Simplified PAC-Bayesian Margin Bounds

David McAllester

Toyota Technological Institute at Chicago
mcallester@tti-c.org

Abstract. The theoretical understanding of support vector machines is largely based on margin bounds for linear classifiers with unit-norm weight vectors and unit-norm feature vectors. Unit-norm margin bounds have been proved previously using fat-shattering arguments and Rademacher complexity. Recently Langford and Shawe-Taylor proved a dimension-independent unit-norm margin bound using a relatively simple PAC-Bayesian argument. Unfortunately, the Langford-Shawe-Taylor bound is stated in a variational form making direct comparison to fat-shattering bounds difficult. This paper provides an explicit solution to the variational problem implicit in the Langford-Shawe-Taylor bound and shows that the PAC-Bayesian margin bounds are significantly tighter. Because a PAC-Bayesian bound is derived from a particular prior distribution over hypotheses, a PAC-Bayesian margin bound also seems to provide insight into the nature of the learning bias underlying the bound.

1 Introduction

Margin bounds play a central role in learning theory. Margin bounds for convex combination weight vectors (unit ℓ_1 norm weight vectors) provide a theoretical foundation for boosting algorithms [15, 9, 8]. Margin bounds for unit-norm weight vectors provide a theoretical foundation for support vector machines [3, 17, 2]. This paper concerns the unit-norm margin bounds underlying support vector machines. Earlier unit-norm margin bounds were proved using fat shattering dimension. This paper, building on results by Langford and Shawe-Taylor [11], gives a PAC-Bayesian unit-norm margin bound that is tighter than known unit-norm margin bounds derived from fat shattering arguments.

Consider a fixed distribution D on pairs $\langle x,\ y\rangle$ with $x \in R^d$ satisfying $||x|| = 1$ and $y \in \{-1, 1\}$. We are interested in finding a weight vector w with $||w|| = 1$ such that the sign of $w \cdot x$ predicts y. For $\gamma > 0$ the error rate of w on distribution D relative to safety margin γ, denoted $\ell_\gamma(w,\ D)$ is defined as follows.

$$\ell_\gamma(w,\ D) = \mathrm{P}_{\langle x,\ y\rangle \sim D}\left[(w \cdot x)y \leq \gamma\right]$$

Let S be a sample of m pairs drawn IID from the distribution D. The sample S can be viewed as an empirical distribution on pairs. We are interested in bounding $\ell_0(w,\ D)$ in terms of $\ell_\gamma(w,\ S)$ and the margin γ. Bartlett and Shawe-Taylor use fat shattering arguments [2] to show that with probability at least $1-\delta$ over the choice of the sample S we have the following simultaneously for all weight vectors w with $||w|| = 1$ and margins $\gamma > 0$.

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + 27.18\sqrt{\frac{\log^2 m + 84}{m\gamma^2}} + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \tag{1}$$

Note that the bound is independent of the dimension d (the number of features and corresponding weights). Intuitively the quantity $1/\gamma^2$ acts like the complexity of the weight vector. Bound (1) has been recently improved using Rademacher complexity — Theorem 21 of [4] implies the following where k is $m\gamma^2$.

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + 8\sqrt{\frac{\ln\frac{4}{\delta}}{k}} + \frac{4}{\sqrt{k}} + \sqrt{\frac{\ln\frac{4}{\delta}}{m}} \tag{2}$$

Bound (2) has the nice scaling property that the bound remains meaningful in a limit where k is held constant while m goes to infinity. Further improvements on (2) are possible within the Rademacher complexity framework [1].

Initial attempts to use PAC-Bayesian arguments to derive unit-norm margin bounds resulted in bounds that depended on d [6]. Here, building on the work of Langford and Shawe-Taylor [11], we use a PAC-Bayesian argument to show that with probability at least $1-\delta$ over the choice of the sample S we have the following simultaneously for all w with $||w|| = 1$ and where $\ln^+(x)$ abbreviates $\max(0,\ \ln x)$.

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + 2\sqrt{\frac{2\left(\ell_\gamma(w, S) + \frac{4}{k}\right)\ln^+\left(\frac{k}{4}\right)}{k}} + \frac{8\left(1 + \ln^+\left(\frac{k}{4}\right)\right)}{k}$$
$$+O\left(\sqrt{\frac{\ln m + \ln\frac{1}{\delta}}{m}}\right) \tag{3}$$

Like (2), bound (3) is meaningful in a limit where k is held constant while m goes to infinity. Bound (3) also interpolates the realizable and unrealizable case — in the case where the training error $\ell_\gamma(w, S)$ is zero the bound is $O((\ln k)/k)$. Note however, that even for $\ell_\gamma(w.\ S) = 1/2$ we have that (3) is an improvement on (1) and (2) for modest values of k.

Bound (3) is derived from a bound given by Langford and Shawe-Taylor [11]. The Langford and Shawe-Taylor bound is tighter than (3) but is given in an implicit form which makes explicit comparison with earlier bounds difficult. Langford and Shawe-Taylor also use PAC-Bayesian analysis to define more refined notions of margin leading to new learning algorithms. The contribution of this paper is to solve the variational problems implicit in the Langford and Shawe-Taylor bound and show clearly how the PAC-Bayesian bounds compare with earlier bounds. PAC-Bayesian bounds seem competitive with the best known bounds derivable by shattering and Rademacher methods.

The derivation of a margin bound from the PAC-Bayesian theorem presents the bias of the learning algorithm in the familiar form of a prior distribution on hypotheses. In particular, the derivation of (3) is based on an isotropic Gaussian prior over the weight vectors. PAC-Bayesian arguments have also been used to give what appears to be the tightest known bounds for Gaussian process classifiers [16] and useful bounds for convex weight vector linear threshold classifiers [9]. In these cases as well, PAC-Bayesian derivations present the bias of the algorithm in the familiar form of a prior distribution.

2 The PAC-Bayesian Theorem

A first version of the PAC-Bayesian theorem appeared in [12]. The improved statement of the theorem given here is due to Langford and the simplified proof in the appendix is due to Seeger [10, 16]. Let D be a distribution on a set Z, let P be a distribution on a set H, and let ℓ be a "loss function" from $H \times Z$ to $[0,1]$. For any distribution W on Z and $h \in H$ let $\ell(h, W)$ be $\mathrm{E}_{z \sim W}\left[\ell(h,z)\right]$. Let S be an IID sample of m elements of Z drawn according to the distribution D. We are interested in using the sample S to select $h \in H$ so as to minimize the loss $\ell(h, D)$. We will treat the sample as a distribution in the standard way so that $\ell(h, S)$ is the (training) loss of h on the sample S and $\ell(h, D)$ is the (generalization) loss of h on D. A common case is where the set Z is of the form $X \times \{0,1\}$, H is a set of functions from X to $\{0,1\}$, and $\ell(h, \langle x, y\rangle)$ is 1 if $h(x) \neq y$ and 0 otherwise. In this case $\ell(h, S)$ is the training error rate of rule h on the sample S and $\ell(h, D)$ is $\mathrm{P}_{\langle x,\, y\rangle \sim D}\left[h(x) \neq y\right]$. We will be interested in Gibbs classifiers, i.e., classifiers which select h stochastically [5]. For any distribution Q on H and distribution W on Z let $\ell(Q, W)$ denote $\mathrm{E}_{h \sim Q, z \sim W}\left[\ell(h,z)\right]$. $\ell(Q, S)$ is the training loss of the Gibbs rule defined by Q and $\ell(Q, D)$ is the generalization loss of this rule. For two distributions Q and P on the same set H the Kullback-Leibler divergence $KL(Q||P)$ is defined to be $\mathrm{E}_{h \sim Q}\left[\ln(dQ(h)/dP(h))\right]$. For $p, q \in [0,1]$ let $KL(p||q)$ denote the Kullback-Leibler divergence from a Bernoulli variable with bias p to a Bernoulli variable with bias q. We have $KL(p||q) = p \ln(p/q) + (1-p)\ln((1-p)/(1-q))$. Let $\forall^\delta S\ \Phi(S, \delta)$ mean that with probability at least $1-\delta$ over the choice of S we have that $\Phi(S,\ \delta)$ holds. The two-sided PAC-Bayesian theorem is the following where P is a fixed "prior" distribution on H and Q ranges over arbitrary "posterior" distributions on H.

$$\forall^\delta S\ \ \forall Q\ \ KL(\ell(Q,S)||\ell(Q,D)) \leq \frac{KL(Q||P) + \ln \frac{2m}{\delta}}{m-1} \tag{4}$$

Note that $\ell(Q, D)$ is the error rate of a Gibbs classifier that first selects a rule stochastically according to the distribution Q and then uses the prediction of that rule. Intuitively, the theorem states that if $KL(Q||P)$ is small then $\ell(Q, D)$ is near $\ell(Q, S)$. Formula (4) bounds the difference between the empirical loss $\ell(Q, S)$ of the Gibbs classifier and its true (generalization) loss $\ell(Q, D)$.

A one-sided version can be stated as follows where $2/\delta$ in the two-sided version becomes $1/\delta$ in the one-sided version.

$$\forall^\delta S \;\; \forall Q \;\; \ell(Q, D) \leq \sup \left\{ \epsilon : KL(\ell(Q, S)||\epsilon) \leq \frac{KL(Q||P) + \ln \frac{m}{\delta}}{m - 1} \right\} \tag{5}$$

For $q > p$ we have that $KL(p||q) \geq (q - p)^2/(2q)$. This inequality implies that if $KL(p||q) \leq x$ then $q \leq p + \sqrt{2px} + 2x$. So (5) implies the following somewhat weaker but perhaps clearer statement.

$$\ell(Q, D) \leq \ell(Q, S) + \sqrt{\frac{2\ell(Q, S)\left(KL(Q||P) + \ln \frac{m}{\delta}\right)}{m - 1}} + \frac{2\left(KL(Q||P) + \ln \frac{m}{\delta}\right)}{m - 1} \tag{6}$$

Note if the empirical loss $\ell(Q, S)$ is small compared to $KL(Q||P)/m$ then the last term dominates (the realizable case). If $\ell(Q, S)$ is large compared to $KL(Q||P)/m$ then the first term dominates. Because the arithmetic mean bounds the geometric mean, we have that in general the bound is $O(\ell(Q, S)+KL(Q||P)/m)$. Proofs of these theorems are given in an appendix.

3 Gibbs Linear Threshold Classifiers

In this section we use the PAC-Bayesian theorem to prove a margin bound for a Gibbs classifier which stochastically selects a linear threshold function. The next section uses similar methods to prove a margin bound for a single (deterministic) linear threshold classifier. In both this section and the next we take Z to be the set of pairs $\langle x, y \rangle$ with $x \in R^d$ satisfying $||x|| = 1$ and $y \in \{-1, 1\}$. Both these sections take H to be weight vectors in R^d and take the prior P on H to be a unit-variance isotropic (the same in all directions) multivariate Gaussian on R^d. For each $\gamma \geq 0$ we define a loss function ℓ_γ as follows.

$$\ell_\gamma(w, \; \langle x, y \rangle) = \begin{cases} 1 \text{ if } y(w \cdot x) \leq \gamma \\ 0 \text{ otherwise} \end{cases}$$

For $w \in R^d$ with $||w|| = 1$ and $\mu > 0$ define the "posterior" $Q(w, \mu)$ by the following density function q where p is the density function of the "prior" P and

Z is a normalizing constant.

$$q(w') = \frac{1}{Z}\begin{cases} p(w') & \text{if } w' \cdot w \geq \mu \\ 0 & \text{otherwise} \end{cases}$$

Note that w' is drawn from a unit-variance multivariate Gaussian and w is a fixed vector with $||w|| = 1$. Since $Q(w,\ \mu)$ is just the prior renormalized on a subset of the space we have the following where $\varPhi(\mu)$ is the probability that a unit-variance Gaussian real-value random variable exceeds μ.

$$KL(Q(w,\mu)||P) = \ln\frac{1}{Z} = \ln\frac{1}{P(w' \cdot w \geq \mu)} = \ln\frac{1}{\varPhi(\mu)}$$

Now for $\gamma > 0$, any hypothesis distribution Q, and any data distribution W, define $\ell_\gamma(Q, W)$ as follows.

$$\ell_\gamma(Q, W) = \mathrm{E}_{w\sim Q, \langle x,\, y\rangle \sim W}\left[\ell_\gamma(w, \langle x, y\rangle)\right]$$

We will often write $\ell_\gamma(w, W)$ where $w \in R^d$ as a notation for $\ell_\gamma(Q, W)$ where Q places all of its weight on w. These quantities are error rates relative to a “safety margin” of γ. The fundamental idea behind the PAC-Bayesian approach to margin bounds is that a small error rate relative to a large safety margin ensures the existence of a posterior distribution (a Gibbs classifier) with a small training error and a small KL-divergence from the prior. We first consider the training error. Langford and Shawe-Taylor prove the following.

Lemma 1 (Langford&Shawe-Taylor). *For $w \in R^d$ with $||w|| = 1$, and for $\mu \geq 0$, we have the following for all $\gamma \geq 0$.*

$$\ell_0(Q(w,\mu), S) \leq \ell_\gamma(w, S) + \varPhi(\gamma\mu)$$

Proof. For $x \in R^d$ with $||x|| = 1$ we let $x_{||}$ be $(w \cdot x)w$ ($x_{||}$ is the component of x parallel to w), and let $x_\perp$ be $x - x_{||}$ ($x_\perp$ is the component of x perpendicular to w). Let $\langle x, y\rangle$ be a tuple in S. We say that $\langle x, y\rangle$ is γ-safe (for w) if $y(w \cdot x) \geq \gamma$. Note that the Gaussian prior on weight vectors has the property that, for any two orthogonal directions, the components of a random vector in those two directions are independent and normally distributed. Fix a γ-safe point $\langle x, y\rangle$ and consider the orthogonal components $w' \cdot x_\perp$ and $w' \cdot x_{||}$ as we select random weight vectors w'. If $w' \cdot w \geq \mu$, and $\langle x, y\rangle$ is γ-safe for w then $y(w' \cdot x_{||}) \geq \gamma\mu$. More specifically we have the following.

$$\begin{aligned}
\mathrm{P}_{w'\sim Q(w,\mu)}\left[y(w' \cdot x) \leq 0\right] &= \mathrm{P}_{w'\sim Q(w,\mu)}\left[-y(w' \cdot x_\perp) \geq y(w' \cdot x_{||})\right] \\
&= \mathrm{P}_{w'\sim Q(w,\mu)}\left[-y(w' \cdot x_\perp) \geq y(x \cdot w)(w' \cdot w)\right] \\
&\leq \mathrm{P}_{w'\sim Q(w,\mu)}\left[-y(w' \cdot x_\perp) \geq \gamma\mu\right] \\
&= \varPhi\left(\frac{\gamma\mu}{||x_\perp||}\right) \\
&\leq \varPhi(\gamma\mu)
\end{aligned}$$

This yields the following.

$$\begin{aligned}\ell_0(Q(w,\mu),S) &= \mathrm{E}_{\langle x,\,y\rangle\sim S}\left[\mathrm{P}_{w'\sim Q(w,\mu)}\left[y(w'\cdot x)\leq 0\right]\right]\\ &\leq \ell_\gamma(w,S)+\mathrm{E}_{\langle x,\,y\rangle\sim S}\left[\mathrm{P}_{w'\sim Q(w,\mu)}\left[y(w'\cdot x)\leq 0\right]\mid y(x\cdot w)\geq\gamma\right]\\ &\leq \ell_\gamma(w,S)+\Phi\left(\gamma\mu\right)\end{aligned}$$

□

Formula (5) (for the loss function ℓ_0) and Lemma 1 together yield the following.

Theorem 1 (Langford&Shawe-Taylor). *With probability at least $1-\delta$ over the choice of the sample we have that the following holds simultaneously for all $w\in R^d$ with $||w||=1$, $\mu\geq 0$ and $\gamma\geq 0$.*

$$\ell_0(Q(w,\mu),D)\leq\sup\left\{\epsilon:\ KL\left(\ell_\gamma(w,\ S)+\Phi\left(\gamma\mu\right)\ ||\ \epsilon\right)\leq\frac{\ln\frac{1}{\Phi(\mu)}+\ln\frac{m}{\delta}}{m-1}\right\}$$

The main contribution of this paper is to give a particular value for μ and then "solve" for the upper bound on $\ell_0(Q(w,\ \mu))$ implicit in Theorem 1. In particular we define $\mu(\gamma)$ as follows.

$$\mu(\gamma)=\frac{\sqrt{2\ln\left(m\gamma^2\right)}}{\gamma}$$

For this choice of μ we have the following.

Theorem 2. *With probability at least $1-\delta$ over the choice of the sample S we have that the following holds simultaneously for all $w\in R^d$ with $||w||=1$ and $\gamma>0$.*

$$\ell_0(Q(w,\mu(\gamma)),D)\leq\sup\left\{\epsilon:\ \begin{array}{c}KL\left(\ell_\gamma(w,\ S)+\frac{1}{m\gamma^2}\ ||\ \epsilon\right)\\ \leq\ \frac{\frac{\ln^+\left(m\gamma^2\right)}{\gamma^2}+\frac{3}{2}\ln m+\ln\frac{1}{\delta}+3}{m-1}\end{array}\right\}$$

Theorem 2 follows from the following of two lemmas.

Lemma 2. *For $\gamma>0$ we have the following.*

$$\ln\frac{1}{\Phi\left(\mu(\gamma)\right)}\leq\frac{\ln^+\left(m\gamma^2\right)}{\gamma^2}+\frac{1}{2}\ln m+3$$

Proof. First, if $\mu(\gamma) \leq 3/2$ we have the following.

$$\ln \frac{1}{\Phi\left(\mu(\gamma)\right)} \leq \ln \frac{1}{\Phi\left(3/2\right)} \leq 3 \tag{7}$$

In this case $\ln\left(m\gamma^2\right)$ might be negative, but the lemma still follows. Now suppose $\mu(\gamma) \geq 3/2$. For $\mu \geq 0$ we have the following well known lower bound on $\Phi\left(\mu\right)$ (see [14]).

$$\Phi\left(\mu\right) \geq \left(1 - \frac{1}{\mu^2}\right) \frac{1}{\sqrt{2\pi}} \frac{1}{\mu} \exp\left(-\frac{\mu^2}{2}\right) \tag{8}$$

For $\mu(\gamma) \geq 3/2$ formula 8 yields the following.

$$\Phi\left(\mu(\gamma)\right) \geq \frac{5}{9} \frac{1}{\sqrt{2\pi}} \frac{1}{\mu(\gamma)} \exp\left(-\mu^2(\gamma)/2\right)$$

This yields the following.

$$\ln \frac{1}{\Phi\left(\mu(\gamma)\right)} \leq 2 + \ln \mu(\gamma) + \frac{\ln(m\gamma^2)}{\gamma^2} \tag{9}$$

We have that $\mu(\gamma)$ goes to zero as γ goes to infinity and goes to negative infinity as γ goes to zero. Furthermore, a simple calculation shows that the derivative is zero at only a single point given by $\gamma = \sqrt{e/m}$ and at this point $\mu(\gamma)$ is positive. These facts imply that this point is a maximum of $\mu(\gamma)$ and we get the following which implies the lemma.

$$\mu(\gamma) \leq \sqrt{\frac{2m}{e}} \tag{10}$$

□

Lemma 3. *For $\gamma > 0$ we have the following.*

$$\Phi\left(\gamma\mu(\gamma)\right) \leq \frac{1}{m\gamma^2}$$

Proof. For $\gamma \leq 1/\sqrt{m}$ we have $1/(m\gamma^2) \geq 1$ and the lemma follows from $\Phi\left(x\right) \leq 1$. For $\gamma \geq 1/\sqrt{m}$ we have $\mu(\gamma) \geq 0$ and the lemma follows from the fact that for $z \geq 0$ we have $\Phi\left(z\right) \leq \exp(-z^2/2)$. □

Theorem 2 now follows from Theorem 1 and Lemmas 2 and 3. Using $KL(p||q) \leq \frac{1}{2}q(q-p)^2$ for $p \leq q$ we get the following corollary of Theorem 2.

Corollary 1. *With probability at least $1-\delta$ over the choice of the sample S we have that the following holds simultaneously for all $w \in R^d$ with $||w|| = 1$ and $\gamma > 0$.*

$$\ell_0(Q(w,\mu(\gamma)),D) \leq \hat{\ell} + \sqrt{2\hat{\ell}\varDelta} + 2\varDelta$$

where

$$\hat{\ell} = \ell_\gamma(w,S) + \frac{1}{m\gamma^2}$$

$$\varDelta = \frac{\frac{\ln^+(m\gamma^2)}{\gamma^2} + \frac{3}{2}\ln m + \ln\frac{1}{\delta} + 3}{m-1}$$

The body of Corollary 1 can be rewritten as follows.

$$\ell_0(Q(w,\mu(\gamma)),D) \leq \ell_\gamma(w,S) + \sqrt{\frac{2\left(\ell_\gamma(w,S) + \frac{1}{k}\right)\ln^+(k)}{k}} + \frac{1+2\ln^+(k)}{k} + O\left(\sqrt{\frac{\ln m + \ln\frac{1}{\delta}}{m}}\right) \tag{11}$$

Note that (11) is vacuously true for $k \leq 1$. For $k \geq 1$ the constants in the big O expression are modest and independent of k (i.e., independent of γ). For large sample size the big O term vanishes and either the error is very near zero or the bound is dominated by the terms involving k. This bound has a nice limiting behavior in a "thermodynamic limit" where $\ell_\gamma(w,S)$ and k are held constant while $m \to \infty$. This thermodynamic limit corresponds to a realistic regime where m is large but $\ell_\gamma(w,S)$ and $1/k$ are still significantly greater than zero. For the realizable case, i.e., when $\ell_\gamma(w,S) = 0$, we get the following.

$$\ell_0(Q(w,\mu(\gamma)),D) \leq \frac{1+\sqrt{2\ln^+(k)} + 2\ln^+(k)}{k} + O\left(\sqrt{\frac{\ln m + \ln\frac{1}{\delta}}{m}}\right) \tag{12}$$

4 Deterministic Linear Classifiers

Theorem 2 gives a margin bound for the loss of a Gibbs classifier — a classifier that stochastically selects the classification rule at classification time. There are two ways of converting Theorem 2 into a margin guarantee on a deterministic linear classification rule. First we observe that the deterministic classification rule defined by the weight vector w corresponds to the majority vote over the distribution $Q(w,\mu)$. More formally we have that $\mathrm{P}_{w' \sim Q(w,\mu)}\left[w' \cdot x \geq 0\right] \geq 1/2$ if and only if $x \cdot w \geq 0$. For any Gibbs classifier, the error rate of the majority vote classifier can be at most twice the error rate of the Gibbs classifier. This is because each error of the majority vote classifier requires that at least half (under the voting measure) of the individual classifiers are making an error and so the error rate of the Gibbs classifier must be at least half the error rate of the majority vote classifier. This general factor of two bound on the error rate of the majority classifier together with (11) yields the following.

$$\ell_0(w, D) \leq 2\ell_\gamma(w, S) + 2\sqrt{\frac{2\left(\ell_\gamma(w, S) + \frac{1}{k}\right)\ln^+(k)}{k}} + \frac{2\left(1 + 2\ln^+(k)\right)}{k}$$
$$+O\left(\sqrt{\frac{\ln m + \ln\frac{1}{\delta}}{m}}\right) \tag{13}$$

Again it is interesting to consider the thermodynamic limit where $\ell_\gamma(w, S)$ and k are held constant as $m \to \infty$. Note that (3) is tighter than (13) in the regime where $1/k$ is small compared to $\ell_\gamma(w, S)$. We now prove (3). We start with the following generalization of Lemma 1.

Lemma 4 (Langford&Shawe-Taylor). *Let W be any distribution on pairs $\langle x, y\rangle$ with $x \in R^d$ satisfying $||x|| = 1$ and $y \in \{-1, 1\}$. Let w be any vector in R^d satisfying $||w|| = 1$. For $\mu \geq 0$ and $\gamma \geq 0$ and any real value β we have the following.*

$$\mathrm{P}_{\langle x, y\rangle \sim W, w' \sim Q(w,\mu)}\left[y(w' \cdot x) \leq \beta\right] \leq \mathrm{P}_{\langle x, y\rangle \sim W}\left[y(w \cdot x) \leq \beta + \gamma\right] + \Phi(\gamma\mu) \tag{14}$$
$$\mathrm{P}_{\langle x, y\rangle \sim W, w' \sim Q(w,\mu)}\left[y(w' \cdot x) > \beta\right] \leq \mathrm{P}_{\langle x, y\rangle \sim W}\left[y(w \cdot x) > \beta - \gamma\right] + \Phi(\gamma\mu) \tag{15}$$

Formula (14) is a generalization of Lemma 1 and the proof of (14) is a straightforward generalization of the proof of Lemma 1. The proof of (15) is similar. Lemma 4 yields the following corollary.

Corollary 2 (Langford&Shawe-Taylor).

$$\ell_{\gamma/2}(Q(w, \mu), S) \leq \ell_\gamma(w, S) + \Phi(\gamma\mu/2) \tag{16}$$
$$\ell_0(w, D) \leq \ell_{\gamma/2}(Q(w, \mu), D) + \Phi(\gamma\mu/2) \tag{17}$$

Proof. Formula (16) is an instance of (14) with $\beta = \gamma/2$ and γ replaced by $\gamma/2$. To prove (17) we construct the following instance of (15) again with $\beta = \gamma/2$ and γ replaced by $\gamma/2$.

$$1 - \ell_{\gamma/2}(Q(w,\ \mu), D) \leq 1 - \ell_0(w, D) + \Phi(\gamma\mu/2)$$

□

To get a bound on $\ell_0(w, D)$ it now suffices to bound $\ell_{\gamma/2}(Q(w, \mu), D)$ in terms of $\ell_{\gamma/2}(Q(w, \mu), S)$. An application of (5) to the loss function ℓ_γ yields the following.

$$\forall\gamma\forall^\delta S \;\; \forall Q \;\; \ell_\gamma(Q, D) \leq \sup\left\{\epsilon : \; KL\left(\ell_\gamma(Q, S)||\epsilon\right) \leq \frac{KL(Q||P) + \ln\frac{m}{\delta}}{m-1}\right\} \tag{18}$$

We now consider discrete values of γ satisfying the statements that $k = m\gamma^2 = i/m$ for $i \in \{1, \; 2, \; \ldots, \; m^2\}$. By a union bound over the m^2 different possible values of γ we get that with probability at least $1-\delta$ over the choice of the sample the following holds for all Q and for all γ satisfying $k \in \{1/m, \; 2/m, \; \ldots, \; m^2/m\}$.

$$\ell_{\gamma/2}(Q, D) \leq \sup\left\{\epsilon : \; KL\left(\ell_{\gamma/2}(Q, S)||\epsilon\right) \leq \frac{KL(Q||P) + \ln\frac{m^3}{\delta}}{m-1}\right\} \tag{19}$$

Formulas (16), (17), and (19) together yield the following variant of a theorem in [11].

Theorem 3 (Langford and Shawe-Taylor). *With probability at least* $1-\delta$ *over the choice of the sample we have the following simultaneously for all* $\mu \geq 0$ *and* $\gamma \in \{1/m, \; 2/m \ldots, \; m/m\}$.

$$\ell_0(w, D) \leq \sup\left\{\epsilon : \; KL\left(\ell_\gamma(w, S) + \Phi\left(\gamma\mu/2\right)||\epsilon - \Phi\left(\gamma\mu/2\right)\right) \leq \frac{\ln\frac{1}{\Phi(\mu)} + \ln\frac{m^3}{\delta}}{m-1}\right\}$$

Again, the main contribution of this paper is to construct more explicit forms of the bounds implicit in Theorems 1 and 3. Using $\mu(\gamma/2)$ in Theorem 3 together with Lemmas 2 and 3 yields the following.

$$\ell_0(w, D) \leq \sup\left\{\epsilon : \begin{array}{c} KL\left(\ell_\gamma(w, S) + \frac{4}{m\gamma^2}||\epsilon - \frac{4}{m\gamma^2}\right) \\ \\ \leq \quad \frac{\frac{4\ln^+\left(m\gamma^2/4\right)}{\gamma^2} + \frac{7}{2}\ln m + \ln\frac{1}{\delta} + 3}{m-1} \end{array}\right\} \tag{20}$$

By the arguments deriving (11) from Theorem 2 we then have the following for the allowed discrete values of k.

$$\begin{aligned} \ell_0(w, D) \leq \ell_\gamma(w, S) &+ 2\sqrt{\frac{2\left(\ell_\gamma(w, S) + \frac{4}{k}\right)\ln^+\left(\frac{k}{4}\right)}{k}} + \frac{8\left(1 + \ln^+\left(\frac{k}{4}\right)\right)}{k} \\ &+O\left(\sqrt{\frac{\ln m + \ln\frac{1}{\delta}}{m}}\right) \end{aligned} \tag{21}$$

To derive (3) for arbitrary γ we first note that (3) is vacuously true for $k \leq 8$. For $k \geq 8$ we have $\gamma \geq 4/\sqrt{m}$. Let α be the largest value with $\alpha \leq \gamma$ such that

$m\alpha^2$ has the form i/m for $i \in \{1, \ldots, m^2\}$. Let k' be $m\alpha$. Note that we have $k' \geq 8$. We now get that (21) holds for α and k' replacing γ and k respectively. Note that $\ell_\alpha(w, S) \leq \ell_\gamma(w, S)$ and $k' \geq k - 1/m$. This give the following for arbitrary γ satisfying $k \geq 8$.

$$\ell_0(w, D) \leq \ell_\gamma(w, S) + 2\sqrt{\frac{2\left(\ell_\gamma(w, S) + \frac{4}{k-1/m}\right)\ln^+\left(\frac{k}{4}\right)}{k - 1/m}} + \frac{8\left(1 + \ln^+\left(\frac{k}{4}\right)\right)}{k - 1/m}$$
$$+O\left(\sqrt{\frac{\ln m + \ln \frac{1}{\delta}}{m}}\right) \tag{22}$$

The difference between k and $k - 1/m$ can then be absorbed into the final term and we get (3).

References

1. Peter Bartlett. Personal communication. , 2003.
2. Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In Bernhard Schlkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.
3. Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the the size of the network. *IEEE Transactions on Information Theory*, March 1998.
4. P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
5. Olivier Catoni. Gibbs estimators. to appear in Probability Theory and Related Fields.
6. Ralph Herbrich and Thore Graepel. A PAC-Bayesian margin bound for linear clasifiers: Why svms work. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.
7. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
8. V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.
9. John Langford, Matthias Seeger, and Nimrod Megiddo. An improved predictive accuracy bound for averaging classifiers. In *ICML2001*, 2001.
10. John Langford and Matthias Seger. Bounds for averaging classifiers. CMU Technical Report CMU-CS-01-102, 2002.
11. John Langford and John Shawe-Taylor. PAC-Bayes and margins. In *Neural Information Processing Systems (NIPS)*, 2002.
12. David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 5:5–21, 2003. A short version appeared as "PAC-Bayesian Model Averaging" in COLT99.

13. David McAllester and Luis Ortiz. Concentration inequalities for the missing mass and for histogram rule error. In *Neural Information Processing systems (NIPS)*, 2002.
14. Harold Ruben. A new asymptotic expansion for the normal probability integral and mill's ratio. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(1):177–179, 1962.
15. Robert Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Machine Learning: Proceedings of the Fourteenth International Conference*, 1997.
16. Matthias Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.
17. John Shawe-Taylor, Peter Bartlett, Robert Williamson, and Martin Anthony. A framework for structural risk minimization. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

5 Appendix: Proofs of PAC-Bayesian Theorems

McAllester's original form of the theorem used a square root rather than an inverse KL divergence. The inverse KL divergence form is due to John Langford. The simple proof based on Jensen's inequality is due to Matthias Seeger. We prove the one-sided version (5). A lower bound version of (5) follows by applying (5) to the loss function $1 - \ell$ and the two-sided version (4) follows by a union bound from the upper and lower bound versions. To prove (5) we first prove the following lemma.

Lemma 5. *Let X be a real valued random variable satisfying*

$$P(X \leq x) \leq e^{-mf(x)}$$

where $f(x)$ is non-negative. For any such X we have the following.

$$\mathrm{E}\left[e^{(m-1)f(X)}\right] \leq m$$

Proof. If $P(X \leq x) \leq e^{-mf(x)}$ then $P(e^{(m-1)f(X)} \geq \nu) \leq \min(1,\ \nu^{-m/(m-1)})$. We can then use the general fact that for W non-negative we have $\mathrm{E}\left[W\right] = \int_0^\infty P(W \geq \nu)d\nu$. This gives the following.

$$\begin{aligned}\mathrm{E}\left[e^{(m-1)f(X)}\right] &\leq 1 + \int_1^\infty \nu^{-m/(m-1)}d\nu \\ &= 1 - (m-1)\left[\nu^{-1/(m-1)}\right]_1^\infty \\ &= m\end{aligned}$$

□

Now let $KL^+(p||q)$ be zero if $p \geq q$ and $KL(p||q)$ if $p \leq q$. Hoeffding [7] proved essentially the following.[1]

Lemma 6 (Hoeffding). *If $X_1, \ldots, X_m$ are IID random variables restricted to the interval $[0,1]$, and $\hat{X}$ is the empirical average $(X_1 + \cdots + X_m)/m$, then for $\epsilon \in [0,1]$ we have the following.*

$$P(\hat{X} \leq \epsilon) \leq e^{-mKL^+(\epsilon||\mathrm{E}[X_i])}$$

Lemma 7.

$$\forall^\delta S \;\; \mathrm{E}_{h\sim P}\left[e^{(m-1)KL^+(\ell(h,S)||\ell(h,\,D))}\right] \leq \frac{m}{\delta}$$

Proof. Lemma 5 and Lemma 6 together imply the following for any fixed $h \in H$.

$$\mathrm{E}_{S\sim D^m}\left[e^{(m-1)KL^+(\ell(h,S)||\ell(h,D))}\right] \leq m$$

This implies the following.

$$\mathrm{E}_{S\sim D^m}\left[\mathrm{E}_{h\sim H}\left[e^{(m-1)KL^+(\ell(h,S)||\ell(h,D))}\right]\right] \leq m$$

The lemma now follows from Markov's inequality. □

We now prove the following shift of measure lemma.

Lemma 8.

$$\mathrm{E}_{x\sim Q}\left[f(x)\right] \leq KL(Q||P) + \ln \mathrm{E}_{x\sim P}\left[e^{f(x)}\right]$$

Proof.

$$\mathrm{E}_{x\sim Q}\left[f(x)\right] = \mathrm{E}_{x\sim Q}\left[\ln e^{f(x)}\right]$$

[1] It is interesting to note that Lemma 6 generalizes to an arbitrary real-valued random variable X. Let P_β be the Gibbs distribution on X at inverse temperature β, let $\mathrm{E}_\beta\left[f(X)\right]$ be the expectation of $f(X)$ under P_β, and let Z_β be the partition function at inverse temperature β.

$$\begin{aligned} P_\beta(X = x) &= \frac{1}{Z_\beta} e^{-\beta x} P(X = x) \\ \mathrm{E}_\beta\left[f(X)\right] &= \frac{1}{Z_\beta}\mathrm{E}\left[f(X)e^{-\beta X}\right] \\ Z_\beta &= \mathrm{E}\left[e^{-\beta X}\right] \end{aligned}$$

Let $DP(x)$ be $P(X \leq x)$ if $x \leq \mathrm{E}\left[X\right]$ and $P(X \geq x)$ if $x \geq \mathrm{E}\left[X\right]$. In general we have $DP(x) \leq \exp(-KL(P_\beta||P))$ where β satisfies $\mathrm{E}_\beta\left[X\right] = x$. This is, in general, the tightest bound provable by Chernoff's exponential moment method [13].

$$
\begin{aligned}
&= \mathrm{E}_{x\sim Q}\left[\ln \frac{dP(x)}{dQ(x)} e^{f(x)} + \ln \frac{dQ(x)}{dP(x)}\right] \\
&= KL(Q||P) + \mathrm{E}_{x\sim Q}\left[\ln \frac{dP(x)}{dQ(x)} e^{f(x)}\right] \\
&\leq KL(Q||P) + \ln \mathrm{E}_{x\sim Q}\left[\frac{dP(x)}{dQ(x)} e^{f(x)}\right] \\
&= KL(Q||P) + \ln \mathrm{E}_{x\sim P}\left[e^{f(x)}\right]
\end{aligned}
$$

□

Formula (5) can now be proved by assuming the body of Lemma 8 (which holds with probability at least $1-\delta$) and then observing the following where the last step follows from Jensen's inequality and strong convexity properties of KL-divergence.

$$
\begin{aligned}
\mathrm{E}_{h\sim Q}\left[(m-1)KL^{+}(\ell(h,S)||\ell(h,D))\right] &\leq KL(Q||P) + \ln \mathrm{E}_{h\sim P}\left[e^{(m-1)KL^{+}(\ell(h,S)||\ell(h,D))}\right] \\
&\leq KL(Q||P) + \ln \frac{m}{\delta} \\
(m-1)KL^{+}(\ell(Q,S)||\ell(Q,D)) &\leq KL(Q||P) + \ln \frac{m}{\delta}
\end{aligned}
$$