Improved Shell Bounds

John Langford
IBM, TJ Watson Research Center
jcl@cs.cmu.edu

Abstract

Shell bounds take into account an empirical observation: for most learning problems the true error rates of most of the hypotheses is far from the optimal hypothesis. Taking advantage of this insight leads to a large (functional) improvement over prior approaches for bounding the true error rate of a learned hypothesis.

Here we followup on the original shell bound paper [5] with a tighter, simpler argument. The new shell bound argument has a corresponding lower bound which is tight up to higher order terms in the confidence of the bound.

1 Introduction

Roughly speaking, shell bounds (usually) provide much tighter true error rate upper bounds on learned hypotheses than conventional approaches. This is done by incorporating *much*[1] more (observable) information into the bound calculation. One particularly interesting aspect of Shell bounds is that the improvement is orthogonal to (and thus cumulative with) other bound construction techniques. This orthogonality implies that improvements in shell bounds might apply to the work of many others in learning theory.

The inspiration behind the work on Shell bounds rests on two pieces of work. In [2] by Haussler, Kearns, Seung, and Tishby, learning theory curves are investigated from an omniscient point of view where the true error rates of various hypotheses are known. The principle improvement in this paper is that our bounds are reduced to *observable* quantities. Put another way, we do not need to know the underlying learning distribution, D. Further work by Scheffer and Joachims [9], made an analysis assuming a known distribution over true error rates. Our analysis does not rely on any assumption about the distribution of true error rates. More specifically, the results here assume only the independence of examples drawn from a fixed (but unknown) distribution. Despite using only observable information and making no extra assumptions, the results here can yield quantitatively useful results [3].

The refined argument starts with the distribution of empirical errors over hypotheses. We subtract a small amount from the empirical error rates to create a pessimistic distribution. With high probability, an expectation made with respect to the pessimistic distribution upper bounds the probability that a large true error hypothesis produces a small empirical error. The bound implied by this observation can be *much* tighter than standard chernoff bound techniques although the

[1] Gathering this information is nontrivial.

quantity of improvement is highly learning problem dependent. See section 4.1 for an example.

As an outline the paper:

1. Presents the improved shell bound.
2. Presents a lower bound on the upper bound.
3. Discuss the implications of the results including examples and asymptotics.

2 The Discrete Shell Bound

The functional improvements on the original shell bound are necesary for practical application of the bound. Heuristically, the improvements come from three sources:

1. Shifting from a Chernoff approximation to a direct calculation of a Binomial tail probability.
2. Removal of unnecessary discretizations.
3. A change in the argument so we bound one expectation rather than many individual expectations.

Aside from practical application, this new shell bound is objectively simpler: fewer quantities need to be defined and the proofs are somewhat shorter and more self-contained. It offers a slightly different intuition which may appeal to some people more than the original.

Throughout this paper, we assume that all of the examples (x, y) we learn from are drawn independently from a distribution D over the feature space, X and output space $Y = \{0, 1\}$. The training set S consists of m such independently drawn samples, and the goal of learning is to find a hypothesis h with a small true error rate, $e(h) \equiv \Pr_{x,y\sim D}(h(x) \neq y)$. The true error rate is unobservable, but we can observe the error rate on the training set, $\hat{e}_S(h) = \Pr_{x,y\sim U(S)}(h(x) \neq y)$ and construct a confidence interval which bounds the true error $e(h)$ with high probability over the draw of the sample set S.

The proof of a shell bound is done in two discrete steps. The first step answers the question, "What could we do if we knew the distribution of true error rates?" and the second step shows how to work using only an observable quantity: the distribution of training error rates.

2.1 Knowledge of learning distribution

Let $B\left(\frac{K}{m}, p\right) = \binom{m}{K} p^K (1-p)^{m-K}$ be the probability that a hypothesis with true error $e(h) = p$ produces K errors on m independent examples.

The full knowledge shell bound works directly with a bound on the probability of observing a confusingly small empirical error. Let

$$P\left(H, \frac{K}{m}, \epsilon\right) \equiv \sum_{h \in H : e(h) > \frac{K}{m} + \epsilon} B\left(\frac{K}{m}, e(h)\right)$$

Intuitively, $P\left(H, \frac{K}{m}, \epsilon\right)$ is a bound on the probability that a hypotheses with a true error rate larger than $\frac{K}{m} + \epsilon$ has an empirical error rate of $\frac{K}{m}$. The contribution to the sum falls off exponentially as the true error, $e(h)$, increases.

Our first step is stating a shell bound which requires unknown information. The purpose of this bound is motivational—it provides insight into why we can expect a large improvement. Later, we remove the unknown information requirements and recover a useful bound.

Lemma 2.1 *(Full knowledge) For all $\delta \in (0,1]$ for all distributions D, for all hypothesis sets H, $\forall \frac{K}{m}$:*

$$\Pr_{S\sim D^m}\left(\forall h \in H: \ \hat{e}_S(h) = \frac{K}{m} \Rightarrow e(h) < \frac{K}{m} + \min\left\{\epsilon : \ P\left(H, \frac{K}{m}, \epsilon\right) \leq \delta\right\}\right) \geq 1-\delta$$

The full knowledge lemma relies on unobservable information—the true error rates of all hypotheses. This lemma is not (quite) trivial because it does not rely upon information about *which* hypothesis has a particular true error. The primary role of this lemma is as a stepping stone providing intuition into the full result.

Proof For every hypothesis with a true error rate of

$$e(h) > \frac{K}{m} + \min\left\{\epsilon : \ P\left(H, \frac{K}{m}, \epsilon\right) \leq \delta\right\}$$

the probability of producing an empirical error of

$$\hat{e}_S(h) = \frac{K}{m}$$

is $B\left(\frac{K}{m}, e(h)\right)$. Using the union bound to sum over all possible hypotheses, we get the result. $\square$

We could easily make this lemma apply regardless of the empirical error by taking a union over the m nontrivial values of K. However, it turns out later that this form is more convenient.

For most balanced machine learning problems (with 50% one class and 50% the other) most hypotheses typically have a true error rate near to 0.5. Given this, and noticing that Binomial tails fall of exponentially, dramatic improvements over a naive Chernoff bound are feasible.

2.2 No knowledge of learning distribution

Applying the full knowledge lemma (2.1) is not practical in almost all learning settings because we do not know the distribution of true error rates. Therefore, it is necessary to construct a theorem which relies upon only observable quantities. Surprisingly, this is possible while introducing only slightly more slack.

First, we need a couple of definitions.

$$\text{Bin}\left(\frac{K}{m}, \ p\right) = \sum_{i=0}^{K} B\left(\frac{i}{m}, \ p\right)$$

is the probability that coin with bias p produces K or fewer heads. This is known as the Binomial distribution.

Next we define $\underline{e}$ as:

$$\underline{e}\left(\frac{K}{m}, \epsilon, \delta, \frac{i}{m}\right) = \max \ \left\{\frac{K}{m} + \epsilon, \ \min\left\{p : \ 1 - \text{Bin}\left(\frac{i}{m}, \ p\right) \geq \delta\right\}\right\}$$

Intuitively, $\underline{e}\left(\frac{K}{m}, \epsilon, \delta, \frac{i}{m}\right)$ is a worst case value of the true error rate which is constrained in two ways.

1. The true error rate is at least $\frac{K}{m}+\epsilon$ (because we do not care if the true error rate is smaller.)
2. The true error rate falls within a confidence interval consistent with with observed empirical error, $\frac{i}{m}$.

Next, we consider the sum of the probabilities that these "worst case true error rates" could produce an unusual event.

$$\hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right) \equiv 2 \sum_{h \in H} B\left(\frac{K}{m}, \underline{e}\left(\frac{K}{m}, \epsilon, \frac{\delta}{2}, \hat{e}_S(h)\right)\right)$$

The quantity $\hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right)$ is an upper bound on the probability that one of the hypotheses with a true error rate of $\bar{e}$ or more could produce K empirical errors. Here, the quantity δ controls the probability that $\hat{P}$, which is based on the observed empirical error distribution, is not large enough to represent the real probability of a too-small observation.

Noting that there are only $m+1$ possible empirical errors, we can first let

$$c\left(\frac{i}{m}\right) = \left|\left\{h \in H : \hat{e}(h) = \frac{i}{m}\right\}\right|$$

be the count of empirical errors at $\frac{i}{m}$. Then we can redefine:

$$\hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right) \equiv 2 \sum_{i=0}^{m} c\left(\frac{i}{m}\right) B\left(\frac{K}{m}, \underline{e}\left(\frac{K}{m}, \epsilon, \frac{\delta}{2}, \frac{i}{m}\right)\right)$$

Later, we prove that with high probability, $\hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right) \geq P\left(H, \frac{K}{m}, \epsilon\right)$. Given that this is so, we can prove a theorem which *only* relies on observable quantities.

Theorem 2.2 *(Observable Shell Bound) For all $\delta > 0$, for all distributions D*

$$\Pr_{S \sim D^m}\left(\forall h \in H \ e(h) < \hat{e}(h) + \min\left\{\epsilon : \ \hat{P}\left(H, \hat{e}(h), \epsilon, \frac{\delta}{2m}\right) \leq \frac{\delta}{2m}\right\}\right) \geq 1-\delta$$

The observable shell bound preserves the important locality property of the full knowledge shell bound. In particular, when most of the true error rates are "far" from the empirical error rate (and m is large enough), we expect to make large (functional) improvements on more naive approaches. This is further discussed in section 4.2.

The proof rests upon a technical lemma which allows us to bound the unobservable "probability of a misleading event" with an observable event.

Lemma 2.3 *(Unobservable bound) For all empirical errors, K, for all $\delta \in (0,1]$:*

$$\Pr_{S \sim D^m}\left(\hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right) \geq P\left(H, \frac{K}{m}, \epsilon\right)\right) \geq 1-\delta$$

This lemma is powerful because it bounds the unobservable right hand side in terms of the observable left hand side.

Proof Let $P(h)$ be any measure on a set of hypotheses.

$$\forall \alpha \in (0,1] \ \text{ with } e(h) \geq \frac{K}{m}+\epsilon \ \Pr_{S \sim D^m}\left(e(h) < \underline{e}\left(\frac{K}{m}, \epsilon, \alpha, \hat{e}_S(h)\right)\right) < \alpha$$

Since this holds for every h, it holds for any distribution on h.

$$\Rightarrow \forall P(h) \Pr_{S\sim D^m, h\sim P}\left(e(h) < \underline{e}\left(\frac{K}{m}, \epsilon, \alpha, \hat{e}_S(h)\right)\right) < \alpha$$

Using the definition of probability, we can rewrite this as:

$$\Rightarrow E_{S\sim D^m} E_{h\sim P} I\left(e(h) < \underline{e}\left(\frac{K}{m}, \epsilon, \alpha, \hat{e}_S(h)\right)\right) < \alpha$$

Now, we can use Markov's inequality. The expectation has the form, $E_{S\sim D^m} f(\alpha) < \alpha$ which implies $\forall\delta \;\; \Pr_{S\sim D^m}\left(f(\alpha) > \frac{\alpha}{\delta}\right) < \delta$. Rewriting, we get:

$$\Rightarrow \forall\delta \Pr_{S\sim D^m}\left(E_{h\sim P} I\left(e(h) < \underline{e}\left(\frac{K}{m}, \epsilon, \alpha, \hat{e}_S(h)\right)\right) > \frac{\alpha}{\delta}\right) < \delta$$

Taking the contrapositive twice, we get:

$$\Rightarrow \Pr_{S\sim D^m}\left(\Pr_{h\sim P}\left(e(h) \geq \underline{e}\left(\frac{K}{m}, \epsilon, \alpha, \hat{e}_S(h)\right)\right) > 1 - \frac{\alpha}{\delta}\right) \geq 1 - \delta$$

Now, rename $P(h) = \frac{1}{N} B\left(\frac{K}{m}, e(h)\right)$ for $e(h) \geq \frac{K}{m} + \epsilon$ and 0 otherwise where N is a normalization constant

$$\Rightarrow \Pr_{S\sim D^m}\left(\frac{\sum_{h\in H: e(h) > \frac{K}{m}+\epsilon \wedge e(h) \geq \underline{e}\left(\frac{K}{m}, \epsilon, \alpha, \hat{e}_S(h)\right)} B\left(\frac{K}{m}, e(h)\right)}{\sum_{h\in H: e(h) > \frac{K}{m}+\epsilon} B\left(\frac{K}{m}, e(h)\right)} > 1 - \frac{\alpha}{\delta}\right) \geq 1 - \delta$$

set $\alpha = \frac{\delta}{2}$ and rearranging we get:

$$\Rightarrow \Pr_{S\sim D^m}\left(2 \sum_{h\in H: e(h) > \frac{K}{m}+\epsilon \wedge e(h) \geq \underline{e}\left(\frac{K}{m}, \epsilon, \frac{\delta}{2}, \hat{e}_S(h)\right)} B\left(\frac{K}{m}, e(h)\right) > \sum_{h\in H: e(h) > \frac{K}{m}+\epsilon} B\left(\frac{K}{m}, e(h)\right)\right) \geq 1 - \delta$$

We can freely increase the sum on the left hand side to sum over all hypotheses since that just increases the inequality. Letting the sum range over all hypotheses, we get:

$$\Rightarrow \Pr_{S\sim D^m}\left(2 \sum_{h\in H} B\left(\frac{K}{m}, \underline{e}\left(\frac{K}{m}, \epsilon, \frac{\delta}{2}, \hat{e}_S(h)\right)\right) > P\left(H, \frac{K}{m}, \epsilon\right)\right) > 1 - \delta$$

$$\Rightarrow \Pr_{S\sim D^m}\left(\hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right) > P\left(H, \frac{K}{m}, \epsilon\right)\right) > 1 - \delta$$

which proves the lemma. $\square$

We now have all the tools required to prove the theorem.

Proof (of theorem 2.2) For all values of K and δ we have that with probability $1 - 2\delta$, when $\hat{e}(h) = \frac{K}{m}$, $\hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right) > P\left(H, \frac{K}{m}, \epsilon\right)$ *and* $e(h) < \frac{K}{m} + \min\left\{\epsilon : \; P\left(H, \frac{K}{m}, \epsilon\right) \leq \delta\right\}$. Using the conjunction, we get that:

$$e(h) < \frac{K}{m} + \min\left\{\epsilon : \; \hat{P}\left(H, \frac{K}{m}, \epsilon, \delta\right) \leq \delta\right\}$$

with probability $1 - 2\delta$. There are m possible nontrivial values of K so the union bound implies that if we substitute $\delta \to \frac{\delta}{2m}$, the result will hold for any observed empirical error, $\hat{e}(h) = \frac{K}{m}$. $\square$

3 Lower Bounds

Since the Shell bound incorporates so much extra information, it can be much tighter than other approaches. One interesting question to ask is: "How tightly can we bound the true error rate using the empirical error rate of all observed hypotheses?" Answering this question provides some idea how much improvement is feasible when we restrict ourselves to bounds which are functions of only empirical error rates across the hypothesis set.

What is a lower bound on the probability that some high true error rate hypothesis has an empirical error of $\frac{K}{m}$? It is easy to analyze this when the distribution of true errors is given.

Theorem 3.1 *(Lower Shell Bound) For all training errors, $\frac{K}{m}$ for all distributions of true error rates $\{e_1, ..., e_n\}$, there exists distributions D and hypothesis sets H such that:*

$$\Pr_{S\sim D^m}\left(\exists h \in H \ \frac{K}{m} = \hat{e}(h) \ \text{and} \ e(h) \geq \frac{K}{m} + \max\left\{\epsilon : \ P\left(H, \frac{K}{m}, \epsilon\right) \geq \delta\right\}\right) \geq \delta - \delta^2$$

Comparing the lower shell bound and the full knowledge lemma 2.1, two distinctions are evident. First, instead of taking a minimum over ϵ, we take a maximum. This distinction is simply related to the discretization introduced by the hypothesis space, and there are discrete value of δ for which the minimum of the full knowledge lemma *is* the maximum in the lower bound. The second distinction is that the disallowed event (a large true error rate) occurs with only probability $\delta - \delta^2$. This represents an improvement by several factors of m over the first shell bound [5].

This lower bound is somewhat more satisfactory than lower bounds for other upper bounds on the true error rate of a learned hypothesis because it holds for *every* distribution of true errors as well as every possible empirical error. It is more typical for lower bounds to hold only for every possible empirical error.

One important question remains: how does this bound compare to the observable shell bound 2.2? In the observable shell bound, the distribution of true errors is replaced with a pessimistic distribution which weakly lower bounds the distribution of true errors. The "size" of this pessimism in terms of the true error bound is equivalent to the size of the pessimism introduced by lower confidence interval on a *test* set of size m. Thus, the gap between the lower shell bound and the upper shell bound are approximately equivalent under the substitution $\delta \to \delta - \delta^2$ and $e_i \to e_i - \frac{1}{\sqrt{2m}}$.

Proof First, we specify the distribution and hypothesis set which is used to realize this claim. Let $X = \{0,1\}^n$ be the feature space and $Y = \{0,1\}$ be the output space. The hypothesis set, H, consists of n hypotheses which predict using the rule, $h_i(x) = x_i$ where i is the index of the vector. The distribution D picks Y with a uniform coin flip and then for each feature flips a coin with bias e_i. A "heads" corresponds to an error, $x_i \neq y$ and a "tails" corresponds to a success, $x_i = y$. By construction, this distribution has the right spectrum of true error rates to satisfy the theorem.

At the optimal value of ϵ, we have:

$$P\left(H, \hat{e}_S(h), \epsilon\right) = \sum_{h \in H : e(h) > \frac{K}{m} + \epsilon} B\left(\frac{K}{m}, e(h)\right) = \delta$$

Since every hypothesis errs independently, the probability that some large true error hypothesis has an empirical error rate of $\frac{K}{m}$ is

$$1 - \prod_{i=1}^{n} \left(1 - B\left(\frac{K}{m}, e(h)\right)\right) \geq \delta - \delta^2$$

□

4 Discussion

Shell bounds can provide large improvements when we can afford to enumerate the hypotheses. One obvious scenario suggests itself—in applied machine learning, it is common to divide the data into three sets: a training set (used by the learning algorithm), a validation set (used for parameter tuning), and a test set (used to evaluate the results). A shell bound could naturally be applied to the validation set, impling that the test set is unnecessary (and so the examples might be added to either the validation set or training set).

4.1 An Example

It is worthwhile to work out a simple example of the application of shell bounds. Suppose that we have a training set of size 100 and observe[2]:

1. 1 hypothesis with a training errror of 0.
2. e^5 hypotheses with a training error of 25.
3. e^{30} hypotheses with a training error of 50.

A shell bound optimization provides a true error bound of $\bar{e} = 0.084$.

In comparison, we might apply a more typical training set bound. To make a fair comparison, suppose that every hypothesis has the same "prior" in an Occam's Razor bound sense[1]. Even with a direct calculation Binomial coefficients (no looseness due to Hoeffding or Chernoff approximations), we get a true error bound of $\bar{e} = 0.268$. A program [4] has been implemented to perform these calculations.

In general, shell bounds have superior performance when the true error rate of most of the hypotheses is much larger than the true error rate of a learned hypothesis. This is typical behavior for some learning problems [8]. When this behavior is not satisfied, a shell bound degrades gracefully to something slightly worse than the Occam's Razor bound with a uniform prior[3].

4.2 Asymptotics

We focused on error bounds for finite sample sizes, but of course we can consider asymptotics. There are two forms of asymptotic limits. For the following, assume that we are using the learning algorithm which picks the hypothesis with the minimal training error.

1. Fix the hypothesis space and take the limit as the number of samples goes to infinity. In this case, the shell bound approximately converges to an Occam's Razor bound over the set of hypotheses with the smallest true

[2] This choice of observations is nontypical, but it is simple to describe and the behavior is typical.

[3] It is possible to work out shell bounds with a nonuniform prior instead.

error rate. If there is only one such hypothesis, the resulting behavior is similar to a *test set* bound with a number of examples m given by the train set. This is, of course, radically tighter than more standard approaches.

2. Increase the hypothesis space size as the number of examples grows, keeping the fraction $\frac{\log |H|}{m}$ constant. In this "local limit", the shell bound may provide a large confidence interval even in the limit as $m \to \infty$. In this limit, as the value of δ (the confidence bound) varies, discontinuous jumps in the true error bound may result. This behavior is "correct" because when hypotheses with a large true error rate (and large empirical errors) are excluded, only those hypotheses with a small true error rate could be confused with the hypoothesis having the smallest true error. Note that these discontinuous jumps in the bound value imply that small improvements in shell bound technology can sometimes result in dramatic decreases in resulting true error bounds.

4.3 Connection with Localized Rademacher Complexity

Localized Rademacher Complexity [10] has been another recent topic of study in learning theory. In Localized Rademacher Complexity, a refined analysis is made taking into account the "size" of subsets of the hypotheses with similar errors. This is similar to the way that shell bounds improves on the Discrete Hypothesis bound.

References

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

[2] David Haussler, Michael Kearns, H. Sebastian Seung, and Naftali Tishby, "Rigorous Learning Curve Bounds from Statistical Mechanics" Machine Learning,25, 1996, pages 195–236.

[3] John Langford, "Quantitatively Tight Sample Complexity Bounds" Thesis, Carnegie Mellon University 2002.

[4] John Langford, Bound calculation program, http://www-2.cs.cmu.edu/~jcl/programs/bound/bound.html

[5] John Langford and David McAllester, "Computable Shell Decomposition Bounds" COLT 2000.

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

[7] Colin McDiarmid, "On the method of bounded differences", In \emph {Surveys in Combinatorics 1989,} pages 148-188. Cambridge University Press, 1989.

[8] Tobias Scheffer. "Average-Case Analysis of Classification Algorithms for Boolean Functions and Decision Trees" Proceedings of the Eleventh International Conference on Algorithmic Learning Theory, 2000.

[9] Tobias Scheffer and Thorsten Joachims, "Expected Error Analysis for Model Selection", ICML 1999.

[10] P. L. Bartlett, O. Bousquet, S. Mendelson, Localized Rademacher complexity, to be presented at the Conference on Computational Learning Theory, Sydney, July 2002.