Journal of Machine Learning Research 5 (2004) 499-517 Submitted 1/03; Revised 8/03; Published 5/04

Computable Shell Decomposition Bounds

John Langford JL@TTI-C.ORG
David McAllester MCALLESTER@TTI-C.ORG
Toyota Technology Institute at Chicago
1427 East 60th Street
Chicago, IL 60637, USA

Editor: Manfred Warmuth

Abstract

Haussler, Kearns, Seung and Tishby introduced the notion of a shell decomposition of the union bound as a means of understanding certain empirical phenomena in learning curves such as phase transitions. Here we use a variant of their ideas to derive an upper bound on the generalization error of a hypothesis computable from its training error and the histogram of training errors for the hypotheses in the class. In most cases this new bound is significantly tighter than traditional bounds computed from the training error and the cardinality of the class. Our results can also be viewed as providing a rigorous foundation for a model selection algorithm proposed by Scheffer and Joachims.

Keywords: Sample complexity, classification, true error bounds, shell bounds

1. Introduction

For an arbitrary finite hypothesis class we consider the hypothesis of minimal training error. We give a new upper bound on the generalization error of this hypothesis computable from the training error of the hypothesis and the histogram of the training errors of the other hypotheses in the class. This new bound is typically much tighter than more traditional upper bounds computed from the training error and cardinality of the class.

As a simple example, suppose that we observe that all but one empirical error in a hypothesis space is $1/2$ and one empirical error is 0. Furthermore, suppose that the sample size is large enough (relative to the size of the hypothesis class) that with high confidence we have that, for all hypotheses in the class, the true (generalization) error of a hypothesis is within $1/5$ of its training error. This implies, that with high confidence, hypotheses with training error near $1/2$ have true error in $[3/10, 7/10]$. Intuitively, we would expect the true error of the hypothesis with minimum empirical error to be very near to 0 rather than simply less than $1/5$ because none of the hypotheses which produced an empirical error of $1/2$ could have a true error close enough to 0 that there exists a significant probability of producing 0 empirical error. The bound presented here validates this intuition. We show that you can ignore hypotheses with training error near $1/2$ in calculating an "effective size" of the class for hypotheses with training error near 0. This new effective class size allows us to calculate a tighter bound on the difference between training error and true error for hypotheses with training error near 0. The new bound is proved using a distribution-dependent application of the union bound similar in spirit to the shell decomposition introduced by Haussler, Kearns, Seung and Tishby (5).

We actually give two upper bounds on generalization error — an uncomputable bound and a computable bound. The uncomputable bound is a function of the unknown distribution of true error rates of the hypotheses in the class. The computable bound is, essentially, the uncomputable bound with the unknown distribution of true errors replaced by the known histogram of training errors. Our main contribution is that this replacement is sound, i.e., the computable version remains, with high confidence, an upper bound on generalization error.

When considering asymptotic properties of learning theory bounds it is important to take limits in which the cardinality (or VC dimension) of the hypothesis class is allowed to grow with the size of the sample. In practice, more data typically justifies a larger hypothesis class. For example, the size of a decision tree is generally proportional the amount of training data available. Here we analyze the asymptotic properties of our bounds by considering an infinite sequence of hypothesis classes $\mathcal{H}_m$, one for each sample size m, such that $\frac{\ln|\mathcal{H}_m|}{m}$ approaches a limit larger than zero. This kind of asymptotic analysis provides a clear account of the improvement achieved by bounds that are functions of error rate distributions rather than simply the size (or VC dimension) of the class.

We give a lower bound on generalization error showing that the uncomputable upper bound is asymptotically as tight as possible — any upper bound on generalization error given as a function of the unknown distribution of true error rates must asymptotically be greater than or equal to our uncomputable upper bound. Our lower bound on generalization error also shows that there is essentially no loss in working with an upper bound computed from the true error distribution rather than expectations computed from this distribution as used by Scheffer and Joachims (12).

Asymptotically, the computable bound is simply the uncomputable bound with the unknown distribution of true errors replaced with the observed histogram of training errors. Unfortunately, we can show that in limits where $\frac{\ln|\mathcal{H}_m|}{m}$ converges to a value greater than zero, the histogram of training errors need not converge to the distribution of true errors — the histogram of training errors is a "smeared out" version of the distribution of true errors. This smearing loosens the bound even in the large-sample asymptotic limit. We give a precise asymptotic characterization of this smearing effect for the case where distinct hypotheses have independent training errors. In spite of the divergence between these bounds, the computable bound is still significantly tighter than classical bounds not involving error distributions.

The computable bound can be used for model selection. In the case of model selection we can assume an infinite sequence of finite model classes $\mathcal{H}_0, \mathcal{H}_1, \ldots$ where each $\mathcal{H}_j$ is a finite class with $\ln|\mathcal{H}_j|$ growing linearly in j. To perform model selection we find the hypothesis of minimal training error in each class and use the computable bound to bound its generalization error. We can then select, among these, the model with the smallest upper bound on generalization error. Scheffer and Joachims propose (without formal justification) replacing the distribution of true errors with the histogram of training errors. Under this replacement, the model selection algorithm based on our computable upper bound is asymptotically identical to the algorithm proposed by Scheffer and Joachims.

The shell decomposition is a distribution-dependent use of the union bound. Distribution-dependent uses of the union bound have been previously exploited in so-called self-bounding algorithms. Freund (4) defines, for a given learning algorithm *and data distribution*, a set S of hypotheses such that with high probability over the sample, the algorithm always returns a hypothesis in that set. Although S is defined in terms of the unknown data distribution, Freund gives a way of computing a set S' from the given algorithm and the sample such that, with high confidence, S' contains S and hence the "effective size" of the hypothesis class is bounded by $|S'|$.

Langford and Blum (8) give a more practical version of this algorithm. Given an algorithm and data distribution they conceptually define a weighting over the possible executions of the algorithm. Although the data distribution is unknown, they give a way of computing a lower bound on the weight of the particular execution of the algorithm generated by the sample at hand. In this paper we consider distribution dependent union bounds defined independent of any particular learning algorithm.

The bounds given in this paper apply to finite concept classes. Of course more sophisticated measures of the complexity of a concept class, such as VC dimension or Rademacher complexity, are possible and can sometimes result in tighter bounds.

However, insight into finite classes remains useful in at least two ways. Finite class analysis is useful as a pedagogical tool, teaching about directions in which to look for the removal of slack from these more sophisticated bounds. Indeed, various localized Rademacher complexity results (1) and the "peeling" technique (13) appear to (roughly) correspond to the orthogonal combination of shell bounds and earlier Rademacher complexity results. One advantage of the shell bounds is the KL-divergence form of the bounds which smoothly interpolates between the linear bounds of the realizable case and the quadratic bounds of the unrealizable case. This realizable-unrealizable interpolation is orthogonal to the shell principle that concepts with large empirical error are unlikely to be confused with concepts with low error rate. The shell bound also supports intuitions that are difficult to achieve in more complex settings. For example, the simple shell bounds clearly exhibit phase transitions in the learning bound, something which does not appear to be well-elucidated for localalized Rademacher bounds. In summary, the simplicity of finite classes (and a shell bound analysis on a finite class) provides a clarity that is difficult to achieve with more complex structure-exploiting bounds.

Finite class analysis is also useful in a more practical sense. In practice a finite VC dimension class usually has a finite parameterization. Given that these real parameters are typically represented as 32 bit floating point numbers, the class becomes finite and the log of the class size is linear in the number of parameters. Since many of the more sophisticated infinite-class techniques are loose by large multiplicative constants, a finite class analysis applied to a VC class discretized to a small number of bits can actually yield *tighter* bounds as shown in Figure 1.

2. Mathematical Preliminaries

For an arbitrary measure on an arbitrary sample space we use the notation[1] $\forall^\delta S\ \Phi[S,\delta]$ to mean that with probability at least $1-\delta$ over the choice of the sample S we have that $\Phi[S,\delta]$ holds. In practice S is the training sample of a learning algorithm. Note that $\forall x\ \forall^\delta S\ \Phi[x,\ S,\ \delta]$ does not imply $\forall^\delta S\ \forall x\ \Phi[x,\ S,\ \delta]$. If X is a finite set, and for all $x \in X$ we have the assertion $\forall \delta > 0\ \forall^\delta S\ \Phi[S,x,\delta]$ then by a standard application of the union bound we have the assertion $\forall \delta > 0\ \forall^\delta S\ \forall x \in X\ \Phi[S,x,\frac{\delta}{|X|}]$. We call this the quantification rule. If $\forall \delta > 0\ \forall^\delta S\ \Phi[S,\delta]$ and $\forall \delta > 0\ \forall^\delta S\ \Psi[S,\ \delta]$ then by a standard application of the union bound we have $\forall \delta > 0\ \forall^\delta S\ \Phi[S,\frac{\delta}{2}] \wedge \Psi[S,\frac{\delta}{2}]$. We call this the conjunction rule.

The KL-divergence of p from q, denoted $D(q||p)$, is $q\ln(\frac{q}{p}) + (1-q)\ln(\frac{1-q}{1-p})$ with $0\ln(\frac{0}{p}) = 0$ and $q\ln(\frac{q}{0}) = \infty$. Let $\hat{p}$ be the fraction of heads in a sequence S of m tosses of a biased coin where

1. This can be read as "for all but δ sets S, the predicate $\Phi[S,\delta]$ holds" or "with probability $1-\delta$ over the draw of S, the predicate $\Phi[S,\delta]$ holds".

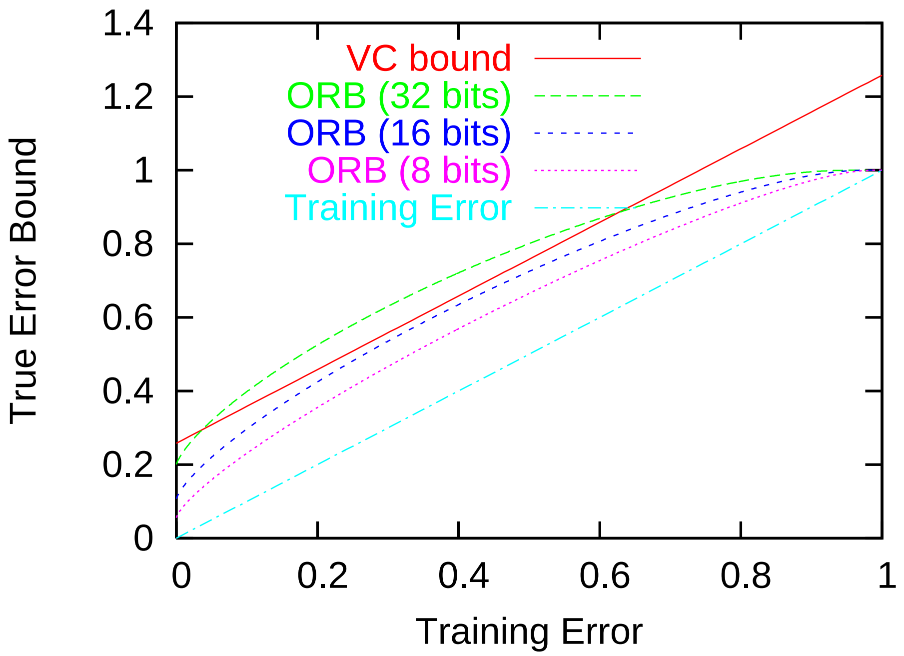

Figure 1: A graph comparing the (infinite hypothesis) VC bound to the finite hypothesis Occam's razor bound. For all curves we use VC dimension $d = 10$, bound failure probability, $\delta = 0.1$, and $m = 1000$ examples. For the VC bound calculation (see (Moore) for details) the formula is true error $\leq$ train error $+ \sqrt{(d \ln \frac{2m}{d} + \ln \frac{4}{\delta})/m}$. For the Occam's Razor Bound (ORB) (see (7) for details) calculation, we use a uniform distribution over the 2^{kd} discrete classifiers which might be representable when we discretize d parameters to $k = 8, 16, 32$ bits per dimension. The basic formula is: KL(train error||true error) $\leq$ $(kd \ln 2 + \ln \frac{1}{\delta})/m$. This graph is approximatly the same for any similar ratio of d/m with smaller values favoring the Occam's Razor Bound.

the probability of heads is p. We have the following inequality given by Chernoff in 1952 (2):

$$\forall q \in [p, 1]: \quad Pr(\hat{p} \geq q) \quad \leq \quad e^{-mD(q||p)}. \tag{1}$$

This bound can be rewritten as follows:

$$\forall \delta > 0 \; \forall^{\delta} S \quad D(\max(\hat{p}, p)||p) \leq \frac{\ln(\frac{1}{\delta})}{m}. \tag{2}$$

To derive (2) from (1) note that $Pr(D(\max(\hat{p}, p)||p) \geq \frac{\ln(\frac{1}{\delta})}{m})$ equals $Pr(\hat{p} \geq q)$ where $q \geq p$ and $D(q||p) = \frac{\ln(\frac{1}{\delta})}{m}$. By (1) we then have that this probability is no larger than $e^{-mD(q||p)} = \delta$. It is just as easy to derive (1) from (2) so the two statements are equivalent. By duality, i.e., by considering the problem defined by replacing p by $1 - p$, we get

$$\forall \delta > 0 \; \forall^{\delta} S \quad D(\min(\hat{p}, p)||p) \leq \frac{\ln(\frac{1}{\delta})}{m}. \tag{3}$$

Conjoining (2) and (3) yields the following corollary of (1):

$$\forall \delta > 0 \; \forall^{\delta} S \quad D(\hat{p}||p) \leq \frac{\ln(\frac{2}{\delta})}{m}. \tag{4}$$

Using the inequality $D(q||p) \geq 2(q - p)^2$ one can show that (4) implies the better known form of the Chernoff bound

$$\forall \delta > 0 \; \forall^{\delta} S \quad |p - \hat{p}| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}. \tag{5}$$

Using the inequality $D(q||p) \geq \frac{(p-q)^2}{2q}$, which holds for $q \leq p$, we can show that (3) implies the following:[2]

$$\forall \delta > 0 \; \forall^{\delta} S \quad p \leq \hat{p} + \sqrt{\frac{2\hat{p}\ln(\frac{1}{\delta})}{m}} + \frac{2\ln(\frac{1}{\delta})}{m}. \tag{6}$$

Note that for small values of $\hat{p}$ formula (6) gives a tighter upper bound on p than does (5). The upper bound on p implicit in (4) is somewhat tighter than the minimum of the bounds given by (5) and (6).

We now consider a formal setting for hypothesis learning. We assume a finite set $\mathcal{H}$ of hypotheses and a space $\mathcal{X}$ of instances. We assume that each hypothesis represents a function from $\mathcal{X}$ to $\{0, 1\}$ where we write $h(x)$ for the value of the function represented by hypothesis h when applied to instance x. We also assume a distribution D on pairs $\langle x, \; y\rangle$ with $x \in \mathcal{X}$ and $y \in \{0, 1\}$. For any hypothesis h we define the error rate of h, denoted $e(h)$, to be $P_{\langle x, \; y\rangle \sim D}(h(x) \neq y)$. For a given sample S of m pairs drawn from D we write $\hat{e}(h)$ to denote the fraction of the pairs $\langle x, \; y\rangle$ in S such that $h(x) \neq y$. Quantifying over $h \in \mathcal{H}$ in (4) yields the following second corollary of (1):

$$\forall^{\delta} S \quad \forall h \in \mathcal{H} \quad D(\hat{e}(h)||e(h)) \leq \frac{\ln|\mathcal{H}| + \ln(\frac{2}{\delta})}{m}. \tag{7}$$

2. A derivation of this formula can be found in (9) or (11). To see the need for the last term consider the case where $\hat{p} = 0$.

By considering bounds on $D(q||p)$ we can derive the more well known corollary of (7),

$$\forall^\delta S \quad \forall h \in \mathcal{H} \quad |e(h) - \hat{e}(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(\frac{2}{\delta})}{2m}}. \tag{8}$$

These two formulas both limit the distance between $\hat{e}(h)$ and $e(h)$. In this paper we work with (7) rather than (8) because it yields an (implicit) upper bound on generalization error that is optimal up to asymptotic equality.

3. The Upper Bound

Our goal now is to improve on (7). Our first step is to divide the hypotheses in $\mathcal{H}$ into m disjoint sets based on their true error rates. More specifically, for $p \in [0, 1]$ define $\lceil\lceil p \rceil\rceil$ to be $\frac{\max(1, \lceil mp \rceil)}{m}$. Note that $\lceil\lceil p \rceil\rceil$ is of the form $\frac{k}{m}$ where either $p = 0$ and $k = 1$ or $p > 0$ and $p \in (\frac{k-1}{m}, \frac{k}{m}]$. In either case we have $\lceil\lceil p \rceil\rceil \in \{\frac{1}{m}, \ldots, \frac{m}{m}\}$ and if $\lceil\lceil p \rceil\rceil = \frac{k}{m}$ then $p \in [\frac{k-1}{m}, \frac{k}{m}]$. Now we define $\mathcal{H}(\frac{k}{m})$ to be the set of $h \in \mathcal{H}$ such that $\lceil\lceil e(h) \rceil\rceil = \frac{k}{m}$. We define $s(\frac{k}{m})$ to be $\ln(\max(1, |\mathcal{H}(\frac{k}{m})|))$. We now have the following lemma.

Lemma 3.1 *With high probability over the draw of S, for all hypotheses, the deviation between the empirical error $\hat{e}(h)$, and true error $e(h)$, of every hypothesis is bounded by $s(\lceil\lceil e(h) \rceil\rceil)$. More precisely,*
$\forall \delta > 0 \; \forall^\delta S \quad \forall h \in \mathcal{H}$,

$$D(\hat{e}(h)||e(h)) \leq \frac{s(\lceil\lceil e(h) \rceil\rceil) + \ln(\frac{2m}{\delta})}{m}.$$

Proof Quantifying over $p \in \{\frac{1}{m}, \ldots, \frac{m}{m}\}$ and $h \in \mathcal{H}(p)$ in (4) gives $\forall \delta > 0$, $\forall^\delta S$, $\forall p \in \{\frac{1}{m}, \ldots, \frac{m}{m}\}$, $\forall h \in \mathcal{H}(p)$,

$$D(\hat{e}(h)||e(h)) \leq \frac{\ln s(p) + \ln(\frac{2m}{\delta})}{m}.$$

But this implies the lemma. ■

Lemma 3.1 imposes a constraint, and hence a bound, on $e(h)$. More specifically, we have the following where `lub` $\{x : \Phi[x]\}$ denotes the least upper bound (the maximum) of the set $\{x : \Phi[x]\}$:

$$e(h) \leq \texttt{lub}\,\{q : D(\hat{e}(h)||q) \leq \frac{s(\lceil\lceil q \rceil\rceil) + \ln(\frac{2m}{\delta})}{m}\}. \tag{9}$$

This is our uncomputable bound. It is uncomputable because the m numbers $s(\frac{1}{m})$, …, $s(\frac{m}{m})$ are unknown. Ignoring this problem, however, we can see that this bound is typically significantly tighter than (7). More specifically, we can rewrite (7) as

$$e(h) \leq \texttt{lub}\,\{q : D(\hat{e}(h)||q) \leq \frac{\ln|\mathcal{H}| + \ln(\frac{2}{\delta})}{m}\}. \tag{10}$$

Since $s(\frac{k}{m}) \leq \ln|\mathcal{H}|$, and since $\frac{\ln m}{m}$ is small for large m, we have that (9) is never significantly looser than (10). Now consider a hypothesis h such that the bound on $e(h)$ given by (7), or equivalently, (10), is significantly less than 1/2. Assuming m is large, the bound given by (9) must also be significantly less than 1/2. But for q significantly less than 1/2 we typically have that $s(\lceil\lceil q\rceil\rceil)$ is significantly smaller than $\ln|\mathcal{H}|$. For example, suppose $\mathcal{H}$ is the set of all decision trees of size $m/10$. For large m, a random decision tree of this size has error rate near 1/2. The set of decision trees with error rate significantly smaller than 1/2 is an exponentially small fraction of the set of all possible trees. So for q small compared to 1/2 we get that $s(\lceil\lceil q\rceil\rceil)$ is significantly smaller than $\ln|H|$. This makes the bound given by (9) significantly tighter than the bound given by (10).

We now show that the distribution of true errors can be replaced, essentially, by the histogram of training errors. We first introduce the following definitions:

$$\hat{\mathcal{H}}\left(\frac{k}{m},\delta\right) \equiv \left\{h \in \mathcal{H}: \ \left|\hat{e}(h)-\frac{k}{m}\right| \leq \frac{1}{m}+\sqrt{\frac{\ln(\frac{16m^2}{\delta})}{2m-1}}\right\},$$

$$\hat{s}\left(\frac{k}{m},\ \delta\right) \ \equiv \ \ln\left(\max\left(1,\ 2\left|\hat{\mathcal{H}}\left(\frac{k}{m},\ \delta\right)\right|\right)\right).$$

The definition of $\hat{s}\left(\frac{k}{m},\ \delta\right)$ is motivated by the following lemma.

Lemma 3.2 *With high probability over the draw of S, for all q, $s(q) \leq \hat{s}(q, 2\delta)$. More precisely,* $\forall\delta > 0$, $\forall^\delta S$, $\forall q \in \{\frac{1}{m},\ \ldots,\ \frac{m}{m}\}$,

$$s(q) \leq \hat{s}(q,\ 2\delta).$$

Before proving Lemma 3.2 we note that by conjoining (9) and Lemma 3.2 we get the following. This is our main result.

Theorem 3.3 *With high probability over the draw of S, for all hypotheses, the deviation between the empirical error $\hat{e}(h)$, and true error $e(h)$, of every hypothesis is bounded by $\hat{s}(\lceil\lceil q\rceil\rceil,\delta)$. More precisely,*
$\forall\delta > 0$, $\forall^\delta S$, $\forall h \in \mathcal{H}$,

$$e(h) \leq \texttt{lub}\left\{q : D(\hat{e}(h)||q) \leq \frac{\hat{s}(\lceil\lceil q\rceil\rceil,\ \delta)+\ln(\frac{4m}{\delta})}{m}\right\}.$$

As for Lemma 3.1, the bound implicit in Theorem 3.3 is typically significantly tighter than the bound in (7) or its equivalent form (10). The argument for the improved tightness of Theorem 3.3 over (10) is similar to the argument for (9). More specifically, consider a hypothesis h for which the bound in (10) is significantly less than 1/2. Since $\hat{s}(\lceil\lceil q\rceil\rceil,\ \delta) \leq \ln|\mathcal{H}|$, the set of values of q satisfying the condition in Theorem 3.3 must all be significantly less than 1/2. But for large m we have that $\sqrt{\frac{\ln(16m^2/\delta)}{2m-1}}$ is small. So if q is significantly less than 1/2 then all hypotheses in $\hat{\mathcal{H}}(\lceil\lceil q\rceil\rceil,\delta)$ have empirical error rates significantly less than 1/2. But for most hypothesis classes, e.g., decision trees, the set of hypotheses with empirical error rates far from 1/2 should be an exponentially small fraction of the class. Hence we get that $\hat{s}(\lceil\lceil q\rceil\rceil,\ \delta)$ is significantly less than $\ln|\mathcal{H}|$ and Theorem 3.3 is tighter than (10).

The remainder of this section is a proof of Lemma 3.2. Our departure point for the proof is the following lemma from (10).

Lemma 3.4 (McAllester 99) *For any measure on any hypothesis class we have the following where* $\mathtt{E}_h f(h)$ *denotes the expectation of* $f(h)$ *under the given measure on* h*:*

$$\forall \delta > 0\, \forall^\delta S \quad \mathtt{E}_h e^{(2m-1)(\hat{e}(h)-e(h))^2} \leq \frac{4m}{\delta}.$$

Intuitively, this lemma states that with high confidence over the choice of the sample most hypotheses have empirical error near their true error. This allows us to prove that $\hat{s}(\lceil\lceil q \rceil\rceil,\ \delta)$ bounds $s(\lceil\lceil q \rceil\rceil)$. More specifically, by considering the uniform distribution on $\mathcal{H}(\frac{k}{m})$, Lemma 3.4 implies

$$\begin{aligned}
\mathtt{E}_{h\sim\mathcal{H}(\frac{k}{m})}\left(e^{(2m-1)(\hat{e}(h)-e(h))^2}\right) &\leq \frac{4m}{\delta} \\
Pr_{h\sim\mathcal{H}(\frac{k}{m})}\left(e^{(2m-1)(\hat{e}(h)-e(h))^2} \geq \frac{8m}{\delta}\right) &\leq \frac{1}{2} \\
Pr_{h\sim\mathcal{H}(\frac{k}{m})}\left(e^{(2m-1)(\hat{e}(h)-e(h))^2} \leq \frac{8m}{\delta}\right) &\geq \frac{1}{2} \\
\left|\left\{h \in \mathcal{H}(\frac{k}{m}) :\ |\hat{e}(h) - e(h)| \leq \sqrt{\frac{\ln(\frac{8m}{\delta})}{2m-1}}\right\}\right| &\geq \frac{1}{2}|\mathcal{H}(\frac{k}{m})| \\
\left|\left\{h \in \mathcal{H}(\frac{k}{m}) :\ |\hat{e}(h) - \frac{k}{m}| \leq \frac{1}{m} + \sqrt{\frac{\ln(\frac{8m}{\delta})}{2m-1}}\right\}\right| &\geq \frac{1}{2}|\mathcal{H}(\frac{k}{m})|
\end{aligned}$$

$$\left|\mathcal{H}(\frac{k}{m})\right| \leq 2\left|\hat{\mathcal{H}}\left(\frac{k}{m},\ 2m\delta\right)\right|.$$

Lemma 3.2 now follows by quantification over $q \in \{\frac{1}{m},\ \ldots,\ \frac{m}{m}\}$. □

4. Asymptotic Analysis and Phase Transitions

This section and the two that follow give an asymptotic analysis of the bounds presented earlier. The asymptotic analysis is stated in Theorem 4.1 and Statement 6.1. To develop the asymptotic analysis, however, a preliminary discussion is needed regarding the phenomenon of phase transitions. The bounds given in (9) and Theorem 3.3 exhibit phase transitions. More specifically, the bounding expression can be discontinuous in δ and m, e.g., arbitrarily small changes in δ can cause large changes in the bound. To see how this happens consider the constraint on the quantity q:

$$D(\hat{e}(h)||q) \leq \frac{s(\lceil\lceil q \rceil\rceil) + \ln(\frac{2m}{\delta})}{m}. \tag{11}$$

The bound given by (9) is the least upper bound of the values of q satisfying (11). Assume that m is sufficiently large that we can think of $\frac{s(\lceil\lceil q \rceil\rceil)}{m}$ as a continuous function of q which we write as $\bar{s}(q)$. We can then rewrite (11) where λ is a quantity not depending on q and $\bar{s}(q)$ does not depend on δ:

$$D(\hat{e}(h)||q) \leq \bar{s}(q) + \lambda. \tag{12}$$

For $q \geq \hat{e}(h)$ we know that $D(\hat{e}(h)||q)$ is a monotonically increasing function of q. It is reasonable to assume that for $q \leq 1/2$ we also have that $\bar{s}(q)$ is a monotonically increasing function of q. But even under these conditions it is possible that the feasible values of q, i.e., those satisfying (12), can be divided into separated regions. Furthermore, increasing λ can cause a new feasible region to come into existence. When this happens the bound, which is the least upper bound of the feasible values, can increase discontinuously. At a more intuitive level, consider a large number of high error concepts and smaller number of lower error concepts. At a certain confidence level the high error concepts can be ruled out. But as the confidence requirement becomes more stringent suddenly (and discontinuously) the high error concepts must be considered. A similar discontinuity can occur in sample size. Phase transitions in shell decomposition bounds are discussed in more detail by Haussler et al. (5).

Phase transitions complicate asymptotic analysis. But asymptotic analysis illuminates the nature of phase transitions. As mentioned in the introduction, in the asymptotic analysis of learning theory bounds it is important that one does not hold $\mathcal{H}$ fixed as the sample size increases. If we hold $\mathcal{H}$ fixed then $\lim_{m\to\infty} \frac{\ln|\mathcal{H}|}{m} = 0$. But this is not what one expects for large samples in practice. As the sample size increases one typically uses larger hypothesis classes. Intuitively, we expect that even for very large m we have that $\frac{\ln|\mathcal{H}|}{m}$ is far from zero.

For the asymptotic analysis of the bound in (9) we assume an infinite sequence of hypothesis classes $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ ... and an infinite sequence of data distributions D_1, D_2, D_3, ... Let $s_m(\frac{k}{m})$ be $s(\frac{k}{m})$ defined relative to $\mathcal{H}_m$ and D_m. In the asymptotic analysis we assume that the sequence of functions $\frac{s_m(\lceil\lceil q\rceil\rceil)}{m}$, viewed as functions of $q \in [0,1]$, converge uniformly to a continuous function $\bar{s}(q)$. This means that for any $\epsilon > 0$ there exists a k such that for all $m > k$ we have

$$\forall q \in [0,1] \;\; |\frac{s_m(\lceil\lceil q\rceil\rceil)}{m} - \bar{s}(q)| \leq \epsilon.$$

Given the functions $\frac{s_m(\lceil\lceil p\rceil\rceil)}{m}$ and their limit function $\bar{s}(p)$, we define the following functions of an empirical error rate $\hat{e}$:

$$\begin{aligned} B_m(\hat{e}) &\equiv \texttt{lub}\left\{q : D(\hat{e}||q) \leq \frac{s_m(\lceil\lceil q\rceil\rceil) + \ln(\frac{2m}{\delta})}{m}\right\}, \\ B(\hat{e}) &\equiv \texttt{lub}\,\{q : D(\hat{e}||q) \leq \bar{s}(q)\}. \end{aligned}$$

The function $B_m(\hat{e})$ corresponds directly to the upper bound in (9). The function $B(\hat{e})$ is intended to be the large m asymptotic limit of $B_m(\hat{e})$. However, phase transitions complicate asymptotic analysis. The bound $B(\hat{e})$ need not be a continuous function of $\hat{e}$. A value of $\hat{e}$ where the bound $B(\hat{e})$ is discontinuous corresponds to a phase transition in the bound. At a phase transition the sequence $B_m(\hat{e})$ need not converge. Away from phase transitions, however, we have the following theorem.

Theorem 4.1 *If the bound $B(\hat{e})$ is continuous at the point $\hat{e}$ (so we are not at a phase transition), and the functions (parameterized by m) $\frac{s_m(\lceil\lceil q\rceil\rceil)}{m}$, viewed as functions of $q \in [0,1]$, converge uniformly to a continuous function $\bar{s}(q)$, then we have*

$$\lim_{m\to\infty} B_m(\hat{e}) = B(\hat{e}).$$

Proof Define the set $F_m(\hat{e})$ as

$$F_m(\hat{e}) \equiv \left\{ q : \; D(\hat{e}||q) \leq \frac{s_m(\lceil\lceil q \rceil\rceil) + \ln(\frac{2m}{\delta})}{m} \right\}.$$

This gives

$$B_m(\hat{e}) = \mathtt{lub}\, F_m(\hat{e}).$$

Similarly, define $F(\hat{e},\, \epsilon)$ and $B(\hat{e},\, \epsilon)$ as

$$\begin{aligned} F(\hat{e},\, \epsilon) &\equiv \{q \in [0,1] : \; D(\hat{e}||q) \leq \bar{s}(q) + \epsilon\}. \\ B(\hat{e},\, \epsilon) &\equiv \mathtt{lub}\, F(\hat{e},\, \epsilon). \end{aligned}$$

We first show that the continuity of $B(\hat{e})$ at the point $\hat{e}$ implies the continuity of $B(\hat{e},\, \epsilon)$ at the point $\langle \hat{e},\, 0\rangle$. We note that there exists a continuous function $f(\hat{e},\, \epsilon)$ with $f(\hat{e},\, 0) = \hat{e}$ and such that for any ϵ sufficiently near 0 we have

$$D(f(\hat{e},\, \epsilon)||q) = D(\hat{e}||q) - \epsilon.$$

We then have

$$B(\hat{e},\, \epsilon) = B(f(x,\, \epsilon)).$$

Since f is continuous, and $B(\hat{e})$ is continuous at the point $\hat{e}$, we get that $B(\hat{e},\, \epsilon)$ is continuous at the point $\langle \hat{e},\, 0\rangle$.

We now prove the lemma. The functions of the form $\frac{s_m(\lceil\lceil q \rceil\rceil) + \ln \frac{2m}{\delta}}{m}$ converge uniformly to the function $\bar{s}(q)$. This implies that for any $\epsilon > 0$ there exists a k such that for all $m > k$ we have

$$F(\hat{e},\, -\epsilon) \subseteq F_m(\hat{e}) \subseteq F(\hat{e},\, \epsilon).$$

But this in turn implies that

$$B(\hat{e},\, -\epsilon) \leq B_m(\hat{e}) \leq B(\hat{e},\, \epsilon). \tag{13}$$

The lemma now follows from the continuity of the function $B(\hat{e},\, \epsilon)$ at the point $\langle \hat{e},\, 0\rangle$. ∎

Theorem 4.1 can be interpreted as saying that for large sample sizes, and for values of $\hat{e}$ other than the special phase transition values, the bound has a well defined value independent of the confidence parameter δ and determined only by a smooth function $\bar{s}(q)$. A similar statement can be made for the bound in Theorem 3.3 — for large m, and at points other than phase transitions, the bound is independent of δ and is determined by a smooth limit curve.

For the asymptotic analysis of Theorem 3.3 we assume an infinite sequence $\mathcal{H}_1,\, \mathcal{H}_2,\, \mathcal{H}_3,\, \ldots$ of hypothesis classes and an infinite sequence $S_1,\, S_2,\, S_3,\, \ldots$ of samples such that sample S_m has size m. Let $\mathcal{H}_m(\frac{k}{m},\, \delta)$ and $\hat{s}_m(\frac{k}{m},\, \delta)$ be $\mathcal{H}(\frac{k}{m},\, \delta)$ and $\hat{s}(\frac{k}{m},\, \delta)$ respectively defined relative to hypothesis class $\mathcal{H}_m$ and sample S_m. Let $U_m(\frac{k}{m})$ be the set of hypotheses in $\mathcal{H}_m$ having an empirical error of exactly $\frac{k}{m}$ in the sample S_m. Let $u_m(\frac{k}{m})$ be $\ln(\max(1,\, |U_m(\frac{k}{m})|)$. In the analysis of Theorem 3.3 we allow that the functions $\frac{u_m(\lceil\lceil q \rceil\rceil)}{m}$ are only locally uniformly convergent to a continuous function $\bar{u}(q)$, i.e., for any $q \in [0,1]$ and any $\epsilon > 0$ there exists an integer k and real number $\gamma > 0$ satisfying

$$\forall m > k,\; \forall p \in (q - \gamma,\; q + \gamma)\; |\frac{u_m(\lceil\lceil p \rceil\rceil)}{m} - \bar{u}(p)| \leq \epsilon.$$

Locally uniform convergence plays a role in the analysis in Section 6.

Theorem 4.2 *If the functions* $\frac{u_m(\lceil\lceil q \rceil\rceil)}{m}$ *converge locally uniformly to a continuous function* $\bar{u}(q)$ *then, for any fixed value of* δ*, the functions* $\frac{\hat{s}_m(\lceil\lceil q \rceil\rceil,\ \delta)}{m}$ *also converge locally uniformly to* $\bar{u}(q)$*. If the convergence of* $\frac{u_m(\lceil\lceil q \rceil\rceil)}{m}$ *is uniform, then so is the convergence of* $\frac{\hat{s}_m(\lceil\lceil q \rceil\rceil,\ \delta)}{m}$*.*

Proof Consider an arbitrary value $q \in [0,1]$ and $\epsilon > 0$. We construct the desired k and γ. More specifically, select k sufficiently large and γ sufficiently small that we have the properties

$$\forall m > k,\ \forall p \in (q - 2\gamma,\ q + 2\gamma)\ \left|\frac{u_m(\lceil\lceil p \rceil\rceil)}{m} - \bar{u}(p)\right| < \frac{\epsilon}{3},$$

$$\forall p \in (q - 2\gamma,\ q + 2\gamma)\ \ |\bar{u}(p) - \bar{u}(q)| \leq \frac{\epsilon}{3},$$

$$\frac{1}{k} + \sqrt{\frac{\ln(\frac{16k^2}{\delta})}{2k-1}} < \gamma,$$

$$\frac{\ln k}{k} \leq \frac{\epsilon}{3}.$$

Consider an $m > k$ and $p \in (q - \gamma,\ q + \gamma)$. It now suffices to show that

$$\left|\frac{\hat{s}_m(\lceil\lceil p \rceil\rceil,\ \delta)}{m} - \bar{u}(p)\right| \leq \epsilon.$$

Because $U_m(\lceil\lceil p \rceil\rceil)$ is a subset of $\mathcal{H}_m(\lceil\lceil p \rceil\rceil,\ \delta)$ we have

$$\frac{\hat{s}_m(\lceil\lceil p \rceil\rceil,\ \delta)}{m} \geq \frac{u_m(\lceil\lceil p \rceil\rceil)}{m} \geq \bar{u}(p) - \frac{\epsilon}{3}.$$

We can also upper bound $\frac{\hat{s}_m(\lceil\lceil p \rceil\rceil,\ \delta)}{m}$ as follows:

$$\begin{aligned}
|\mathcal{H}_m(\lceil\lceil p \rceil\rceil, \delta)| &\leq \sum_{|\frac{k}{m} - p| \leq \gamma} \left|U_m\left(\frac{k}{m}\right)\right| \\
&\leq \sum_{|\frac{k}{m} - p| \leq \gamma} e^{u_m(\frac{k}{m})} \\
&\leq \sum_{|\frac{k}{m} - p| \leq \gamma} e^{m(\bar{u}(\frac{k}{m}) + \frac{\epsilon}{3})} \\
&\leq \sum_{|\frac{k}{m} - p| \leq \gamma} e^{m(\bar{u}(p) + \frac{2\epsilon}{3})} \\
&\leq m e^{m(\bar{u}(p) + \frac{2\epsilon}{3})} \\
\frac{\hat{s}(\lceil\lceil p \rceil\rceil,\ \delta)}{m} &\leq \bar{u}(p) + \frac{2\epsilon}{3} + \frac{\ln m}{m} \\
&\leq \bar{u}(p) + \epsilon.
\end{aligned}$$

A similar argument shows that if $\frac{u_m(\lceil\lceil q\rceil\rceil)}{m}$ converges uniformly to $\bar{u}(q)$ then so does $\frac{u_m(\lceil\lceil q\rceil\rceil)}{m}$. ■

Given quantities $\frac{\hat{s}_i(\lceil\lceil q\rceil\rceil,\,\delta)}{m}$ that converge uniformly to $\bar{u}(q)$ the remainder of the analysis is identical to that for the asymptotic analysis of (9). We define the upper bounds

$$\begin{aligned} \hat{B}_m(\hat{e}) &\equiv \texttt{lub}\left\{q : D(\hat{e}||q) \leq \frac{\hat{s}_m(\lceil\lceil q\rceil\rceil,\ \delta) + \ln\left(\frac{4m}{\delta}\right)}{m}\right\}. \\ \hat{B}(\hat{e}) &\equiv \texttt{lub}\left\{q : D(\hat{e}||q) \leq \bar{u}(q)\right\}. \end{aligned}$$

Again we say that $\hat{e}$ is at a phase transition if the function $\hat{B}(\hat{e})$ is discontinuous at the value $\hat{e}$. We then get the following whose proof is identical to that of Theorem 4.1.

Theorem 4.3 *If the bound $\hat{B}(\hat{e})$ is continuous at the point $\hat{e}$ (so we are not at a phase transition), and the functions $\frac{u_m(\lceil\lceil q\rceil\rceil)}{m}$ converge uniformly to $\bar{u}(q)$, then we have that*

$$\lim_{m\to\infty} \hat{B}_m(\hat{e}) = \hat{B}(\hat{e}).$$

5. Asymptotic Optimality of (9)

Formula (9) can be viewed as providing an upper bound on $e(h)$ as a function of $\hat{e}(h)$ and the function s. In this section we show that for any curve s and value $\hat{e}$ there exists a hypothesis class and data distribution such that the upper bound in (9) is realized up to asymptotic equality. Up to asymptotic equality, (9) is the tightest possible bound computable from $\hat{e}(h)$ and the m numbers $s(\frac{1}{m}),\ \ldots,\ s(\frac{m}{m})$.

The classical VC dimensions bounds are nearly optimal over bounds computable from the chosen hypothesis error rate $\hat{e}(h^*)$ and the class $\mathcal{H}$. The m numbers $s(\frac{1}{m}),\ \ldots,\ s(\frac{m}{m})$ depend on both $\mathcal{H}$ and the data distribution. Hence the bound in (9) uses information about the distribution and hence can be tighter than classical VC bounds. A similar statement applies to the bound in Theorem (3.3) computed from the empirically observable numbers $\hat{s}(\frac{1}{m}),\ \ldots,\ \hat{s}(\frac{m}{m})$. In this case, the bound uses more information from the sample than just $\hat{e}(h)$. The optimality theorem given here also differs from the traditional lower bound results for VC dimension in that here the lower bounds match the upper bounds up to asymptotic equality.

The departure point for our optimality analysis is the following lemma from (3).

Lemma 5.1 (Cover and Thomas) *If $\hat{p}$ is the fraction of heads out of m tosses of a coin where the true probability of heads is p then for $q \geq p$ we have*

$$Pr(\hat{p} \geq q) \geq \frac{1}{m+1} e^{-mD(q||p)}.$$

This lower bound on $Pr(\hat{p} \geq q)$ is very close to Chernoff's 1952 upper bound (1). The tightness of (9) is a direct reflection of the tightness (1). To exploit Lemma 5.1 we need to construct hypothesis classes and data distributions where distinct hypotheses have independent training errors. More specifically, we say that a set of hypotheses $\{h_1,\ \ldots,\ h_n\}$ has independent training errors if the random variables $\hat{e}(h_1),\ \ldots,\ \hat{e}(h_n)$ are independent.

By an argument similar to the derivation of (3) from (1) we can prove the from Lemma 5.1 that

$$Pr\left(D(\min(\hat{p},p)||p) \geq \frac{\ln(\frac{1}{\delta}) - \ln(m+1)}{m}\right) \geq \delta. \tag{14}$$

Lemma 5.2 *Let X be any finite set, S a random variable, and $\Theta[S,x,\delta]$ a formula such that for every $x \in X$ and $\delta > 0$ we have $Pr(\Theta[S,x,\delta]) \geq \delta$, and $Pr(\forall x \in X\ \Theta[S,\ x,\ \delta]) = \prod_{x\in X} Pr(\Theta[S,\ x,\ \delta])$. We then have $\forall\delta > 0\ \forall^\delta S\ \ \exists x \in X\ \ \Theta[S, x, \frac{\ln(\frac{1}{\delta})}{|X|}]$.*

Proof

$$\begin{array}{rcl} Pr(\Theta[S,x,\frac{\ln(\frac{1}{\delta})}{|X|}]) & \geq & \frac{\ln(\frac{1}{\delta})}{|X|} \\ Pr(\neg\Theta[S,x,\frac{\ln(\frac{1}{\delta})}{|X|}]) & \leq & 1 - \frac{\ln(\frac{1}{\delta})}{|X|} \\ & \leq & e^{-\frac{\ln(\frac{1}{\delta})}{|X|}} \\ Pr(\forall x \in X \neg\Theta[S,\ x,\frac{\ln(\frac{1}{\delta})}{|X|}]) & \leq & e^{-\ln(\frac{1}{\delta})} = \delta. \end{array}$$

■

Now define $h^*(\frac{k}{m})$ to be the hypothesis of minimal training error in the set $\mathcal{H}(\frac{k}{m})$. Let `glb` $\{x : \Phi[x]\}$ denote the greatest lower bound (the minimum) of the set $\{x :\ \Phi[x]\}$. We now have the following lemma.

Lemma 5.3 *If the hypotheses in the class $\mathcal{H}(\lceil\lceil q\rceil\rceil)$ are independent then $\forall\delta > 0$, $\forall^\delta S$, $\forall q \in \{\frac{1}{m},\ \ldots,\ \frac{m}{m}\}$,*

$$\hat{e}(h^*(q)) \leq \texttt{glb}\ \left\{\hat{e} :\ \ D(\min(\hat{e},\ q - \tfrac{1}{m})||q) \leq \tfrac{s(q) - \ln(m+1) - \ln(\ln(\frac{m}{\delta}))}{m}\ \right\}.$$

Proof To prove Lemma 5.3 let q be a fixed rational number of the form $\frac{k}{m}$. Assuming independent hypotheses we can applying Lemma 5.2 to (14) to get $\forall\delta > 0$, $\forall^\delta S$, $\exists h \in \mathcal{H}(\frac{k}{m})$,

$$D(\min(\hat{e}(h), e(h))||e(h)) \geq \frac{s(q) - \ln(m+1) - \ln(\ln(\frac{1}{\delta}))}{m}$$

Let w be the hypothesis in $\mathcal{H}(q)$ satisfying this formula. We now have $\hat{e}(h^*(q)) \leq \hat{e}(w)$ and $q - \frac{1}{m} \leq e(w) \leq q$. These two conditions imply $\forall\delta > 0$, $\forall^\delta S$,

$$D(\min(\hat{e}(h^*(q)), q - \tfrac{1}{m})||q) \geq \tfrac{s(q) - \ln(m+1) - \ln(\ln\frac{1}{\delta})}{m}.$$

This implies that

$$\hat{e}(h^*(q)) \leq \texttt{glb}\ \left\{\hat{e} :\ \ D(\min(\hat{e},\ q - \tfrac{1}{m})||q) \leq \tfrac{s(q) - \ln(m+1) - \ln(\ln(\frac{1}{\delta}))}{m}.\ \right\}$$

Lemma 5.3 now follows by quantification over $q \in \{\frac{1}{m},\ \ldots,\ \frac{m}{m}\}$.

■

For $q \in [0,1]$ we have that Lemma 3.1 implies that

$$\hat{e}(h^*(\lceil\lceil q \rceil\rceil)) \geq \texttt{glb}\left\{\hat{e}: \; D\left(\hat{e}||\lceil\lceil q\rceil\rceil - \frac{1}{m}\right) \leq \frac{s(\lceil\lceil q\rceil\rceil)+\ln\left(\frac{2m}{\delta}\right)}{m}. \right\}$$

We now have upper and lower bounds on the quantity $\hat{e}(h^*(\lceil\lceil q\rceil\rceil))$ which agree up to asymptotic equality — in a large m limit where $\frac{s_m(\lceil\lceil q\rceil\rceil)}{m}$ converges (pointwise) to a continuous function $\bar{s}(q)$ we have that the upper and lower bound on $\hat{e}(h^*(\lceil\lceil q\rceil\rceil))$ both converge (pointwise) to

$$\hat{e}(h^*(q)) = \texttt{glb}\,\{\hat{e} : D(\hat{e}||q) \leq \bar{s}(q)\}.$$

This asymptotic value of $\hat{e}(h^*(q))$ is a continuous function of q. Since q is held fixed in calculating the bounds on $\hat{e}(\lceil\lceil q\rceil\rceil)$, phase transitions are not an issue and uniform convergence of the functions $\frac{s_m(\lceil\lceil q\rceil\rceil)}{m}$ is not required. Note that for large m and independent hypotheses we get that $\hat{e}(h^*(q))$ is determined as a function of the true error rate q and $\frac{s(\lceil\lceil q\rceil\rceil)}{m}$.

The following lemma states that any limit function $\bar{s}(p)$ is consistent with the possibility that hypotheses are independent. This, together with Lemma 5.3 implies that no uniform bound on $e(h)$ as a function of $\hat{e}(h)$ and $|\mathcal{H}(\frac{1}{m})|, \ldots, |\mathcal{H}(\frac{m}{m})|$ can be asymptotically tighter than (9).

Theorem 5.4 *Let $\bar{s}(p)$ be any continuous function of $p \in [0,1]$. There exists an infinite sequence of hypothesis spaces $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$, $\ldots$, and sequence of data distributions D_1, D_2, D_3, $\ldots$ such that each class $\mathcal{H}_m$ has independent hypotheses for data distribution D_m and such that $\frac{s_m(\lceil\lceil p\rceil\rceil)}{m}$ converges (pointwise) to $\bar{s}(p)$.*

Proof First we show that if $|\mathcal{H}_m(\frac{i}{m})| = e^{m\bar{s}(\frac{i}{m})}$ then the functions $\frac{s_m(\lceil\lceil p\rceil\rceil)}{m}$ converge (pointwise) to $\bar{s}(p)$. Assume $|\mathcal{H}_m(\frac{i}{m})| = e^{m\bar{s}(\frac{i}{m})}$. In this case we have

$$\frac{s_m(\lceil\lceil p\rceil\rceil)}{m} = \bar{s}(\lceil\lceil p\rceil\rceil).$$

Since $\bar{s}(p)$ is continuous, for any fixed value of p we get that $\frac{s_m(\lceil\lceil p\rceil\rceil)}{m}$ converges to $\bar{s}(p)$.

Recall that D_m is a probability distribution on pairs $\langle x, y\rangle$ with $y \in \{0,1\}$ and $x \in X_m$ for some set X_m. We take $\mathcal{H}_m$ to be a disjoint union of sets $\mathcal{H}_m(\frac{k}{m})$ where $|\mathcal{H}_m(\frac{k}{m})|$ is selected as above. Let $f_1, \ldots, f_N$ be the elements of $\mathcal{H}_m$ with $N = |\mathcal{H}_m|$. Let X_m be the set of all N-bit bit strings and define $f_i(x)$ to be the value of ith bit of the bit vector x. Now define the distribution D_m on pairs $\langle x, y\rangle$ by selecting y to be 1 with probability 1/2 and then selecting each bit of x independently where the ith bit is selected to disagree with y with probability $\frac{k}{m}$ where k is such that $f_i \in \mathcal{H}_m(\frac{k}{m})$. ■

6. Relating $\hat{s}$ and s

In this section we show that in large m limits of the type discussed in Section 4 the histogram of empirical errors need not converge to the histogram of true errors. So even in the large m asymptotic limit, the bound given by Theorem 3.3 is significantly weaker than the bound given by (9).

To show that $\hat{s}(\lceil\lceil q \rceil\rceil,\ \delta)$ can be asymptotically different from $s(\lceil\lceil q \rceil\rceil)$ we consider the case of independent hypotheses. More specifically, given a continuous function $\bar{s}(p)$ we construct an infinite sequence of hypothesis spaces $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$, ... and an infinite sequence of data distributions D_1, D_2, D_3, ... using the construction in the proof of Theorem 5.4. We note that if $\bar{s}(p)$ is differentiable with bounded derivative then the functions $\frac{s_m(\lceil\lceil p \rceil\rceil)}{m}$ converge uniformly to $\bar{s}(p)$.

For a given infinite sequence data distributions we generate an infinite sample sequence S_1, S_2, S_3, ..., by selecting S_m to consists of m pairs $\langle x,\ y \rangle$ drawn IID from distribution D_m. For a given sample sequence and $h \in \mathcal{H}_m$ we define $\hat{e}_m(h)$ and $\hat{s}_m(\frac{k}{m},\ \delta)$ in a manner similar to $\hat{e}(h)$ and $\hat{s}(\frac{k}{m},\ \delta)$ but for sample S_m. The main result of this section is the following.

Conjecture 6.1 *If each $\mathcal{H}_m$ has independent hypotheses under data distribution D_m, and the functions $\frac{s_m(\lceil\lceil p \rceil\rceil)}{m}$ converge uniformly to a continuous function $\bar{s}(p)$, then for any $\delta > 0$ and $p \in [0,1]$, we have, with probability 1 over the generation of the sample sequence, that*

$$\lim_{m\to\infty} \frac{\hat{s}_m(\lceil\lceil p \rceil\rceil, \delta)}{m} = \sup_{q\in[0,1]} \bar{s}(q) - D(p||q).$$

We call this a conjecture rather than a theorem because the proof has not been worked out to a high level of rigor. Nonetheless, we believe the proof sketch given below can be expanded to a fully rigorous argument.

Before giving the proof sketch we note that the limiting value of $\frac{\hat{s}_m(\lceil\lceil p \rceil\rceil,\ \delta)}{m}$ is independent of δ. This is consistent with Theorem 4.2. Define

$$\bar{\hat{s}}(p) \equiv \sup_{q\in[0,1]} \bar{s}(q) - D(p||q).$$

Note that $\bar{\hat{s}}(p) \geq \bar{s}(p)$. This gives an asymptotic version of Lemma 3.2. But since $D(p||q)$ can be locally approximated as $c(p-q)^2$ (up to its second order Taylor expansion), if $\bar{s}(p)$ is increasing at the point p then we also get that $\bar{\hat{s}}(p)$ is strictly larger than $\bar{s}(p)$.

Proof Outline: To prove Statement 6.1 we first define $\mathcal{H}_m(p,\ q)$ for $p, q \in \{\frac{1}{m},\ \ldots,\ \frac{m}{m}\}$ to be the set of all $h \in \mathcal{H}_m(q)$ such that $\hat{e}_m(h) = p$. Intuitively, $\mathcal{H}_m(p,\ q)$ is the set of concepts with true error rate near q that have empirical error rate p. Ignoring factors that are only polynomial in m, the probability of a hypothesis with true error rate q having empirical error rate p can be written as (approximately) $e^{-mD(p||q)}$. So the expected size of $\mathcal{H}_m(p,\ q)$ can be written as $|\mathcal{H}_m(q)|e^{-mD(p||q)}$, or alternatively, (approximately) as $e^{m\bar{s}(q)}e^{-mD(p||q)}$ or $e^{m(\bar{s}(q)-D(p||q))}$. More formally, we have, for any fixed value of p and q,

$$\lim_{m\to\infty} \frac{\ln(\max(1,\ \mathtt{E}(|\mathcal{H}_m(\lceil\lceil p \rceil\rceil,\ \lceil\lceil q \rceil\rceil)|)))}{m}$$
$$= \max(0,\ \bar{s}(q) - D(p||q)).$$

We now show that the expectation can be eliminated from the above limit. First, consider distinct values of p and q such that $\bar{s}(q) - D(p||q) > 0$. Since p and q are distinct, the probability that a fixed hypothesis in $\mathcal{H}_m(\lceil\lceil q \rceil\rceil)$ is in $\mathcal{H}_m(\lceil\lceil p \rceil\rceil,\ \lceil\lceil q \rceil\rceil)$ declines exponentially in m. Since $\bar{s}(q) - D(p||q) > 0$ the expected size of $\mathcal{H}_m(\lceil\lceil p \rceil\rceil,\ \lceil\lceil q \rceil\rceil)$ grows exponentially in m. Since the hypotheses are independent, the distribution of possible values of $|\mathcal{H}_m(\lceil\lceil p \rceil\rceil,\ \lceil\lceil q \rceil\rceil)|$ becomes essentially a Poisson mass distribution with an expected number of arrivals growing exponentially in m. The probability that $|\mathcal{H}_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q\rceil\rceil)|$ deviates from its expectation by as much as a factor of 2 declines exponentially in m. We say that a sample sequence is safe after k if for all $m > k$ we have that $|\mathcal{H}_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q\rceil\rceil)|$ is within a factor of 2 of its expectation. Since the probability of being unsafe at m declines exponentially in m, for any δ there exists a k such that with probability at least $1-\delta$ the sample sequence is safe after k. So for any $\delta > 0$ we have that with probability at least $1-\delta$ the sequence is safe after some k. But since this holds for all $\delta > 0$, with probability 1 such a k must exist:

$$\lim_{m\to\infty} \frac{\ln(\max(1, |\mathcal{H}_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q\rceil\rceil)|))}{m}$$
$$= \bar{s}(q) - D(p||q).$$

We now define

$$s_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q\rceil\rceil) \equiv \ln(\max(1,\ |\mathcal{H}_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q\rceil\rceil)|)).$$

It is also possible to show for $p = q$ we have that with probability 1 we have that $\frac{s_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q\rceil\rceil)}{m}$ approaches $\bar{s}(p)$ and that for distinct p and q with $\bar{s}(q) - D(p||q) \leq 0$ we have that $\frac{s_m(\lceil\lceil q\rceil\rceil,\ \lceil\lceil q\rceil\rceil)}{m}$ approaches 0. Putting these together yields that, with probability 1, we have

$$\lim_{m\to\infty} \frac{s_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q\rceil\rceil)}{m} = \max(0,\ \bar{s}(q) - D(p||q)). \tag{15}$$

Define $U_m(\frac{k}{m})$ and $u_m(\frac{k}{m})$ as in Section 4. We now have the following equality:

$$U_m(p) = \cup_{q\in\{\frac{1}{m},\ \ldots,\ \frac{m}{m}\}} \mathcal{H}_m(p,\ q).$$

We now show that with probability 1, $\frac{u_m(p)}{m}$ approaches $\bar{\hat{s}}(p)$. First, consider a $p \in [0,1]$ such that $\bar{\hat{s}}(p) > 0$. Let Since $\bar{s}(q) - D(q||p)$ is a continuous function, and $[0,1]$ is a compact set, $\sup_{q\in[0,1]} \bar{s}(q) - D(p||q)$ must be realized at some value $q^* \in [0,1]$. Let q^* be such that $\bar{s}(q^*) - D(p||q^*)$ equals $\bar{\hat{s}}(p)$. We have that $u_m(\lceil\lceil p\rceil\rceil) \geq s_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q^*\rceil\rceil)$. This, together with (15), implies that

$$\liminf_{m\to\infty} \frac{u_m(\lceil\lceil p\rceil\rceil)}{m} \geq \bar{\hat{s}}(p).$$

The sample sequence is "safe" at m and $\frac{k}{m}$ if $|\mathcal{H}_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil \frac{k}{m}\rceil\rceil)|$ does not exceed twice the expectation of $|\mathcal{H}_m(\lceil\lceil p\rceil\rceil,\ \lceil\lceil q^*\rceil\rceil)|$. Assuming uniform convergence of $\frac{s_m(\lceil\lceil p\rceil\rceil)}{m}$, the probability of not being safe at m and $\frac{k}{m}$ declines exponentially in m at a rate at least as fast as the rate of decline of the probability of not being safe at m and $\lceil\lceil q^*\rceil\rceil$. By the union bound this implies that for a given m the probability that there exists an unsafe $\frac{k}{m}$ also declines exponentially. We say that the sequence is safe after N if it is safe for all m and $\frac{k}{m}$ with $m > N$. The probability of not being being safe after N also declines exponentially with N. By an argument similar to that given above, this implies that with probability 1 over the choice of the sequence there exists a N such that the sequence is safe after N. But if we are safe at m then $|U_m(\lceil\lceil p\rceil\rceil)| \leq 2mE|\mathcal{H}_m(p,\ \lceil\lceil q^*\rceil\rceil)|$. This implies that

$$\limsup_{m\to\infty} \frac{u_m(\lceil\lceil p\rceil\rceil)}{m} \leq \bar{\hat{s}}(p).$$

Putting the two bounds together we get

$$\lim_{m\to\infty} \frac{u_m(\lceil\lceil p\rceil\rceil)}{m} = \bar{\hat{s}}(p).$$

The above argument establishes (to some level of rigor) pointwise convergence of $\frac{u_m([[p]])}{m}$ to $\bar{\hat{s}}(p)$. It is also possible to establish a convergence rate that is a continuous function of p. This implies that the convergence of $\frac{u_m([[p]])}{m}$ can be made locally uniform. Theorem 4.2 then implies the desired result. □

7. Improvements

Theorem 3.3 has been improved in various ways (6):

- Removing the discretization of true errors,
- Using one-sided bounds,
- Using nonuniform union bounds over discrete values of the form $\frac{k}{m}$,
- Tightening the Chernoff bound using direct calculation of Binomial coefficients, and
- Improving Lemma 3.4.

These improvements allow the removal of all but one $\ln(m)$ terms from the statement of the bound. However, they do not improve the asymptotic equations given by Theorem 4.1 and Statement 6.1.

A practical difficulty with the bound in Theorem 3.3 is that it is usually impossible to enumerate the elements of an exponentially large hypothesis class and hence impractical to compute the histogram of training errors for the hypotheses in the class. In practice the values of $s(\frac{k}{m})$ might be estimated using some form of Monte-Carlo Markov chain sampling over the hypotheses. For certain hypothesis spaces it might also be possible to directly calculate the empirical error distribution without evaluating every hypothesis. For example, this can be done with "partition rules" which, given a fixed partition of the input space, make predictions which are constant on each partition. If there are n elements in the partition then there are 2^n partition rules. For a fixed partition, the histogram of empirical errors for the 2^n partition rules can be computed in polynomial time. Note that the class of decision trees is a union of partition rules where the structure of a tree defines a partition and the labels at the leaves of the tree define a particular partition rule relative to that partition. Taking advantage of this, it is suprisingly easy to compute a shell bound for small decision trees (6).

8. Discussion and Conclusion

Traditional PAC bounds are stated in terms of the training error and class size or VC dimension. The computable bound given here is sometimes much tighter because it exploits the additional information in the histogram of training errors. The uncomputable bound uses the additional (unavailable) information in the distribution of true errors. Any distribution of true errors can be realized in a case with independent hypotheses. We have shown that in such cases this uncomputable bound is asymptotically equal to actual generalization error. Hence this is the tightest possible bound, up to asymptotic equality, over all bounds expressed as functions of $\hat{e}(h^*)$ and the distribution of true errors. We have also shown that the use of the histogram of empirical errors results in a bound that, while still tighter than traditional bounds, is looser than the uncomputable bound even in the large sample asymptotic limit.

One of the goals of learning theory is to give generalization guarantees that are predictive of actual generalization error. It is well known that the actual generalization error can exhibit phase transitions — as the sample size increases the expected generalization error can jump essentially discontinuously in sample size. So accurate true error bounds should also exhibit phase transitions. Shell bounds exhibit these phase transitions while other bounds such as VC dimension results do not.

The phase transitions can also be interpreted as a statement about the bound as a function of the confidence parameter δ. As the value of δ is varied the bound may shift essentially discontinuously. To put this another way, let h^* be the hypothesis of minimal training error on a large sample. Near a phase transition in true generalization error (as opposed to a phase transition in the bound) we may have that with probability $1-\delta$ the true error of h^* is near its training error but with probability $\delta/2$, say, the true error of h^* can be far from its training error. More traditional bounds do not exhibit this kind of sensitivity to δ. Bounds that exhibit phase transitions seem to bring the theoretical analysis of generalization closer to the actual phenomenon.

Acknowledgments

Yoav Freund, Avrim Blum, and Tobias Scheffer all provided useful discussion in forming this paper.

References

[1] P. Bartlett, O. Bousquet and S. Mendelson. Localized rademacher complexities. *Proceedings of the 15th annual conference on Computational Learning Theory*, 44-58 (2002).

[2] H. Chernoff. A measure of asymptotic efficiency for test of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics*, 23:493-507, 1952.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, Wiley, 1991.

[4] Y. Freund, Self bounding algorithms, *Computational Learning Theory (COLT)*, 1998.

[5] D. Haussler, M. Kearns, H. Sebastian Seung, and N. tishby, Rigorous learning curve bounds from statistical mechanics, *Machine Learning* issue 25, 195-236, 1996.

[6] J. Langford, Quantitatively Tight sample complexity bounds, Thesis, Carnegie Mellon 2002.

[7] J. Langford, Practical prediction theory for classification, ICML 2003 tutorial, avaliable at http://hunch.net/ jl/projects/prediction_bounds/tutorial/tutorial.ps.

[8] J. Langford and A. Blum, Microchoice and self-bounding algorithms, *Computational Learning Theory (COLT)*, 1999.

[9] Y. Mansour and D. McAllester, Generalization bounds for decision trees, *Computational Learning Theory (COLT)*, 2000.

[Moore] A. Moore, VC dimension for characterizing classifiers, Tutorial at http://www-2.cs.cmu.edu/ awm/tutorials/vcdim08.pdf

[10] D. McAllester, Pac-Bayesian model averaging, *Computational Learning Theory (COLT)*, 1999.

[11] D. McAllester and R. Schapire, On the convergence rate of good-turing estimators, *Computational Learning Theory (COLT)*, 2000.

[12] T. Scheffer and T. Joachims, Expected error analysis for model selection, *International Conference on Machine Learning (ICML)*, 1999.

[13] S. van de Geer, *Empirical Process in M-Estimation*, Cambridge University Press, 1999.