178	**0.37998**⋆	**0.29157**	**0.29069** *0.29007*	**0.29163**	**0.29162**	**0.28997**⋆

Table 2: **PAC-Bayesian bounds and averaged test error rates for $Q^*_{\phi,\mathbf{KL}}$** We compare bound values $B^*_{\phi,\text{KL}}$ and average test error rates $T_{\phi,\text{KL}}$ of optimal posteriors due to five distance functions, ϕ: KL-divergence kl, its Pinsker's approximation ϕ_{P} and a sixth degree polynomial approximation ϕ_{CH}; linear ϕ_{lin} and squared distances ϕ_{sq} for $H = 1990$ SVM classifiers. For large sample size ($m \geq 1028$), $\mathcal{I}^K_{\text{lin}}(m)$ cannot be computed due to storage limitations for floating point numbers and in that case, $B^*_{\text{lin,KL}}$ is denoted by NaN. $Q^*_{\text{sq, KL}}$ was determined using: $2\sqrt{m}$ (in regular font) and $\mathcal{I}^K_{\text{sq}}(m)$ (in italicized font). $\mathcal{I}^K_{\text{sq}}(m)$ cannot be computed for $m \geq 1028$ due to storage limitations. For such cases, we report the values computed using $2\sqrt{m}$ alone. ⋆ refers to values obtained using fixed point(FP) equation because `Ipopt` solver does not converge. Lowest 10% bound values and test error rates for each dataset are denoted in bold face. kl has the tightest bound and lowest 10% error rate for most datasets, but is computationally expensive and has multiple local minima. Between ϕ_{P} and ϕ_{CH}, the latter has lower test error values but a slightly complicated bound evaluation. ϕ_{sq} and ϕ_{P} are related by a scaling ($\phi_{\text{P}} = 2\phi_{\text{sq}}$). ϕ_{P} provides a lower bound value than that of ϕ_{sq}, but both have comparable test set performances with differences of at most 3%. ϕ_{lin} has second lowest bound value for all datasets (except where $m \geq 1028$, namely, Spambase, Mushroom and Waveform, where $B^*_{\text{lin},KL}$ cannot be computed) and also has the lowest 10% test error rates for most datasets. All 5 ϕs have lowest 10% test error values on most datasets considered, except for Bupa dataset and two almost separable datasets, Banknote and Mushroom, where ϕ_{lin} and ϕ_{kl} do better.

H **Dataset**	**50**		**200**		**500**		**1000**		**1990**	
(Validation set size, v)	$B^{FP}_{\text{kl, KL}}$	$B^{solver}_{\text{kl, KL}}$	$B^{FP}_{\text{kl, KL}}$	$B^{solver}_{\text{kl, KL}}$	$B^{FP}_{\text{kl, KL}}$	$B^{solver}_{\text{kl, KL}}$	$B^{FP}_{\text{kl, KL}}$	$B^{solver}_{\text{kl, KL}}$	$B^{FP}_{\text{kl, KL}}$	$B^{solver}_{\text{kl, KL}}$
Spambase ($v = 1840$)	0.14726	0.14726	0.14942	0.14942	0.15157	0.27004(**E**)	0.15202	0.29484(**E**)	0.15332	0.31452(**E**)
Bupa ($v = 138$)	0.20833	0.20833	0.22006	0.22006	0.22750	0.43732(**E**)	0.23300	0.50867(**E**)	0.23851	0.57682(**E**)

Table 3: Comparing bound values due to fixed point solution, $B^{KKT}_{\text{kl, KL}}$, and bound values due to solver output, $B^{solver}_{\text{kl, KL}}$, for bound minimization problem (7) involving KL-distance function with KL-divergence measure. We observe that the fixed point equation always converges to a solution, even when the `Ipopt` solver is not able to identify a solution (denoted by '**E**' (Unknown Error)). Other examples of solver failure are in Table 5 in Suppl. file (eg. '**R**' (Restoration Phase Failed) or '**M**' (Maximum Number of Iterations Exceeded)).

References

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *NeurIPS*, pages 9–16, 2006.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *AISTATS*, pages 435–444, 2016.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Alexei A Fedotov, Peter Harremoës, and Flemming Topsøe. Refinements of Pinsker's inequality. *IEEE Trans. Info. Theory*, 49(6):1491–1498, 2003. doi: 10.1109/TIT.2003.811927.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *ICML*, pages 353–360, 2009. doi: 10.1145/1553374.1553419.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *J. Stat. Soft.*, 11(9):1–20, 2004. doi: 10.18637/jss.v011.i09.

John Langford. Tutorial on practical prediction theory for classification. *JMLR*, 6(Mar):273–306, 2005.

Andreas Maurer. A note on the PAC Bayesian theorem. *CoRR*, cs.LG/0411099, 2004. URL http://arxiv.org/abs/cs.LG/0411099.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003. doi: 10.1023/A:1021840411064.

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

Puja Sahu and Nandyala Hemachandra. Some new PAC-Bayesian bounds and their use in selection of regularization parameter for linear SVMs. In *CODS-COMAD*, pages 240–248, 2018. doi: 10.1145/3152494.3152514.

Matthias Seeger. The proof of McAllester's PAC-Bayesian theorem. In *NeurIPS*, 2002.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *ALT*, pages 466–492, 2017.

Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.*, 106(1):25–57, 2006. doi: 10.1007/s10107-004-0559-y.