Journal of Machine Learning Research 4 (2003) 839-860 Submitted 4/03; Published 10/03

Generalization Error Bounds for Bayesian Mixture Algorithms

Ron Meir RMEIR@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Haifa 32000, Israel

Tong Zhang TZHANG@WATSON.IBM.COM
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA

Editors: Thore Graepel and Ralf Herbrich

Abstract

Bayesian approaches to learning and estimation have played a significant role in the Statistics literature over many years. While they are often provably optimal in a frequentist setting, and lead to excellent performance in practical applications, there have not been many precise characterizations of their performance for finite sample sizes under general conditions. In this paper we consider the class of Bayesian mixture algorithms, where an estimator is formed by constructing a data-dependent mixture over some hypothesis space. Similarly to what is observed in practice, our results demonstrate that mixture approaches are particularly robust, and allow for the construction of highly complex estimators, while avoiding undesirable overfitting effects. Our results, while being data-dependent in nature, are insensitive to the underlying model assumptions, and apply whether or not these hold. At a technical level, the approach applies to unbounded functions, constrained only by certain moment conditions. Finally, the bounds derived can be directly applied to non-Bayesian mixture approaches such as Boosting and Bagging.

1. Introduction and Motivation

The standard approach to Computational Learning Theory is usually formulated within the so-called frequentist approach to Statistics. Within this paradigm one is interested in constructing an estimator, based on a finite sample, which possesses a small loss (generalization error). While many algorithms have been constructed and analyzed within this context, it is not clear how these approaches relate to standard optimality criteria within the frequentist framework. Two classic optimality criteria within the latter approach are *minimaxity* and *admissibility*, which characterize optimality of estimators in a rigorous and precise fashion (Robert, 2001). Minimaxity essentially measures the performance of the *best* estimator for the *worst* possible distribution from some set of distributions. Admissibility is related to the extent to which an estimator uniformly dominates all other estimators. We refer the reader to Robert (2001) for precise definitions of these notions, as they play no role in the sequel. Except for some special cases (e.g., Yang, 1999), it is not known whether any of the approaches used within the Machine Learning community lead to optimality in either of the above senses of the word. On the other hand, it is known that under certain regularity conditions, Bayesian estimators lead to either minimax or admissible estimators, and thus to well-defined optimality in the classical (frequentist) sense. In fact, it can be shown that Bayes estimators, or limits thereof, are essentially the only estimators which can achieve optimality in the above senses (Robert, 2001). This optimality feature provides strong motivation for the study of Bayesian approaches in a *frequentist* setting.

While Bayesian approaches have been widely studied, there have not been generally applicable finite-sample bounds in the frequentist framework. Recently, several approaches have attempted to address this problem. In this paper we establish finite sample data-dependent bounds for Bayesian mixture methods, which together with the above optimality properties suggest that these approaches should become even more widely used.

Consider the problem of supervised learning where we attempt to construct an estimator based on a finite sample of pairs of examples $S = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, each pair drawn independently at random according to an unknown distribution $\mu(X,Y)$. Let $\mathcal{A}$ be a learning algorithm which, based on the sample S, selects a hypothesis (estimator) h from some set of hypotheses $\mathcal{H}$. Denoting by $\ell(y,h(\mathbf{x}))$ the instantaneous loss of the hypothesis h, we wish to assess the true loss

$$L(h) = \mathbf{E}_{X,Y}\ell(Y,h(X)) \qquad ((X,Y) \sim \mu).$$

In particular, the objective is to provide *algorithm* and *data-dependent* bounds of the following form. For any $h \in \mathcal{H}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,

$$L(h) \leq \Lambda(h,S) + \Delta(h,S,\delta),$$

where $\Lambda(h,S)$ is some empirical assessment of the true loss, and $\Delta(h,S,\delta)$ is a complexity term. For example, in the classic Vapnik-Chervonenkis framework (Vapnik and Chervonenkis, 1971), $\Lambda(h,S)$ is the empirical error $(1/n)\sum_{i=1}^n \ell(Y_i,h(X_i))$, and $\Delta(h,S,\delta)$ depends on the VC-dimension of $\mathcal{H}$ but is independent of both the hypothesis h and the sample S. By algorithm and data-dependent bounds we mean bounds where the complexity term depends on both the hypothesis (chosen by the algorithm $\mathcal{A}$) and the sample S.

The main contribution of the present work is the extension of the PAC-Bayesian framework of McAllester (1999, 2003) to a rather unified setting for Bayesian mixture methods, where different regularization criteria may be incorporated, and their effect on the performance can be easily assessed. Furthermore, it is also essential that the bounds obtained are *dimension-independent*, since otherwise they yield useless results when applied to methods based on high-dimensional mappings, such as kernel machines. Similar results can also be obtained using the covering number analysis by Zhang (2002a). However the approach presented in the current paper, which relies on the direct computation of the Rademacher complexity, is more direct and gives better bounds in many cases. The analysis is also easier to generalize than the corresponding covering number approach. Moreover, our analysis applies directly to other non-Bayesian mixture approaches such as Bagging and Boosting. On a technical level, our results remove a common limitation of many of the bounds in the learning community, namely their assumption of the boundedness of the underlying loss functions. This latter assumption is usually inappropriate for regression, and is often inapplicable to classification problems, where the $0-1$ loss function is replaced by a convex upper bound (see Section 6.4).

The remainder of the paper is organized as follows. We begin in Section 2 with a description of the decision theoretic framework for Bayesian learning. We then move on in Section 3 to discuss mixture distributions, and recall some basic properties of convex functions. Section 4 presents a new uniform convergence result for unbounded loss functions, and Section 5 then establishes bounds on the (Rademacher) complexity of classes of functions defined by convex constraints. Section 6 applies these general results to several cases of interest, establishing data-dependent bounds. We conclude in Section 7 and present some technical details in the appendix.

Before moving to the body of the paper, we make several comments concerning notation. Unless otherwise specified, the natural base of the logarithm is used. We denote random variables by upper-case letters and their realizations by lower case letters. Expectations with respect to a random variable X are denoted by $\mathbf{E}_X$. Vectors will be denoted using boldface.

2. A Decision Theoretic Bayesian Framework

In the decision theoretic Bayesian setting we consider three spaces. An input space $\mathcal{X}$, an action space $\mathcal{A}$ and an output space $\mathcal{Y}$. Consider a (deterministic) action $a = a(\mathbf{x})$ performed upon observing input $\mathbf{x}$, and let the loss function $\ell : \mathcal{Y} \times \mathcal{A} \mapsto \mathbb{R}$, be given by $\ell(y, a(\mathbf{x}))$. Let μ be a probability measure defined over $\mathcal{X} \times \mathcal{Y}$. The *Bayes optimal* decision rule $a \doteq a_\mu$ is given by minimizing $\mathbf{E}_{X,Y}\ell(Y, a(X))$, namely

$$\mathbf{E}_{X,Y}\ell(Y, a_\mu(X)) \leq \inf_{a\in\mathcal{A}} \mathbf{E}_{X,Y}\ell(Y, a(X)) \qquad ((X,Y) \sim \mu),$$

where, for ease of notation, we suppress the μ-dependence in the expectation.

In general, we do not have access to μ, but rather observe a sample $S = \{(X_i, Y_i)\}_{i=1}^n$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$. Let $a = a(\mathbf{x}, S)$ be an action selected based on the sample S and the current input $\mathbf{x}$. We refer to such a sample-dependent action as an *algorithm*. The *sample dependent* loss of a is given by

$$R(\mu, a, S) = \mathbf{E}_{X,Y}\ell(Y, a(X, S)).$$

We are interested in the expected loss of an algorithm averaged over samples S,

$$R(\mu, a) = \mathbf{E}_S R(\mu, a, S) = \int R(\mu, a, S) d\mu(S),$$

where the expectation is taken with respect to the sample S drawn i.i.d. from the probability measure μ. If we consider a family of measures μ, which possesses some underlying *prior distribution* $\pi(\mu)$, then we can construct the averaged risk function with respect to the prior as,

$$r(\pi, a) = \mathbf{E}_\pi R(\mu, a) = \int d\mu(S) d\pi(\mu) \int R(\mu, a, S) d\pi(\mu|S),$$

where

$$d\pi(\mu|S) = \frac{d\mu(S) d\pi(\mu)}{\int_\mu' d\mu'(S) d\pi(\mu')}$$

is the *posterior distribution* on the μ family, which induces a posterior distribution on the sample space as $\pi_S = E_{\pi(\mu|S)}\mu$. An action (algorithm) $a \doteq a_B$ minimizing the Bayes risk $r(\pi, a)$ is referred to as a *Bayes algorithm*, namely

$$r(\pi, a_B) \leq \inf_{a\in\mathcal{A}} r(\pi, a).$$

In fact, for a given prior, and a given sample S, the optimal algorithm should return the Bayes optimal predictor with respect to the posterior measure π_S.

For many important practical problems, the optimal Bayes predictor is a linear functional of the underlying probability measure. For example, if the loss function is quadratic, namely $\ell(y, a(\mathbf{x})) =$

$(y-a(\mathbf{x}))^2$, then the optimal Bayes predictor $a_\mu(\mathbf{x})$ is the conditional mean of y, namely $\mathbf{E}[Y|\mathbf{x}]$. For binary classification problems, we can let the predictor be the conditional probability $a_\mu(\mathbf{x}) = \mu(Y = 1|\mathbf{x})$ (the optimal classification decision rule then corresponds to a test of whether $a_\mu(\mathbf{x}) > 0.5$), which is also a linear functional of μ. Clearly if the Bayes predictor is a linear functional of the probability measure, then the optimal Bayes algorithm with respect to the prior π is given by

$$a_B(\mathbf{x}, S) = \int_\mu a_\mu(\mathbf{x}) d\pi(\mu|S) = \frac{\int_\mu a_\mu(\mathbf{x}) d\mu(S) d\pi(\mu)}{\int_\mu d\mu(S) d\pi(\mu)}. \tag{1}$$

In this case, an optimal Bayesian algorithm can be regarded as the predictor constructed by averaging over all predictors with respect to a data-dependent posterior $\pi(\mu|S)$. We refer to such methods as *Bayesian mixture methods*. While the Bayes estimator $a_B(\mathbf{x}, S)$ is optimal with respect to the Bayes risk $r(\pi, a)$, it can be shown, that under appropriate conditions (and an appropriate prior) it is also a minimax and admissible estimator (Robert, 2001).

In general, a_μ is unknown. Rather we may have some prior information about possible models for a_μ. In view of (1) we consider a hypothesis space $\mathcal{H}$, and an algorithm based on a mixture of hypotheses $h \in \mathcal{H}$. This should be contrasted with classical approaches where an algorithm selects a single hypothesis h from $\mathcal{H}$. For simplicity, we consider a countable hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots\}$, and a probability distribution $\{q_j\}_{j=1}^\infty$ over $\mathcal{H}$, namely $q_j \geq 0$ and $\sum_j q_j = 1$.[1] We introduce the vector notation $\mathbf{q} = (q_1, q_2, \ldots)$ and $\mathbf{h} = (h_1, h_2, \ldots)$, and define the *probability simplex*

$$\Pi = \left\{ \mathbf{q} : \; q_j \geq 0, \; \sum_j q_j = 1 \right\}.$$

Further, denote

$$f_q(\mathbf{x}) \stackrel{\triangle}{=} \langle \mathbf{q}, \mathbf{h}(\mathbf{x}) \rangle = \sum_{j=1}^\infty q_j h_j(\mathbf{x}) \qquad (\mathbf{q} \in \Pi).$$

Observe that in general $f_q(\mathbf{x})$ may be a great deal more complex that any single hypothesis h_j. For example, if $h_j(\mathbf{x})$ are non-polynomial ridge functions, the composite predictor f corresponds to a two-layer neural network with universal approximation power (Leshno et al., 1993).

A main feature of this work is the establishment of data-dependent bounds on $L(f_q)$, the loss of the Bayes mixture algorithm. There has been a flurry of recent activity concerning data-dependent bounds (a non-exhaustive list includes Bartlett et al., 2002b, Bousquet and Chapelle, 2002, Koltchinksii and Panchenko, 2002, Shawe-Taylor et al., 1998, Zhang, 2001). In a related vein, McAllester (1999, 2003) provided a data-dependent bound for the so-called Gibbs algorithm, which selects a hypothesis at random from $\mathcal{H}$ based on the posterior distribution $\pi(h|S)$. Essentially, this result provides a bound on the average error $\sum_j q_j L(h_j)$ rather than a bound on the error of the *averaged hypothesis*, $L(\sum_j q_j h_j)$, which may be much smaller. Later, Langford et al. (2001) extended this result to a mixture of classifiers using a margin-based loss function. A more general result can also be obtained using the covering number approach described by Zhang (2002a). Finally, Herbrich and Graepel (2001) showed that under certain conditions the bounds for the Gibbs classifier can be extended to a Bayesian mixture classifier. However, their bound contained an explicit dependence on the dimension (see Thm. 3 in Herbrich and Graepel, 2001).

Although the approach pioneered by McAllester (1999, 2003) came to be known as PAC-Bayes, this term is somewhat misleading since an optimal Bayesian method (in the decision theoretic framework outline above) does not average over loss functions but rather over hypotheses. In this regard,

1. The assumption that the hypothesis space is countable can be removed. We retain it, however, for ease of presentation.

the learning behavior of a true Bayesian method is not addressed in the PAC-Bayes analysis. In this paper, we attempt to narrow the discrepancy by analyzing Bayesian mixture methods, where we consider a predictor that is the average of a family of predictors with respect to a data-dependent posterior distribution. Bayesian mixtures can often be regarded as a good approximation to truly optimal Bayesian methods. In fact, we have argued above that they are equivalent for many important practical problems.

3. Mixture Algorithms with Convex Constraints

A learning algorithm within the Bayesian mixture framework uses the sample S to select a distribution $\mathbf{q}$ over $\mathcal{H}$ and then constructs a mixture hypothesis f_q. In order to constrain the class of mixtures used in forming the mixture f_q we impose constraints on the mixture vector $\mathbf{q}$. Let $g(\mathbf{q})$ be a non-negative convex function of $\mathbf{q}$ and define for any positive A,

$$\begin{aligned} \Omega_A &= \{\mathbf{q} \in \Pi : \ g(\mathbf{q}) \le A\}, \\ \mathcal{F}_A &= \left\{ f_q : \ f_q(\mathbf{x}) = \langle \mathbf{q}, \mathbf{h}(\mathbf{x}) \rangle, \ \mathbf{q} \in \Omega_A \right\}. \end{aligned} \tag{2}$$

In subsequent sections we will consider different choices for $g(\mathbf{q})$, which essentially acts as a regularization term. Finally, for any mixture f_q we define the loss by

$$L(f_q) = \mathbf{E}_{X,Y} \ell(Y, f_q(\mathbf{X}))$$

and the empirical loss incurred on the sample by

$$\hat{L}(f_q) = (1/n) \sum_{i=1}^{n} \ell(Y_i, f_q(X_i)).$$

In the sequel we use the notation $\hat{\mathbf{E}}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$, and $\mathbf{E}_S$ stands for an average over the sample S with respect to the distribution $\mu(S)$.

For future reference, we formalize our assumptions concerning $g(\mathbf{q})$.

Assumption 1 *The constraint function $g(\mathbf{q})$ is convex and non-negative.*

An important tool which is used extensively in this paper is the theory of convex duality (Rockafellar, 1970, Boyd and Vandenberghe, 2002). We begin by discussing some issues and introduce several useful results.

3.1 Some Results on Convex Functions and Duality

Let $f(\mathbf{x})$ denote a convex function, namely f is defined over a convex domain K and for any $0 \le \theta \le 1$ and $\mathbf{x}, \mathbf{y} \in K$

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \le \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}).$$

Definition 1 *For a function f, we define*

$$u_f(\mathbf{x}) = \sup_{\mathbf{r} \in K} \left[\frac{f(\mathbf{r}+\mathbf{x}) + f(\mathbf{r}-\mathbf{x})}{2} - f(\mathbf{r}) \right].$$

The following result follows directly from a Taylor expansion.

Lemma 2 *Assume f possesses continuous first order derivatives. Then for all $q > 1$:*

$$u_f(\mathbf{x}) \leq \sup_{\mathbf{r},\theta\in(0,1)} \theta^{1-q} \frac{d}{d\theta}\left\{\frac{f(\mathbf{r}+\theta\mathbf{x}) - f(\mathbf{r}-\theta\mathbf{x})}{2q}\right\}.$$

Moreover, if f possesses continuous second order derivatives, then

$$u_f(\mathbf{x}) \leq \frac{1}{2} \sup_{\mathbf{r},|\theta|\leq 1} \frac{d^2}{d\theta^2} f(\mathbf{r}+\theta\mathbf{x}).$$

Proof For any $\theta \in \mathbb{R}$, let $s(\theta) = [f(\mathbf{r}+\theta\mathbf{x}) + f(\mathbf{r}-\theta\mathbf{x})]/2 - f(\mathbf{r})$. Observe that $s(0) = s'(0) = 0$. From the generalized mean value Theorem (e.g., Theorem 5.15 in Apostol 1957) it is known that for two functions h and g, which are continuously differentiable over $[0,1]$, $[h(\theta) - h(\theta_0)]g'(\theta_1) = [g(\theta) - g(\theta_0)]f'(\theta_1)$, for any $\theta, \theta_0 \in [0,1]$ and some $\theta_1 \in [\theta_0, \theta]$. Replacing h by s and setting $g(\theta) = \theta^q$, $q \geq 1$, we infer that there exists a $\theta_1 \in (0,1)$ such that $s(1) = s'(\theta_1)/(q\theta_1^{q-1})$. If s is continuously second order differentiable, then a second order Taylor expansion with remainder shows that there exists a $\theta_2 \in (0,1)$ such that $s(1) = s''(\theta_2)/2$. ☐

For any function f defined over a domain K we define the conjugate f^* by

$$f^*(\mathbf{y}) = \sup_{\mathbf{x}\in K} \left(\langle \mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x})\right),$$

noting that $f^*(\cdot)$ is always convex (irrespective of the convexity of $f(\cdot)$). The domain of f^* consists of all values of $\mathbf{y}$ for which the supremum is finite, namely the values of $\mathbf{y}$ for which $\langle \mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x})$ is bounded from above on K.

A simple consequence of the definition of f^* is the so called *Fenchel inequality*, which states that for all $\mathbf{x}$ and $\mathbf{y}$

$$\langle \mathbf{y}, \mathbf{x}\rangle \leq f(\mathbf{x}) + f^*(\mathbf{y}). \tag{3}$$

4. A Concentration Inequality for Unbounded Functions

In general, loss functions used in applications cannot be bounded a-priori. The starting point for our analysis is a concentration result similar to Theorem 1 of Koltchinksii and Panchenko (2002) (see also Theorem 8 of Bartlett and Mendelson, 2002). The main advantage of the current formulation is that the functions in $\mathcal{F}$ are not assumed to be bounded. This is particularly useful in the context of regression. The proof is given in the appendix.

Theorem 3 *Let $\mathcal{F}$ be a class of functions mapping from a domain $\mathcal{X}$ to $\mathbb{R}$, and let $\{X_i\}_{i=1}^n$ be independently selected according to a probability measure P. Assume that there exists a positive number $M(\mathcal{F})$ such that for all $\lambda > 0$:*

$$\log \mathbf{E}_X \sup_{f\in\mathcal{F}} \cosh(2\lambda f(X)) \leq \lambda^2 M(\mathcal{F})^2/2.$$

Then, for any integer n and $0 < \delta < 1$, with probability at least $1-\delta$ over samples of length n, every $f \in \mathcal{F}$ satisfies

$$\mathbf{E}f(X) \leq \hat{\mathbf{E}}_n f(X) + \mathbf{E}_S \sup_{f\in\mathcal{F}} \left\{\mathbf{E}f(X) - \hat{\mathbf{E}}_n f(X)\right\} + M(\mathcal{F})\sqrt{\frac{2\log(1/\delta)}{n}}.$$

We note that the dependence of M on $\mathcal{F}$ is made explicit, as it will play a role in the sequel. The bound can be slightly improved when the functions in $\mathcal{F}$ are bounded.

Corollary 4 *Let the conditions of Theorem 3 hold, and in addition assume that*

$$\sup_{f,\mathbf{x},\mathbf{x}'} |f(\mathbf{x}) - f(\mathbf{x}')| \le M(\mathcal{F}).$$

Then

$$\mathbf{E}f(X) \le \hat{\mathbf{E}}_n f(X) + \mathbf{E}_S \sup_{f\in\mathcal{F}} \left\{\mathbf{E}f(X) - \hat{\mathbf{E}}_n f(X)\right\} + M(\mathcal{F})\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof In the proof of Lemma 17 in the Appendix, note that $\sup_{\mathbf{x}_1,\mathbf{x}_1'} |c'(\mathbf{x}_1) - c'(\mathbf{x}_1')| \le \lambda \sup_{\mathbf{x},\mathbf{x}'} |f(\mathbf{x}) - f(\mathbf{x}')| \le \lambda M$. Now instead of bounding $E_{X_1}\exp(c'(X_1) - E_{X_1'}c'(X_1'))$ using the symmetrization argument as in Lemma 17, we may apply Chernoff's bound which leads to $\log E_{X_1}\exp(c'(X_1) - E_{X_1'}c'(X_1')) \le \lambda^2 M^2/8$. □

In spite of the slightly improved bound in the case of bounded functions, we will use the bound of Theorem 3 for generality.

A great deal of recent work has dealt with Rademacher complexity based bounds. Denote by $\{\sigma_i\}_{i=1}^n$ independent Bernoulli random variables assuming the values ± 1 with equal probability. For a set of n data points $X^n = \{X_i\}_{i=1}^n \in \mathcal{X}^n$, we define the data-dependent Rademacher complexity of $\mathcal{F}$ as

$$\hat{R}_n(\mathcal{F}) = \mathbf{E}_\sigma \left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \,|\, X^n\right],$$

where $\sigma = (\sigma_1,\ldots,\sigma_n)$. The expectation of $\hat{R}_n(\mathcal{F})$ with respect to X^n will be denoted by $R_n(\mathcal{F})$. Note that $\hat{R}_n(\mathcal{F})$ differs from the standard Rademacher complexity $\hat{R}_n(\mathcal{F})$ which is defined using the absolute value $|(1/n)\sum_{i=1}^n \sigma_i f(X_i)|$ in the argument of the supremum (van der Vaart and Wellner, 1996). The current version of Rademacher complexity has the merit that it vanishes for function classes consisting of single constant function, and is always dominated by the standard Rademacher complexity. Both definitions agree for function classes which are closed under negation, namely classes $\mathcal{F}$ for which $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$.

Using standard symmetrization arguments (for example, Lemma 2.3.1 of van der Vaart and Wellner, 1996) one can show that

$$\mathbf{E}_{X^n} \sup_{f\in\mathcal{F}} \left\{\mathbf{E}f(X) - \hat{\mathbf{E}}_n f(X)\right\} \le 2R_n(\mathcal{F}).$$

It is often convenient to use the Rademacher average due to the following Lemma.

Lemma 5 *Let $\{g_i(\theta)\}$ and $\{h_i(\theta)\}$ be sets of functions defined for all θ in some domain Θ. If for all i, θ, θ', $|g_i(\theta) - g_i(\theta')| \le |h_i(\theta) - h_i(\theta')|$, then for any function $c(\mathbf{x},\theta)$, $x \in \mathcal{X}_1$, and probability distribution over $\mathcal{X}$,*

$$\mathbf{E}_\sigma \mathbf{E}_X \sup_{\theta\in\Theta} \left\{c(X,\theta) + \sum_{i=1}^n \sigma_i g_i(\theta)\right\} \le \mathbf{E}_\sigma \mathbf{E}_X \sup_{\theta\in\Theta} \left\{c(X,\theta) + \sum_{i=1}^n \sigma_i h_i(\theta)\right\}.$$

Proof By induction. The result holds for $n=0$. Then when $n=k+1$

$$
\begin{aligned}
&\mathbf{E}_{\sigma_1,\ldots,\sigma_{k+1}}\mathbf{E}_X \sup_{\theta}\left\{c(X,\theta)+\sum_{i=1}^{k+1}\sigma_i g_i(\theta)\right\} \\
&=\mathbf{E}_{\sigma_1,\ldots,\sigma_k}\mathbf{E}_X \sup_{\theta_1,\theta_2}\left\{\frac{c(X,\theta_1)+c(X,\theta_2)}{2}+\sum_{i=1}^{k}\sigma_i\frac{g_i(\theta_1)+g_i(\theta_2)}{2}+\frac{g_{k+1}(\theta_1)-g_{k+1}(\theta_2)}{2}\right\} \\
&=\mathbf{E}_{\sigma_1,\ldots,\sigma_k}\mathbf{E}_X \sup_{\theta_1,\theta_2}\left\{\frac{c(X,\theta_1)+c(X,\theta_2)}{2}+\sum_{i=1}^{k}\sigma_i\frac{g_i(\theta_1)+g_i(\theta_2)}{2}+\frac{|g_{k+1}(\theta_1)-g_{k+1}(\theta_2)|}{2}\right\} \\
&\leq\mathbf{E}_{\sigma_1,\ldots,\sigma_k}\mathbf{E}_X \sup_{\theta_1,\theta_2}\left\{\frac{c(X,\theta_1)+c(X,\theta_2)}{2}+\sum_{i=1}^{k}\sigma_i\frac{g_i(\theta_1)+g_i(\theta_2)}{2}+\frac{|h_{k+1}(\theta_1)-h_{k+1}(\theta_2)|}{2}\right\} \\
&=\mathbf{E}_{\sigma_1,\ldots,\sigma_k}\mathbf{E}_{\sigma_{k+1}}\mathbf{E}_X \sup_{\theta}\left\{c(X,\theta)+\sigma_{k+1}h_{k+1}(\theta)+\sum_{i=1}^{k}\sigma_i g_i(\theta)\right\} \\
&\leq\mathbf{E}_{\sigma_1,\ldots,\sigma_k}\mathbf{E}_{\sigma_{k+1}}\mathbf{E}_X \sup_{\theta}\left\{c(X,\theta)+\sigma_{k+1}h_{k+1}(\theta)+\sum_{i=1}^{k}\sigma_i h_i(\theta)\right\}.
\end{aligned}
$$

The last inequality follows from the induction hypothesis. □

Remark 6 The above lemma is a refined (and *symmetric*) version of the Rademacher process comparison theorem (Theorem 4.12 of Ledoux and Talgrand, 1991). The proof presented here is also simpler.

Let $\{\phi_i\}$ be a set of functions, each characterized by a Lipschitz constant γ_i, namely $|\phi_i(\theta)-\phi_i(\theta')|\leq\gamma_i|\theta-\theta'|$. The following consequence is immediate from Lemma 5.

Theorem 7 *Let $\{\phi_i\}_{i=1}^n$ be functions with Lipschitz constants γ_i, then*

$$\mathbf{E}_\sigma\left\{\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sigma_i\phi_i(f(\mathbf{x}_i))\right\}\leq\mathbf{E}_\sigma\left\{\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sigma_i\gamma_i f(\mathbf{x}_i)\right\}.$$

Let $\ell(y,f(\mathbf{x}))$ be a loss function and set $\phi_i(f(\mathbf{x}_i))=(\phi_i\circ f)(y_i,\mathbf{x}_i)=\ell(y_i,f(\mathbf{x}_i))$. Assume that $\phi_i(f(X_i))$ is Lipschitz with constant κ, namely $|\phi_i(f(\mathbf{x}_i))-\phi_i(f'(\mathbf{x}_i)|\leq\kappa|f(\mathbf{x}_i)-f'(\mathbf{x}_i)|$ for all i. Let $\mathcal{L}_\mathcal{F}$ consist of functions from $\mathcal{Y}\times\mathcal{X}$, defined by $\mathcal{L}_\mathcal{F}=\{g:\ g=\phi\circ f,\ f\in\mathcal{F}\}$, where ϕ is Lipschitz with constant κ. Then we find from Theorem 7 that $R_n(\mathcal{L}_\mathcal{F})\leq\kappa R_n(\mathcal{F})$. We note in passing that by using Theorem 7 we gain a factor of 2 compared to the bound in Corollary 3.17 of Ledoux and Talgrand (1991) and do away with their requirement that $\phi_i(0)=0$.

Setting $L(f)=\mathbf{E}_{X,Y}\ell(Y,f(X))$ and $\hat{L}(f)=\hat{\mathbf{E}}_n\ell(Y,f(X))$, we obtain the following bound for the expected loss.

Theorem 8 *Let $\mathcal{F}$ be a class of functions mapping from a domain $\mathcal{X}$ to $\mathbb{R}$, and let $\{(X_i,Y_i)\}_{i=1}^n$, $X_i\in\mathcal{X}$, $Y_i\in\mathbb{R}$, be independently selected according to a probability measure P. Assume there exists a positive real number $M(\mathcal{F})$ such that for all positive λ*

$$\log\mathbf{E}_{X,Y}\sup_{f\in\mathcal{F}}\cosh(2\lambda\ell(Y,f(X)))\leq\lambda^2M(\mathcal{F})^2/2,$$

where for every $f \in \mathcal{F}$, $\phi_i(f(X_i)) = (\phi \circ f)(Y_i, X_i) = \ell(Y_i, f(X_i))$ is Lipschitz with constant $\kappa(\mathcal{F})$. Then with probability at least $1-\delta$ over samples of length n, every $f \in \mathcal{F}$ satisfies

$$L(f) \leq \hat{L}(f) + 2\kappa(\mathcal{F}) R_n(\mathcal{F}) + M(\mathcal{F})\sqrt{\frac{2\log(1/\delta)}{n}}.$$

5. The Rademacher Complexity for Classes Defined by Convex Constraints

We consider the class of functions $\mathcal{F}_A$ defined in (2) through a convex constraint function $g(\mathbf{q})$. We wish to compute the Rademacher complexity $R_n(\mathcal{F}_A)$. Denoting by g^* the conjugate function to g, we have from (3) that for all $\mathbf{q}$ and $\mathbf{z}$

$$\langle \mathbf{q}, \mathbf{z} \rangle \leq g(\mathbf{q}) + g^*(\mathbf{z}).$$

Setting $\mathbf{z} = (\lambda/n)\sum_{i=1}^n \sigma_i \mathbf{h}(X_i)$, we conclude that for any positive λ

$$\mathbf{E}_\sigma \sup_{\mathbf{q}\in\Omega_A} \left\{ \frac{1}{n}\sum_{i=1}^n \sigma_i \langle \mathbf{q}, \mathbf{h}(X_i)\rangle \right\} \leq \frac{1}{\lambda}\left\{ A + \mathbf{E}_\sigma g^*\left((\lambda/n)\sum_{i=1}^n \sigma_i \mathbf{h}(X_i) \right) \right\}.$$

Since this inequality holds for every $\lambda > 0$, we obtain the following upper bound on the Rademacher complexity,

$$\hat{R}_n(\mathcal{F}_A) \leq \inf_{\lambda \geq 0} \left\{ \frac{A}{\lambda} + \frac{1}{\lambda}\mathbf{E}_\sigma g^*\left((\lambda/n)\sum_{i=1}^n \sigma_i \mathbf{h}(X_i) \right) \right\}. \qquad (4)$$

We note that a similar use of convex duality was made in a related context by Seeger (2002)

In general, it may be difficult to compute the expectation of g^* with respect to σ. For this purpose we make use of the following Lemma. Note that $g(\mathbf{q}) \geq 0$ implies that $g^*(0) = \sup_{\mathbf{q}\in\Omega_A}\{-g(\mathbf{q})\} \leq 0$.

Lemma 9 *For any $a > 0$ and convex function f such that $f(0) \leq 0$,*

$$\mathbf{E}_\sigma f\left(a\sum_{i=1}^n \sigma_i \mathbf{h}(\mathbf{x}_i) \right) \leq \sum_{i=1}^n u_f(a\mathbf{h}(\mathbf{x}_i)). \qquad (5)$$

Proof We prove the claim by induction. For $n = 1$ we have

$$\begin{aligned} \mathbf{E}_\sigma f(a\sigma\mathbf{h}(\mathbf{x}_1)) &= \frac{1}{2}[f(a\mathbf{h}(\mathbf{x}_1)) + f(-a\mathbf{h}(\mathbf{x}_1))] - f(0) + f(0) \\ &\leq u_f(a\mathbf{h}(\mathbf{x}_1)), \end{aligned}$$

where we have used $f(0) \leq 0$. Next, assume the claim holds for n and let $\sigma_n = \{\sigma_1, \ldots, \sigma_n\}$. We have

$$
\begin{aligned}
& \mathbf{E}_{\sigma_n} \mathbf{E}_{\sigma_{n+1}} f\left(a \sum_{i=1}^{n+1} \sigma_i \mathbf{h}(\mathbf{x}_i)\right) \\
&= \frac{1}{2} \mathbf{E}_{\sigma_n} \left[f\left(a \sum_{i=1}^{n} \sigma_i \mathbf{h}(\mathbf{x}_i) + a\mathbf{h}(\mathbf{x}_{n+1})\right) + f\left(a \sum_{i=1}^{n} \sigma_i \mathbf{h}(\mathbf{x}_i) - a\mathbf{h}(\mathbf{x}_{n+1})\right) \right] \\
&= \frac{1}{2} \mathbf{E}_{\sigma_n} \left[f\left(a \sum_{i=1}^{n} \sigma_i \mathbf{h}(\mathbf{x}_i) + a\mathbf{h}(\mathbf{x}_{n+1})\right) + f\left(a \sum_{i=1}^{n} \sigma_i \mathbf{h}(\mathbf{x}_i) - a\mathbf{h}(\mathbf{x}_{n+1})\right) \right] \\
& - \mathbf{E}_{\sigma} f\left(a \sum_{i=1}^{n} \sigma_i \mathbf{h}(\mathbf{x}_i)\right) + \mathbf{E}_{\sigma} f\left(a \sum_{i=1}^{n} \sigma_i \mathbf{h}(\mathbf{x}_i)\right) \\
& \leq u_f(a\mathbf{h}(\mathbf{x}_{n+1})) + \sum_{i=1}^{n} u_f(a\mathbf{h}(\mathbf{x}_i)),
\end{aligned}
$$

where the last step used the definition of u_f and the induction hypothesis. □

Using (4) and Lemma 9 we find that

$$
\hat{R}_n(\mathcal{F}_A) \leq \inf_{\lambda \geq 0} \left\{ \frac{A}{\lambda} + \frac{1}{\lambda} \sum_{i=1}^{n} u_{g^*}((\lambda/n)\mathbf{h}(X_i)) \right\}. \tag{6}
$$

6. Data-dependent Bounds

Consider the loss bound derived in Theorem 8. This bound requires prior knowledge of the constant A, characterizing the class $\mathcal{F}_A$. In general, we would like to be able to establish a bound which is *data-dependent*, namely does not assume any such a-priori knowledge. We begin by rewriting the bound of Theorem 8 in a slightly different form. For any $f_q = \langle \mathbf{q}, \mathbf{h} \rangle$, $\mathbf{q} \in \Omega_A$, with probability at least $1 - \delta$

$$
L(f_q) \leq \hat{L}(f_q) + 2\kappa(A)\Upsilon(A) + M(A)\sqrt{\frac{2\log(1/\delta)}{n}}, \tag{7}
$$

where we slightly abuse notation, setting $\kappa(A) = \kappa(\mathcal{F}_A)$, $M(A) = M(\mathcal{F}_A)$ and where

$$
\Upsilon(A) = R_n(\mathcal{F}_A).
$$

Observe that $\Upsilon(A)$ is monotonically increasing in A. Either (4) or (6) may be used to upper bound $\Upsilon(A)$. For example, using (4) we have that

$$
\Upsilon(A) \leq \mathbf{E}_S \inf_{\lambda \geq 0} \left\{ \frac{A}{\lambda} + \frac{1}{\lambda} \mathbf{E}_{\sigma} g^* \left((\lambda/n) \sum_{i=1}^{n} \sigma_i \mathbf{h}(X_i) \right) \right\}.
$$

Eliminating the dependence on A in (7) leads to the following fully data-dependent bound.

Theorem 10 *Let the assumptions of Theorem 8 hold. Consider two parameters $g_0 > 0$ and $s > 1$, and let $\tilde{g}(\mathbf{q}) = s\max(g(\mathbf{q}), g_0)$. Then with probability at least $1 - \delta$ for all f_q, $\mathbf{q} \in \Pi$,*

$$
L(f_q) \leq \hat{L}(f_q) + 2\kappa(\tilde{g}(\mathbf{q}))\Upsilon(\tilde{g}(\mathbf{q})) + M(\tilde{g}(\mathbf{q}))\sqrt{\frac{4\log\log_s(s\tilde{g}(\mathbf{q})/g_0) + 2\log(1/\delta)}{n}}.
$$

Proof First, observe that $\tilde{g}(\mathbf{q})/g_0 \geq s$, so that the final term is always well-defined. Let $\{A_i\}_{i=1}^{\infty}$ and $\{p_i\}_{i=1}^{\infty}$ be a sets of positive numbers such that $\sum_i p_i = 1$. From Theorem 8 and the multiple-testing Lemma (essentially a slightly refined union bound) we have that with probability at least $1-\delta$ for all A_i and $\mathbf{q} \in \Omega_{A_i}$,

$$L(f_q) \leq \hat{L}(f_q) + 2\kappa(A_i)\Upsilon(A_i) + M(A_i)\sqrt{\frac{2\log(1/p_i\delta)}{n}}. \tag{8}$$

Next, pick $A_i = g_0 s^i$ and $p_i = 1/i(i+1),\ i = 1,2,\ldots$ (note that $\sum_i p_i = 1$). For each $\mathbf{q}$ let $i_{\mathbf{q}}$ denote the smallest index for which $A_{i_{\mathbf{q}}} \geq g(\mathbf{q})$. We have $i_{\mathbf{q}} \leq \log_s(\tilde{g}(\mathbf{q})/g_0)$, and $A_{i_{\mathbf{q}}} \leq \tilde{g}(\mathbf{q})$. Substituting $p_{i_{\mathbf{q}}} = 1/i_{\mathbf{q}}(1+i_{\mathbf{q}})$ we have that $\log(1/p_{i_{\mathbf{q}}}) \leq 2\log(i_{\mathbf{q}}+1) \leq 2\log\log_s(s\tilde{g}(\mathbf{q})/g_0)$. Combing these bounds with (8), and keeping in mind the monotonicity of $\Upsilon(A)$, we have that with probability at least $1-\delta$ for all $\mathbf{q}$

$$L(f_q) \leq \hat{L}(f_q) + 2\kappa(\tilde{g}(\mathbf{q}))\Upsilon(\tilde{g}(\mathbf{q})) + M(\tilde{g}(\mathbf{q}))\sqrt{\frac{4\log\log_s(s\tilde{g}(\mathbf{q})/g_0) + 2\log(1/\delta)}{n}},$$

which concludes the proof. □

Note that the parameter g_0 essentially 'sets the scale' for $g(\mathbf{q})$. For example, if g_0 is selected so that $g(\mathbf{q}) \leq g_0$ for all $\mathbf{q}$, we get a data-independent bound, where q_0 replaces $g(\mathbf{q})$. We also observe that the bounds derived in Theorem 10 are *data-dependent* and can thus be used in order to select the optimal posterior distribution $\mathbf{q}$. We comment on this further in Section 6.1.

We observe that the bounds in Theorem 10 yields rates which are $O\left(n^{-1/2}\right)$. More recent techniques based on more refined concentration inequalities (e.g. Boucheron et al. 2003, Bartlett et al. 2002a, Mannor et al. 2003) are sometimes able to achieve faster rates of convergence under favorable circumstances. For example, faster rates are possible if the empirical error is small. We leave the extension of our results to these situations to future work.

6.1 Entropic Constraints

Assume a *data-independent* prior distribution ν is assigned to all hypotheses in $\mathcal{H}$, namely $\nu_j \geq 0$ and $\sum_j \nu_j = 1$, where $\nu_j = \nu(h_j)$. We set $g(\mathbf{q})$ to be the Kullback-Leibler divergence of $\mathbf{q}$ from ν.

$$g(\mathbf{q}) = D(\mathbf{q}\|\nu) \quad ; \quad D(\mathbf{q}\|\nu) = \sum_j q_j \log(q_j/\nu_j).$$

In this case, the conjugate function g^* can be explicitly calculated yielding

$$g^*(\mathbf{z}) = \log \sum_j \nu_j e^{z_j}.$$

Note that

$$\frac{d^2}{d\theta^2} g^*(\mathbf{z} + \theta\mathbf{z}') \leq \frac{\sum_j \nu_j {z'_j}^2 e^{z_j + \theta z'_j}}{\sum_j \nu_j e^{z_j + \theta z'_j}}.$$

It is easy to see that

$$\sup_{\mathbf{z},\mathbf{z}',\theta} \frac{d^2}{d\theta^2} g^*(\mathbf{z} + \theta\mathbf{z}') \leq \|\mathbf{z}'\|_\infty^2.$$

Using Lemma 2, we have $u_{g^*}(\mathbf{h}(\mathbf{x}_i)) \leq \|\mathbf{h}(\mathbf{x}_i)\|_\infty^2/2$, and (6) can then be applied. However, a slightly better bound can be obtained with a more refined derivation. Using (4) we can derive an upper bound on the Rademacher complexity, captured in the following Lemma.

Lemma 11 *The empirical Rademacher complexity of $\mathcal{F}_A$ using $g(\mathbf{q}) = D(\mathbf{q}\|\nu)$ is upper bounded as follows:*

$$\hat{R}_n(\mathcal{F}_A) \leq \left(\sqrt{\frac{2A}{n}}\right) \sup_j \sqrt{\frac{1}{n}\sum_{i=1}^n h_j(X_i)^2} \ .$$

Proof From (4) and the expression for g^* we have that for any $\lambda > 0$

$$\sup_{\mathbf{q}\in\Omega_A} \left\{ \frac{1}{n}\sum_{i=1}^n \sigma_i \langle \mathbf{q}, \mathbf{h}(X_i)\rangle \right\} \leq \frac{1}{\lambda}\left\{ A + \log \sum_j \nu_j \exp\left[\frac{\lambda}{n}\sum_i \sigma_i h_j(X_i)\right]\right\}.$$

Taking the expectation with respect to $\sigma = (\sigma_1, \ldots, \sigma_n)$, and using the Chernoff bound $\mathbf{E}_\sigma \left\{\exp\left(\sum_i \sigma_i a_i\right)\right\} \leq \exp\left(\sum_i a_i^2/2\right)$, we have that for any $\lambda \geq 0$

$$\begin{aligned}
\hat{R}_n(\mathcal{F}_A) &\leq \frac{1}{\lambda}\left\{ A + \mathbf{E}_\sigma \log \sum_j \nu_j \exp\left[\frac{\lambda}{n}\sum_i \sigma_i h_j(X_i)\right]\right\} \\
&\overset{(a)}{\leq} \frac{1}{\lambda}\left\{ A + \sup_j \log \mathbf{E}_\sigma \exp\left[\frac{\lambda}{n}\sum_i \sigma_i h_j(X_i)\right]\right\} \\
&\overset{(b)}{\leq} \frac{1}{\lambda}\left\{ A + \sup_j \log \exp\left[\frac{\lambda^2}{n^2}\sum_i \frac{h_j(X_i)^2}{2}\right]\right\} \\
&= \frac{A}{\lambda} + \frac{\lambda}{2n^2}\sup_j \sum_i h_j(X_i)^2,
\end{aligned}$$

where (a) made use of Jensen's inequality and (b) used Chernoff's bound. Minimizing the r.h.s. with respect to λ, we obtain the desired result. $\square$

Using this result in Theorem 10 we obtain the main result of this section.

Theorem 12 *Let the conditions of Theorem 10 hold, and set*

$$\tilde{g}(\mathbf{q}) = s\max(D(\mathbf{q}\|\nu), g_0) \quad ; \quad \Delta_{\mathcal{H}} = \sqrt{\frac{1}{n}\mathbf{E}_S \sup_j \sum_{i=1}^n h_j(X_i)^2}.$$

Then for all f_q, $\mathbf{q} \in \Pi$, with probability at least $1-\delta$,

$$L(f_q) \leq \hat{L}(f_q) + 2\Delta_{\mathcal{H}}\kappa(\tilde{g}(\mathbf{q}))\sqrt{\frac{2\tilde{g}(\mathbf{q})}{n}} + M(\tilde{g}(\mathbf{q}))\sqrt{\frac{4\log\log_s(s\tilde{g}(\mathbf{q})/g_0) + 2\log(1/\delta)}{n}} \qquad (9)$$

Note that if the functions h_j are uniformly bounded, say $|h_j(\mathbf{x})| \leq c$, then $\Delta_{\mathcal{H}} \leq c$.

It is instructive to compare the results of Theorem 12 to those obtained by McAllester (2003) using the Gibbs algorithm. The latter algorithm selects a hypothesis h at random from the posterior distribution $\mathbf{q}$ and forms a prediction based on h. McAllester (2003) establishes the following bound on the expected performance of the randomized predictor. With probability at least $1-\delta$ for all $\mathbf{q} \in \Pi$

$$\mathbf{E}_{h\sim\mathbf{q}} L(h) \leq \mathbf{E}_{h\sim\mathbf{q}} \hat{L}(h) + \sqrt{\frac{D(\mathbf{q}\|\nu) + \ln(1/\delta) + \ln n + 2}{2n-1}}. \qquad (10)$$

When the hypotheses and losses are bounded in value (as assumed in McAllester, 2003), we see that, up to small numerical constants, the leading terms in the complexity penalties in (9) and (10) are very similar. While the bound in (10) contains an extra logarithmic term in n, the bound in (9) contains an extra term of order $\log\log D(\mathbf{q}\|\nu)$. Note, however, that the term $\mathbf{E}_{h\sim\mathbf{q}}\hat{L}(h)$ can be significantly larger than the term $\hat{L}(f_q)$, since the mixture hypothesis $f_q = \mathbf{E}_{h\sim\mathbf{q}}h$ can be far more complex than a single hypothesis h. A more detailed numerical comparison of the two bounds is left for future work. We comment that a similar bound to (10), based on the the margin loss, was established by Langford et al. (2001) for a mixture of classifiers.

Finally, as mentioned following Theorem 10, these data-dependent bounds can be used in order to select an optimal posterior distribution $\mathbf{q}$. While $D(\mathbf{q}\|\nu)$ is convex in $\mathbf{q}$, this is not the case for $\sqrt{D(\mathbf{q}\|\nu)}$. However, one may formulate the optimization problem as a constrained optimization problem of the form

$$\begin{array}{ll} \min\limits_{\mathbf{q}\in\Pi} & D(\mathbf{q}\|\nu) \\ \text{s.t.} & \hat{L}(f_q) \le a, \end{array}$$

for some parameter a which can be optimized in order to obtain the best bound. If $\hat{L}(f_q)$ is a convex function of $\mathbf{q}$ (for example, if a quadratic loss is used), we obtain a convex programming problem which can be solved using standard approaches (e.g., Boyd and Vandenberghe, 2002). We note that this approach is very similar to the so-called *maximum entropy discrimination* proposed by Jaakkola et al. (1999). Finally, if $\ell(y, f_q(\mathbf{x}))$ is convex in $\mathbf{q}$, we may use Jensen's inequality to upper bound $\hat{L}(f_q) = \hat{L}(\langle\mathbf{q},\mathbf{h}\rangle)$ by $\sum_j q_j L(h_j)$. In the latter case, McAllester (2003) has shown that an exact solution in the form of a Gibbs distribution can be obtained. This solution may in principle be used as a starting point for numerical optimization algorithms for solving the current problem.

6.2 Norm-Based Constraints

In Section 6.1 we used an entropic term to constrain the distributions $\mathbf{q}$ relative to some prior distribution $\mathbf{p}$. In many cases we do not have prior information provided in terms of a prior $\mathbf{p}$. Instead, we may believe that sparser solutions are more appropriate, which in principle would require us to use a constraint of the form $\|\mathbf{q}\|_p$ with p close to zero. While our results below do not hold for the case $p = 0$, they indicate in principle how to take into account other types of norms. Moreover, it is not hard to use our approach to derive bounds for support vector machines, in which case we can replace the L_1 constraint $\sum_j q_j = 1$ by the L_2 constraint.

We begin with the simple case where $g(\mathbf{q}) = (1/2)\|\mathbf{q}\|_2^2$, namely the L_2 norm is used. In this case, we simplify the notation by using $\|\mathbf{q}\| \doteq \|\mathbf{q}\|_2$. It is then easy to see that $g^*(\mathbf{z}) = (1/2)\|\mathbf{z}\|^2$. A simple calculation yields

$$\mathbf{E}_\sigma g^*\left((\lambda/n)\sum_{i=1}^n \sigma_i \mathbf{h}(X_i)\right) = \frac{\lambda^2}{2n^2}\sum_{i=1}^n \|\mathbf{h}(X_i)\|^2.$$

Substituting this result in (4), and minimizing over λ, we find that

$$\hat{R}_n(\mathcal{F}_A) \le \sqrt{\frac{2A}{n}\left(\frac{1}{n}\sum_{i=1}^n \|\mathbf{h}(X_i)\|^2\right)}.$$

Using Theorem 10, and Jensen's inequality $\mathbf{E}\sqrt{X} \le \sqrt{\mathbf{E}[X]}$, $X \ge 0$, we obtain the following bound.

Theorem 13 *Let the conditions of Theorem 10 hold, and set*

$$\tilde{g}(\mathbf{q}) = s\max((1/2)\|\mathbf{q}\|^2, g_0), \qquad \Delta_{\mathcal{H}} = \sqrt{\frac{1}{n}\mathbf{E}_S\sum_{i=1}^n \|\mathbf{h}(X_i)\|^2}$$

Then for all f_q, $\mathbf{q} \in \Pi$, *with probability at least* $1-\delta$,

$$L(f_q) \leq \hat{L}(f_q) + 2\Delta_{\mathcal{H}}\kappa(\tilde{g}(\mathbf{q}))\sqrt{\frac{2\tilde{g}(\mathbf{q})}{n}} + M(\tilde{g}(\mathbf{q}))\sqrt{\frac{4\log\log_s(s\tilde{g}(\mathbf{q})/g_0) + 2\log(1/\delta)}{n}}.$$

Consider next the case of general p and q such that $1/q + 1/p = 1$, $p \in (1,\infty)$. Let $p' = \max(p,2)$ and $q' = \min(q,2)$, and consider p-norm regularization $g(\mathbf{q}) = \frac{1}{p'}\|\mathbf{q}\|_p^{p'}$ and its associated conjugate function $g^*(\mathbf{z})$, namely

$$g(\mathbf{q}) = \frac{1}{p'}\|\mathbf{q}\|_p^{p'} \qquad ; \qquad g^*(\mathbf{z}) = \frac{1}{q'}\|\mathbf{z}\|_q^{q'} .$$

Note that if $p \leq 2$ then $q \geq 2$ and $q' = p' = 2$, while if $p > 2$ then $q < 2$, $q' = q, p' = p$.

In the present case, the average over σ required in (4) is rather cumbersome, and we resort to using (6) instead. The Rademacher averaging result for p-norm regularization is known in the Geometric theory of Banach spaces (type structure of the Banach space), for example, see Ledoux and Talgrand (1991), and follows from Khinchine's inequality. It can also be derived from the general techniques developed in this work, where we use the following bound on u_{g^*} in (6).

Lemma 14 *The following bound is valid,*

$$u_{g^*}(\mathbf{h}(\mathbf{x})) \leq \frac{\max(1,q-1)}{q'}\|\mathbf{h}(\mathbf{x})\|_q^{q'}.$$

Proof When $q \geq 2$ (implying $q' = 2$), we have that $g^*(\mathbf{z}+\theta\mathbf{z}') = (1/2)\left(\sum_j |z_j + \theta z'_j|^q\right)^{2/q}$. A direct computation of the second order derivatives required in Lemma 2, and use of the condition $q \geq 2$, yields

$$\begin{aligned}
\frac{d^2}{d\theta^2} g^*(\mathbf{z}+\theta\mathbf{z}') &\leq (q-1)\|\mathbf{z}+\theta\mathbf{z}'\|_q^{2-q}\sum_j |z_j+\theta z'_j|^{q-2} {z'_j}^2 \\
&\leq (q-1)\|\mathbf{z}+\theta\mathbf{z}'\|_q^{2-q}\|\mathbf{z}+\theta\mathbf{z}'\|_q^{q-2}\|\mathbf{z}'\|_q^2 \\
&= (q-1)\|\mathbf{z}'\|_q^2,
\end{aligned}$$

where the second inequality follows from Hölder's inequality with the dual pair $(q/(q-2), q/2)$.

When $q < 2$ (implying $q' = q$), we have $g^*(\mathbf{z}+\theta\mathbf{z}') = (1/q)\sum_j |z_j + \theta z'_j|^q$ and use the first part of Lemma 2.

$$\begin{aligned}
|\theta|^{1-q}\left|\frac{d}{d\theta}\left\{\frac{g^*(\mathbf{z}+\theta\mathbf{z}') - g^*(\mathbf{z}-\theta\mathbf{z}')}{2q}\right\}\right| &\leq |\theta|^{1-q}\sum_j \left|\frac{|z_j+\theta z'_j|^{q-1} - |z_j - \theta z'_j|^{q-1}}{2q} z'_j\right| \\
&\leq |\theta|^{1-q}\sum_j \left|\frac{|2\theta z'_j|^{q-1} z'_j}{2q}\right| \\
&= \frac{2^{q-2}}{q}\|\mathbf{z}'\|_q^q.
\end{aligned}$$

where the inequality $||a|^{q-1} - |b|^{q-1}| \leq |a-b|^{q-1}$ was used in the second inequality. Use of Lemma 2 and the observation that $\max(1,q-1)/q' = (q-1)/2$ if $q \geq 2$, and $\max(1,q-1)/q' = 1/2$ if $q < 2$ establishes the claim. □

From (4) we obtain a bound on the Rademacher complexity of $\mathcal{F}_A$.

$$\begin{aligned}\hat{R}_n(\mathcal{F}_A) &\leq \inf_{\lambda \geq 0}\left\{\frac{A}{\lambda}+\frac{\max(1,q-1)}{\lambda q'}\left(\frac{\lambda}{n}\right)^{q'}\sum_{i=1}^{n}\|\mathbf{h}(X_i)\|_q^{q'}\right\} \\ &= \frac{C_q}{n}A^{1-1/q'}\left(\sum_{i=1}^{n}\|\mathbf{h}(X_i)\|_q^{q'}\right)^{1/q'}, \qquad (11)\end{aligned}$$

where $C_q=(1-1/q')^{1/q'-1}\max(1,q-1)^{1/q'}$. Combining (11) with Theorem 10, and using Jensen's inequality $\mathbf{E}[X^{1/q'}] \leq (\mathbf{E}[X])^{1/q'}$, we obtain the following result.

Theorem 15 *Let the conditions of Theorem 10 hold, and set*

$$\tilde{g}(\mathbf{q}) = s\max\left((1/p')\|\mathbf{q}\|_p^{p'}, g_0\right) \quad ; \quad \Delta_{\mathcal{H},q}=\left[(1/n)\mathbf{E}_S\sum_{i=1}^{n}\|\mathbf{h}(X_i)\|_q^{q'}\right]^{1/q'} .$$

Then for all f_q, $\mathbf{q}\in\Pi$, with probability at least $1-\delta$,

$$L(f_q) \leq \hat{L}(f_q) + \frac{2C_q\Delta_{\mathcal{H},q}\kappa(\tilde{g}(\mathbf{q}))\left(\tilde{g}(\mathbf{q})\right)^{1/p'}}{n^{1/p'}} + M(\tilde{g}(\mathbf{q}))\sqrt{\frac{4\log\log_s(s\tilde{g}(\mathbf{q})/g_0)+2\log(1/\delta)}{n}} .$$

where $C_q=(1-1/q')^{1/q'-1}\max(q-1,1)^{1/q'}$.

6.3 Oracle Inequalities

Up to this point we have obtained data-dependent bounds which can be used for the purpose of model selection. In general, one is interested in knowing how the empirical estimator compares to the best possible mixture estimator, which can only be known if the underlying probability distribution is known. Such bounds are referred to as *oracle inequalities*. Let $\hat{\mathbf{q}}$ be an empirically derived posterior distribution. In particular, we establish an oracle inequality which relates the loss $L(\langle\hat{\mathbf{q}},\mathbf{h}\rangle)$ to the minimal loss $\inf_{\mathbf{q}\in\Pi}L(\langle\mathbf{q},\mathbf{h}\rangle)$.

We recall from Theorem 10 that with probability at least $1-\delta$ for all f_q, $\mathbf{q}\in\Pi$,

$$L(f_q) \leq \hat{L}(f_q) + \Delta_n(\mathcal{H},\mathbf{q},\delta), \qquad (12)$$

where

$$\Delta_n(\mathcal{H},\mathbf{q},\delta) = 2\kappa(\tilde{g}(\mathbf{q}))\Upsilon(\tilde{g}(\mathbf{q})) + M(\tilde{g}(\mathbf{q}))\sqrt{\frac{4\log\log_s(s\tilde{g}(\mathbf{q})/g_0)+2\log(1/\delta)}{n}}.$$

As in structural risk minimization (Vapnik, 1998), we select $\hat{\mathbf{q}}$ based on a complexity regularization criterion

$$\hat{\mathbf{q}} = \operatorname*{argmin}_{\mathbf{q}\in\Pi}\left\{\hat{L}(f_q)+\Delta_n(\mathcal{H},\mathbf{q},\delta)\right\}.$$

From (12), with probability at least $1-\delta/2$

$$L(f_{\hat{q}}) \leq \hat{L}(f_{\hat{q}}) + \Delta_n(\mathcal{H},\hat{\mathbf{q}},\delta/2).$$

By the optimality of the selection of $\hat{\mathbf{q}}$

$$\hat{L}(f_{\hat{q}}) + \Delta_n(\mathcal{H}, \hat{\mathbf{q}}, \delta/2) \leq \hat{L}(f_{\bar{q}}) + \Delta_n(\mathcal{H}, \bar{\mathbf{q}}, \delta/2),$$

where $\bar{\mathbf{q}}$ is an arbitrary hypothesis that does not depend on the data. We may apply Theorem 3 to $-L(f_{\bar{q}})$ and obtain that with probability greater than $1 - \delta/2$

$$\hat{L}(f_{\bar{q}}) < L(f_{\bar{q}}) + M(g(\bar{\mathbf{q}}))\sqrt{\frac{2\log(2/\delta)}{n}} \leq L(f_{\bar{q}}) + \Delta_n(\mathcal{H}, \bar{\mathbf{q}}, \delta/2).$$

Note that in this case the function class $\mathcal{F}$ consists of the single element $f_{\bar{q}}$, so that the term leading to the Rademacher complexity vanishes. Therefore, with probability at least $1 - \delta$,

$$\hat{L}(f_{\hat{q}}) + \Delta_n(\mathcal{H}, \hat{\mathbf{q}}, \delta/2) \leq L(f_{\bar{q}}) + 2\Delta_n(\mathcal{H}, \bar{\mathbf{q}}, \delta/2).$$

Since $\bar{\mathbf{q}}$ is arbitrary, we obtain the following result.

Theorem 16 *Under the same conditions as in Theorem 10, with probability at least* $1 - \delta$

$$L(f_{\hat{q}}) \leq \inf_{\mathbf{q} \in \Pi} [L(f_q) + 2\Delta_n(\mathcal{H}, \mathbf{q}, \delta/2)].$$

Note that if $\Delta_n(\mathcal{H}, \mathbf{q}, \delta/2)$ can be uniformly bounded, say $2\Delta_n(\mathcal{H}, \mathbf{q}, \delta/2) \leq c_n(\delta)$ independently of $\mathbf{q}$, we find that with probability at least $1 - \delta$, $L(f_{\hat{q}}) \leq \inf_{\mathbf{q} \in \Pi} L(f_q) + c_n(\delta)$.

6.4 Binary Classification

So far we have mainly been concerned with regression. The case of binary classification can easily be incorporated into the present framework. Let $S = \{(X_i, Y_i)\}_{i=1}^n$ be a sample where $X_i \in \mathcal{X}$ and $Y_i \in \{-1, +1\}$. Consider a soft classifier $f(\mathbf{x})$ and define the $0-1$ loss as $\ell_{0-1}(y, f(\mathbf{x})) = I(yf(\mathbf{x}) \leq 0)$. Let $\phi(yf(\mathbf{x}))$ be a Lipschitz function with Lipschitz constant $\kappa(\mathcal{F})$, which dominates the $0-1$ loss, namely $\ell_{0-1}(y, f(\mathbf{x})) \leq \phi(yf(\mathbf{x}))$. It is then not hard to conclude that under the same conditions as those in Theorem 8 we find that for all $f \in \mathcal{F}$, with probability at least $1 - \delta$,

$$\mathbf{P}\{Yf(X) \leq 0\} \leq \hat{\mathbf{E}}_n \phi(Yf(X)) + 2\kappa(\mathcal{F}) R_n(\mathcal{F}) + M(\mathcal{F})\sqrt{\frac{2\log(1/\delta)}{n}}.$$

One can then proceed to develop data-dependent bounds for this problem along the lines of Theorem 10. Note that several possible choices for $\phi(f(\mathbf{x}), y)$ have been proposed in the literature. A proof of the Bayes consistency of algorithms based on these dominating functions can be found in work by Lugosi and Vayatis (2002), Mannor et al. (2002) and Zhang (2003). An extension to multi-category classification has recently been proposed by Desyatnikov and Meir (2003).

7. Conclusion

We have developed a general procedure for establishing data-dependent bounds for mixture based approaches to regression and classification. As discussed in Section 1, Bayesian mixture approaches possess several desirable attributes from a frequentist perspective. However, in opposition to many Bayesian approaches, our results hold independently of the correctness of the model assumptions.

The approach pursued can effectively use many forms of prior knowledge, which may be incorporated through the selection of appropriate constraint functions. Additionally, the results apply to general mixture based approaches such as Bagging and Boosting. At a technical level, we have replaced the boundedness assumptions, prevalent in the Learning Theory literature, with more general moment constraint conditions.

Several open issues remain for future research. First, it would be be interesting to combine the current approach with recent methods based on local Rademacher complexities (e.g., Bartlett et al., 2002a), which are sometimes able to attain faster convergence rates. Second, a particularly interesting question relates to using the data itself to learn an appropriate constraint function, or perhaps several constraint functions. Finally, it is clearly important to conduct careful numerical studies of the bounds. Related work by Seeger (2002) demonstrated the tightness of similar bounds in the context of Gaussian processes, and their relevance to real-world problems. Preliminary studies indicate similar behavior for our bounds, but a systematic numerical investigation still needs to be done.

In this paper we have been concerned solely with mixture based Bayesian solutions. As pointed out in Section 1, general optimal Bayesian solutions are not always of a mixture form. In this context, it would be particularly interesting to establish finite sample bounds for optimal Bayesian procedures, which, under appropriate conditions, would provide tight upper bounds on the performance of *any* learning algorithm, and not only those based on selecting hypotheses from some class of hypotheses.

Given the suggested connections established in this work between the frequentist and Bayesian approaches, we would like to conclude with the following quote from Lehmann and Casella (1998).

> *"The strengths of combining the Bayesian and frequentist approaches are evident. The Bayes approach provides a clear methodology for constructing estimators, while the frequentist approach provides the methodology for evaluation."*

Although we have restricted ourselves to Bayesian *mixture* algorithms, which are not necessarily optimal in general, we hope that this paper has made some steps towards strengthening this claim.

Acknowledgments The work of R.M. was partially supported by the Technion V.P.R. fund for the promotion of sponsored research. Support from the Ollendorff center of the department of Electrical Engineering at the Technion is also acknowledged.

Appendix A. Examples of Convex Functions and their Conjugates

We provide several examples of convex functions and their conjugates. Further examples can be found in Boyd and Vandenberghe (2002) and Zhang (2002b).

We use $g(\mathbf{u})$ to denote a convex function with variable $\mathbf{u}$, while $g^*(\mathbf{v})$ denotes its conjugate with dual variable $\mathbf{v}$. The ℓ_p norm of a vector $\mathbf{u}$ is given by $\|\mathbf{u}\|_p = \left(\sum_j |u_j|^p\right)^{1/p}$.

- Let K be a symmetric positive-definite matrix. Then

$$g(\mathbf{u}) = \frac{1}{2}\langle \mathbf{u}, K\mathbf{u}\rangle \quad ; \quad g^*(\mathbf{v}) = \frac{1}{2}\langle \mathbf{v}, K^{-1}\mathbf{v}\rangle.$$

- Let $p, p', q, q' \geq 1$ be real numbers obeying $1/p + 1/q = 1$ and $1/p' + 1/q' = 1$. Then

$$g(\mathbf{u}) = \frac{1}{p'}\|\mathbf{u}\|_p^{p'} \quad ; \quad g^*(\mathbf{v}) = \frac{1}{q'}\|\mathbf{v}\|_q^{q'}.$$

- Assume $u_j \geq 0$ and $\mu_j > 0$. Then

$$g(\mathbf{u}) = \sum_j u_j \log \frac{u_j}{e\mu_j} \quad ; \quad g^*(\mathbf{v}) = \sum_j \mu_j \exp(v_j).$$

Appendix B. Proof of Theorem 3

We first prove the following lemma.

Lemma 17 *Consider real-valued functions $c_i : \Theta \times \mathcal{X}_i \to \mathbb{R}$, $i = 1, 2$. Define $c(\mathbf{x}_1, \mathbf{x}_2) = \sup_{\theta \in \Theta}(c_1(\theta, \mathbf{x}_1) + c_2(\theta, \mathbf{x}_2))$. Let $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$ be two independent random variables. Then*

$$\log \mathbf{E}_{X_1} \exp(\mathbf{E}_{X_2} c(X_1, X_2)) \leq \mathbf{E}_{X_1, X_2} c(X_1, X_2) + \log \mathbf{E}_{X_1} \sup_{\theta \in \Theta} \cosh(2(c_1(\theta, X_1)).$$

Proof Let

$$c'(X_1) = \mathbf{E}_{X_2}[c(X_1, X_2) - \sup_{\theta \in \Theta} c_2(\theta, X_2)].$$

It is clear that

$$\inf_{\theta \in \Theta} c_1(\theta, X_1) \leq c'(X_1) \leq \sup_{\theta \in \Theta} c_1(\theta, X_1).$$

Therefore using Jensen's inequality and symmetrization, we obtain

$$\begin{aligned}
\mathbf{E}_{X_1} \exp\left\{c'(X_1) - \mathbf{E}_{X_1'} c'(X_1')\right\} &\overset{(a)}{\leq} \mathbf{E}_{X_1, X_1'} \exp\left\{c'(X_1) - c'(X_1')\right\} \\
&\overset{(b)}{\leq} \mathbf{E}_{X_1, X_1'} \frac{1}{2}[\exp(2c'(X_1)) + \exp(-2c'(X_1'))] \\
&\overset{(c)}{=} \mathbf{E}_{X_1} \cosh(2c'(X_1)) \\
&\leq \mathbf{E}_{X_1} \sup_{\theta \in \Theta} \cosh(2c_1(\theta, X_1)),
\end{aligned}$$

where (a) and (b) used Jensen's inequality and (c) applied a symmetrization argument. □

Let $Z^n = \{Z_1, \ldots, Z_n\}$, $Z_i \in \mathcal{Z}$, be independently drawn from a distribution P, and let $\mathcal{F}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. Set

$$\hat{A}_{\mathcal{F}}(Z^n) = \sup_{f \in \mathcal{F}} \left[n\mathbf{E}_Z f(Z) - \sum_{i=1}^n f(Z_i) \right].$$

Lemma 18 *For all positive λ*

$$\log \mathbf{E}_{Z^n} \exp\left\{\lambda \hat{A}_{\mathcal{F}}(Z^n)\right\} \leq \lambda \mathbf{E}_{Z^n} \hat{A}_{\mathcal{F}}(Z^n) + n \log \mathbf{E}_Z \sup_{f \in \mathcal{F}} \cosh(2\lambda(f(Z)).$$

Proof The lemma follows by recursively applying Lemma 17 for $k = n, n-1, \ldots, 1$, and identifying the function f with the parameter θ. For each value of k we set

$$X_1 = Z_k \quad ; \quad X_2 = \{Z_{k+1}, \ldots, Z_n\},$$

where we assume that $\{Z_1, \ldots, Z_{k-1}\}$ are fixed. Moreover, set

$$\begin{aligned} c_1(\theta, X_1) &= -\lambda f(Z_k) \\ c_2(\theta, X_2) &= n\lambda \mathbf{E}_Z f(Z) - \sum_{i \neq k} \lambda f(Z_i), \end{aligned}$$

and note that $c(X_1, X_2) = \lambda \hat{A}_{\mathcal{F}}(Z^n)$. We simplify the notation by using $Z_k^l = \{Z_k, \ldots, Z_l\}$ for any positive integers k and l, $l \geq k$. From Lemma 17 we have (for fixed Z_1^k),

$$\log \mathbf{E}_{Z_k} \exp\left\{\mathbf{E}_{Z_{k+1}^n} \lambda \hat{A}_{\mathcal{F}}(Z^n)\right\} \leq \mathbf{E}_{Z_k^n} \lambda \hat{A}_{\mathcal{F}}(Z^n) + \log \mathbf{E}_Z \sup_{f \in \mathcal{F}} \cosh\left(2\lambda f(Z)\right),$$

which, upon exponentiation, is rewritten as

$$\mathbf{E}_{Z_k} \exp\left\{\mathbf{E}_{Z_{k+1}^n} \lambda \hat{A}_{\mathcal{F}}(Z^n)\right\} \leq \exp\left\{\mathbf{E}_{Z_k^n} \lambda \hat{A}_{\mathcal{F}}(Z^n) + \log \mathbf{E}_Z \sup_{f \in \mathcal{F}} \cosh\left(2\lambda f(Z)\right)\right\}.$$

Taking expectations with respect to Z_1^{k-1} on both sides of the inequality, followed by applying the logarithm function, we find that

$$\log \mathbf{E}_{Z_1^k} e^{\mathbf{E}_{Z_{k+1}^n} \lambda \hat{A}_{\mathcal{F}}(Z^n)} \leq \log \mathbf{E}_{Z_1^{k-1}} e^{\mathbf{E}_{Z_k^n} \lambda \hat{A}_{\mathcal{F}}(Z^n)} + \log \mathbf{E}_Z \sup_{f \in \mathcal{F}} \cosh\left(2\lambda f(Z)\right). \tag{13}$$

Summing both sides of (13) over $k = n, n-1, \ldots, 1$ we obtain

$$\begin{aligned} &\log \mathbf{E}_{Z_1^n} e^{\lambda \hat{A}_{\mathcal{F}}(Z^n)} + \log \mathbf{E}_{Z_1^{n-1}} e^{\mathbf{E}_{Z_n} \lambda \hat{A}_{\mathcal{F}}(Z^n)} + \cdots + \log \mathbf{E}_{Z_1} e^{\mathbf{E}_{Z_2^n} \lambda \hat{A}_{\mathcal{F}}(Z^n)} \\ &\leq \log \mathbf{E}_{Z_1^{n-1}} e^{\mathbf{E}_{Z_n} \lambda \hat{A}_{\mathcal{F}}(Z^n)} + \log \mathbf{E}_{Z_1^{n-2}} e^{\mathbf{E}_{Z_{n-1}^n} \lambda \hat{A}_{\mathcal{F}}(Z^n)} + \cdots + \log e^{\mathbf{E}_{Z_1^n} \lambda \hat{A}_{\mathcal{F}}(Z^n)} \\ &+ n \log \mathbf{E}_Z \sup_{f \in \mathcal{F}} \cosh\left(2\lambda f(Z)\right). \end{aligned}$$

Upon subtracting identical terms from both sides of the inequality we find that

$$\log \mathbf{E}_{Z_1^n} e^{\lambda \hat{A}_{\mathcal{F}}(Z^n)} \leq \lambda \mathbf{E}_{Z_1^n} \hat{A}_{\mathcal{F}}(Z^n) + n \log \mathbf{E}_Z \sup_{f \in \mathcal{F}} \cosh\left(2\lambda f(Z)\right),$$

which establishes the claim. □

Let $X^n = \{X_1, \ldots, X_n\}$, and set

$$\delta = \mathbf{P}\left\{\sup_{f \in \mathcal{F}} \left[n \mathbf{E}_X f(X) - \sum_{i=1}^n f(X_i)\right] \geq \mathbf{E}_{X^n} \sup_{f \in \mathcal{F}} \left[n \mathbf{E}_X f(X) - \sum_{i=1}^n f(X_i)\right] + n\varepsilon\right\}.$$

From Chernoff's inequality, $\mathbf{P}\{X \geq x\} \leq \inf_\lambda \{\exp(-\lambda x) \mathbf{E} \exp(\lambda x) : \lambda \geq 0\}$, we have for all non-negative λ

$$\delta \leq e^{-\lambda \mathbf{E}_{X^n} \hat{A}_{\mathcal{F}}(X^n) - \lambda n \varepsilon} \mathbf{E}_{X^n} e^{\lambda \hat{A}_{\mathcal{F}}(X^n)}.$$

Taking logarithms of both sides of the inequality, we find that

$$\begin{aligned}\log\delta &\leq -\lambda\mathbf{E}_{X^n}\hat{A}_{\mathcal{F}}(X^n) - \lambda n\varepsilon + \log\mathbf{E}_{X^n} e^{\lambda\hat{A}_{\mathcal{F}}(X^n)} \\ &\overset{(a)}{\leq} -\lambda n\varepsilon + n\log\mathbf{E}_X \sup_{f\in\mathcal{F}}\cosh(2\lambda f(X)) \\ &\overset{(b)}{\leq} -\lambda n\varepsilon + \frac{n}{2}\lambda^2 M^2,\end{aligned}$$

where Lemma 18 was used in (a) and the assumption of Theorem 3 was used in (b).

Since $\lambda \geq 0$ is arbitrary, we conclude that

$$\log\delta \leq \inf_{\lambda\geq 0}\left[\frac{n}{2}\lambda^2 M^2 - \lambda n\varepsilon\right] = -\frac{n\varepsilon^2}{2M^2}.$$

We thus obtain with probability of at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\left\{\mathbf{E}f(X) - \hat{\mathbf{E}}f(X)\right\} \leq \mathbf{E}_{X^n}\sup_{f\in\mathcal{F}}\left\{\mathbf{E}f(X) - \hat{\mathbf{E}}_n f(X)\right\} + M\sqrt{\frac{2\log(1/\delta)}{n}}. \qquad \square$$

References

T.M. Apostol. *Mathematical Analysis*. Addison-Wesley, 1957.

P. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexity. In *Proceedings of the Annual Conference on Computational Learning Theory*, volume 2375 of *LNAI*, Sydney, February 2002a. Springer.

P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002b.

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 2, 2003. To appear.

O. Bousquet and A. Chapelle. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

S. Boyd and L. Vandenberghe. *Convex Optimization*. 2002.
Available at http://www.stanford.edu/∼boyd/cvxbook.html.

I. Desyatnikov and R. Meir. Data-dependent bounds for multi-category classification based on convex losses. Unpublished, 2003.

R. Herbrich and T. Graepel. A PAC-Bayesian margin bound for linear classifiers; why SVMs work. In *Advances in Neural Information Processing Systems 13*, pages 224–230, Cambridge, MA, 2001. MIT Press.

T.S. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In M. Kearns, editor, *Advances in Neural Information Processing Systems*, volume 11, 1999.

V. Koltchinksii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1), 2002.

J. Langford, M. Seeger, and N. Megiddo. An improved predictive accuracy bound for averaging classifiers. In *Proceeding of the Eighteenth International Conference on Machine Learning*, pages 290–297, 2001.

M. Ledoux and M. Talgrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Press, New York, 1991.

E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Verlag, New York, second edition, 1998.

M. Leshno, V. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6:861–867, 1993.

G. Lugosi and N. Vayatis. A consistent strategy for boosting algorithms. In *Proceedings of the Annual Conference on Computational Learning Theory*, volume 2375 of *LNAI*, pages 303–318, Sydney, February 2002. Springer.

S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the fifteenth Annual conference on Computational learning theory*, volume 2375 of *LNAI*, pages 319–333, Sydney, 2002. Springer.

S. Mannor, R. Meir, and T. Zhang. Greedy algorithms for classification—consistency, convergence rates, and adaptivity. Technical Report CCIT 420, Department of Electrical Engineering, Technion, 2003.

D. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

C. P. Robert. *The Bayesian Choice: A Decision Theoretic Motivation*. Springer Verlag, New York, second edition, 2001.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, N.J., 1970.

M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *JMLR*, 3:233–269, 2002.

J. Shawe-Taylor, P. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transaction on Information Theory*, 44:1926–1940, 1998.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Verlag, New York, 1996.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

V. N. Vapnik. *Statistical Learning Theory*. Wiley Interscience, New York, 1998.

Y. Yang. Minimax nonparametric classification - part I: rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

T. Zhang. Generalization performance of some learning problems in Hilbert functional space. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2001. MIT Press.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002a.

T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46:91–129, 2002b.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 2003. To appear.