

DATA COMPRESSION & LEARNING

AVRIM BLUM AND JOHN LANGFORD

ABSTRACT. Every bound on classification errors for supervised learning in the PAC setting has a tightness related to the communication complexity of the labels given the unlabeled data. By an alteration to the transductive setting, we construct a single bound from which the results of all other bounds are derivable as particular choices of communication language.

1. INTRODUCTION

One of the most basic results about learning in the PAC model is the “Occam’s razor” statement which says, roughly, that if we can explain the labels of a set of m training examples by a hypothesis of size (in bits) of only $k \ll m$, then we can be confident that this hypothesis generalizes well to future data. This bound is intrinsically related to the communication complexity of the labels given the unlabeled data. Suppose Alice and Bob are each given the m unlabeled examples, but only Alice is given the m labels. The number of bits Alice must use to specify the hypothesis to Bob is the communication complexity of the labels given the unlabeled data. This pattern is typical. All common bounds on the true error rate of a learned hypothesis can be described in terms of a communication complexity.

What if we allow more general forms of data compression? In particular, a general compression-decompression procedure for this communication game 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. That is, instead of being a function from X to Y (which is then run m times by the receiver) a string could be viewed as an arbitrary function from X^m to Y^m .

Here, we provide PAC confidence bounds for this more general form of compression as well. This more general notion of compressibility includes VC bounds, the PAC-Bayes bound, the Occam’s Razor bound, and the Compression bound, as a special case. The proof of the result here, in fact, follows very much along the lines of the standard VC-bound argument with its double-sample trick.

The theory of this more general form of compression is stated purely in terms of observable quantities. In particular, no notion of a “true error rate” is required. This alteration makes the theory more adaptable since it can be reinterpreted to apply to nonindependent data.

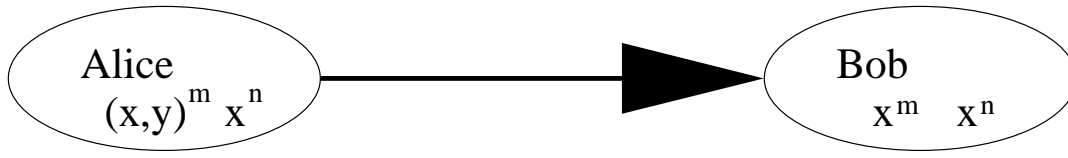
2. DEFINITIONS

We are considering the standard PAC setting. Examples are drawn from an instance space X and labeled by some (possibly randomized) target function $f : X \rightarrow Y$. We can also think of this as drawing labeled examples $z = (x, y)$ from a probability distribution D over $Z = X \times Y$, but sometimes it is convenient to separate out the two parts.

We consider a compression game where Alice wants to send Bob some information using as few bits as possible. Alice has available a training set S_{train} consisting of $|S_{\text{train}}| = m$ labeled examples drawn independently from D . Alice also has a test set of $|S_{\text{test}}| = n$ unlabeled examples drawn independently from D . Bob has available just the unlabeled versions of the test set and the train set. Pictorially, the available information is:

Date: March 5, 2003.

At least, every bound we have closely examined.



Alice's goal is to communicate labels to Bob using as few bits as possible. Alice encodes with a function $A : (X \times Y)^m \times X^n \rightarrow \{0, 1\}^*$ and Bob decodes with a function $B : X^{m+n} \times \{0, 1\}^* \rightarrow Y^{m+n}$. Note that the encode/decode process may be lossy. Given any compression/decompression procedure (A, B) , we can view the transmitted string σ as representing a function $\sigma : X^{m+n} \rightarrow Y^{m+n}$. 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} , $|A(S_{\text{train}}, S_{\text{test}})|$ is the number of bits sent to Bob in order to label the train and test sets. Let $\hat{e}(\sigma, S_{\text{test}}) = \frac{1}{n} \sum_{(x,y) \in S_{\text{test}}} I(y_\sigma \neq y)$ 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 train set.

3. MAIN RESULT

3.1. Simple Implications. Our first statement is an easy approximation of the deeper result. It is meant to provide some intuition.

Corollary 3.1. *(The approximate realizable case) For all encoding algorithms, A :*

$$\Pr_{S, S' \sim D^{m+n}} \left(\hat{e}(A(S, S'), S) = 0 \wedge \hat{e}(A(S, S'), S') \leq \frac{|A(S, S')| \ln 2 + \ln \frac{1}{\delta}}{n \ln \left(1 + \frac{m}{n}\right)} \right) > 1 - \delta$$

The corollary has a bound which is linear in the communication complexity, $|A(S, S')|$. The exact bound is controlled by the size of the sample set m , test set n , and the required confidence, δ .

Proof. Let $\sigma = A(S, S')$. Apply the Observable bound theorem for the realizable ($\hat{e}(\sigma, S) = 0$) case to get:

$$\begin{aligned} & \frac{\binom{n}{n\hat{e}(\sigma, S')}}{\binom{m+n}{n\hat{e}(\sigma, S')}} 2^{|\sigma|} \leq \delta \\ \Rightarrow & \frac{n!(m+n-n\hat{e}(\sigma, S'))!}{(n-n\hat{e}(\sigma, S'))!(m+n)!} 2^{|\sigma|} \leq \delta \\ \Rightarrow & \frac{n(n-1)\dots(n-n\hat{e}(\sigma, S')+1)}{(m+n)(m+n-1)\dots(m+n-n\hat{e}(\sigma, S')+1)} 2^{|\sigma|} \leq \delta \end{aligned}$$

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

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

Taking the \ln of both sides, we get:

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

□

3.2. Main Theorem. The approximate bound is a consequence of the following observable bound.

Theorem 3.2. (*Observable bound*) For all encoding algorithms, $A(S, S')$:

$$\Pr_{S, S' \sim D^{m+n}} \left(\frac{\binom{n}{n\hat{e}(A(S, S'), S')} \binom{m}{m\hat{e}(A(S, S'), S)}}{\binom{n+m}{n\hat{e}(A(S, S'), S') + m\hat{e}(A(S, S'), S)}} 2^{|A(S, S')|} \geq \delta \right) > 1 - \delta$$

Proof. Assume for the moment that we have a fixed vector of labeled examples, $(x, y)^{m+n}$. 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 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:

$$\Pr_{S, S' \sim \pi(m, n)}(t \text{ errors in test set and } e - t \text{ errors in train set} | e \text{ errors}) = \frac{\binom{n}{t} \binom{m}{e-t}}{\binom{m+n}{e}}$$

Set $\phi(\sigma, S, S') = \text{true}$ if the total number of errors is $e \equiv m\hat{e}(\sigma, S) + n\hat{e}(\sigma, S')$ and the number of test errors $t \equiv n\hat{e}(\sigma, S')$ satisfies:

$$\frac{\binom{n}{t} \binom{m}{e-t}}{\binom{m+n}{e}} < 2^{-|\sigma|} \delta$$

We have:

$$\forall (x, y)^{m+n} \forall \sigma \Pr_{S, S' \sim \pi(m, n)}(\phi(\sigma, S, S')) \leq 2^{-|\sigma|} \delta$$

The union bound implies:

$$\forall (x, y)^{m+n} \Pr_{S, S' \sim \pi(m, n)}(\exists \sigma : \phi(\sigma, S, S')) \leq \delta$$

Let A be some algorithm generating any string σ dependent upon the train and test sets S and S' . We have:

$$\forall A \forall (x, y)^{m+n} \Pr_{S, S' \sim \pi(m, n)}(\phi(A(S, S'), S, S')) \leq \delta$$

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

$$\forall A E_{(x, y)^{m+n} \sim D^{m+n}} \Pr_{S, S' \sim \pi(m, n)}(\phi(A(S, S'), S, S')) \leq \delta$$

$$\Rightarrow \forall A \Pr_{(x, y)^{m+n} \sim D^{m+n}}(\phi(A(S, S'), S, S')) \leq \delta$$

Negating this, we know that for every algorithm A with probability $1 - \delta$ we have:

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

Since $\hat{e}(A(S, S'), S')$ is the only unknown this is an implicit constraint on the value of $\hat{e}(A(S, S'), S')$, the error rate on the test set. \square

4. RELATIONSHIP WITH OTHER INDUCTIVE BOUNDS

4.1. Transduction \rightarrow Induction. Most bounds on the error rate in learning theory are inductive bounds. Any transductive bound can be turned into an inductive bound by using some fixed size of the test set and noting that the any learned hypothesis is independent of the hypothesis. One particularly interesting choice of “fixed size” is the limit as $n \rightarrow \infty$.

Corollary 4.1. (*Realizable Inductive bound*) For all encoding algorithms, A , in the limit as the size of the test set approaches ∞ , we have:

$$\lim_{n \rightarrow \infty} \Pr_{S, S' \sim D^{m+n}} \left(\hat{e}(A(S, S'), S) = 0 \wedge \hat{e}(A(S, S'), S') \leq \frac{|A(S, S')| \ln 2 + \ln \frac{1}{\delta}}{m} \right) > 1 - \delta$$

Proof. Apply the asymptotically tight approximation, $n \ln \left(1 + \frac{m}{n}\right) = n \frac{m}{n} = m$ to corollary 3.1. \square

This result is handy in making direct comparisons to inductive bounds.

4.2. The Occam's Razor bound. For any fixed classifier c there is a probability $p \equiv e_D(c)$ of making an error when an example is drawn from D . What is the probability of observing k heads out of m coin flips? This is the Binomial and so naturally many bounds presented are fundamentally dependent upon the cumulative distribution of a Binomial.

Definition 4.2. (Binomial Tail Distribution)

$$\text{Bin} \left(\frac{k}{m}, p \right) \equiv \Pr_{X_1, \dots, X_m \sim p^m} \left(\sum_{i=1}^m X_i \leq k | p \right) = \sum_{j=1}^k \binom{m}{j} p^j (1-p)^{m-j}$$

= the probability that m coins with bias p produce k or fewer heads.

For the learning problem, we always choose $p \equiv e_D(c)$ and $X_i = \text{error on the } i\text{th example}$. With these definitions, we can interpret the Binomial tail as the probability of an empirical error less than or equal to $\frac{k}{m}$.

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

Definition 4.3. (Binomial Tail Inversion)

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

= the largest true error rate such that the probability of observing $\frac{k}{m}$ or fewer "heads" is at least δ .

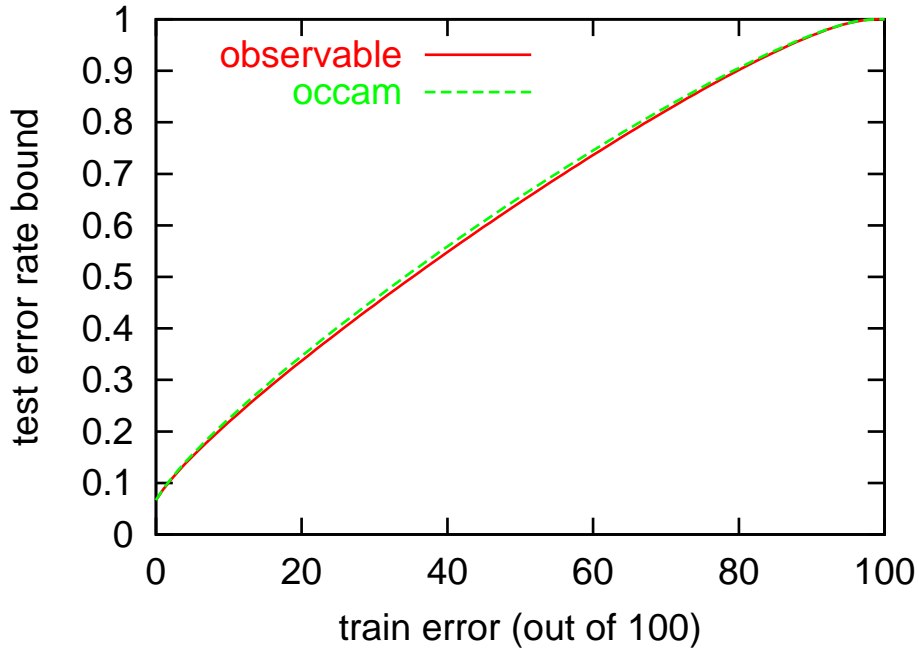
The Occam's Razor Bound applies to any measure P over a set of classifiers, c

Theorem 4.4. [1] (*Occam's Razor Bound*) For all "priors" $p(c)$ over the classifiers, c , for all $\delta \in (0, 1]$:

$$\forall p(c) \Pr_{S \sim D^m} (\exists c : e(c) \geq \overline{\text{Bin}}(\hat{e}_S(c), \delta p(c))) \leq \delta$$

The Occam's Razor bound is similar to the observable bound with a specific language for labels: namely, the language which consists of a choice of hypothesis. In fact, in the realizable case, they are exactly the same bound.

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



In conclusion, the Occam’s Razor bound is essentially a specialization of the observable bound where the description language is given by the prior on the hypotheses.

4.3. The PAC-Bayes bound. The PAC-Bayes bound is similar to the Occam’s Razor bound, except that it works with a “posterior” as well as a “prior” over hypotheses.

Theorem 4.5. (*PAC-Bayes bound*) [6] $\forall P \sim c$

$$\Pr_{S \sim D^m} \left(\exists Q \sim c \text{ } KL(E_{c \sim Q} \hat{e}(c) \| E_{c \sim Q} e(c)) \geq \frac{KL(Q \| P) + \ln \frac{m+1}{\delta}}{m} \right) \leq \delta$$

The PAC-Bayes bound is generally superior to the Occam’s Razor bound because redundant classifiers do not necessarily worsen the bound.

For any Q , the “bits back” [3] argument is sufficient to show that the communication complexity of the labels is only $KL(Q \| P)$. For specific choices of Q , we can show this result directly. Consider $Q(K)$ satisfying $\min_Q KL(Q \| P)$ subject to $E_{c \sim Q} \hat{e}(c) = K$.

4.3.1. The realizable case. In the realizable case, the optimal distribution Q is $Q(c) = \frac{P(c)}{\int P(c) I(\hat{e}(c)=0) dc}$ which implies $KL(Q \| P) = \ln \frac{1}{\int P(c) I(\hat{e}(c)=0) dc}$. This is the description length of the labels given the “prior” P induces on labelings. In particular, let:

$$R(y^m) = \int_c P(c) \prod_i I(c(x_i) = y_i) dc$$

The description language is simply given by $R(y^m)$, and the analysis is identical to the Occam’s Razor bound. It is interesting to note that we improve on the PAC-Bayes bound by an additive constant of $\frac{\ln(m+1)}{m}$.

4.3.2. The agnostic case. For the unrealizable setting, suppose that our average satisfies $E_{c \sim Q} \hat{e}(c) = K$. The distribution minimizing $KL(Q \| P)$ has the form: $Q^*(c) = \frac{P(c) e^{-\beta \hat{e}(c)}}{Z(\beta)}$ for some constant β . Markov’s inequality tells us: $\Pr_{c \sim Q} (\hat{e}(c) \leq K + \frac{1}{m}) \geq \frac{1}{K+1}$. Let $Q_{\leq K}(c) = \frac{1}{N} Q^*(c)$ if $\hat{e}(c) \leq K + \frac{1}{m}$ and 0 otherwise. Then, $KL(Q_{\leq K} \| P) \leq KL(Q^* \| P) + \ln(K+1)$.

This argument can also be made by using the distribution: $Q(c) = \frac{P(c)}{\int P(c) I(\hat{e}(c) \leq K) dc}$ for $\hat{e}(c) \leq K$ and 0 otherwise. The KL divergence is then: $\ln \frac{1}{\int P(c) I(\hat{e}(c) \leq K) dc}$ = the prior probability of a labeling with K

or fewer errors. We have $\ln \frac{1}{\int P(c) I(\hat{e}(c) \leq K) dc} \leq \text{KL}(Q^* \| P) + \ln(K + 1)$ so, we can send a codeword σ with length bounded by $\text{KL}(Q^* \| P) + \ln(K + 1)$ having less than K errors.

4.4. 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 would not be necessary, then the “unnecessary” examples would be independent of the hypothesis chosen. We don’t know in advance, so it is necessary to worsen our 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.

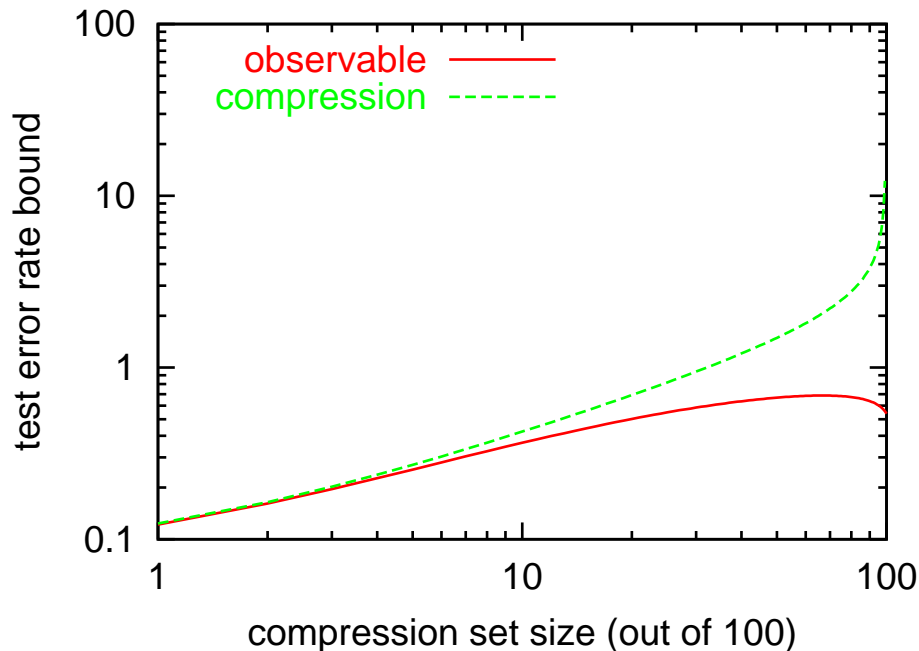
Theorem 4.6. (*Compression Bound*) [5][2]

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

Given the compression bound, a bound can be found on the number of test errors using the test error bound lemma.

The compression bound is similar to the observable 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, all labels are transferred.

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



Once again, the observable bound does better. One odd artifact appears here and is related to the choice of language: the bound actually improves as the compression set size increases beyond a certain point. This is “real” and is due to the fact that $\binom{m}{|A(S)|}$ decreases as $|A(S)|$ increases.

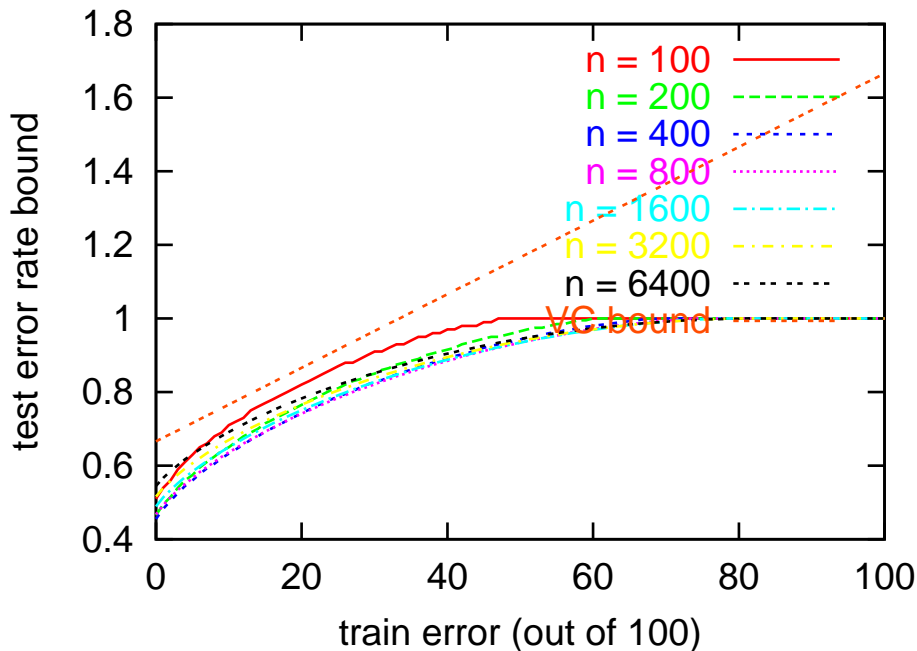
4.5. VC bounds. A typical VC [7] result has the form:

$$\Pr_{S \sim D^m} \left(\exists c \in H \quad e(c) \leq \hat{e}_S(c) + \sqrt{\frac{d \left(1 + \ln \frac{2m}{d}\right) + \ln \frac{4}{\delta}}{m}} \right) > 1 - \delta$$

VC bounds are unique amongst the bounds we compare with here because the encoding of the labels $|A(S, S')|$ grows with S' . Let $\Pi(m+n)$ be the number of shatterings of the m training points and n test points. Using this in the observable bound, we get:

$$\Pr_{S, S' \sim D^{m+n}} \left(\frac{\binom{n}{n\hat{e}(A(S, S'), S')} \binom{m}{m\hat{e}(A(S, S'), S)}}{\binom{n+m}{n\hat{e}(A(S, S'), S') + m\hat{e}(A(S, S'), S)}} \Pi(m+n) \geq \delta \right) > 1 - \delta$$

This quantity does not necessarily result in a bound for arbitrarily large values of n . However, since the choice of hypothesis depends upon only the m examples in the train set, we can construct an inductive bound by using the transductive bound for some finite value of n . One obvious choice is $n = m$, as suggested by the VC proof, although this is not necessarily optimal. Picking various values of n , and using Sauer's Lemma ($\Pi(m+n) \leq (e^{\frac{m+n}{d}})^d$), we can apply the observable bound and compare it to the VC result.



Even for the naive choice, $n = 100$, the observable bound provides a better bound on the true error rate of the learned classifier than this form of the VC bound. There are actually many forms of the VC bound, all of which are dependent in some way upon the growth function $\Pi(m+n)$. The tightest possible bound we can construct using the observable bound would use the distribution of the growth function, $\Pr_{S' \sim D^n} (\Pi(m+n))$ at some value of n (probably not m). This is similar to (but not the same as) the “VC-entropy” in [8].

5. DISCUSSION

There are a few things we have accomplished here.

- (1) Bound unification. All of the major bounds can be thought of as applications of the observable bound.
- (2) Generalization. We can now bound the error of (for example) hypotheses drawn from a hypothesis space with infinite VC dimension when the distribution of the data turns out to be fortuitous.

- (3) Simplification. Proving bounds has become an exercise in showing that we have some language for labels.
- (4) Tightening. Several of the inductive bounds are tightened (in application) by the argument via the transductive bound.

It is worth noting that there are a few effects which are *not* captured by the observable 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 observable 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] Sally Floyd and Manfred Warmuth, "Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension", Machine Learning, Vol.21 (3), pp. 269-304.
- [3] G. E. Hinton and D. van Camp, "Keeping neural networks simple by minimizing the description length of the weights", COLT 1993.
- [4] John Langford "Quantitatively Tight Sample Complexity Bounds", Carnegie Mellon Thesis, 2002.
- [5] Nick Littlestone and Manfred Warmuth, "Relating Data Compression and Learnability", Unpublished.
- [6] David McAllester, "PAC-Bayesian Model Averaging" COLT 1999.
- [7] 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.
- [8] Vladimir N. Vapnik, "Statistical Learning Theory", Wiley, December 1999.