PAC-MDL bounds

Avrim Blum*[1] and John Langford[2]

[1] Computer Science Department, Carnegie Mellon University avrim@cs.cmu.edu
[2] IBM, Watson Research Center jcl@cs.cmu.edu

Abstract. We point out that a number of standard sample complexity bounds (VC-dimension, PAC-Bayes, and others) are all related to the number of bits required to communicate the labels given the unlabeled data for a natural communication game. Motivated by this observation, we give a general sample complexity bound based on this game that allows us to unify these different bounds in one common framework.

> *One Bound to rule them all, One Bound to find them,*
> *One Bound to bring them all and in the darkness bind them.*
> –J.R.R. Tolkien (roughly)[3]

1 Introduction

One of the most basic results about learning in the PAC model is the "Occam's razor" theorem [1] that states that if we can explain the labels of a set of m training examples by a hypothesis that can be described using only $k \ll m$ bits, then we can be confident that this hypothesis generalizes well to future data. One way to view this statement is to consider a setting in which two players Alice and Bob are each given the m examples, but only Alice is given the m labels, and Alice must communicate these labels to Bob. In this case, the result tells us that compression implies learning *for the description language of hypotheses* (that is, Alice communicates the m labels by sending a hypothesis h to Bob, and Bob then reconstitutes the labels by evaluating h on each example).

What if we allow more general procedures for label transmission? In particular, a label transmission procedure is simply an agreement on how a string of bits σ, together with a list of m unlabeled examples, should produce a list of m labels. Instead of being a function from X to Y (which is then run m times by the receiver), in general, a compressed string could represent any function from X^m to Y^m.

We start by pointing out that a number of standard sample complexity bounds can be viewed as stating that "compression implies learning" for different description languages in the above game. Motivated by this observation,

* This work is supported in part by NSF grants CCR-0105488 and NSF-ITR CCR-0122581.

[3] This quote is intended to describe the motivation for this line of work rather than our current state — our hope here is to have made some progress in this direction.

we give a general bound showing that for any description language, if Alice can communicate the labels of the training data in a small number of bits, then Alice can be confident in her ability to predict well on new data (see Section 2 for a formal description). We then show how this statement allows us to derive these other bounds as special cases. In particular, besides the Occam's razor bound, the standard bounds we consider include:

1. The PAC-Bayes bound [7], which states for any "prior" $P(c)$ and any "posterior" $Q(c)$, a bound on the true error rate of a stochastic classifier as a function of the KL divergence between P and Q.
2. The VC bound [8] (of which there are many variants), which states that any classifier chosen from a class of VC dimension d (or a fixed covering number, or a VC entropy) has a true error bound related to d.
3. The Compression bound [6][3], which states that any classifier learned using only a small subset of the training examples has a true error bound related to performance on unused examples.[4]

1.1 Viewing sample complexity bounds as label compression

Consider the case of VC-dimension. Suppose Alice and Bob agree on a hypothesis class H of VC-dimension d, and Alice is able to find a hypothesis in H that is consistent with the training data. This means that in the above communication game, Alice can send the labels to Bob using only $O(d \log m)$ bits. That is because H makes only $O(m^d)$ different partitions of the data and, since Bob has the unlabeled sample, Alice can just send the index of the appropriate one in some canonical ordering.[5] For Structural Risk-Minimization, in which a set of nested hypothesis classes are used, Alice first uses $O(\log d)$ bits to send the hypothesis class, a low-order additive term. Thus, the statement that we can have confidence in consistent hypotheses from a class of low VC-dimension can be thought of as an instantiation of the statement that label compression implies learning, for this particular transmission language. In fact, not surprisingly, the proof for this generalized form of Occam's razor involves a similar random-partitioning trick as used in the standard proof of VC-dimension bounds. The bound also works for lossy compression, so we can recover the VC guarantees in the unrealizable case as well.

As another example, consider PAC-Bayes bounds (the realizable case for simplicity). In this case, instead of agreeing on a hypothesis *class*, Alice and Bob agree in advance on a "prior" $P(c)$ over hypotheses. Notice, now, that given any list of m unlabeled examples $x^m = (x_1, \dots, x_m)$, this prior on hypotheses induces

[4] Unlike the other bounds considered we can not, quite, reproduce the Compression bound with our results.

[5] One might be concerned that the *computational* task of finding the index of this partition could be quite expensive. But since this is just for the purpose of a sample-complexity bound that depends on the number of bits transmitted, Alice does not need to actually produce this index so long as she knows an upper bound on its length.

a prior over labelings $P(y^m) = \sum_{c:c(x_1)=y_1,c(x_2)=y_2,\ldots,c(x_m)=y_m} P(c)$. This means that any particular labeling y^m can be described in roughly $\log(1/P(y^m))$ bits. Thus, if the data is such that the "version space" of consistent classifiers has high total probability mass under P (i.e., $P(y^m)$ is large), then this labeling can be transmitted in a small number of bits.

1.2 Stating a PAC-MDL bound

Our goal, then, is to state a kind of Occam's razor bound, which we call a PAC-MDL bound, that holds for this general communication game. One technical "detail" we need to address in order to give such a bound is a method for interpreting an arbitrary compression procedure as a way of predicting on new examples. In particular, even though the string σ is a function from X^m to Y^m, it is not necessarily in itself a good prediction rule on new data. For example, in the "VC-dimension language" considered above, Alice might send to Bob a string indicating that she is using the 57th linear separator in some canonical ordering — but on some new data set, the 57th linear separator might look totally different depending on how this canonical ordering is defined and perform very badly.

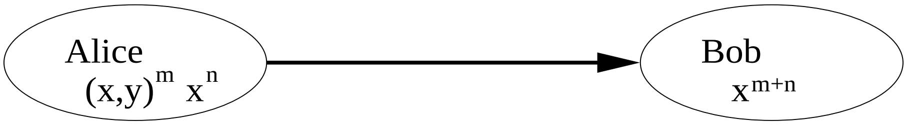
To address this issue without introducing a lot of excess baggage, we define the bound in a transductive setting. In particular, we assume that Alice is given both a set of labeled training data and a set of unlabeled test data. Bob has the all the data without labels, given in some canonical order (e.g., lexicographic) so that he does not know which are training examples and which are test examples. Alice is required to send a string that uncompresses to a labeling on the entire data set. The conclusion is that if this string is small and uncompresses to a labeling with low error on the training set, then Alice can be confident in the labels it yields on the test set.

Working in the transductive setting somewhat complicates arguments like those above for viewing standard bounds in the context of label compression. The VC-dimension case does not change by much: if there are m training examples and n test examples, then Alice sends $O(d\log(m+n))$ bits to transmit the labeling given by some function in the class. However, for the PAC-Bayes language, the different consistent hypotheses may well not agree over the test data. We could just send the shortest string corresponding to a labeling that agrees with the training data, but we use a somewhat different argument in order to get the best bounds.

In Sections 3.1 and 5 we show how these transductive bounds can be used to imply standard *inductive* sample-complexity bounds. The high-level idea of the argument is the same as one that appears in standard VC-dimension arguments. In particular, to take the contrapositive, if there were a significant probability of having a hypothesis with low empirical error but high true error, then there must also be a significant probability of having a hypothesis with low empirical error on the training set and high error on a similarly-sized test set, violating the transductive bounds.

2 The setup

We consider the following compression game between two players Alice and Bob. Alice has available a training set S_{train} consisting of m labeled examples drawn independently from D. Alice also has a test set S_{test} of n unlabeled examples drawn independently from D. Bob has available just the unlabeled versions of the test set and the training set, sorted together in some canonical order (e.g., lexicographic) so that he does not know which are training and which are test examples. Pictorially, the available information is:



Alice's goal is to communicate labels to Bob using as few bits as possible. Alice encodes with a function $A : (X \times Y)^m \times X^n \to \{0,1\}^*$ and Bob decodes with a function $B : X^{m+n} \times \{0,1\}^* \to Y^{m+n}$. We assume the string σ sent is self-terminating, i.e., is given in a prefix-free code. Given any compression/decompression procedure (A, B), we can view the transmitted string σ as representing a function $\sigma : \; X^{m+n} \to Y^{m+n}$. Formally, we define the labeling given by σ to be the labeling produced by running σ on the $m+n$ examples in sorted order (the order in which Bob has the examples). Thus, we can view the encoded information as something more general than just a hypothesis from X to Y.

For a compression algorithm A, labeled training set S_{train} and unlabeled test set S_{test}, the output $\sigma = A(S_{\text{train}}, S_{\text{test}})$ has a specific number of bits $|\sigma|$ sent to Bob in order to label the training and test sets. Let $y(x)$ be the correct labeling of some example x and $y_\sigma(x)$ be the labeling that Bob computes. We can then define $\hat{e}(\sigma, S_{\text{test}}) = \frac{1}{n} \sum_{x \in S_{\text{test}}} I(y_\sigma(x) \neq y(x))$ to be the rate of errors on the test set of the labeling induced by σ. Similarly, $\hat{e}(\sigma, S_{\text{train}})$ is the rate of errors on the training set.

3 Realizable Case

We begin by considering the realizable case: that is, the setting in which the string σ provides a lossless compression of the labels on the training set. We then use our bounds to derive the realizable case of a number of standard sample complexity bounds.

Theorem 1 *(Realizable PAC-MDL bound) For all description languages (i.e., methods of describing the labels via strings σ), for all $\delta > 0$:*

$$\mathbf{Pr}_{S \sim D^m, S' \sim D^n} \left(\forall \sigma : \;\; \hat{e}(\sigma, S) > 0 \text{ or } \hat{e}(\sigma, S') \leq \frac{|\sigma| \ln 2 + \ln \frac{1}{\delta}}{n \ln \left(1 + \frac{m}{n}\right)} \right) > 1 - \delta$$

In other words, it is unlikely there will exist a short string σ that uncompresses perfectly on the training data and yet has high error on the test set. This statement has a bound which is linear in the description complexity, $|\sigma|$. The exact bound is controlled by the size of the sample set m, test set n, and the required confidence, δ.

Proof. Rather than give a proof from first principles, we simplify our more general Theorem 6 (Section 4) for the realizable case of $\hat{e}(\sigma, S) = 0$, using a prior $P(\sigma) = 2^{-|\sigma|}$. Since the training error is 0 in this setting, Theorem 6 simplifies as:

With probability at least $1 - \delta$, all σ with $\hat{e}(\sigma, S) = 0$ satisfy:

$$\frac{\binom{n}{n\hat{e}(\sigma,S')}}{\binom{m+n}{n\hat{e}(\sigma,S')}} \geq 2^{-|\sigma|}\delta$$

$$\Rightarrow \frac{n(n-1)...(n-n\hat{e}(\sigma,S')+1)}{(m+n)(m+n-1)...(m+n-n\hat{e}(\sigma,S')+1)} \geq 2^{-|\sigma|}\delta$$

Using the crude approximation: $\frac{n(n-1)...(n-t+1)}{(m+n)(m+n-1)...(m+n-t+1)} \leq \left(\frac{n}{m+n}\right)^t$ we get:

$$\left(\frac{n}{m+n}\right)^{n\hat{e}(\sigma,S')} \geq 2^{-|\sigma|}\delta$$

Taking the ln of both sides, we get:

$$n\hat{e}(\sigma,S')\ln\left(1+\frac{m}{n}\right) \leq \ln 2^{|\sigma|} + \ln\frac{1}{\delta}$$

$$\Rightarrow \hat{e}(\sigma,S') \leq \frac{\ln 2^{|\sigma|} + \ln\frac{1}{\delta}}{n\ln\left(1+\frac{m}{n}\right)} \leq \frac{|\sigma|\ln 2 + \ln\frac{1}{\delta}}{n\ln\left(1+\frac{m}{n}\right)}$$

□

3.1 Comparison to standard bounds: qualitative results

In the transductive setting discussed above, we have a training set and test set and desire a bound on the number of mistakes on the test set based on observable quantities (like the error rate on the training set and the number of bits needed to communicate the labels). In the inductive setting, which most standard bounds are concerned with, we simply have a training set and desire a bound on the future error rate of a resulting prediction procedure. In this section we show how the PAC-MDL bound can be used to derive standard bounds at the qualitative level. By "qualitative" we mean that we are not concerned with exact constants and we will only focus on the realizable case. Later, we discuss the agnostic case and quantitative comparisons.

Let us begin by considering two special cases of Theorem 1: the limit as $n \to \infty$, and the case of $n = m$.

Corollary 2 *For all description languages over strings σ,*

$$\lim_{n\to\infty} \mathbf{Pr}_{S\sim D^m, S'\sim D^n}\left(\forall\sigma\ \ \hat{e}(\sigma,S) > 0 \ \textit{ or } \hat{e}(\sigma,S') \le \frac{|\sigma|\ln 2 + \ln\frac{1}{\delta}}{m}\right) > 1-\delta$$

Proof. Apply the asymptotically tight approximation, $n\ln\left(1+\frac{m}{n}\right) = n\frac{m}{n} = m$ to Theorem 1. □

Corollary 3 *For all description languages over strings σ,*

$$\mathbf{Pr}_{S\sim D^m, S'\sim D^m}\left(\forall\sigma\ \ \hat{e}(\sigma,S) > 0 \ \textit{ or } \hat{e}(\sigma,S') \le \frac{|\sigma| + \log_2\frac{1}{\delta}}{m}\right) > 1-\delta$$

Proof. Just plug in $n = m$ into Theorem 1. □

One first observation is that Corollary 2 immediately implies the standard Occam's Razor bound, because in that case $|\sigma|$ is just the size in bits of the hypothesis, and as $n \to \infty$, $\hat{e}(\sigma, S')$ approaches the true error with probability 1.

3.2 VC-bounds

For the case of VC-dimension, the number of bits sent *does* depend on the number of test examples, so we instead use Corollary 3, together with the fact that any hypothesis violating the VC bound (having zero empirical error but high true error) likely has a high error rate on a randomly chosen test set. Notice that this argument is also used in the standard proof of VC-bounds.

Theorem 4 *Let H be a hypothesis class of VC-dimension d. Then,*

$$\mathbf{Pr}_{S\sim D^m}\left(\exists c \in H : \hat{e}(c,S) = 0 \ \textit{ and } e(c) > \frac{1}{m}\left(d\log_2(2m) + \log_2\frac{4}{\delta}\right)\right) \le \delta.$$

Proof. Suppose this were false. That is, suppose there were a chance greater than δ that a hypothesis in H existed with zero empirical error but true error greater than the above bound. Then, since the mean and median of the binomial are within 1, this implies that

$$\mathbf{Pr}_{S\sim D^m, S'\sim D^m}\left(\exists c \in H : \hat{e}(c,S) = 0 \text{ and } \hat{e}(c,S') > \frac{1}{m}\left(d\log_2(2m) + \log_2\frac{2}{\delta}\right)\right)$$
$$> \ \delta/2.$$

But, now we appeal to Sauer's lemma which states that the number of ways to label any m examples using hypotheses in a class of VC-dimension d is at most m^d. Thus, for the setting of Corollary 3 where we have $2m$ examples, the labeling given by any $c \in H$ can be transmitted in only a number of bits $|\sigma| = d\log(2m)$ by sending its index in some canonical ordering. But then this violates Corollary 3 using "δ" of $\delta/2$. □

3.3 PAC-Bayes bounds

We can also reconstruct PAC-Bayes bounds (again, focusing on the realizable case for simplicity). PAC-Bayes bounds state that for any "prior" $P(c)$ over hypotheses, if the learning algorithm finds a set of consistent hypotheses U with large probability mass, then we can be confident that the stochastic prediction strategy that randomizes over concepts in U has low true error. To view this in our setting, we consider a communication protocol in which Alice and Bob agree on the prior P, and also agree on a common random string. This common random string is then used to repeatedly sample from the prior P, and the message σ that Alice sends to Bob is just the index of the first consistent hypothesis produced by this sampling. Notice that (as with the Occam bounds) the string σ does not depend on the test data and so we are able to use Corollary 2. In the statement below, let $P(U) = \sum_{c \in U} P(c)$.

Theorem 5 *Let $P(c)$ be any distribution over classifiers. Then,*

$$\mathbf{Pr}_{S \sim D^m}\left(\exists U : \hat{e}(U,S) = 0 \textit{ and } e(U) > \frac{1}{m}\left(\ln\frac{m}{P(U)} + \ln\ln\frac{2}{\delta} + \ln\frac{2}{\delta} + 1\right)\right) \leq \delta,$$

where $e(U)$ is the expected true error of a random hypothesis drawn from U (according to $P|_U$), and likewise $\hat{e}(U,S)$ is the expected empirical error of a random hypothesis from U on set S.

The argument used here is (essentially) a derandomization of the bits-back argument used by Hinton and van Camp [4]. A similar argument also appears in the Slepian-Wolf theorem [2] of information theory.

Proof. Suppose this were false. That is, there was a chance greater than δ that a bad set U existed (one satisfying the conditions of the theorem). For any such set U, let $U' = \{c \in U : e(c) \geq e(U) - 1/m\}$. Since the error rate of any classifier is at most 1, we know $P(U') \geq P(U)/m$. Now, for any fixed such U', if we repeatedly sample from P, the expected number of samples until we first pick a classifier in U' is at most $\frac{m}{P(U)}$. Furthermore, the chance that we do not pick such a classifer in $\frac{m}{P(U)} \ln\frac{2}{\delta}$ samples is at most $\delta/2$. This means that over the sample S and the common random oracle between Alice and Bob, there is a greater than $\delta/2$ chance that a consistent classifier $c \in U'$ exists whose description length $|\sigma_c|$ satisfies

$$|\sigma_c| \leq \log_2\left(\frac{m}{P(U)}\ln\frac{2}{\delta}\right).$$

Now, by definition of U', we have

$$\begin{aligned} e(c) &\geq e(U) - 1/m \\ &> \frac{1}{m}\left(\ln\frac{m}{P(U)} + \ln\ln\frac{2}{\delta} + \ln\frac{2}{\delta}\right) \\ &\geq \frac{1}{m}\left(|\sigma_c|\ln 2 + \ln\frac{2}{\delta}\right). \end{aligned}$$

Putting this all together, there is a probability greater than $\delta/2$ over the sample and common random oracle that such a classifier exists, violating Corollary 2 for "δ" of $\delta/2$. □

4 The Agnostic Case

To discuss the agnostic case, it will be convenient to make the following definitions. First, imagine we have a bucket with m red balls and n blue balls, and we draw $a+b$ of them without replacement. Let us define $\text{Bucket}(m,n,a,b)$ to be the probability that we get at least b blue balls. That is,

$$\text{Bucket}(m,n,a,b) = \sum_{t=b}^{a+b} \frac{\binom{n}{t}\binom{m}{a+b-t}}{\binom{n+m}{a+b}}.$$

Notice that Bucket is a decreasing function of b. Now, for a given value of δ, let $b_{\max}\left(n, \frac{a}{m}, \delta\right)$ be the largest value of b such that $\text{Bucket}(m,n,a,b) \geq \delta$. In other words, for any $b > b_{\max}\left(n, \frac{a}{m}, \delta\right)$, if we pick $a+b$ balls out of the $m+n$, the chance we get more than b blue balls is less than δ. We now give our main result.

Theorem 6 *(PAC-MDL bound) Let $P(\sigma)$ be any probability distribution over strings σ (a "prior"). With probability at least $1-\delta$ over the draw of the train and test sets $S, S' \sim D^{m+n}$:*

$$\forall\sigma \quad n\hat{e}(\sigma, S') \leq b_{\max}\left(n, \hat{e}(\sigma, S), P(\sigma)\delta\right).$$

Intuitively, this theorem statement can be thought of as "with high probability, the test set error rate is not too much larger than the the training error rate."

Proof. Assume for the moment that we have a fixed vector of labeled examples, $(x,y)^{m+n}$ so that we can use an argument similar to the "double sample trick" in VC bounds. Let $S, S' \sim \pi(m,n)$ denote a binary partition drawn uniformly from the set of $\binom{m+n}{n}$ possible binary partitions into sets of size m, n. For any particular string, σ, there is a total number of errors e_σ on the labeled data. What we wish to disallow is the possibility that most of these errors are in the test set. We know that:

$$\mathbf{Pr}_{S,S'\sim\pi(m,n)}(t \text{ test set errors and } e_\sigma - t \text{ train set errors}) = \frac{\binom{n}{t}\binom{m}{e_\sigma - t}}{\binom{m+n}{e_\sigma}}$$

In order to construct a confidence interval we must choose some set of "bad events" to exclude. In our case, this set of bad events consists of all test errors $t \geq \min\{b : \text{Bucket}(m,n,e_\sigma - b, b) < P(\sigma)\delta\}$. Notice that there is at most a $P(\sigma)\delta$ probability that the number of test errors is such that:

$$\text{Bucket}(m,n,m\hat{e}(\sigma,S), n\hat{e}(\sigma,S')) < P(\sigma)\delta.$$

Therefore, we have:

$$\forall (x,y)^{m+n} \forall \sigma \;\; \mathbf{Pr}_{S,S' \sim \pi(m,n)}(\mathrm{Bucket}(m,n,m\hat{e}(\sigma,S),n\hat{e}(\sigma,S')) < P(\sigma)\delta) \leq P(\sigma)\delta$$

(note that this is the step which requires that Bob not know which samples are in the train set and which are in the test set) The union bound implies:

$$\forall (x,y)^{m+n} \;\; \mathbf{Pr}_{S,S' \sim \pi(m,n)}(\exists \sigma : \; \mathrm{Bucket}(m,n,m\hat{e}(\sigma,S),n\hat{e}(\sigma,S')) < P(\sigma)\delta) \leq \delta$$

Taking the expectation over draws, $(x,y)^{m+n} \sim D^{m+n}$, we get:

$$E_{(x,y)^{m+n} \sim D^{m+n}} \mathbf{Pr}_{S,S' \sim \pi(m,n)}(\exists \sigma \;\; \mathrm{Bucket}(m,n,m\hat{e}(\sigma,S),n\hat{e}(\sigma,S')) < P(\sigma)\delta) \leq \delta$$

$$\Rightarrow \mathbf{Pr}_{(x,y)^{m+n} \sim D^{m+n}}(\exists \sigma \;\; \mathrm{Bucket}(m,n,m\hat{e}(\sigma,S),n\hat{e}(\sigma,S')) < P(\sigma)\delta) \leq \delta$$

Negating this statement, we get:

$$\Rightarrow \mathbf{Pr}_{(x,y)^{m+n} \sim D^{m+n}}(\forall \sigma \;\; \mathrm{Bucket}(m,n,m\hat{e}(\sigma,S),n\hat{e}(\sigma,S')) \geq P(\sigma)\delta) > 1-\delta$$

To construct a bound we must take a worst case over unknown quantities. In this case, the unknown quantity is the number of test errors, or equivalently the total number of errors. Taking the worst case over the number of test errors, we get the theorem. □

4.1 Lower bound

Here, we show that there exists no other bound which is a significant improvement on the PAC-MDL bound for the communication game we consider.

Theorem 7 *(PAC-MDL lower bound) For all "priors" over descriptions, $P(\sigma)$ there exists transductive classifiers σ and distributions D so that with probability* $\frac{\delta}{(n+1)(m+1)} - \frac{\delta^2}{(n+1)^2(m+1)^2}$

$$\exists \sigma : \quad n\hat{e}(\sigma,S') > b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right).$$

Furthermore, if any σ satisfies the clause, it is the σ with smallest train error.

Intuitively, this theorem says "the probabability that one of the transductive classifiers has a large test error rate is within a $(n+1)(m+1)$ fraction of δ".

Proof. Let D be the distribution which is uniform on any domain X and always chooses $Y=0$.

Pick any fixed train error rate e_{train}. For each possible string, σ, pick e_σ according to:

$$e_\sigma = e_{\text{train}} + b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right).$$

Define F_e to be the set of all functions from X^{m+n} to Y^{m+n} such that for every input we have exactly e 1's in the output. We now define σ to be a random element in F_{e_σ}. This choice implies that every description decodes to e_σ total errors. Since errors in the test set and errors in the train set have a constant total, any string with a too-large number of test errors has a number of train errors less than e_{train}.

We have:

$$\forall\sigma \ \mathbf{Pr}_{S,S'\sim D^{m+n}}\left(\hat{e}(\sigma,S') > b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right)\right)$$

$$\geq \text{Bucket}(m, n, e_{\text{train}} - 1, e_\sigma - e_{\text{train}} + 1)$$

Noting that $\binom{n}{k+1} > \frac{1}{n}\binom{n}{k}$ and $\binom{n}{k-1} > \frac{1}{n}\binom{n}{k}$ We get:

$$\forall\sigma \ \mathbf{Pr}_{S,S'\sim D^{m+n}}\left(\hat{e}(\sigma,S') > b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right)\right) \geq \frac{P(\sigma)\delta}{(n+1)(m+1)}$$

Negating, we get:

$$\forall\sigma \ \mathbf{Pr}_{S,S'\sim D^{m+n}}\left(\hat{e}(\sigma,S') \leq b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right)\right) \leq 1 - \frac{P(\sigma)\delta}{(n+1)(m+1)}$$

Using independence from the construction of the σ, we get:

$$\mathbf{Pr}_{S,S'\sim D^{m+n}}\left(\forall\sigma \ \hat{e}(\sigma,S') \leq b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right)\right) \leq \prod_{\sigma}(1 - \frac{P(\sigma)\delta}{(n+1)(m+1)})$$

Approximating the product, we get:

$$\mathbf{Pr}_{S,S'\sim D^{m+n}}\left(\forall\sigma \ \hat{e}(\sigma,S') \leq b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right)\right)$$

$$\leq 1 - (\frac{\delta}{(n+1)(m+1)} - \frac{\delta^2}{(n+1)^2(m+1)^2})$$

And negating again, we get:

$$\mathbf{Pr}_{S,S'\sim D^{m+n}}\left(\exists\sigma \ \hat{e}(\sigma,S') > b_{\max}\left(n, \hat{e}(\sigma,S), P(\sigma)\delta\right)\right)$$

$$\geq \frac{\delta}{(n+1)(m+1)} - \frac{\delta^2}{(n+1)^2(m+1)^2}$$

□

5 Comparison to standard bounds: quantitative results

The goal of this section is making a quantitatively tight comparison of the PAC-MDL bound to various other bounds in the inductive setting where other bounds are typically stated. For a fair quantitative comparison, we state bounds in their tightest forms, and for this we work directly with the Binomial tail and its inverse.

Definition 8 *(Binomial Tail Distribution)*

$$Bin\left(\frac{k}{m}, p\right) \equiv \mathbf{Pr}_{X_1,\ldots X_m \sim p^m}\left(\sum_{i=1}^{m} X_i \leq k\right) = \sum_{j=1}^{k} \binom{m}{j} p^j (1-p)^{m-j}$$

= *the probability a given classifier of true error p has empirical error at most* $\frac{k}{m}$ *on a sample of size* m.

Since we are interested in calculating a bound on the true error rate given a confidence, δ, and an empirical error, $\hat{e}_S(c)$, it is handy to define the inversion of a Binomial tail.

Definition 9 *(Binomial Tail Inversion)*

$$\overline{Bin}\left(\frac{k}{m}, \delta\right) \equiv \max_{p}\left\{p : \; Bin\left(\frac{k}{m}, p\right) = \delta\right\}$$

= *the largest true error rate such that the probability of having empirical error* $\leq \frac{k}{m}$ *is at least* δ.

We next state each of the bounds in their tightest form and show how each bound is related to the description complexity of the labels given the unlabeled data.

5.1 The Occam's Razor bound

The Occam's Razor Bound applies to any measure P over a set of classifiers, c

Theorem 10 *[1] (Occam's Razor Bound) For all "priors"* $p(c)$ *over the classifiers, c, for all* $\delta \in (0, 1]$*:*

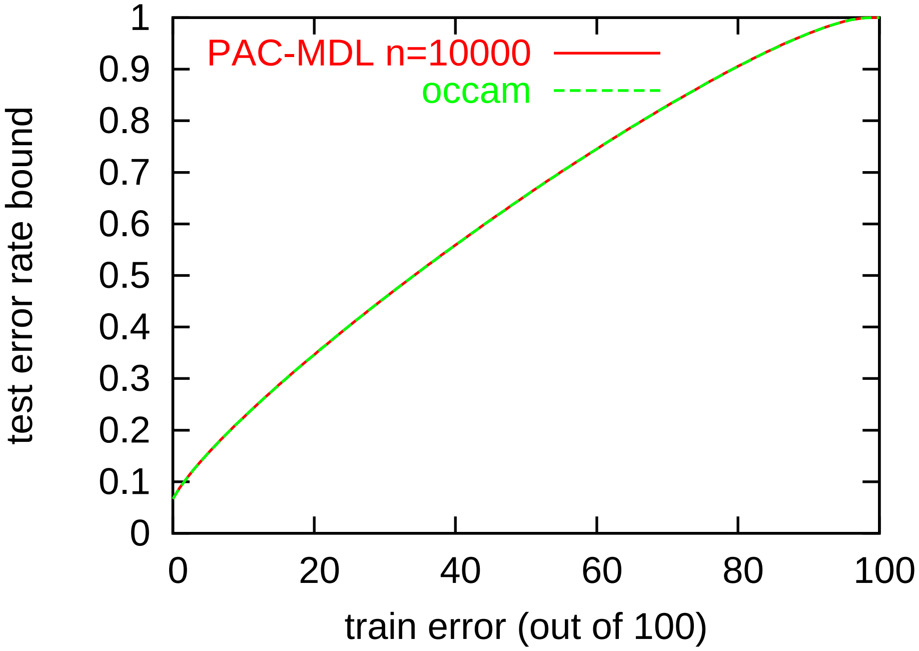
$$\forall p(c) \; \mathbf{Pr}_{S \sim D^m}\left(\exists c : \; e(c) \geq \overline{Bin}\left(\hat{e}_S(c), \delta p(c)\right)\right) \leq \delta$$

The prior $p(c)$ can be interpreted as a language for describing classifiers, c. Given any prior $p(c)$, we can construct a language $L(c)$ which uses code words of length approximately $2^{-|L(c)|} = p(c)$.

The appropriate method for constructing the Occam's Razor bound from the PAC-MDL bound uses the limit as $n \to \infty$, similar to Corollary 2 (although really using Theorem 6 in order to cover the agnostic case).

Using the PAC-MDL bound, we can construct an inductive bound on the true error rate given any "prior" on classifiers, $p(c)$. How does this compare with the bound constructed with the Occam's Razor bound directly?

For the agnostic case, we can do a numerical calculation with $m_{\text{train}} = 100$ training examples, a confidence of $\delta p(c) = 0.001$, a near-infinite (size 10000) test set, and a varying training error. Then, the two true error bounds yield the following graph:



Quantitatively, the Occam's Razor bound is essentially a specialization of the PAC-MDL bound where the description language is given by the prior on the classifiers.

5.2 The Compression Bound

The compression bound works from the observation that examples which do not affect the output hypothesis are "sort of" test examples. In particular, if we knew in advance which examples are unnecessary, then the "unnecessary" examples would be independent of the chosen hypothesis and a test set bound would apply. We don't know in advance, so it is necessary to worsen the results by some factor.

Let $|A(S)|$ be the number of examples used by the learning algorithm and $\bar{A}(S)$ be the set of examples not used by the learning algorithm.

Theorem 11 *(Compression Bound) [6][3]*

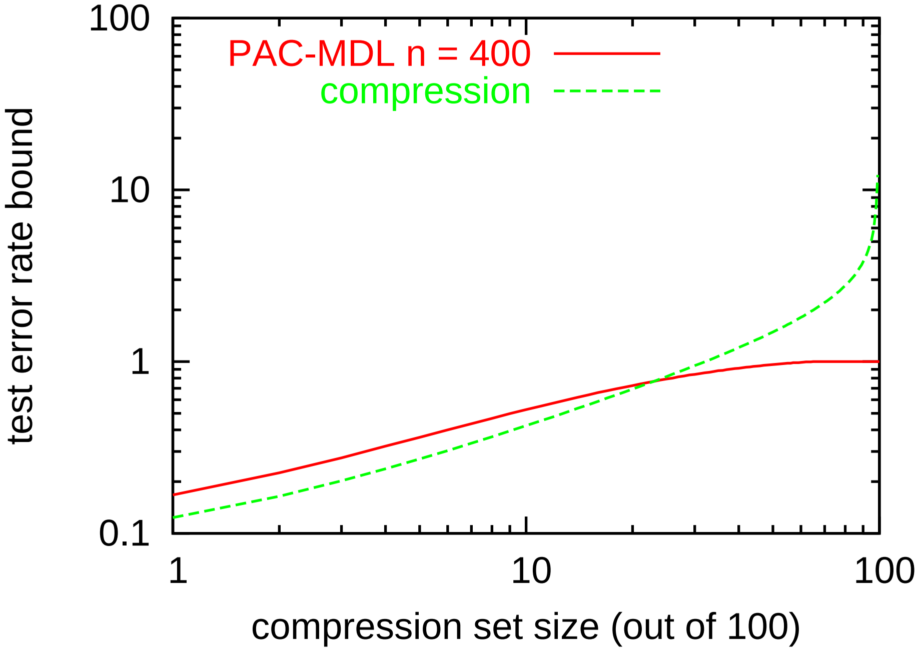
$$\mathbf{Pr}_{S\sim D^m}\left(e(c) \geq \overline{Bin}\left(\hat{e}_{\bar{A}(S)}(c), \frac{\delta}{\binom{m}{|A(S)|}(m+1)}\right)\right) \leq \delta$$

(Note: we have improved the compression bound here to work in the agnostic setting)

The language here is quite simple. First, specify the size of the compression set using $\log(m+1)$ bits, then specify the compression set using $\log \binom{m}{|A(S)|}$ bits, and then specify the labels using $|A(S)|$ bits. Given the labels of the compression set, the decoding end can run the learning algorithm and produce labels for all of the examples.

The compression bound is similar (but not identical) to the PAC-MDL bound with a specific language: the language which states the labels of the critical subset of $|A(S)|$ labels. Given the critical subset, it is possible to learn the hypothesis, and given the hypothesis, a labeling of all the data exists.

The number of bits required to specify the critical subset of labels is $\log_2(m+n+1)$ (to specify the size of the subset) plus $\log_2 \binom{m+n}{|A(S)|}$ (to specify the particular subset) plus $|A(S)|$ (to specify the labels of the subset). We compare these two bounds (with $\delta = 0.05$ on a training set of size 100) and find the following graph:



The PAC-MDL bound does somewhat worse because we are forced to choose from a larger set, of size $m+n$ rather than n.

6 Discussion

There are a few things we have accomplished here.

1. Bound unification. All of the major bounds can be thought of as (approximate or exact) applications of the PAC-MDL bound.
2. Generalization. We have a better understanding of the transductive setting since the PAC-MDL bound is naturally transductive. This allows us to bound the error rate on a given test set in many situations where it might be much more difficult to state an inductive bound.
3. Simplification. Proving bounds becomes an exercise in showing that we have some language for labels.

The qualitative and quantitative comparisons with the inductive bounds were made in the inductive setting where a "home court advantage" exists for naturally inductive bounds. If instead, we compare in a transductive setting, with a specific train and test set, the advantage shifts towards the naturally transductive PAC-MDL bound.

Information theory has a few implications for the PAC-MDL bound. If we let $H(Y|X)$ be the conditional entropy of the label given the unlabeled data, we know that asymptotically $H(Y|X)$ bits per label are required transfer the labels. Thus, for the realizable case, we can (asymptotically) find a code where $|\sigma|/(m+n) = H(Y|X)$.

It is worth noting that there are a few effects which are *not* captured by the PAC-MDL bound. Results which depend upon the distribution of observed empirical errors, such as shell bounds, are not captured. It may be possible to create an PAC-MDL bound which handles this as well, but that is an item for future work.

References

1. A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth. "Occam's Razor." Information Processing Letters 24: 377-380, 1987.
2. Thomas Cover and Joy Thomas, "Elements of Information Theory" Wiley, New York 1991.
3. Sally Floyd and Manfred Warmuth, "Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension", Machine Learning, Vol.21 (3), pp. 269–304.
4. G. E. Hinton and D. van Camp, "Keeping neural networks simple by minimizing the description length of the weights", COLT 1993.
5. John Langford "Quantitatively Tight Sample Complexity Bounds", Carnegie Mellon Thesis, 2002.
6. Nick Littlestone and Manfred Warmuth, "Relating Data Compression and Learnability", Unpublished manuscript. June 10, 1986.
7. David McAllester, "PAC-Bayesian Model Averaging" COLT 1999.
8. V. N. Vapnik and A. Y. Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities." Theory of Probab. and its Applications, 16(2):264-280, 1971.
9. Vladimir N. Vapnik, "Statistical Learning Theory", Wiley, December 1999.