Bounds for Averaging Classifiers

John Langford [a] **Matthias Seeger** [b]

January 2001

CMU-CS-01-102

[a] *jcl@cs.cmu.edu*
http://www.cs.cmu.edu/~jcl
[b] Institute for Adaptive and Neural Computation
University of Edinburgh
seeger@dai.ed.ac.uk
http://www.dai.ed.ac.uk/homes/seeger/

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We present a generalized PAC bound for averaging classifiers which applies to base hypotheses with a bounded real valued output. In addition, we discuss several methods for quantitatively tightening the bound. In the process, a tightened version of the PAC-Bayes bound [5] is proved.

Keywords: PAC bound, Maximum entropy discrimination, averaging hypotheses

1 Introduction

This paper is the technical companion for an accompanying ICML submission. As such, we are concerned here with the details of what can and can not be proved rather then the implications of this work to the field of machine learning.

The bounds presented in this paper are a qualitative improvement on the margin bound of Schapire, Freund, Bartlett, and Lee [6]. The qualititative improvement essentially suggests a new optimization criterion: optimize for a large margin *and* for a uniform average over as many hypotheses as possible.

The layout of this paper is as follows:

1. Discussion of the relationship with prior relevant results.
2. Statement and proof of the generalized averaging bound.
3. Discussion of various techniques for improving on the bound.
4. Statement and proof of a tightened version of the PAC-Bayes bound necessary for improving the generalized averaging bound.

2 The setting and Important earlier results

2.1 The setting

We first explain the setting, which is the same as the one used in [6].

An input space $\mathcal{X}$ is given, where the members of $\mathcal{X}$ are also referred to as *examples*. The set $\mathcal{X} \times \{-1, 1\}$ is the space of *labeled* examples. A *base hypothesis* h is a mapping from the input space $\mathcal{X}$ into $\{-1, 1\}$. A (possibly infinite) space $\mathcal{H}$ (the hypothesis space) is given and the goal is to construct an *averaging classifier* $c : \mathcal{X} \to \{-1, 1\}$ as a weighted average of base hypotheses:

$$c(x) = \operatorname{sign} \sum_{j=1}^{k} q_j h_j(x) \quad (x \in \mathcal{X}) \ ,$$

where $q_j \geq 0$, $j = 1, \ldots, k$, and $\sum_{j=1}^{k} q_i = 1$. The fundamental assumption here is that labeled training examples are drawn independently, with replacement, from some probability distribution D over $\mathcal{X} \times \{-1, 1\}$. In all the theorems we discuss, D is assumed to be unknown to the procedure which constructs the classifier, and the results hold for all D. Probabilities and expectations over D will be denoted by the subscript D; for example, the true error of an averaging classifier is denoted by:

$$e_D(c) = E_D[I_{\{c(x) \neq y\}}] = E_{(x,y)\sim D}[I_{\{c(x) \neq y\}}]. \tag{1}$$

Here, $I_{\{\cdot\}}$ is the indicator function, which is 1 if its argument evaluates to *true* and 0 otherwise. Probabilities with respect to D are written as $\Pr_D$.

With S, we denote a sample $\{(x_i, y_i) \mid i = 1, \ldots, m\}$ drawn independently and identically distributed (i.i.d.) from $D(x, y)$. The i.i.d assumption is the one fundamental assumption we make in this work. The subscript S denotes empirical expectation or probability over S, for example the empirical error of an averaging classifier is given by:

$$e_S(c) = E_S[I_{\{c(x) \neq y\}}] = \frac{1}{m} \sum_{i=1}^{m} I_{\{c(x_i) \neq y\}} \tag{2}$$

Probabilities with respect to S are written as $\Pr_S$.

2.2 Quantities used in the bound

The basic learning model needs to be augmented with a few definitions for the analysis.

Given a subset $\{h_1, \ldots, h_k\} \subseteq \mathcal{H}$, the set $\{q_1, \ldots, q_k\}$ can be interpreted as a probability distribution over the set $\mathcal{H}$.[1] This distribution will be denoted by Q. We will also often use the unsigned version of the classifier:

$$f(x) = E_{h \sim Q}[h(x)] = \sum_{j=1}^{k} q_j h_j(x) \ .$$

It is important to note that we make no assumption about how the weights $q_1, \ldots, q_k$ are obtained, so our results are applicable to many algorithms.

The derived bounds depend on the powerful concept of the *margin*, $t(x, y)$, of a labeled example with respect to a classifier, namely,

$$t(x, y) = y \sum_{j=1}^{k} q_j h_j(x) = y f(x) \ .$$

The margin is a quantitative measure of how decided the average is. Obviously, $-1 \leq t(x, y) \leq 1$. If $t(x, y) = 1$ (resp. $t(x, y) = -1$), then all the base hypotheses classify correctly (resp. incorrectly). When $t(x, y)$ is close to zero, the classifier is, in some sense, undecided. Note that $c(x) = y$ iff $t(x, y) > 0$.

2.3 Earlier results

The improved averaging bound arises from improving one critical step in the proof of the original margin bound, which we state here for reference. We denote by $\Pr_D$ the probability measure of the distribution D defined above. For any set S of examples, we denote by $\Pr_S$ the uniform probability distribution over the set S.

[1] In the case of finite or countably infinite $\mathcal{H}$, this is achieved by assigning all hypotheses outside the subset the weight zero. If $\mathcal{H}$ is finite, we will usually work with $\mathcal{H} = \{h_1, \ldots, h_k\}$ for simplicity. For uncountable spaces, we define Q as $\sum_j q_j \delta(h, h_j)$, where $\delta(h, h_j)$ is the delta distribution centered on h_j.

Theorem 1 (Margin Bound [6]) *Let $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S from D we have that for all distributions $Q = (q_1, \ldots, q_k)$ over the finite hypothesis space $\mathcal{H}$ and all $\theta \in (0,1]$:*

$$\Pr_D[yf(x) \leq 0] \leq \Pr_S[yf(x) \leq \theta] + O\left(\sqrt{\frac{\theta^{-2}\ln|\mathcal{H}|\log m + \ln\delta^{-1}}{m}}\right), \tag{3}$$

where $f(x) = E_{h\sim Q}[h(x)] = \sum_j q_j h_j(x)$.

Here, the notation $b(m) = O(a(m))$ means there exists a constant C such that $b(m) \leq C \cdot a(m)$ for all m. This margin bound implies that if most training examples have a large margin θ (i.e. $t(x,y) > \theta$ for most $(x,y) \in S$) and the hypothesis space is not too large, then the generalization error cannot be large.

To improve on this bound, we employ a PAC-Bayes bound from McAllester [5]. In the PAC-Bayes setting, a classifier is also defined by a distribution Q over the hypothesis space. However, each classification[2], is carried out according to a hypothesis sampled from Q rather than by the averaging classifier c defined by Q. We are interested in the gap between the *expected* generalization error and the *expected* empirical error, where both expectations are taken with respect to Q. We need to introduce the *relative entropy* (or *Kullback-Leibler divergence*; e.g. [2]):

$$\mathrm{D}(Q\|P) = E_{h\sim Q}\left[\ln\frac{q(h)}{p(h)}\right], \tag{4}$$

where q, p denote the probability densities of the distributions Q, P. If $\mathcal{H}$ is finite, we have

$$\mathrm{D}(Q\|P) = \sum_{j=1}^{k} q_j \ln\frac{q_j}{p_j}, \tag{5}$$

where $Q = (q_1, \ldots, q_k)$, $P = (p_1, \ldots, p_k)$. The relative entropy is an asymmetric distance measure between probability distributions, with $\mathrm{D}(Q\|P) = 0$ iff $Q = P$ almost everywhere.

Theorem 2 (PAC-Bayes [5]) *Let $l(h,(x,y))$ be a binary loss function, P any prior distribution over $\mathcal{H}$ and $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S from D we have that for all distributions Q over the hypothesis space $\mathcal{H}$:*

$$\Pr_{D,Q}[l(h,(x,y)) = 1] \leq \Pr_{S,Q}[l(h,(x,y)) = 1] + \sqrt{\frac{D(Q\|P) + \ln\delta^{-1} + \ln m + 2}{2m-1}} \tag{6}$$

[2] Such classifiers, also called "discriminants", are called *Gibbs classifiers*(e.g. [4]).

An example of a loss function is the well-known *zero-one loss* $l(h,(x,y)) = I_{\{h(x)\neq y\}}$.

We can tighten this bound by employing a more accurate tail bound on the Binomial distribution, which leads us to the following theorem.

Theorem 3 *(PAC-Bayes Relative Entropy bound) Let P be any prior distribution over $\mathcal{H}$ and $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S from D we have that for all distributions Q over the hypothesis space $\mathcal{H}$:*

$$D(Ber_{S,Q} \| Ber_{D,Q}) \leq \frac{D(Q\|P) + \ln \frac{2m}{\delta}}{m-1}$$

where $Ber_{S,Q} = 1$ with probability $\Pr_{S,Q}(l(h,(x,y)) = 1)$ and 0 otherwise, and $Ber_{D,Q} = 1$ with probability $\Pr_{D,Q}(l(h,(x,y)) = 1)$ and 0 otherwise.

This theorem gives a constraint on the KL divergence between the average empirical and average true errors rather then the standard l_1 distance. This bound is always at least as tight as the original PAC-Bayes bound [5] and sometimes much tighter, such as when the average empirical error is near 0. A proof is given in section 5.

This theorem holds for finite and infinite hypothesis spaces. The PAC-Bayes theorem guarantees a tighter bound (except at low order) than earlier results such as the following Occam's razor theorem.

Theorem 4 (Occam's Razor [1]) *Let P be a distribution over a hypothesis space $\mathcal{H}$ and $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S from D, for all hypotheses $h \in \mathcal{H}$:*

$$\Pr_D[h(x) \neq y] \leq \Pr_S[h(x) \neq y] + \sqrt{\frac{\ln(1/p(h)) + \ln \delta^{-1}}{2m}}\ . \tag{7}$$

Note that, for finite $\mathcal{H}$ and up to low order terms, theorem 4 is a special case of theorem 3, where we choose delta distributions $Q = (0,\ldots,0,1,0,\ldots,0)$ in theorem 3 and upper bound the KL divergence with the hoeffding bound. The essence of our improvement of the standard margin bound comes from the application of the PAC-Bayes bound instead of the Occam's razor bound within the standard proof of the margin bound.

3 A generalized averaging bound

In this section, we state and prove our main result, a PAC-Bayes generalization error bound for averaging *bounded real-valued* hypotheses. Averaging binary hypotheses, with which we have been concerned with so far, is a special case for which we do not have to sacrifice any accuracy for the generalization.

Within this section, let $A > 0$, and let $\tilde{\mathcal{H}}$ be a set of real-valued hypotheses $h : \mathcal{X} \to [-A, A]$. If Q denotes a distribution over $\tilde{\mathcal{H}}$, we focus on the average classifier $c(x) = \mathrm{sign} f(x)$ where $f(x) = E_{h\sim Q}[h(x)]$.

Theorem 5 (Relative Chernoff Main theorem) *Let P be any probability distribution over $\tilde{\mathcal{H}}$ and let $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S of D we have that for all $\theta \in (0,1]$ and for every distribution Q over $\tilde{\mathcal{H}}$:*

$$D(\mathrm{Pr}_S\left[yf(x) \leq \theta A\right] || \mathrm{Pr}_D\left[yf(x) \leq 0\right]) \\ \leq O\left(\frac{\theta^{-2} D(Q\|P)\ln m + \ln m + \ln \delta^{-1}}{m}\right) \tag{8}$$

where $f(x) = E_{h\sim Q}[h(x)]$.

The main theorem uses a KL-divergence based pseudodistance which is a bit hard to understand intuitively. In order to gain intuition, we can relax the tightness of the proof with an inequality.

$$D(p||q) \geq 2(p-q)^2$$

This relaxation gives us an immediate corollary.

Corollary 1 *Let P be any probability distribution over $\tilde{\mathcal{H}}$ and let $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S of D we have that for all $\theta \in (0,1]$ and for every distribution Q over $\tilde{\mathcal{H}}$:*

$$\mathrm{Pr}_D\left[yf(x) \leq 0\right] \leq \mathrm{Pr}_S\left[yf(x) \leq \theta A\right] \\ + O\left(\sqrt{\frac{\theta^{-2} D(Q\|P)\ln m + \ln m + \ln \delta^{-1}}{m}}\right) \tag{9}$$

where $f(x) = E_{h\sim Q}[h(x)]$.

The proof of the theorem is given in 3.1. Note that the theorems are stated in an asymptotic fashion which may not be very useful in practical applications. Section 4 gives some ideas of how to tighten the result, and the nonasymptotic form, given by the inequalities (26) can be used directly in practice.

The continuous form of the improved averaging bound applies to averages over continuous hypothesis spaces. Note that in this setting, the average needs to be an integral over an uncountably-infinite set of hypotheses or the KL-divergence does not converge. In practice, this is not a significant problem because machine learning algorithms over large hypothesis spaces typically have some parameter stability. In other words, a small shift in the parameters of the learned model produces a small change in the prediction of the hypothesis. With hypothesis stability, we can convert any average over a finite set of hypotheses into an average over an infinite set of hypotheses without significantly altering the predictions of the average.

3.1 Proof of main theorem

The proof has the same structure as the original margin bound proof 1 with one step replaced by the application of the Relative Chernoff PAC-Bayes theorem 3.

First of all, it is clear that we only have to prove the theorem for the case $A = 1$. Let N be any natural number; later, the choice of N will be optimized. In the first part of the proof, we regard θ and N as fixed. Later we generalize this so that they may depend on the sample S.

We construct the distribution Q_N as follows. Draw N hypotheses $h_i \sim Q$ and N variables $u_i \sim U([-1, 1])$ (here, $U(I)$ denotes the uniform distribution over the interval I) such that $h_1, u_1, \ldots, h_N, u_N$ are mutually independent. Q_N might therefore be viewed as the product distribution

$$(Q \times U([-1, 1]))^N . \tag{10}$$

Define the binary valued functions $\phi_i : \mathcal{X} \to \{-1, +1\}$ by

$$\phi_i(x;\, h_i, u_i) = 2I_{\{u_i \le h_i(x)\}} - 1. \tag{11}$$

Given the h_i and u_i we define

$$g(x) = \frac{1}{N} \sum_{i=1}^{N} \phi_i(x;\, h_i, u_i). \tag{12}$$

The set of all such functions g is

$$\tilde{\mathcal{H}}_N = \left\{ \frac{1}{N} \sum_{i=1}^{N} \phi_i(x;\, h_i, u_i) \,\middle|\, h_i \in \tilde{\mathcal{H}}, u_i \in [-1, 1] \right\}. \tag{13}$$

Q_N therefore induces a distribution over $\tilde{\mathcal{H}}_N$, this will be denoted by $g \sim Q_N$. Note that for fixed x, y, the $y\phi_i(x;\, h_i, u_i)$ are i.i.d. Bernoulli variables with mean

$$\begin{aligned} E_{h_i,u_i}\left[y\phi_i(x;\, h_i, u_i)\right] &= yE_{h_i}\Big[(+1)P_{u_i}\Big[u_i \le h_i(x) \,\Big|\, h_i\Big] \\ &\quad + (-1)P_{u_i}\Big[u_i > h_i(x) \,\Big|\, h_i\Big]\Big] \\ &= yE_{h_i}\left[\frac{1}{2}(1 + h_i(x)) - \frac{1}{2}(1 - h_i(x))\right] \\ &= yE_{h_i}\left[h_i(x)\right] = yf(x), \end{aligned} \tag{14}$$

therefore $E_{g \sim Q_N}[yg(x)] = yf(x)$. Since $yg(x)$ is the average over N i.i.d. Bernoulli variables, Hoeffding's bound (see [3], p.122) applies. Thus, for every $x \in \mathcal{X}$, $y \in \{-1, +1\}$, the probability with respect to the sampling of $g \sim Q_N$ satisfies

$$\mathrm{Pr}_{g \sim Q_N}[y(g(x) - f(x)) > \epsilon] \le e^{-\frac{1}{2}N\epsilon^2} \tag{15}$$

For every $\theta \in (0,1]$ and for every (fixed) $g \in \tilde{\mathcal{H}}_N$, the following simple inequality holds:

$$
\begin{aligned}
&\Pr_D[yf(x) \leq 0] \\
&= \Pr_D[yg(x) \leq \tfrac{1}{2}\theta,\, yf(x) \leq 0] \\
&+ \Pr_D[yg(x) > \tfrac{1}{2}\theta,\, yf(x) \leq 0] \\
|\, yf(x) \leq 0] &\leq \Pr_D[yg(x) \leq \tfrac{1}{2}\theta] + \Pr_D[yg(x) > \tfrac{1}{2}\theta \,|\, yf(x) \leq 0].
\end{aligned}
\tag{16}
$$

Note that the left-hand side does not depend on g. By taking the expectation over $g \sim Q_N$ (and exchanging the order of expectations in the second term on the right-hand side), we arrive at

$$
\begin{aligned}
\Pr_D[yf(x) \leq 0] \leq\ & E_{g\sim Q_N}\left[\Pr_D[yg(x) \leq \tfrac{1}{2}\theta]\right] \\
&+ E_D\left[P_{g\sim Q_N}[yg(x) > \tfrac{1}{2}\theta \,|\, yf(x) \leq 0]\right].
\end{aligned}
\tag{17}
$$

As discussed above, we are now ready to apply Hoeffding's inequality (15) with $\epsilon = \theta/2$. For any fixed (x,y) we have

$$
P_{g\sim Q_N}[yg(x) > \tfrac{1}{2}\theta \,|\, yf(x) \leq 0] \leq e^{-\frac{1}{8}N\theta^2}, \tag{18}
$$

so

$$
\Pr_D[yf(x) \leq 0] \leq E_{g\sim Q_N}\left[\Pr_D[yg(x) \leq \tfrac{1}{2}\theta]\right] + e^{-\frac{1}{8}N\theta^2}. \tag{19}
$$

We would like to apply the PAC-Bayes theorem 3 to the right-hand side. For simplicity we stated theorem 3 for the common *zero-one loss* $I_{\{h(x)\neq y\}}$, but it holds more generally for arbitrary binary loss functions. Here we use the loss function $I_{\{yg(x)\leq\theta/2\}}$. Recall that theorem 3 applies for any fixed hypothesis space and "prior" distribution. The hypothesis space here will be $\tilde{\mathcal{H}}_N$. We use as the "prior" the distribution P_N over $\tilde{\mathcal{H}}_N$, which is constructed from the prior P over $\tilde{\mathcal{H}}$ exactly as Q_N is constructed from Q (see (12)). It is easy to see that $\mathrm{D}(Q_N\|P_N) = N\mathrm{D}(Q\|P)$.[3]

It follows from Theorem 3 that with probability at least $1-\delta$ over random choices of S, for every Q,

$$
\mathrm{D}(Ber_{S,Q_N}\|Ber_{D,Q_N}) \leq \frac{\mathrm{D}(Q_N\|P_N) + \ln\frac{2m}{\delta}}{m-1} \tag{20}
$$

where $Ber_{S,Q_N} = 1$ with probability $E_{g\sim Q_N}\left[\Pr_D[yg(x) \leq \frac{1}{2}\theta]\right]$ and 0 otherwise, and $Ber_{D,Q_N} = 1$ with probability $E_{g\sim Q_N}\left[\Pr_S[yg(x) \leq \frac{1}{2}\theta]\right]$ and 0 otherwise.

[3] Note that this reveals a tradeoff between N and $\mathrm{D}(Q\|P)$. Namely, for large N, $g \sim Q_N$ will be a close approximation to the averaging classifier f, which keeps (18) small, but if $\mathrm{D}(Q\|P)$ is not very small, Q_N will be rather far from P_N in terms of relative entropy, as a consequence of the strict factorized forms of the two distributions (they are constructed using i.i.d. samples of size N).

By the same argument as in (16), for every $g \in \tilde{\mathcal{H}}_N$:

$$
\begin{aligned}
&\Pr_S[yg(x) \leq \tfrac{1}{2}\theta] \\
&\leq \Pr_S[yg(x) \leq \tfrac{1}{2}\theta,\ yf(x) > \theta] + \Pr_S[yf(x) \leq \theta] \\
&\leq \Pr_S[yg(x) \leq \tfrac{1}{2}\theta \,|\, yf(x) > \theta] + \Pr_S[yf(x) \leq \theta].
\end{aligned} \tag{21}
$$

Again, we take expectations over $g \sim Q_N$ on both sides, interchange the order of the expectations and apply Hoeffding's inequality (15) with $\epsilon = \theta/2$:

$$
E_S\left[P_{g\sim Q_N}\left[yg(x) \leq \tfrac{1}{2}\theta \,|\, yf(x) > \theta\right]\right] \leq e^{-\frac{1}{8}N\theta^2}, \tag{22}
$$

to arrive at

$$
E_S\left[P_{g\sim Q_N}[yg(x) \leq \tfrac{1}{2}\theta]\right] \leq e^{-\frac{1}{8}N\theta^2} + \Pr_S\left[yf(x) \leq \theta\right]. \tag{23}
$$

Combining (19), (20) and (23), we conclude that with probability at least $1-\delta$, for every Q

$$
\mathrm{D}(q_S \| p_D) \leq \frac{N\mathrm{D}(Q\|P) + \ln\frac{2m}{\delta}}{m-1} \tag{24}
$$

where $q_S = 1$ with probability $e^{-\frac{1}{8}N\theta^2} + \Pr_S\left[yf(x) \leq \theta\right]$ and 0 otherwise, and $p_D = 1$ with probability $\Pr_D[yf(x) \leq 0] - e^{-\frac{1}{8}N\theta^2}$ and 0 otherwise.

This bound holds for any fixed N and θ, which is not yet what we need here, since we want to allow these to depend on the data S. We apply a standard technique to resolve this problem. In essence, the bound we proved so far is a statement about certain events, parameterized by N and θ, namely the probability of each event is smaller than δ. However, we need to prove that the probability of the *union* of all these events is smaller than δ. To this end, we first observe that this union is contained in the union over only a *countable* number of events. Note that if $g \in \tilde{\mathcal{H}}_N$ (see (12)), then $g(x) \in \{(2k-N)/N | k = 0, 1, \ldots, N\}$. Thus, even with all the possible (positive) values of θ, there are no more than $N+1$ events of the form $\{yg(x) \leq \theta/2\}$. Denote by $k(\theta, N)$ the largest integer k such that $k/N \leq \theta/2$. We observe that for every $\theta > 0$, every $g \in \tilde{\mathcal{H}}_N$ and every distribution over (x, y):

$$
Pr\left[yg(x) \leq \theta/2\right] = Pr\left[yg(x) \leq k(\theta, N)/N\right]. \tag{25}
$$

This means that the middle step in the proof above, i.e. the application of theorem 3, depends on (N, θ) only through (N, k). Since the other steps, i.e. the applications of Hoeffding's inequality, are true with probability one, we see that we can restrict ourselves to the union of countably many events, indexed by (N, k). Now, we "allocate" parts of the confidence quantity δ to each of these events, namely (N, k) receives $\delta_{N,k} = \delta/(N(N+1)^2)$, $N = 1, 2, \ldots;\ k = 0, \ldots, N$. It follows easily that the union of all these events has probability at most $\sum_{N,k} \delta_{N,k} = \delta$. Therefore we have proved that with probability at least

$1-\delta$ over random choices of S it holds true that for *all* N and *all* $\theta \in (0,1]$,

$$\begin{aligned} \mathrm{D}(q_S \| p_D) &\le \frac{N\mathrm{D}(Q\|P) + \ln \frac{2m}{\delta_{N,k}}}{m-1} \\ &\le \frac{N\mathrm{D}(Q\|P) + \ln \frac{2m}{\delta} + 3\ln N + 1}{m-1} \end{aligned} \tag{26}$$

where $k = k(\theta, N)$. We can now choose N such as to minimize this bound. N may depend on θ, Q and the sample S.

The asymptotic bound stated in the theorem can be derived by choosing N (with respect to θ and Q) so as to *approximately* minimize the bound we have derived above. If $c \ge 1$ is such that $D(Q\|P) = O(m^c)$, we can choose

$$N = \left\lceil 8\theta^{-2} \ln \frac{m^c}{\mathrm{D}(Q\|P)} \right\rceil .$$

This choice gives us:

$$e^{-\frac{N\theta^2}{8}} = \frac{D(Q\|P)}{m^c}$$

Which implies we have an equation of the form:

$$\begin{aligned} &D(q\|p) + D(q + \frac{D(Q\|P)}{m^c} \| p - \frac{D(Q\|P)}{m^c}) - D(q\|p) \\ &\le O\left(\frac{\theta^{-2} D(Q\|P) \ln m + \ln \frac{1}{\delta} + \ln m}{m} \right) \end{aligned} \tag{27}$$

In order to prove the theorem, we must show that the second and third terms together are of a similar size to the last term. The second and third terms have the form:

$(q+k)\ln \frac{q+k}{p-k} + (1-q-k)\ln \frac{1-q-k}{1-p+k} - q\ln \frac{q}{p} - (1-q)\ln \frac{1-q}{1-p}$

If $p - q > 2k$ and $k < \frac{1}{2}$ then we have:

$= \Theta[(q+k)\ln \frac{q+k+\frac{k}{p}}{p} + (1-q-k)\ln \frac{1-q-k-\frac{k}{1-p}}{1-p} - q\ln \frac{q}{p} - (1-q)\ln \frac{1-q}{1-p}]$

$= \Theta[(q+k)[\ln \frac{q}{p} + \frac{k+\frac{k}{p}}{\frac{q}{p}} + (1-q-k)\ln \frac{1-q}{1-p} - [\frac{k+\frac{k}{1-p}}{\frac{1-q}{1-p}}] - q\ln \frac{q}{p} - (1-q)\ln \frac{1-q}{1-p}]$

$= \Theta[(q+k)[\ln \frac{q}{p} + k(\frac{p+1}{q})] + (1-q-k)[\ln \frac{1-q}{1-p} - k(\frac{2-p}{1-q})] - q\ln \frac{q}{p} - (1-q)\ln \frac{1-q}{1-p}]$

$= \Theta[k(1+p) - k(2-p)]$

$= \Theta[k]$

This completes the proof of the main theorem.

4 Methods for tightening

The previous section showed a bound in asymptotic form which is good for understanding the tradeoffs between the number of examples (m), the size of the hypothesis space ($|H|$), the margin (θ) and the entropy of the average ($H(Q)$). However, it is not a good form for those interested in quantitative application of the bound to specific problems. We state improvements which aid in the development of a quantitatively applicable bound. We can tighten the bound above through several techniques:

1. Making direct use of the tail distribution for the Binomial.
2. Parameterizing and then optimizing the parameterization of arbitrary choices within the proof.
3. Tighter argument within the proof.

4.1 Binomial Tail Bounds

The direct use of the tail distribution of the Binomial will rely upon the cumulative distribution of the Binomial. Let

$$Bin(m, p, \hat{p}) = \sum_{i=0}^{\frac{i}{m} < \hat{p}} \left(\begin{array}{c} m \\ i \end{array} \right) p^i (1-p)^i$$

be the cumulative distribution of a Binomial distribution with true error p up to empirical error $\hat{p}$. We then have the following equality:

$$\Pr_{Bin(m,p)}(p - \epsilon > \hat{p}) = Bin(m, p, p - \epsilon)$$

This inequality can be much tighter then the corresponding Hoeffding bound:

$$\Pr_{Bin(m,p)}(p - \epsilon > \hat{p}) \leq e^{-2\epsilon^2 m}$$

In using the tail probabilities, it will often be the case that we want a constant probability of failure (δ) and want to solve for the smallest ϵ which has a probability of failure of δ or less. This bound can't be stated concisely in closed form but can be found moderately quickly by doing a binary search over ϵ to solve for the ϵ with our desired δ. The binary search solves for the value of p s.t. $Bin(m, p, \hat{p}) = \delta$. We have the equality:

$$\Pr_{Bin(m,p)}(\hat{p} | p \geq \max \bar{p} : Bin(m, \bar{p}, \hat{p}) \geq \delta) \leq \delta$$

This equality implies that with probability $1 - \delta$, we have: $p \leq \max \bar{p} : Bin(m, \bar{p}, \hat{p}) \geq \delta$.

The computational cost of calculating the Binomial tail distribution is often too large so it is possible to use an intermediate approximation known as the relative entropy chernoff bound:

$$\Pr_{Bin(m,p)}(p - \epsilon > \hat{p}) \leq e^{-m\mathrm{D}(p-\epsilon \| p)} \tag{28}$$

$$\Pr_{Bin(m,p)}(p + \epsilon > \hat{p}) \leq e^{-m\mathrm{D}(p+\epsilon \| p)} \tag{29}$$

where $\mathrm{D}(q \| p)$ is the KL divergence between a coin with a bias of q and a coin with a bias of p.

4.2 Extra parameterizations

In the (improved) margin bound proof, we arbitrarily decided to work with the margin of the randomly produced function $g(x)$ at $\frac{\theta}{2}$. This is a good heuristic, but not the optimal choice when we use the improved tail bounds. Since the decision of the margin for the random function $g(x)$ is a parameter of the proof, we are free to optimize it.

4.3 PAC-Bayes bound vs. the standard Occam's razor bound

In the finite case, there are some low order terms in the PAC-Bayes bound which can make it worse then the Occam's razor bound when the posterior is over a small set of hypotheses. Ideally, we would improve the low order terms in the PAC-Bayes bound to remove this discrepancy, but this appears difficult. Instead, we can apply another simple technique: When there are two PAC bounds either of which may be tighter, we can assign $\delta \rightarrow \frac{\delta}{2}$(slightly worsening the bound) to each bound and use the minimum of the two bounds. Since the probability of failure for each bound is $\frac{\delta}{2}$, the total probability of failure of the minimum is δ.

4.4 Optimizing N

The optimal value of N is a function of $\theta, m, \mathrm{D}(Q\|P)$, and δ. All of these are known in advance except for $\mathrm{D}(Q\|P)$. If we can estimate in advance the value of $\mathrm{D}(Q\|P)$, then it becomes possible to optimize the value of N in a data-independent manner. Consequently, it becomes unnecessary to stratify over the possible values of N and we need only stratify over the values of θ in proving the bound. The effect of this improvement is reducing $1/\delta_{N,k} = 1/(N(N+1)^2)$ to $1/\delta_{N,k} = 1/(N(N+1))$ giving us a small improvement in the low order terms of the improved averaging bound.

5 Improving the PAC-Bayes bound

In order to fully benefit from the improved Chernoff bound we need to prove 3 using the Chernoff relative entropy bound. The retrofit of the PAC-Bayes bound is not a simple substitution of the Hoeffding inequality with the Chernoff relative entropy bound so a proof is given.

The proof of the improved PAC-bayes theorem (3) relies upon two lemmas. The first is Lemma 22 from [5] which is given by:

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

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

then

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

The second lemma we will need to prove ourselves. It is basically an improved version of Lemma 17 from [5].

Lemma 2

$$\forall \delta > 0\, \forall^{\delta} S \quad E_{h\sim P} e^{(m-1)D(\hat{e}(h)\|e(h))} \leq \frac{2m}{\delta}$$

First, for any given hypothesis h we prove the following.

$$E_S[e^{(m-1)\mathrm{D}(\hat{e}(h)\|e(h))}] \leq 2m \tag{30}$$

Lemma 2 follows from (30) by taking an expectation over selecting h according to any distribution P over h, reversing the two expectations, and applying Markov's inequality. We now show that (30) follows from (28) and (29). More specifically, we maximize $\int_0^1 e^{(m-1)\mathrm{D}(x\|p)} f(x)dx$ over all functions $f(x)$ satisfying the following for all $q_1 \geq e(h)$ and $q_2 \leq e(h)$.

$$\int_{q_1}^{1} f(x)dx \leq e^{-m\mathrm{D}(q_1\|e(h))}$$

$$\int_{0}^{q_2} f(x)dx \leq e^{-m\mathrm{D}(q_2\|e(h))}$$

The value of $E_S[e^{(m-1)\mathrm{D}(\hat{e}(h)\|e(h))}]$ must be less than this maximum. The integral $\int_0^1 e^{(m-1)\mathrm{D}(x\|e(h))} f(x)dx$ is maximized when $f(x)$ is as "spread out" as possible, i.e., when the above inequalities are replaced by equalities. This gives the following.

$$f(x) = \begin{cases} m\frac{\partial \mathrm{D}(x\|p)}{\partial x} e^{-m\mathrm{D}(x\|p)} & \text{for } x \geq p \\ \\ -m\frac{\partial \mathrm{D}(x\|p)}{\partial x} e^{-m\mathrm{D}(x\|p)} & \text{for } x \leq p \end{cases}$$

Which, in turn, gives the following.

$$\begin{aligned} E_S[e^{(m-1)\mathrm{D}(\hat{e}(h)\|e(h))}] &\leq \int_0^1 e^{(m-1)\mathrm{D}(x\|e(h))} f(x)dx \\ &= \int_0^{e(h)} -m\frac{\partial \mathrm{D}(x\|e(h))}{\partial x} e^{-\mathrm{D}(x\|e(h))} dx \\ &+ \int_{e(h)}^1 m\frac{\partial \mathrm{D}(x\|e(h))}{\partial x} e^{-\mathrm{D}(x\|e(h))} dx \\ &\leq 2m \end{aligned}$$

Now we have the necessary lemmas to finish the proof of 3.

By Jensen's inequality, we have:

$$\mathrm{D}(E_{h\sim Q}\hat{e}(h)\|E_{h\sim Q}e(h)) \leq E_{h\sim Q}\mathrm{D}(\hat{e}(h)\|e(h))$$

Furthermore, according to Lemma 2 we can apply Lemma 1 with $K = \frac{2m}{\delta}$ and $\beta = m-1$ and $y_i = \mathrm{D}(\hat{e}(h)\|\hat{e}(h)+\epsilon)$ to get:

$$E_{h\sim Q}\mathrm{D}(\hat{e}(h)\|e(h)) = \sum_{i=1}^{n} Q_i \mathrm{D}(\hat{e}(h)\|\hat{e}(h)+\epsilon) \leq \frac{\mathrm{D}(Q\|P) + \ln \frac{2m}{\delta}}{m-1}$$

and we are done.

6 Conclusion and Future Work

The improved averaging bound is not yet as tight as it could be and it appears there are several possible theoretical improvements.

1. Remove the low order terms from the bound to make it more quantitatively applicable.
2. Improve the argument to take into account the *distribution* of the margin rather then the margin at some point.
3. Prove a lower bound which corresponds to the upper bound given here. Since no good lower bound yet exists, we do not know that large improvements in the upper bound are not possible.

References

[1] A. Blumer, A. Ehrenfucht, D. Haussler, and M. K. Warmuth, "Occam's Razor" Information Processing Letters, 24:377-380, April 1987

[2] Thomas Cover and Joy Thomas, "Elements of Information Theory" Wiley, New York 1991

[3] L. Devroye, L. Györfi and G. Lugosi "Applications of Mathematics: Stochastic Modelling and Applied Probability" Springer 1996

[4] David Haussler, Michael Kearns, and Robert Schapire, "Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension", Machine Learning 1994 14:83–113

[5] David McAllester, "PAC-Bayesian Model Averaging" COLT 1999

[6] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee, "Boosting the Margin: A new explanation for the effectiveness of voting methods" The Annals of Statistics, 26(5):1651-1686, 1998.