Some PAC-Bayesian Theorems

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

September 28, 1998

Abstract

This paper gives PAC guarantees for "Bayesian" algorithms — algorithms that optimize risk minimization expressions involving a prior probability and a likelihood for the training data. PAC-Bayesian algorithms are motivated by a desire to provide an informative prior encoding information about the expected experimental setting but still having PAC performance guarantees over all IID settings. The PAC-Bayesian theorems given here apply to an arbitrary prior measure on an arbitrary concept space. These theorems provide an alternative to the use of VC dimension in proving PAC bounds for parameterized concepts.

1 INTRODUCTION

Much of modern learning theory can be divided into two seemingly separate areas — Bayesian inference and PAC learning. Both areas study learning algorithms which take as input training data and produce as output a concept or model which can then be tested on test data. In both areas learning algorithms are associated with correctness theorems. PAC correctness theorems provide performance guarantees which hold whenever the training and test data are drawn independently from an identical distribution (IID). Bayesian correctness theorems apply whenever the training and test data are generated according to the given prior. For an experimental setting where training and test data are generated according to some probability distribution other than the prior, no guarantee is proved.

The difference between the two areas can be viewed as a generality/performance tradeoff. We define an "experimental setting" to be a probability distribution over training and test data. A PAC performance guarantee applies to a broad class of experimental settings. A Bayesian correctness theorem applies to only experimental settings consistent with the prior used in the algorithm. However, in this restricted class of settings the Bayesian learning algorithm can be optimal and will generally outperform PAC learning algorithms.

There has been considerable work in the area of structural risk minimization (SRM). 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 learned and the "goodness of fit", "description length", or "likelihood" of the training data. Under this interpretation of SRM, Bayesian algorithms which 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, Mansour, Ng, and Ron in [3]. They note that SRM algorithms for which one can prove PAC guarantees assign larger weight to concept complexity (prior probability) than do classical Bayesian MAP, or equivalently, minimum description length (MDL) algorithms. They give experimental evidence that Bayesian and MDL algorithms tend to over fit in experimental settings where the Bayesian assumptions fail. Algorithms associated with PAC theorems guarantee a certain performance on test data and so provide some theoretical insurance against over fitting. This supports the idea of a performance/generality trade off. Bayesian algorithms are highly effective when the Bayesian assumptions hold but can over fit when those assumptions fail. PAC algorithms avoid over-fitting in all IID experimental settings but can not be tuned with the kind of detailed information that an informative Bayesian prior provides. The PAC-Bayesian theorems and algorithms described in this paper attempt to get the best of both PAC and Bayesian approaches by combining the ability to be tuned with an informative prior with PAC guarantees that hold in all IID experimental settings.

The PAC-Bayesian learning theorems presented here are related to a recent theorem by Shawe-Taylor and Williamson [7]. They show that if one can find a ball of sufficient volume in a parameterized concept space then the center of that ball has low error rate. The theorems presented here, on the other hand, do not make any assumptions about the nature of the concept space — the theorems apply to any prior measure on any space of concepts. Here there is no assumption that the concept space has geometric structure — there is no notion of "ball" or "center", only the notion of a set of a certain measure. The error rate of a set is taken to be the average error rate of its members. Ignoring the fact that the Shawe-Taylor and Williamson result is about the center of a ball while the result here is about the average over a set, the bounds given here are simpler and significantly tighter — they have smaller constants and eliminate a factor of $\log(m)$.

2 NOTATIONAL CONVENTIONS

For a countable distribution P the notation $P(x)$ denotes the probability of value x under the distribution P. We use "distributed variables", i.e., variables implicitly associated with probability distributions.[1] The notation $E_x\ f(x)$ denotes the expectation of $f(x)$ when the distributed variable x is selected according to its associated distribution. $P_x\ \Phi(x)$ denotes the probability of $\Phi(x)$ when x is selected according to its distribution. When the intended random variable (the intended measure space) is clear from context, the notion $P(U)$ will abbreviate $P_x(x \in U)$. The notation $P_{x \in U}\Phi(x)$ denotes $(P_x\ x \in U \wedge \Phi(x))/P_x(x \in U)$. Similarly, $E_{x\in U}\ f(x) = (E_x\ I_{x\in U} f(x))/P_x(x \in U)$ where $I_{x\in U}$ is 1 if $x \in U$ and 0 otherwise. For $\delta > 0$, the notation $\forall^\delta x\ \ \Phi(x)$ means $P_x\ \Phi(x) \geq 1 - \delta$.

3 TWO PAC-BAYESIAN THEOREMS

This section presents the two main theorems of the paper. Each theorem is preceded by a preliminary theorem that has appeared in similar forms in earlier papers. In each case the preliminary theorem is a uniform convergence

[1] Distributed variables should be distinguished from random variables. A random variable is a function of a distributed variable, e.g., if x is a distributed variable then $f(x)$, $g(x)$, and $h(x)$ are different random variables whose values are determined by the value of x. Traditionally $f(x)$, $g(x)$ and $h(x)$ are written as f, g, and h.

theorem quantifying over all concepts in a countable concept class while the main theorem is a uniform convergence theorem quantifying over all measurable subsets of an arbitrary measure space of concepts.

The first preliminary theorem is a variant of a theorem given by Shawe-Taylor et al. [6]. As Shawe-Taylor et al. note, the first theorem is closely related to a theorem due to Linial, Mansour and Rivest [4]. Linial, Mansour and Rivest consider realizable PAC concept learning where the concept space is countably infinite with infinite VC dimension. In their analysis they make use of a particular assignment of weights to concepts where the sum of the weights equals 1. Shawe-Taylor et al. realized that any weighting summing to 1 suffices for the analysis and reinterpreted this weighting as a kind of Bayesian prior.

Assume a given distribution P over a countably infinite class of concepts such that $P(c) > 0$ for all concepts c in the class. Also assume a fixed distribution over instances, a given target concept t in the class of concepts, and some way of associating each concept with a set of instances (the set of instances accepted by that concept). We write $x \in c$ to indicate that x is an instance of the concept c. As usual, the error rate $\epsilon(c)$ of concept c is defined for the given target t to be the probability over the choice of x that c disagrees with t on x. We say that a concept is consistent with a set of instances if it agrees with the target concept on those instances.

Preliminary Theorem 1 *For any probability distribution P assigning nonzero probability to every concept in a countable concept class containing a target concept t, and any probability distribution on instances, we have, for any $\delta > 0$, that with probability at least $1 - \delta$ over the selection of a sample of m instances, the following holds for all concepts c agreeing with t on that sample.*

$$\epsilon(c) \leq \frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{m}$$

The proof is a straightforward union bound argument. One observes that the probability that a concept c with error rate ϵ is consistent with a sample of m instances is at most $e^{-\epsilon m}$. If ϵ is larger than that allowed by the above bound, the probability that c is consistent with the sample is no larger than $P(c)\delta$. The probability that some c violating the bound is consistent with the sample is then bounded by $\sum_c P(c)\delta = \delta$.

This theorem suggests a simple PAC-Bayesian learning algorithm. In particular, given a training sample one selects a concept c that minimizes the given upper bound on $\epsilon(c)$. This will be a concept, among those consistent with the training data, that maximizes $P(c)$, and hence a concept with maximum posterior probability.

Consider an experimental setting where training and test data are generated by first selecting a target concept at random using some particular distribution. For this setting we can tune the performance of the algorithm associated with preliminary theorem 1 by adjusting the prior used in the algorithm to be similar to the distribution in the experimental setting. In this way we can give the algorithm knowledge of the expected experimental setting while preserving a performance guarantee that holds in all IID settings.

The algorithm associated with preliminary theorem 1 selects a concept with maximum posterior probability (MAP). From a Bayesian perspective, MAP algorithms have serious drawbacks. Consider an experimental setting where the target concept is selected according to the prior used in the algorithm. The optimal learning algorithm in this setting takes as input a sample S and outputs the concept that accepts an instance x if $P(x \in t|S) \geq \frac{1}{2}$. We can compute this posterior probability as follows where U is the set of concepts consistent with the sample S.

$$P(x \in t|S) = \frac{\sum_{c \in U: x \in c} P(c)}{\sum_{c \in U} P(c)}$$

The optimal concept in this setting takes a weighted vote over all concepts consistent with the sample where the weight of each concept is proportional to its prior probability.

We now give a "mixture" version of preliminary theorem 1. For the mixture theorem we allow any (possibly uncountable) measure space of concepts. For a fixed target concept t and any measurable subset U of the concept space we define $\epsilon(U)$ to be $E_{c \in U}\ \epsilon(c)$. Given a concept set U we can define a prediction process by selecting a concept c from U (with probability proportional to the prior) and then using c to predict whether x is an instance of the target concept. The error rate $\epsilon(U)$ is the error rate of this stochastic prediction process. We now have the following mixture theorem.

Theorem 1 *For any measure on any concept space and any measure on a space of instances we have, for $\delta > 0$, that with probability at least $1 - \delta$*

over the choice of a sample of m instances all measurable subsets U of the concepts such that every element of U is consistent with the sample and with $P(U) > 0$ satisfies the following.

$$\epsilon(U) \leq \frac{\ln \frac{1}{P(U)} + \ln \frac{1}{\delta} + 2 \ln m + 1}{m}$$

Although theorem 1 holds for arbitrary concept measures, it is useful to consider the special case of enumerable concept classes. Note that if U is a singleton set then theorem 1 gives essentially the same bound as the corresponding preliminary theorem. However, theorem 1 is significantly stronger than the preliminary theorem. Theorem 1 makes a uniform claim about all subsets of concepts.

Theorem 1 justifies a learning algorithm which selects a set U so as to minimize the above upper bound on the error rate. The optimal set U is precisely the set of all concepts consistent with the sample. In practice, however, the set of concepts consistent with the sample may be difficult to find, or may be infinite. So the algorithm may only be able to construct a subset of this optimal set. The above bound provides a performance guarantee for any such subset. As with the preliminary theorem, the performance of theorem 1 can be tuned by setting the prior distribution on concepts to be similar to the distribution appearing in the expected experimental setting.

Although theorem 1 is much stronger than the preliminary theorem it follows almost immediately from a general quantifier reversal principle.

Quantifier Reversal Lemma: *Let x and y be random variables and let δ range over real numbers. Let $\Phi(x,\ y,\ \delta)$ be any measurable formula such that for any x and y we have $\{\delta \in (0,\ 1] :\ \Phi(x,\ y,\ \delta)\} = (0,\ \delta_{max}]$ for some δ_{max}. If*

$$\forall x\, \forall \delta > 0\, \forall^{\delta} y\ \Phi(x,\ y,\ \delta)$$

then for any $\delta > 0$ and $0 < \beta < 1$ we have

$$\forall^{\delta} y\ \ \forall \alpha > 0\ \ \forall^{\alpha} x\ \ \Phi(x,\ y,\ (\alpha\beta\delta)^{\frac{1}{1-\beta}})$$

The proof of the quantifier reversal lemma is given in the next section. The quantifier reversal lemma immediately yields a proof of theorem 1. Let $C(S)$ be the set of concepts that agree with the target concept on the sample

S. By a standard argument, if $\epsilon(c) > \frac{\ln \frac{1}{\delta}}{m}$ then $P_S\ c \in C(S) \leq \delta$. This can be rewritten as follows.

$$P_S(c \in C(S) \wedge \epsilon(c) > \frac{\ln \frac{1}{\delta}}{m}) \leq \delta$$

$$\forall\, c\ \ \forall\, \delta > 0\ \ \forall^{\delta} S\ \ [c \in C(S)\ \text{ implies }\ \epsilon(c) \leq \frac{\ln \frac{1}{\delta}}{m}]$$

Now by the quantifier reversal lemma we get the following.

$$\forall^{\delta} S\ \ \forall \alpha > 0\ \ \forall^{\alpha} c\ \ [c \in C(S)\ \text{ implies }\ \epsilon(c) \leq \frac{\ln \frac{1}{\alpha\beta\delta}}{(1-\beta)m}]$$

Now consider any sample S satisfying this condition and any set U of concepts such that each concept in U is consistent with S. We can now instantiate α with $\frac{P(U)}{m}$ yielding the following.

$$\forall^{\frac{P(U)}{m}} c\ \ [c \in C(S)\ \text{ implies }\ \epsilon(c) \leq \frac{\ln \frac{1}{P(U)} + \ln \frac{1}{\delta\beta} + \ln m}{(1-\beta)m}]$$

We now have that all but a fraction of $\frac{1}{m}$ of the concepts in U must have error rate bounded by this formula. Since the remaining fraction of U can have error rate of at most 1, averaging over U yields the following.

$$\epsilon(U) \leq (1 - \frac{1}{m})\frac{\ln \frac{1}{P(U)} + \ln \frac{1}{\delta\beta} + \ln m}{(1-\beta)m} + \frac{1}{m}$$

The final result is obtained by taking $\beta = \frac{1}{m}$.

Next we consider the unrealizable case. Here we assume only that for each concept c and instance x there is a loss $l(c,\ x) \in [0,\ 1]$. Concept learning is a special case where we can take $l(c,\ x)$ to be 0 if c agrees with the target and 1 if c disagrees with the target. But now we do not assume the target concept is in the concept class. From a Bayesian perspective it is more interesting to consider the case of bounded log loss where each concept c represents a probability distribution over instances. We let $P(x|c)$ denote the probability of instance x under the distribution defined by c. We must assume that there is some (very small) minimum probability $\epsilon > 0$ such that for all concepts c and instances x we have $P(x|c) \geq \epsilon$. We can then take $l(c,\ x)$ to be $(-\log P(x|c))/(-\log \epsilon)$. This ensures that $l(c,\ x) \in [0,\ 1]$.

Again we have a preliminary theorem making a uniform statement over a countable set of concepts and then a main theorem making an analogous statement for any subset of any measure space of concepts. For the preliminary theorem we again assume a distribution P over a countable concept space assigning nonzero probability to each concept. Given any loss function l such that $l(c,\ x) \in [0,\ 1]$ and a fixed distribution on instances we define $\bar{l}(c)$ to be $E_x\ l(c,\ x)$. Given a sample S we define $\hat{l}(c,\ S)$ to be $\frac{1}{m}\sum_{x\in S} l(c,\ x)$. Note that in the case of log loss we have that $\hat{l}(c,\ S)$ is a function of $P(S|c)$.

Preliminary Theorem 2 *For any probability distribution P assigning nonzero probability to each concept in a countable concept class, any probability measure on instances, and any loss function l mapping a concept and an instance to $[0,\ 1]$, we have, for $\delta > 0$, that with probability at least $1-\delta$ over the selection of an IID sample S of m instances all concepts c satisfy the following.*

$$\bar{l}(c) \leq \hat{l}(c,\ S) + \sqrt{\frac{\ln\frac{1}{P(c)} + \ln\frac{1}{\delta}}{2m}}$$

Essentially the same result can be found in a variety of places, e.g., [1, 2, 5]. As with preliminary theorem 1, the proof is a simple application of the union bound over the set of concepts but using the Chernoff bound for the probability that a given concept violates the theorem.

Preliminary theorem 2 is associated with a learning algorithm which selects a concept minimizing the stated upper bound on error rate. In the case of log loss we have that $\hat{l}(c,\ S)$ a function of $P(S|c)$. So the algorithm selects a concept c minimizing a function of $P(c)$ and $P(S|c)$. This is analogous to a Bayesian learning algorithm that selects a concept of maximum posterior probability (a MAP algorithm). As with a Bayesian MAP algorithm, the performance can be tuned to a particular experimental setting by selecting an appropriate prior.

Bayesian concept mixtures are generally superior to a single MAP concept. Theorem 2 is a mixture version of preliminary theorem 2. For any measurable set U of concepts we define $\bar{l}(U)$ to be $E_{c\in U}\ \bar{l}(c)$. For any sample S we define $\hat{l}(U,\ S)$ to be $E_{c\in U}\ \hat{l}(c,\ S)$.

Theorem 2 *For any probability measure on a space of concepts, any probability measure on a space of instances, and any measurable loss function l*

mapping a concept and an instance to $[0,\ 1]$*, we have, for* $\delta > 0$*, that with probability at least* $1-\delta$ *over the selection of an IID sample* S *of* m *instances all measurable subsets* U *of the concept space with* $P(U) > 0$ *satisfy the following.*

$$\bar{l}(U) \leq \hat{l}(U,\ S) + \sqrt{\frac{\ln \frac{1}{P(U)} + \ln \frac{1}{\delta} + 2\ln m}{2m}} + \frac{1}{m}$$

It is again instructive to consider theorem 2 for the case of a countable concept class. If U is a singleton set then the bound in theorem 2 is essentially the same as the bound in the corresponding preliminary theorem. However, theorem 2 is much stronger — it makes a uniform statement about all subsets of concepts.

Again the theorem can be associated with an algorithm which selects a set U minimizing the given upper bound. In the case of log loss the procedure finds a set U minimizing a function of $P(U)$ and $P(S|U)$. Hence it can be viewed as a kind of MAP procedure over *sets* of concepts.

Theorem 2 is proved from the quantifier reversal lemma in a manner similar to that of theorem 1. Using the Chernoff bound for an individual concept c we get the following.

$$\forall c\ \ \forall \delta > 0\ \ \forall^{\delta} S\ \ \bar{l}(c) \leq \hat{l}(c,\ S) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

The quantifier reversal lemma then implies the following.

$$\forall^{\delta} S\ \ \forall \alpha > 0\ \ \forall^{\alpha} c\ \ \bar{l}(c) \leq \hat{l}(c,\ S) + \sqrt{\frac{\ln \frac{1}{\beta\delta\alpha}}{(1-\beta)2m}}$$

Again the result follows by setting $\beta = \frac{1}{m}$ and $\alpha = \frac{P(U)}{m}$.

4 THE QUANTIFIER REVERSAL LEMMA

First we prove a preliminary lemma. Let f be any measurable function from the space of values of y the reals and let g be any measurable antimonotone function from the open interval $(0,\ 1)$ to the reals, i.e., g is such that $x \geq y$ implies $g(x) \leq g(y)$. We show that if $\forall^{\delta} y\quad f(y) \geq \delta$ then

$E_y\ g(f(y)) \leq \int_0^1 g(z)dz$. To see this first note that without loss of generality we can assume that singleton sets have zero measure — we can always replace the space of values of y by a cross product of that space with the unit interval. If singleton sets have zero measure then for any natural number $n > 0$ it is possible to divide the values of y into n disjoint sets $U_0, \ldots, U_{n-1}$ such that each U_i has measure $\frac{1}{n}$ and for all $i > 0$ we have that $y \in U_i$ implies $f(y) \geq \frac{i}{n}$. So we have the following.

$$E_y\ g(f(y)) \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-1} g(\frac{i}{n}) = \int_0^1 g(z)dz$$

We now prove the quantifier reversal lemma. Let x and y be any two random variables and let $\Phi(x,\ y,\ \delta)$ be any measurable formula where δ ranges over real numbers where Φ satisfies the condition that for any x and y we have $\{\delta \in (0,\ 1] :\ \Phi(x,\ y,\ \delta)\} = (0,\ \delta_{max}]$ for some δ_{max}. Define $f(x,\ y)$ to be 0 if this set is empty and the unique value of δ_{max} otherwise. For $0 < \delta \leq 1$ we have that $\Phi(x,\ y,\ \delta)$ can be written as $f(x,\ y) \geq \delta$. Now assume that for any x and $\delta > 0$ we have $\forall^\delta y\ \ f(x,\ y) \geq \delta$. For $0 < \beta < 1$ the preceding lemma implies the following.

$$E_y\ [f(x,\ y)]^{\beta-1} \leq \int_0^1 z^{\beta-1}dz = \frac{1}{\beta}$$

Taking the expectation over x, reversing quantifiers, and applying Markov's inequality gives the following.

$$E_yE_x\ [f(x,\ y)]^{\beta-1} \leq \frac{1}{\beta}$$

$$\forall^\delta y\ \ E_x[f(x,\ y)]^{\beta-1} \leq \frac{1}{\beta\delta}$$

$$\forall^\delta y\ \ \forall\alpha > 0\ \ \forall^\alpha x\ \ [f(x,\ y)]^{\beta-1} \leq \frac{1}{\alpha\beta\delta}$$

$$\forall^\delta y\ \ \forall\alpha > 0\ \ \forall^\alpha x\ \ f(x,\ y) \geq (\alpha\beta\delta)^{\frac{1}{1-\beta}}$$

$$\forall^\delta y\ \ \forall\alpha > 0\ \ \forall^\alpha x\ \ \Phi(x,\ y,\ (\alpha\beta\delta)^{\frac{1}{1-\beta}})$$

5 DISCUSSION

The PAC-Bayesian theorems presented here justify learning algorithms which can take advantage of informative priors while preserving a PAC guarantee in all IID experimental settings. While it seems that these theorems represent progress, there are some open questions. Theorem 4 can be viewed as optimizing a function of $P(U)$ and $P(S|U)$ — it is a kind of MAP procedure over *sets* of concepts. From a Bayesian perspective it would be more satisfying to have some form of PAC-Bayesian posterior distribution over concepts. Whether such a distribution can be formulated, and whether it can improve the performance of the learning algorithm, remains open.

Acknowledgments: Yoav Freund, Micheal Kearns, John Langford, Yishai Mansour, Rob Schapire and Yoram Singer provided useful comments and suggestions for this paper.

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] 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.

[4] Nathan Linial, Yishay Mansour, and Ronald Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Information and Computation*, 90:33–49, 1991.

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

[6] John Shawe-Taylor, Peter Bartlett, Robert Williamson, and Martin Anthony. A framework for structural risk minimization. In *Proceedings of the Ninth Annual conference on Computational Learning Theory*, pages 68–76. ACM Press, 1996.

[7] 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.