PAC-Bayesian Model Averaging

David A. McAllester
AT&T Shannon Labs
180 Park Avenue
Florham Park, NJ 07932-0971
dmac@research.att.com

December 13, 1999

Abstract

PAC-Bayesian learning methods combine the informative priors of Bayesian methods with distribution-free PAC guarantees. Building on earlier methods for PAC-Bayesian model selection, this paper presents a method for PAC-Bayesian model averaging. The main result is a bound on generalization error of an arbitrary weighted mixture of concepts that depends on the empirical error of that mixture and the KL-divergence of the mixture from the prior. A simple characterization is also given for the error bound achieved by the optimal weighting.

1 INTRODUCTION

A PAC-Bayesian approach to machine learning attempts to combine the advantages of both PAC and Bayesian approaches [18, 13]. The Bayesian approach has the advantage of using arbitrary domain knowledge in the form of a Bayesian prior. The PAC approach has the advantage that one can prove guarantees for generalization error without assuming the truth of the prior. A PAC-Bayesian approach basis the bias of the learning algorithm on an arbitrary prior distribution, thus allowing the incorporation of domain knowledge, and yet provides a guarantee on generalization error that is independent of any truth of the prior.

PAC-Bayesian approaches are related to structural risk minimization (SRM) [10]. Here we interpret this broadly as describing any learning algorithm optimizing a tradeoff between the "complexity", "structure", or "prior probability" of the concept or model and the "goodness of fit", "description length", or "likelihood" of the training data. Under this interpretation of SRM, Bayesian algorithms that select a concept of maximum posterior probability (MAP algorithms) are viewed as a kind of SRM algorithm. Various approaches to SRM are compared both theoretically and experimentally by Kearns et al. in [10]. They give experimental evidence that Bayesian and MDL algorithms tend to over fit in experimental settings where the Bayesian assumptions fail. A PAC-Bayesian approach uses a prior distribution analogous to that used in MAP or MDL but provides a theoretical guarantee against over fitting independent of the truth of the prior. Earlier work on PAC-Bayesian algorithms has focused on model selection — selecting either a single concept or a uniformly weighted set of concepts. Here we consider nonuniform model averaging, i.e., selecting a weighted mixture of the concepts.

Model averaging is empirically important in certain applications. For example, in statistical language modeling for speech recognition one "smooths" a trigram model with a bigram model and smooths the bigram model with a unigram model. This smoothing is essential for minimizing the cross entropy between, say, the model and a test corpus of newspaper sentences. It turns out that smoothing in statistical language modeling is more naturally formulated as model averaging than as model selection. A smoothed language model is very large — it contains a full trigram model, a full bigram model and a full unigram model as parts. If one uses MDL to select the structure of a language model, selecting model parameters with maximum likelihood, the resulting structure is much smaller than that of a smoothed trigram model. Furthermore, the MDL model performs quite badly. A smoothed trigram model can be theoretically derived as a compact representation of a Bayesian mixture of an exponential number of (smaller) suffix tree models [16].

Model averaging can also be applied to decision trees that produce probabilities at their leaves rather than hard classifications. A common method of constructing decision trees is to first build an overly large tree which over fits the training data and then prune the tree in some way so as to get a smaller tree that does not over fit the data [17, 9]. For trees with probabilities at leaves, an alternative is to construct a weighted mixture of the subtrees of the original over fit tree. It is possible to construct a concise representation of a weighting over exponentially many different subtrees [3, 15, 8].

Weighted model mixtures are also widely used in constructing algorithms with on-line guarantees. In particular, the weighted majority algorithm and its variants can be proved to compete well with the best expert [11, 4, 6, 5]. The weighting used in on-line weighted majority algorithms is analogous to a Bayesian posterior distribution. In spite of the common use of posterior distributions, however, existing on-line algorithms do not provide analogous structural risk minimization guarantees — they are not guaranteed to perform well relative to an optimal SRM tradeoff. In order to select an appropriate value for a certain learning parameter, existing weighted majority algorithms require a-priori knowledge of the prior probability (or structural size) of the optimal concept. Learning algorithms based on the PAC bounds presented here require no such a-priori knowledge.

Section 2 gives the first main result of this paper, theorem 2. This theorem gives a bound on the generalization error of weighted mixtures. A weighted mixture which gives too much weight to models with low prior probability will over fit the data in the sense that its provable loss bound is large. A weighting that puts too much weight on models that fit the training data poorly but have high prior probability under fits the data in the sense that its provable loss bound is also large. A weighting that minimizes the provable loss bound is formally analogous to the classical Bayesian posterior distribution.

Section 3 gives the second main result of this paper which is stated informally in equation (7) and more precisely in theorem 14. It gives a simple characterization of the error bound for the optimal posterior distribution. The informal characterization can be justified using approximations expected to hold for large sample size. The precise characterization is proved in a certain large sample limit similar to the thermodynamic limits studied in the statistical physics model of machine learning [7]. Section 3 also gives two different posterior distributions which both achieve optimal behavior in the large sample limit.

2 The First Main Result

In this section we state the main result of this paper. Proofs are given in later sections. We assume a prior probability measure P on a concept class $\mathcal{C}$ and a sampling distribution D on a set of instances $\mathcal{X}$. We allow both $\mathcal{C}$ and $\mathcal{X}$ to be uncountably infinite. We also assume a measurable loss function l such that for any concept c and instance x we have $l(c,\ x) \in [0,\ 1]$. For example, we might have that concepts are predicates on instances and there is a target concept c_t such that $l(c,\ x)$ is 0 if $c(x) = c_t(x)$ and 1 otherwise. We define $l(c)$ to be the expectation over sampling an instance x of $l(c,x)$, i.e., $E_{x\sim D}l(c,\ x)$. We let S range over samples of m instances each drawn independently according to distribution D. We define $\hat{l}(c,\ S)$ to be $\frac{1}{m}\sum_{x\in S} l(c,\ x)$. If Q is a probability measure on concepts then $l(Q)$ denotes $E_{c\sim Q}l(c)$ and $\hat{l}(Q,\ S)$ denotes $E_{c\sim Q}\hat{l}(c_i,\ S)$. The notation $\forall^\delta S\ \Phi(S)$ signifies that the probability over the generation of the sample S of $\Phi(S)$ is at least $1-\delta$.

We first review PAC-Bayesian model selection. The departure point for the current paper is the following "naive bound" from [13].[1] This bound applies to countable concept classes where concept c_i has prior probability P_i.

$$\forall^\delta S\ \ \forall c_i\ \ l(c_i) \leq \hat{l}(c_i,\ S) + \sqrt{\frac{\ln\frac{1}{P_i} + \ln\frac{1}{\delta}}{2m}} \tag{1}$$

This theorem states that for most draws of a sample S we have that, for all concepts c_i, the generalization loss of c_i is bounded by the measured loss of c_i plus a penalty term that gets larger as the prior probability of c_i declines. The bound yields a model selection algorithm where one selects a concept c^* so as to minimize the upper bound.

The proof of (1) is straightforward. By the Chernoff bound we have that for any fixed concept c_i the probability that c_i violates the theorem is no larger than $P_i\delta$. By the union bound the probability that some c_i violates the theorem is then no larger than $\sum P_i\delta = \delta\sum P_i = \delta$.

Formula (1) can be related to Bayesian model selection or MDL. We assume that each concept has an associated distribution on instances and let $P(x|c)$ denote the probability of x under the distribution associated with c. We let $L(c,\ x)$ denote the classical log loss of c on x, i.e., the quantity $\ln\frac{1}{P(x|c)}$. For a sample S of size m we define $\hat{L}(c,\ S)$ to be $\frac{1}{m}\sum_{x\in S} L(c,\ x)$. The concept maximizing the Bayesian posterior is the same as the concept minimizing $\ln\frac{1}{P_iP(S|c_i)}$ which is the same as the concept minimizing $\hat{L}(c_i,\ S) + \frac{\ln\frac{1}{P_i}}{m}$. In the MDL interpretation $\frac{\ln\frac{1}{P_i}}{m}$ is proportional to the complexity of the concept c_i and $\hat{L}(c_i,\ S)$ is proportional to the complexity of the sample given c_i. Formula (1) defines a different trade-off between the "concept complexity" $\frac{\ln\frac{1}{P_i}}{m}$ and the "sample complexity" $\hat{l}(c_i,\ S)$.

The main result in [13] is a generalization of (1) to a uniform statement over all *subsets* of an arbitrary (uncountable) concept class. The first

[1] Results similar to (1) have appeared repeatedly in the literature, e.g., [1, 2, 12].

main result of this paper is a generalization of (1) to a uniform statement over *distributions* on an arbitrary concept class. The new bound involves the Kullback-Leibler divergence, denoted $D(Q||P)$, from distribution Q to distribution P. The quantity $D(Q||P)$ is defined to be $E_{c\sim Q} \ln \left.\frac{dQ}{dP}\right|_c$. The following is the first main result of this paper and is proved in section 4.

Theorem 2

$$\forall^\delta S \;\; \forall Q \;\; l(Q) \leq \hat{l}(Q,\, S) + \sqrt{\frac{D(Q||P) + \ln \frac{1}{\delta} + \ln m + 2}{2m-1}}$$

In the case of a countable concept class where Q is concentrated on the single concept c_i the quantity $D(Q||P)$ equals $\ln \frac{1}{P_i}$ and, for large m, theorem 2 is essentially the same as (1). The main result in [13] can also be viewed as the special case of theorem 2 where Q is simply a re-normalization of the prior on a given set of concepts. Theorem 2 corresponds to a model averaging learning algorithm where one selects the distribution minimizing the bound. The construction of a distribution optimizing the bound is discussed in section 3.

An earlier version of this paper [14] proved a version of theorem 2 as a corollary of a tighter bound which, for large m, has the form $\hat{l}(Q,\, S) + E_{c\sim Q}\sqrt{\ln(dQ/dP|_c)/2m}$. This bound is tighter than $\hat{l}(Q,\, S)+\sqrt{D(Q||P)/2m}$ by the slack in Jensen's inequality. However, this tighter bound is more difficult to prove and seems unlikely to improve equation (7) given in section 3.

We now consider the problem of constructing a posterior distribution minimizing the bound in theorem 2. More specifically, for any posterior distribution Q we define $B(Q)$ to be the bound in theorem 2, i.e., we have the following.

$$B(Q) \equiv \hat{l}(Q,\, S) + \sqrt{\frac{D(Q||P) + \ln \frac{1}{\delta} + \ln m + 2}{2m-1}} \tag{3}$$

By abuse of notation we will write $B(Q^*)$ to denote $\inf_Q B(Q)$ (in the case of infinite concept classes is not obvious that this infimum is achieved by any particular distribution). For any value of $\beta \geq 0$ we define Q^β to be the posterior distribution defined as follows where Z is a normalizing constant.

$$\left.\frac{dQ^\beta}{dP}\right|_c = \frac{1}{Z} e^{-\beta \hat{l}(c,\, S)}$$

A straightforward application of the Kuhn-Tucker conditions (Lagrange multipliers) yields the following theorem concerning the optimal posterior distribution for a finite concept class.

Theorem 4 *For a finite concept class there exists $\beta \geq 0$ satisfying the following conditions.*

$$\begin{aligned} B(Q^*) &= B(Q^\beta) && (5) \\ \beta &= 2\sqrt{(2m-1)(D(Q^\beta||P) + \ln(1/\delta) + \ln m + 2)} && (6) \end{aligned}$$

Unfortunately, there can be multiple local minima in $B(Q^\beta)$ as a function of β and even multiple local minima satisfying (6). A simpler characterization of an optimal posterior is the second main result of this paper and is discussed in section 3.

3 The Second Main Result

The second main result of this paper characterizes the bound achieved by the optimal distribution Q under certain natural assumptions. More specifically, we consider the case of an uncountably infinite concept class (possibly of infinite VC dimension) and, for a given sample S of size m, we assume there exists a continuous probability density function p such that for $\hat{l} \in [0, 1]$ we have the following.

$$p(\hat{l}) = \lim_{\Delta\hat{l} \to 0} \frac{P(\{c: \ \hat{l}(c, \ S) \in [\hat{l}, \ \hat{l} + \Delta\hat{l}]\})}{\Delta\hat{l}}$$

The second main result of the paper can be summarized as the following approximate equation where $B(Q)$ and $B(Q^*)$ are defined as in section 2.

$$B(Q^*) \approx \min_{\hat{l}} \ \hat{l} + \sqrt{\frac{\ln(1/p(\hat{l}))}{2m}} \tag{7}$$

Note that for m large relative to $\ln(1/\delta)$ this can be viewed as a continuous version of naive bound (1).

Equation (7) is derived below in two ways. First a derivation is given using informal approximations which should hold for large m. Second, the equation is proved rigorously in a certain large m thermodynamic limit similar to limits used in the statistical physics model of machine learning [7]. Before doing either of these, however, we rigorously prove the following inequality which holds without any large m assumptions.

Theorem 8

$$B(Q^*) \geq \min_{\hat{l}} \;\; \hat{l} + \sqrt{\frac{\ln(1/p(\hat{l}))}{2m}}$$

Proof: Let Q be an arbitrary distribution on concepts. We show that $B(Q)$ can be no smaller than the above expression. Let W be the subset of the values $\hat{l} \in [0,1]$ such that $p(\hat{l}) > 0$. Let α denote the size of W as measured by the uniform measure on $[0,\ 1]$. Note that $\alpha \leq 1$. We now define a concept distribution U as follows.

$$\left.\frac{dU}{dP}\right|_c = \begin{cases} \frac{1}{\alpha p(\hat{l}(c,\ S))} & \text{if } p(\hat{l}(c,\ S)) > 0 \\ 0 & \text{otherwise} \end{cases}$$

The total measure of U can be written as follows.

$$\int_{\hat{l}\in W} (\left.\frac{dU}{dP}\right|_{\hat{l}})(\left.\frac{dP}{d\hat{l}}\right|_{\hat{l}}) d\hat{l} = \int_{\hat{l}\in W} \frac{1}{\alpha p(\hat{l})} p(\hat{l}) d\hat{l} = 1$$

Hence U is a probability measure on concepts. We now have the following.

$$\begin{aligned} D(Q||P) &= E_{c\sim Q} \ln \frac{dQ}{dP} \\ &= E_{c\sim Q}\left[\ln\frac{dU}{dP} + \ln\frac{dQ}{dU}\right] \\ &= E_{c\sim Q} \ln(1/p(\hat{l}(c,\ S))) + \ln(1/\alpha) + D(Q||U) \\ &\geq E_{c\sim Q} \ln(1/p(\hat{l}(c,\ S))) \end{aligned}$$

This implies the following where the third line follows from Jensen's inequality.

$$\begin{aligned} B(Q) &\geq \hat{l}(Q,\ S) + \sqrt{\frac{D(Q||P)}{2m}} \\ &\geq \left[E_{c\sim Q}\;\; \hat{l}(c,\ S)\right] + \sqrt{\frac{E_{c\sim Q} \ln(1/p(\hat{l}(c,\ S)))}{2m}} \\ &\geq E_{c\sim Q}\left[\hat{l}(c,\ S) + \sqrt{\frac{\ln(1/p(\hat{l}(c,\ S)))}{2m}}\right] \\ &\geq \min_{\hat{l}} \;\; \hat{l} + \sqrt{\frac{\ln(1/p(\hat{l}))}{2m}} \end{aligned}$$

$\square$

Let $\hat{l}^*$ be a value of $\hat{l}$ minimizing $\hat{l} + \sqrt{\ln(1/p(\hat{l}))/(2m)}$ (since $[0, 1]$ is a compact interval, such a value must exist). Next we argue that for large m there exists a distribution Q with $B(Q)$ close to $\hat{l}^* + \sqrt{\ln(1/p(\hat{l}^*))/(2m)}$. More specifically, for $\hat{l} \in [0, 1]$ and $\Delta > 0$ we define $C(\hat{l},\ \Delta)$ to be the set of concepts c such that $\hat{l}(c,\ S) \in [\hat{l},\ \hat{l} + \Delta]$. We then define the distribution $Q(\hat{l},\ \Delta)$ to be the re-normalization of the prior restricted to the set $C(\hat{l},\ \Delta)$. More formally we define $Q(\hat{l},\ \Delta)$ by the following.

$$\left.\frac{dQ}{dP}\right|_c = \begin{cases} \frac{1}{P(C(\hat{l},\ \Delta))} & \text{if } c \in C(\hat{l},\ \Delta) \\ & \\ 0 & \text{otherwise} \end{cases}$$

This definition gives $D(Q(\hat{l}, \Delta)||P) = \ln(1/P(C(\hat{l},\ \Delta)))$. By the mean value theorem for integrals there must exist $\hat{l}^+ \in [\hat{l},\ \hat{l}+\Delta]$ such that $P(C(\hat{l},\ \Delta)) = p(\hat{l}^+)\Delta$. This gives $\ln(1/P(Q(\hat{l},\ \Delta))) = \ln(1/p(\hat{l}^+)) + \ln(1/\Delta)$. By another application of the mean value theorem we have the following for some $\hat{l}^{++} \in [\hat{l},\ \hat{l} + \Delta]$.

$$B(Q(\hat{l},\ \Delta)) = \hat{l}^{++} + \sqrt{\frac{\ln(1/p(\hat{l}^+)) + \ln(1/\Delta) + \ln(1/\delta) + \ln m + 2}{2m - 1}}$$

Now inserting $\hat{l}^*$ for $\hat{l}$ and $1/m$ for Δ yields the following.

$$B(Q(\hat{l}^*,\ 1/m)) = \hat{l}^{++} + \sqrt{\frac{\ln(1/p(\hat{l}^+)) + \ln(1/\delta) + 2\ln m + 2}{2m - 1}} \tag{9}$$

We now assume that m is large compared to $\ln(1/\delta) + 2\ln m + 2$ and that p is sufficiently smooth so that the following assumption holds.

Assumption 10 *For* $|\Delta| \leq 1/m$ *we have* $\ln(1/p(\hat{l}+\Delta))/2m \approx \ln(1/p(\hat{l}))/2m$.

Under these assumptions we have the following.

$$B(Q(\hat{l}^*,\ 1/m)) \approx \hat{l}^* + \sqrt{\frac{\ln(1/p(\hat{l}^*))}{2m}} \tag{11}$$

Assumption 10 rules out the possibility that $\ln(1/p(\hat{l}^*))$ can be significantly less than 0 and ensures that the above expression can never be significantly smaller than $\hat{l}^*$.

Of course the derivation of (11) from (9) lacks rigor. It is based on assumption 10 which is itself not a rigorous statement. We now construct a rigorous version of assumption 10 and derive a rigorous version of (11) from (9).

Assumption 10 is formalized by taking a certain "thermodynamic limit". The basic idea is to work with scale-invariant quantities — quantities that remain order one even for very large values of m. In particular we have the following scale-invariant function of $\hat{l} \in [0,\ 1]$.

$$\gamma(\hat{l}) = \frac{\ln(1/\min(1,\ p(\hat{l})))}{2m} \tag{12}$$

Consider a sequence of learning problems of ever increasing sample size. In language modeling, for example, as the sample size increases the underlying distribution changes as new words are continually introduced into the language. The size of the optimal language model increases with the sample size and the function γ can remain roughly constant — the size in bits of the optimal model remains some fixed fraction of the sample size. A similar phenomenon can arise in decision trees where the size of the optimal tree can remain a fixed fraction of the size of the training data. So it seems reasonable to work with a model where γ is held fixed as m increases.

To build a formal model where γ is held fixed as m increases we "invert" equation 12 and define p as a function of γ and m. To invert 12 first note that we have that $\gamma(\hat{l}) \geq 0$ for all $\hat{l}$ and $\gamma(\hat{l}) = 0$ for some non-zero fraction of values of $\hat{l}$ (if $\gamma(\hat{l}) > 0$ for all but a set of measure zero then then $p(\hat{l}) < 1$ for all but a set of measure zero and this is impossible). We now define p in terms of γ and m as follows where α is selected so as to normalize the distribution.

$$p(\hat{l}) \equiv \begin{cases} e^{-2m\gamma(\hat{l})} & \text{if } \gamma(\hat{l}) > 0 \\ \alpha & \text{if } \gamma(\hat{l}) = 0 \end{cases} \tag{13}$$

To show that (12) holds for this definition of p we need to show that for $\gamma(\hat{l}) = 0$ we have $p(\hat{l}) \geq 1$. This will be true provided we have $\alpha \geq 1$. But note that if $\alpha < 1$ then $p(\hat{l}) < 1$ for all $\hat{l}$ which violates the requirement that α is selected so as to normalize p. So we have that definition (13) satisfies equation (12).

We can also show that for all values of $\hat{l}$ the following limit equation holds.

$$\gamma(\hat{l}) = \lim_{m\to\infty} \frac{\ln(1/p(\hat{l}))}{2m}$$

This equation is trivial for $\gamma(\hat{l}) > 0$. To show the equation for $\gamma(\hat{l}) = 0$ let β be the fraction of values of $\hat{l}$ for which $\gamma(\hat{l}) = 0$. As m increases we have that the constant α in (13) approaches $1/\beta$. But this implies that $\ln(1/p(\hat{l}))/2m$ approaches 0.

It is also useful to note that when γ is continuous we have the following.

$$\lim_{m\to\infty} \frac{\ln(1/p(\hat{l}+1/m))}{2m} = \lim_{m\to\infty} \gamma(\hat{l}+1/m) = \gamma(\hat{l}) = \lim_{m\to\infty} \frac{\ln(1/p(\hat{l}))}{2m}$$

This is a rigorous version of assumption 10. Taking the limit as $m \to \infty$ of (9) yields the following rigorous version of (11).

Theorem 14 *For any continuous function γ on $[0,1]$ with $\gamma(\hat{l}) \geq 0$ for all $\hat{l}$, and $\gamma(\hat{l}) = 0$ for a nonzero fraction of values of $\hat{l}$, for $\hat{l}^*$ a value minimizing $\hat{l} + \sqrt{\gamma(\hat{l})}$, and for p as defined by (13), we have the following.*

$$\lim_{m\to\infty} B(Q(\hat{l}^*,\ 1/m)) = \hat{l}^* + \sqrt{\gamma(\hat{l}^*)} = \lim_{m\to\infty} \left[\hat{l}^* + \sqrt{\frac{\ln(1/p(\hat{l}^*))}{2m}}\right]$$

Theorems 8 and 14 now imply the following rigorous version of (7).

Theorem 15 *Under the same conditions as theorem 14 we have the following.*

$$\lim_{m\to\infty} B(Q^*) = \hat{l}^* + \sqrt{\gamma(\hat{l}^*)}$$

Theorem 14 shows that, in the limit defined by (13), $Q(\hat{l}^*,\ 1/m)$ is an optimal posterior distribution. However, this distribution seems difficult to work with in practice. It turns out that other posterior distributions are also optimal in the limit defined by (13). The final result of this section shows that, as suggested by theorem 4, for $\gamma(\hat{l}^*) > 0$ there exists an optimal posterior of the form Q^β. The proof of this theorem is given in an appendix.

Theorem 16 *Under the same conditions as theorem 14 we have that if $\gamma(\hat{l}^*) > 0$ then the following holds where $\beta^* = 2\sqrt{2m\ln(1/p(\hat{l}^*))}$.*

$$\lim_{m\to\infty} B(Q^{\beta^*}) = \hat{l}^* + \sqrt{\gamma(\hat{l}^*)}$$

4 Proof of Theorem 2

The departure point for the proof of theorem 2 is the following where S is a sample of size m and $\Delta(c)$ abbreviates $|l(c) - \hat{l}(c,\ S)|$.

Lemma 17

$$\forall^\delta S \quad E_{c\sim P}\ e^{(2m-1)\Delta(c)^2} \quad \leq \frac{4m}{\delta}$$

Proof: It suffices to prove the following.

$$E_S\ \ E_{c\sim P}\ \ e^{(2m-1)\Delta(c)^2} \leq 4m \tag{18}$$

Lemma 17 follows from (18) by an application of Markov's inequality. To prove (18) it suffices to prove the following for any individual given concept.

$$E_S\ \ e^{(2m-1)\Delta^2} \leq 4m \tag{19}$$

By the Chernoff bound we have $P(\Delta \geq x) \leq 2e^{-2mx^2}$. We now consider the density function $f(\Delta)$ maximizing $\int_0^\infty e^{(2m-1)\Delta^2} f(\Delta) d\Delta$ subject to the constraint that $\int_x^\infty f(\Delta) d\Delta \leq 2e^{-2mx^2}$. The maximum occurs when we have $\int_x^\infty f(\Delta) d\Delta = 2e^{-2mx^2}$ which is realized when $f(\Delta) = 8m\Delta e^{-2m\Delta^2}$. So we have the following

$$\begin{aligned}
E_S\ e^{(2m-1)\Delta^2} &\leq \int_0^\infty e^{(2m-1)\Delta^2} f(\Delta) d\Delta \\
&= \int_0^\infty 8m\Delta e^{(2m-1)\Delta^2} e^{-2m\Delta^2} d\Delta \\
&= \int_0^\infty 8m\Delta e^{-\Delta^2} d\Delta \\
&= 4m
\end{aligned}$$

$\square$

To prove theorem 2 we consider selecting a sample S. Lemma 17 implies that with probability at least $1 - \delta$ over the selection of a sample S we have the following.

$$E_{c\sim P}\ \ e^{(2m-1)\Delta(c)^2} \leq \frac{4m}{\delta} \tag{20}$$

To prove theorem 2 it now suffices to show that the constraint (20) on the function $\Delta(c)$ implies the body of theorem 2. We are interested in computing an upper bound on the quantity $l(Q) - \hat{l}(Q,\ S)$. Note that $l(Q) - \hat{l}(Q,\ S) \leq E_{c\sim Q}|l(c_i) - \hat{l}(c_i,\ S)| = E_{c\sim Q}\Delta(c)$. We now prove the following lemma.

Lemma 21 *For $\beta > 0$, $K > 0$, and $Q, P, \Delta \in R^n$ satisfying $P_i \geq 0$, $Q_i \geq 0$, $\Delta_i \geq 0$, and $\sum_{i=1}^n Q^i = 1$, we have that if*

$$\sum_{i=1}^n P_i e^{\beta \Delta_i^2} \leq K$$

then

$$\sum_{i=1}^n Q_i \Delta_i \leq \sqrt{\frac{D(Q||P) + \ln K}{\beta}}.$$

Before proving lemma 21 we note that lemmas 21 and 17 together imply theorem 2. To see this consider a sample satisfying 20 and an arbitrary posterior probability measure Q on concepts. It is possible to define three infinite sequences of vectors Q^1, Q^2, Q^3, ..., P^1, P^2, P^3 ..., and Δ^1, Δ^2, Δ^3, ... such that Q^n, P^n, and Δ^n satisfy the conditions of lemma 21 with $K = 4m/\delta$ and $\beta = 2m - 1$ and satisfying the following.

$$E_{c \sim Q} \Delta(c) = \lim_{n \to \infty} \sum_{i=1}^n Q_i^n \Delta_i^n$$

$$D(Q||P) = \lim_{n \to \infty} \sum_{i=1}^n Q_i^n \ln \frac{Q_i^n}{P_i^n}$$

By taking the limit of the conclusion of lemma 21 we then get $E_{c \sim Q} \Delta(c) \leq \sqrt{(D(Q||P) + \ln(1/\delta) + \ln m + 2)/(2m-1)}$.

To prove lemma 21 it suffices to consider only those values of i for which $Q_i > 0$ — dropping the indices where $Q_i = 0$ does not change the value of $\sum_{i=1}^n Q_i \Delta_i$ while enlarging the feasible set by weakening the constraint (20). Furthermore, if $P_i = 0$ at some point where $Q_i > 0$ then $D(Q||P) = \infty$ and the theorem is immediate. So we can assume without loss of generality that $Q_i > 0$ and $P_i > 0$ for all i.

By Jensen's inequality we have $(\sum_{i=1}^n Q_i \Delta_i)^2 \leq \sum_{i=1}^n Q_i \Delta_i^2$. So it now suffices to prove that $\sum_{i=1}^n Q_i \Delta_i^2 \leq (D(Q||P) + \ln K)/\beta$. This is a consequence of the following lemma.[2]

[2]The original version of this paper [14] proved a bound of approximately the form $\hat{l}(Q) + \sum_{i=1}^n Q_i \sqrt{\ln(Q_i/P_i)/(2m)}$ by maximizing $\sum_{i=1}^n Q_i \Delta_i$ subject to constraint 20. A version of theorem 2, which is of the form $\hat{l}(Q,\ S) + \sqrt{(\sum_{i=1}^n Q_i \ln(Q_i/P_i))/(2m)}$, was then proved from this bound by an application of Jensen's inequality. The idea of maximizing $\sum_{i=1}^n Q_i \Delta_i^2$ and achieving theorem 2 directly is due to Robert Schapire.

Lemma 22 *For $\beta > 0$, $K > 0$, and $Q, P, y \in R^n$ satisfying $P_i > 0$, $Q_i > 0$, and $\sum_{i=1}^n Q_i = 1$, if*

$$\sum_{i=1}^n P_i e^{\beta y_i} \leq K \tag{23}$$

then

$$\sum_{i=1}^n Q_i y_i \leq \frac{D(Q||P) + \ln K}{\beta}$$

To prove lemma 22 we take P and Q as given and use the Kuhn-Tucker conditions to find a vector y maximizing $\sum_{i=1}^n Q_i y_i$ subject to the constraint (23).

Lemma 24 (Kuhn-Tucker) *If C and $f_1, \ldots, f_n$ are functions from R^n to R, y is a maximum of $C(y)$ over the set satisfying $f_1(y) \leq 0, \ldots, f_n(y) \leq 0$, and C and each f_i are continuous and differentiable at y, then either $\nabla C = 0$ (at y), or there exists some f_i with $f_i(y) = 0$ and $\nabla f_i = 0$ (at y), or there exists a nonempty subset of the constraints $f_{i_1}(y) \leq 0, \ldots, f_{i_k}(0) \leq 0$ such that $f_{i_j}(y) = 0$ for $1 \leq j \leq k$ and positive coefficients $\lambda_1, \ldots, \lambda_k$ such that $\nabla C = \lambda_1 \nabla f_{i_1} + \ldots + \lambda_n \nabla f_{i_n}$ (at y).*

Note that lemma 22 allows y_i to be negative. The first step in proving lemma 22 is to show that without loss of generality we can work with a closed and compact feasible set. For $K > 0$ it is not difficult to show that there exists a feasible point, i.e., a vector y such that $\sum_{i=1}^n P_i e^{\beta y_i} \leq K$. Let C_0 denote an arbitrary feasible value, i.e., $\sum_{i=1}^n Q_i y_i$ for some feasible point y. Without loss of generality we need only consider points y satisfying $\sum_{i=1}^n Q_i y_i \geq C_0 - 1$. So we now have a constrained optimization problem with objective function $\sum_{i=1}^n Q_i y_i$ and feasible set defined by the following constraints.

$$\sum_{i=1}^n P_i e^{\beta y_i} \leq K \tag{25}$$

$$\sum_{i=1}^n Q_i y_i \geq C_0 - 1 \tag{26}$$

Constraint (25) implies an upper bound on each y_i and constraint (26) then implies a lower bound on each y_i. Hence the feasible set is closed and compact.

We now note that any continuous objective function on a closed and compact feasible set must be bounded and must achieve its maximum value on some point in the set. A constraint of the form $f(y) \leq 0$ will be called active at y if $f(y) = 0$. For an objective function whose gradient is nonzero everywhere, at least one constraint must be active at the maximum. Since C_0 is a feasible value of the objective function, constraint (26) can not be active at the maximum. So by the Kuhn-Tucker lemma, the point y achieving the maximum value must satisfy the following.

$$Q_i = \lambda P_i \beta e^{\beta y}$$

Which implies the following.

$$y_i = \frac{\ln(\frac{Q_i}{\lambda P_i \beta})}{\beta}$$

Since constraint (25) must be active at the maximum, we have the following.

$$\sum_{i=1}^{n} P_i e^{\beta y_i} = \sum_{i=1}^{n} \frac{Q_i}{\lambda \beta} = \frac{1}{\lambda \beta} = K$$

So we get $\lambda = 1/(\beta K)$ and the following.

$$\sum_{i=1}^{n} Q_i y_i = \sum_{i=1}^{n} Q_i \frac{\ln \frac{Q_i}{P_i} + \ln K}{\beta} = \frac{D(Q||P) + \ln K}{\beta}$$

Since this is the maximum value of $\sum_{i=1}^{n} Q_i y_i$, the lemma is proved.

5 Conclusion

Model averaging has important applications in a variety of machine learning applications including language modeling and decision trees. The results of this paper show that model averaging, even over uncountable concept classes, can be given a theoretical PAC justification independent of Bayesian assumptions. Whether analogous results for infinite concept classes can be obtained for on-line algorithms remains an open question.

Acknowledgments: I would like to give special thanks to Manfred Warmuth for inspiring this paper and emphasizing the analogy between the PAC and on-line settings. I would also like to give special thanks to Robert Schapire for simplifying and strengthening theorem 2. Avrim Blum, Yoav Freund, Michael Kearns, John Langford, Yishay Mansour, and Yoram Singer also provided useful comments and suggestions.

References

[1] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic Publishers, 1991.

[2] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[3] Wray Buntine. Learning classification trees. *Statistics and Computing*, 2:63–73, 1992.

[4] Nicol Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *JACM*, 44(3):427–485, 1997.

[5] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior.* to appear.

[6] Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *COLT-97*, 1997.

[7] D. Haussler, M. Kearns, H.S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25:195–236, 1996. Also Appeareed in COLT-94.

[8] David P. Helmbold and Robert E. Schapire. Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, 27(1):51–68, 1997.

[9] Michael Kearns and Yishai Mansour. A fast, bottom-up decision tree pruning algorithm with near-optimal generalization. In *Proceedings of the 15th International Conference on Machine Learning.* Morgan Kaufmann, 1998.

[10] Michael Kearns, Yishai Mansour, Andrew Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth ACM Conference on Computational Learning Theory*, pages 21–30. ACM Press, 1995.

[11] N. Littlestone and Manfred Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. An extended abstract appeared in COLT-89.

[12] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Transactions on Information Theory*, 42:48–54, 1996.

[13] David McAllester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, 1998.

[14] David McAllester. Pac-bayesian model averaging. In *COLT-99*, 1999.

[15] Jonathan J. Oliver and David Hand. On pruning and averaging decision trees. In *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[16] F.C. Pereira and Y. Singer. An efficient extension to mixture techniques for prediction and decision trees. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 114–121, 1997. Long version to appear in Machine Learning.

[17] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[18] John Shawe-Taylor and Robert Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*. ACM Press, 1997.

6 Appendix: Proof of Theorem 16

Let $\hat{l}^*$ be a value of $\hat{l}$ achieving the minimum value of $\hat{l} + \sqrt{\gamma(\hat{l})}$. Let γ^* denote $\gamma(\hat{l}^*)$ and let b^* denote $\hat{l}^* + \sqrt{\gamma^*}$. Assume that $\gamma^* > 0$. We must show that $\lim_{m\to\infty} B(Q^{\beta^*}) = b^*$ where $\beta^* = 2\sqrt{2m\ln(1/p(\hat{l}^*))}$.

For any value $\beta > 0$, the posterior distribution Q^β induces a density function on the values of $\hat{l}$. More specifically we define $q(\hat{l})$ as follows where η equals $\beta/(2m)$ and where $\gamma'(\hat{l})$ equals $\ln(1/p(\hat{l}))/(2m)$.

$$q(\hat{l}) \quad = \quad \lim_{\Delta\to 0} \frac{Q^{\beta^*}(C(\hat{l},\, \Delta))}{\Delta}$$

$$\begin{aligned} &= p(\hat{l})\frac{dQ^{\beta^*}}{dP} \\ &= p(\hat{l})(1/Z)e^{-\beta^*\hat{l}} \\ &= (1/Z)e^{-2m(\gamma'(\hat{l})+\eta\hat{l})} \end{aligned}$$

Note that for $\gamma(\hat{l}) = 0$ we have $\gamma'(\hat{l}) \neq \gamma(\hat{l})$ but that $\lim_{m\to\infty} \gamma'(\hat{l}) = 0 = \gamma(\hat{l})$. The last line above suggests that as m increases the posterior distribution will becomes arbitrarily highly focused on those values $\hat{l}$ minimizing $\gamma(\hat{l}) + \eta\hat{l}$. The parameter η should be selected so that the distribution becomes focused on $\hat{l}^*$. By the definition of $\hat{l}^*$ we have $\hat{l} + \sqrt{\gamma(\hat{l})} \geq b^*$ for all $\hat{l} \in [0,\ 1]$. This implies the following.

$$\gamma(\hat{l}) \geq \begin{cases} (\hat{l} - b^*)^2 & \text{if } \hat{l} < b^* \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

Note that $\hat{l}^* = b^* - \sqrt{\gamma^*}$. Since we have assumed $\gamma^* > 0$ we must have $\hat{l}^* < b^*$. The parameter η^* is selected so that there exists a line of the form $\gamma + \eta^*\hat{l} = c$ that is tangent to the curve $\gamma = (\hat{l} - b^*)^2$ at the point $\langle \hat{l}^*,\ \gamma^* \rangle$. This gives $\eta^* = 2(b^* - \hat{l}^*) = 2\sqrt{\gamma^*}$ which yields $\beta^* = 2\sqrt{2m\ln(1/p(\hat{l}^*))}$. We then have that $\hat{l}^*$ is the unique value of $\hat{l}$ minimizing $\gamma(\hat{l}) + \eta^*\hat{l}$. Let A be the quantity $\gamma^* + \eta^*\hat{l}^*$. We now have the following for all $\hat{l} \in [0, 1]$.

$$\gamma(\hat{l}) + \eta^*\hat{l} \geq A \tag{28}$$

We now show that for sufficiently large m we can replace $\gamma(\hat{l})$ with $\gamma'(\hat{l})$ in (28). Note that (27) implies that for $\hat{l} < b^*$ we have $\gamma(\hat{l}) > 0$ and hence $\gamma'(\hat{l}) = \gamma(\hat{l})$. For $\hat{l} \geq b^*$ we have $\eta^*\hat{l} \geq \eta^* b^* = 2\gamma^* + \eta^*\hat{l}^* = A + \gamma^*$. Although $\gamma'(\hat{l})$ is negative, it approaches zero as m goes to infinity. Hence for sufficiently large m we have the following for all $\hat{l}$.

$$\gamma'(\hat{l}) + \eta^*\hat{l} \geq A \tag{29}$$

Note that we have $Z = \int_0^1 e^{-2m(\gamma'(\hat{l})+\eta^*\hat{l})}\ d\hat{l}$. Bounding this by (29) we get that for sufficiently large m we have $Z \leq e^{-2mA}$, which yields the following.

$$\lim_{m\to\infty} \frac{\ln(1/Z)}{2m} \geq A \tag{30}$$

We will now show that (30) is tight. We define $Z_{1/m}$ as follows.

$$Z_{1/m} = \int_{\hat{l}^*-1/m}^{\hat{l}^*+1/m} e^{-2m(\gamma'(\hat{l})+\eta^*\hat{l})}\ d\hat{l}$$

For sufficiently large m we have that $\gamma'(\hat{l}) = \gamma(\hat{l})$ for all $\hat{l} \in [\hat{l}^* - 1/m,\ \hat{l}^* + 1/m]$. Furthermore, by the mean value theorem we have that there exists $\hat{l}^+ \in [\hat{l}^* - 1/m,\ \hat{l}^* + 1/m]$ satisfying the following.

$$Z_{1/m} = \frac{2}{m} e^{-2m(\gamma(\hat{l}^+) + \eta^* \hat{l}^+)}$$

So we get the following.

$$\frac{\ln(1/Z_{1/m})}{2m} = \gamma(\hat{l}^+) + \eta^* \hat{l}^+ + \frac{\ln(m/2)}{2m}$$

Taking the limit as $m \to \infty$ now gives the following.

$$\lim_{m\to\infty} \frac{\ln(1/Z_{1/m})}{2m} = A \tag{31}$$

We now have that (30) and (31) imply the following.

$$\lim_{m\to\infty} \frac{\ln(1/Z)}{2m} = A \tag{32}$$

We now show that the posterior becomes highly peaked around $\hat{l}^*$. We write $q(\hat{l})$ as $q_m(\hat{l})$ to emphasize the fact that the distribution changes with increasing values of m. For a function $f(\hat{l})$ on $\hat{l} \in [0, 1]$ we write $E_{\hat{l} \sim q_m} f(\hat{l})$ to denote the expectation of $f(\hat{l})$ over selecting $\hat{l}$ according to the distribution q_m. The distribution q_m becomes highly peaked in the sense of the following lemma.

Lemma 33 *For any continuous function f from $[0, 1]$ to the reals we have the following.*

$$\lim_{m\to\infty} E_{\hat{l}\sim q_m} f(\hat{l}) = f(\hat{l}^*)$$

Before proving lemma 33 we note that lemma 33 together with (32) imply theorem 16. In particular, given (32) and lemma 33, we have the following.

$$\begin{aligned}
\lim_{m\to\infty} B(Q^{\beta^*}) &= \lim_{m\to\infty} \left[E_{\hat{l}\sim q_m} \hat{l} + \sqrt{\frac{E_{\hat{l}\sim q_m} \ln(q(\hat{l})/p(\hat{l}))}{2m}} \right] \\
&= \hat{l}^* + \sqrt{\lim_{m\to\infty} E_{\hat{l}\sim q_m} \frac{\ln(1/p(\hat{l}))}{2m} + \lim_{m\to\infty} E_{\hat{l}\sim q_m} \frac{\ln q(\hat{l})}{2m}}
\end{aligned}$$

$$
\begin{aligned}
&= \hat{l}^* + \sqrt{\frac{\ln(1/p(\hat{l}^*))}{2m} - \lim_{m\to\infty} E_{\hat{l}\sim q_m}\left[\gamma'(\hat{l}) + \eta^*\hat{l}\right] + \lim_{m\to\infty}\frac{\ln(1/Z)}{2m}} \\
&= \hat{l}^* + \sqrt{\gamma(\hat{l}^*) - [\gamma(\hat{l}^*) + \eta^*\hat{l}^*] + A} \\
&= \hat{l}^* + \sqrt{\gamma(\hat{l}^*)}
\end{aligned}
$$

It now remains only to prove lemma 33. Define the distribution q_m^ϵ to be a scaled version of q_m restricted to the interval $[\hat{l}-\epsilon,\ \hat{l}+\epsilon]$. More specifically we have the following.

$$
q_m^\epsilon(\hat{l}) = \begin{cases} (1/Z_\epsilon)e^{-2m(\gamma'(\hat{l})+\eta^*\hat{l})} & \text{if } \hat{l} \in [\hat{l}-\epsilon,\ \hat{l}+\epsilon] \\ 0 & \text{otherwise} \end{cases}
$$

Here Z_ϵ is defined as follows.

$$
Z_\epsilon = \int_{\hat{l}^*-\epsilon}^{\hat{l}^*+\epsilon} e^{-2m(\gamma'(\hat{l})+\eta^*\hat{l})}\, d\hat{l}
$$

Clearly $Z_\epsilon \leq Z$. Furthermore, for $1/m \leq \epsilon$ we have $Z_\epsilon \geq Z_{1/m}$. These observations plus (30) and (31) imply the following.

$$
\lim_{m\to\infty} \frac{\ln(1/Z_\epsilon)}{2m} = A \tag{34}
$$

We now consider values of $\hat{l}$ not in the interval $[\hat{l}-\epsilon,\ \hat{l}+\epsilon]$. Define I^- and I^+ as follows.

$$
I^- = \int_0^{\hat{l}^*-\epsilon} e^{-2m(\gamma'(\hat{l})+\eta^*\hat{l})}\, d\hat{l}
$$

$$
I^+ = \int_{\hat{l}^*+\epsilon}^{1} e^{-2m(\gamma'(\hat{l})+\eta^*\hat{l})}\, d\hat{l}
$$

We will show that, for fixed ϵ, these quantities become exponentially smaller than Z_ϵ. First we consider I^+. We can assume without loss of generality that $\hat{l}^* + \epsilon < b^*$. Considering the line tangent to the curve $\gamma = (\hat{l} - b^*)^2$ at $\hat{l} = \hat{l}^* + \epsilon$. For $\hat{l} > \hat{l}^* + \epsilon$ we must have that $\gamma(\hat{l})$ is above this line. Furthermore, even for $\gamma(\hat{l})$ on this line we have that $\gamma(\hat{l}) + \eta^*\hat{l} - A$ must be strictly positive and must grow linearly with $\hat{l}$. Hence there exist constants $a, c > 0$ such that for $\hat{l} > \hat{l}^* + \epsilon$ we have $\gamma(\hat{l}) + \eta^*\hat{l} > A + a + cz$ where

$z = \hat{l} - (\hat{l}^* + \epsilon)$. This gives the following.

$$\begin{aligned} I^+ &\leq e^{-2m(A+a)} \int_0^\infty e^{-2mcz}\, dz \\ &= \frac{1}{2mc} e^{-2m(A+a)} \end{aligned}$$

So for sufficiently large m we have the following.

$$\frac{\ln(1/I^+)}{2m} > A + a$$

This implies that as m increases I^+ becomes exponentially smaller than Z_ϵ. A similar analysis holds for I^-. These observations imply the following.

$$\lim_{m\to\infty} \left[E_{\hat{l}\sim q_m} f(\hat{l}) - E_{\hat{l}\sim q_m^\epsilon} f(\hat{l}) \right] = 0 \tag{35}$$

Let $f([\hat{l}^* - \epsilon,\ \hat{l}^* + \epsilon])$ be the set of values of the form $f(\hat{l}^+)$ with $\hat{l}^+ \in [\hat{l}^* - \epsilon,\ \hat{l}^* + \epsilon]$. by the mean value theorem we have the following.

$$E_{\hat{l}\sim q_m^\epsilon} f(\hat{l}) \in f([\hat{l}^* - \epsilon,\ \hat{l}^* + \epsilon]) \tag{36}$$

Since f is continuous the set $f([\hat{l}^* - \epsilon,\ \hat{l}^* + \epsilon])$ must be an interval around $f(\hat{l}^*)$ whose width goes to zero as ϵ goes to zero. So (35) and (36) imply that as m increases the quantity $E_{\hat{l}\sim q_m} f(\hat{l})$ enters ever tighter intervals around $f(\hat{l}^*)$. This implies lemma 33.