Report Number: WUCS-92-33

1992-09-22

On the Sample Complexity of Weakly Learning

Sally A. Goldman, Michael J. Kearns, and Robert E. Schapire

In this paper, we study the sample complexity of weak learning. That is, we ask how much data must be collected from an unknown distribution in order to extract a small but significant advantage in prediction. We show that it is important to distinguish between those learning algorithms that output deterministic hypotheses and those that output randomized hypotheses. We prove that in the weak learning model, any algorithm using deterministic hypotheses to weakly learn a class of Vapnik-Chervonenkis dimension d(n) requires (omega) (square root of d(n) examples.

On the Sample Complexity of Weakly Learning

Sally A. Goldman, Michael J. Kearns and Robert E. Schapire

WUCS-92-33

September, 1992

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
St. Louis MO 63130-4899

On the Sample Complexity of Weakly Learning

Sally A. Goldman
Department of Computer Science
Washington University
St. Louis, MO 63130
sg@cs.wustl.edu

Michael J. Kearns
AT&T Bell Laboratories
Murray Hill, NJ 07974
mkearns@research.att.com

Robert E. Schapire
AT&T Bell Laboratories
Murray Hill, NJ 07974
schapire@research.att.com

WUCS-92-33

September 22, 1992

Abstract

In this paper, we study the sample complexity of weak learning. That is, we ask how much data must be collected from an unknown distribution in order to extract a small but significant advantage in prediction. We show that it is important to distinguish between those learning algorithms that output *deterministic hypotheses* and those that output *randomized hypotheses.* We prove that in the weak learning model, any algorithm using deterministic hypotheses to weakly learn a class of Vapnik-Chervonenkis dimension $d(n)$ requires $\Omega(\sqrt{d(n)})$ examples. In contrast, when randomized hypotheses are allowed, we show that $\Theta(1)$ examples suffice in some cases. We then show that there exists an efficient algorithm using deterministic hypotheses that weakly learns against *any* distribution on a set of size $d(n)$ with only $O(d(n)^{2/3})$ examples. Thus for the class of symmetric Boolean functions over n variables, where the strong learning sample complexity is $\Theta(n)$, the sample complexity for weak learning using deterministic hypotheses is $\Omega(\sqrt{n})$ and $O(n^{2/3})$, and the sample complexity for weak learning using randomized hypotheses is $\Theta(1)$. Next we prove the existence of classes for which the distribution-free sample size required to obtain a slight advantage in prediction over random guessing is essentially equal to that required to obtain arbitrary accuracy. Finally, for a class of small circuits, namely all parity functions of subsets of n Boolean variables, we prove a weak learning sample complexity of $\Theta(n)$. This bound holds even if the weak learning algorithm is allowed to replace random sampling with membership queries, and the target distribution is uniform on $\{0,1\}^n$.

Most of this research was carried out while all three authors were at MIT Laboratory for Computer Science with support provided by ARO Grant DAAL03-86-K-0171, DARPA Contract N00014-89-J-1988, NSF Grant CCR-88914428, and a grant from the Siemens Corporation. S. Goldman is currently supported in part by a G.E. Foundation Junior Faculty Grant and NSF Grant CCR-9110108.

1 Introduction

In this paper, we study the sample complexity of weak learning. More precisely, we are interested in the number of examples required for the distribution-free learning of a parameterized concept class C over $\{0,1\}^n$ when the hypothesis output by the learning algorithm need only have accuracy $\frac{1}{2} + \frac{1}{p(n)}$ for some polynomial $p(n)$. Thus, the hypothesis must perform only slightly better than random guessing. Viewed more fundamentally, we are asking how much data must be collected from an unknown distribution in order to extract a small but significant advantage in prediction. This *weak learning* model is derived from the distribution-free "probably approximately correct" (or PAC) model introduced by Valiant, in which the learning algorithm must output a hypothesis with accuracy $1 - \epsilon$ for any small $0 < \epsilon \leq 1/2$. We refer to Valiant's original model as *strong learning.*

Our motivation for studying the sample complexity of weak learning comes from several sources. First, in the strong learning model it is assumed that learning algorithms have access to an unlimited supply of labeled examples drawn according to the unknown target distribution. Given this unlimited supply of examples, the goal of a learning algorithm is to discover almost *all* information about the target concept with respect to the target distribution (i.e., to be able to correctly classify all but a fraction ϵ of the examples with respect to the target distribution). While much of the research in the strong learning model has aimed at achieving this goal in polynomial time, many results have addressed the question of the number of examples required.

In practice, however, we often find that there is a limited supply of examples. Research involving archeological evidence or protein sequences are typical settings in which the available data are severely limited. Furthermore, in such settings one is rarely expecting to obtain a highly accurate theory explaining all the evidence; indeed, a theory that provides even the slightest bias may provide valuable clues and guidance for further investigations. Thus we are motivated to ask, what is the minimum number of examples required to obtain *some* information about the target concept? An understanding of weak learning sample complexity may be important in applications in which the number of available examples falls short of the number required to obtain overwhelming accuracy in prediction, but suffices to obtain a significant advantage over guessing.

A second motivation for our study is the recent result of Schapire [14] showing that a concept class is weakly learnable in polynomial time if and only if it is strongly learnable in polynomial time. Are the sample sizes required for weak learning and strong learning always polynomially related? Some of our results give a negative answer to this question, and we investigate conditions under which the weak learning sample complexity is significantly smaller than the strong learning sample complexity.

A third motivation is that the nature of the weak learning model forces us to find distributions with large support sets in order to prove good lower bounds on sample size. One objection to the sample size lower bounds in the strong learning model is that these bounds are typically obtained for a distribution over a small support set. Since, as our results will show, such lower bounds break down for the weak learning model, we must look for hard distributions over large support sets, such as the uniform distribution. In addition to involving what are perhaps more natural distributions, these results may be of some interest to researchers in cryptography, where one is often interested in functions that are unpredictable (in the weak learning sense) on the uniform distribution. Whereas cryptography has been primarily and naturally interested in functions that are unpredictable in a computationally bounded setting (such as quadratic residues), some of our results may be interpreted as an investigation of this same problem in an information-theoretic setting.

We now give a summary of our results. Although our lower bounds on weak learning sample size are information-theoretic (that is, they hold regardless of computation time), we are primarily concerned with polynomial-time learning, and all example-efficient algorithms we give run in polynomial time. We begin by observing that if the Vapnik-Chervonenkis dimension of a concept class C_n is super-polynomial in n, then the lower bound proofs for the strong learning model [2] are easily adapted to give super-polynomial lower bounds on the sample size required for weak learning. Thus, we focus on classes C_n whose Vapnik-Chervonenkis dimension is polynomial in n.

We note that the sample size lower bound for the strong learning model breaks down for the weak learning model: namely, if a class C_n has Vapnik-Chervonenkis dimension polynomial in n, and the target distribution is uniform over a shattered set, then *one* example suffices to obtain a weak learning hypothesis. The hypothesis uses the obvious technique of correctly classifying the known point, and flipping a coin for the classification of any other point. This simple hypothesis is *randomized* (this should not be confused with the learning algorithm itself being randomized, which we always assume may be the case).

This example raises the natural question of the relative power of deterministic hypotheses and randomized hypotheses. In the strong learning model, the sample size lower bounds hold regardless of whether the hypothesis is deterministic or randomized. However, we show that in the weak learning model it is important to distinguish between those learning algorithms that output *deterministic hypotheses* and those that output *randomized hypotheses*. Namely, we prove that in the weak learning model, any algorithm using deterministic hypotheses to learn a class of Vapnik-Chervonenkis dimension $d(n)$ requires $\Omega(\sqrt{d(n)})$ examples; the hard distribution is again uniform over a shattered set. We then give an efficient algorithm using deterministic hypotheses that weakly learns against *any* distribution on a shattered set (or more generally, any distribution on any set of size $d(n)$, which we assume is polynomial in n) with only $O(d(n)^{2/3})$ examples. This is a provable decrease from the number of examples required for strong learning against the same class of distributions. The algorithm uses a simple sampling technique for converting any weak learning algorithm using randomized hypotheses into one using deterministic hypotheses.

Furthermore, for some classes, such as symmetric functions over $\{0,1\}^n$, any distribution can be reduced to a distribution over a shattered set. Thus, for symmetric functions we obtain an interesting separation of the sample sizes required in the various distribution-free settings: the strong learning sample size is $\Theta(n)$, the sample size required for weak learning with deterministic hypotheses is $\Omega(\sqrt{n})$ and $O(n^{2/3})$, and the sample size required for weak learning with randomized hypotheses is $\Theta(1)$. These bounds are given for fixed ϵ and δ; the dependence on these parameters is described in the technical sections.

These results show that the sample complexity for weak learning may be considerably smaller than for strong learning, and that the power of using randomized hypotheses for weak learning may be dramatic. The results so far leave open the possibility that any concept class of polynomial Vapnik-Chervonenkis dimension can be weakly learned using randomized hypotheses with only a *constant* number of examples (for fixed δ).

We show that this is not the case by proving the existence of classes C_n whose Vapnik-Chervonenkis dimension is $\Theta(n)$ and whose weak learning sample complexity is $\Theta(n)$ (regardless of the hypotheses used). In contrast to the results described above, this shows that there are classes for which the distribution-free sample size required to obtain a slight advantage in prediction over random guessing is essentially the same as that required to obtain arbitrary accuracy. However, we use a probabilistic construction to obtain this result, and the resulting class C_n, while having small Vapnik-Chervonenkis dimension, does not have small (size polynomial in n) circuits, and thus is not learnable in polynomial time by results of Schapire [14]. Are there classes of small circuits, learnable in polynomial time, whose weak learning sample complexity is as large as their strong learning sample complexity?

By defining a combinatorial property of concept classes that is sufficient to imply large weak learning sample complexity, and then demonstrating a class of small circuits possessing this property, we are able to answer this question in the affirmative. The class of circuits is simply all parity functions of subsets of n Boolean variables, which we prove has weak learning sample complexity $\Theta(n)$. We show that this holds even if the weak learning algorithm is allowed to choose the examples itself (that is, the learning algorithm may replace random sampling with membership queries), and the target distribution is uniform.

The sufficient property used is a first step towards characterizing weak learning sample complexity in the same way that the Vapnik-Chervonenkis dimension gives a combinatorial characterization of strong learning sample complexity. A necessary and sufficient characterization of weak learning sample complexity remains an interesting open problem.

2 Definitions

We begin by describing the distribution-free learning model introduced by Valiant [16]. The learner is attempting to infer an unknown *target concept* c chosen from some known *concept class* C. In this paper, $C = \bigcup_{n \geq 1} C_n$ is parameterized by the number of variables n, and each $c \in C_n$ is a subset of the domain $\{0,1\}^n$. The learner is given access to labeled (positive and negative) examples of the target concept, drawn randomly according to some unknown *target distribution* D over $\{0,1\}^n$. The learner is also given as input $0 < \epsilon, \delta < 1$. The learner's goal is to output with probability at least $1 - \delta$ a *hypothesis* h that has probability at most ϵ of disagreeing with c on a randomly drawn example from D (thus, the hypothesis has *accuracy* at least $1 - \epsilon$, or is ϵ*-good*). If such a learning algorithm A exists (that is, an algorithm A meeting the goal for any $n \geq 1$, any target concept $c \in C_n$, any target distribution D, and any ϵ, δ), we say that C is *strongly learnable in the distribution-free model.* In this setting polynomial time means polynomial in n, $1/\epsilon$ and $1/\delta$. The *support set* of a distribution D is the set of all x such that $D(x) > 0$.

In the related *weak learning* model [12], we drop the demand for accuracy $1-\epsilon$ and simply ask that the hypothesis h have accuracy at least $\frac{1}{2} + \frac{1}{p(n)}$ for some polynomial $p(n)$. Thus we ask only for a small correlation in the underlying distribution. In this setting polynomial time means polynomial in n and $1/\delta$. The *weak sample complexity* for a parameterized concept class C is a function of n and δ that denotes the minimum number of examples required to weakly learn any $c \in C_n$.

We will see shortly that it is important to distinguish between the cases where the learning algorithm A outputs deterministic and randomized hypotheses. This should not be confused with the learning algorithm itself, which we always assume may be randomized. A *deterministic hypothesis* over $\{0,1\}^n$ is a function $h : \{0,1\}^n \to \{0,1\}$. A *randomized hypothesis* over $\{0,1\}^n$ is a function $h : \{0,1\}^n \times \{0,1\}^{p(n)} \to \{0,1\}$, where $p(n)$ is some fixed polynomial. On input $x \in \{0,1\}^n$, the randomized hypothesis h is evaluated by choosing a random string $r \in \{0,1\}^{p(n)}$ uniformly and then computing $h(x,r)$. Here, the accuracy of h with respect to the target distribution is the probability of agreement with the target, where the probability is now taken over both the random draw of $x \in \{0,1\}^n$ according to D and the random string r.

We also need the following definitions. A finite set $Y \subseteq \{0,1\}^n$ is *shattered* by C_n if we have $\{c \cap Y \mid c \in C_n\} = 2^Y$. The *Vapnik-Chervonenkis dimension* of C_n, denoted $\mathrm{VCD}(C_n)$, is defined to be the largest d such that some set of cardinality d is shattered by C_n.

Finally, to compute the sample sizes needed for several of our algorithms we use the following versions of Chernoff bounds. The first bound stated, Hoeffding's inequality [11], will be used whenever $p \geq 1/4$. However, when $p < 1/4$ the last two bounds as stated by Angluin and Valiant [1] give better bounds. (See also Chernoff [3], and Erdös and Spencer [5].)

Lemma 1 (Chernoff Bounds) *Let $X_1, \ldots, X_m$ be a sequence of m independent Bernoulli trials, each succeeding with probability p. Let $S = X_1 + \cdots + X_m$ be the random variable describing the total number of successes. Then for $\beta \le p$ and $0 \le \gamma \le 1$, the following inequalities hold:*

$$\begin{aligned} \Pr[S \le \beta m] &\le e^{-2m(p-\beta)^2} \\ \Pr[S \le mp(1-\gamma)] &\le e^{-\gamma^2 mp/2} \\ \Pr[S \ge mp(1+\gamma)] &\le e^{-\gamma^2 mp/3} \end{aligned}$$

3 Previous Work

In the strong learning model, a major contribution to the understanding of sample complexity was made by Blumer et al. [2]. Building on the work of Vapnik and Chervonenkis [17], they proved that the number of examples required for strongly learning a concept class C_n is $\Omega(\text{VCD}(C_n))$ (ignoring dependence on ϵ and δ). Furthermore, they prove that the general technique of finding a consistent hypothesis, when feasible, always results in a (possibly super-polynomial time) learning algorithm using $O(\text{VCD}(C_n))$ examples. Thus, for strong learning the sample complexity is characterized by the Vapnik-Chervonenkis dimension.

In the weak learning model there are no previous lower bounds on sample size, and the only upper bounds are those already provided by results in the strong learning model. However, in the case that $\text{VCD}(C_n)$ is super-polynomial in n, it is easy to adapt the lower bound of Blumer et al. to give super-polynomial lower bounds on the sample size for weak learning. Since we are primarily concerned with classes learnable from a polynomial number of examples in polynomial time, we restrict our attention to classes with dimension polynomial in n.

4 Simple Bounds

In this section we look at two initial results on the sample complexity of weak learning. In the polynomial-time setting, Schapire [14] proved that a concept class C can be weakly learned in polynomial time if and only if it can be strongly learned in polynomial time. More precisely, he gives an efficient strong learning algorithm for C that uses an efficient weak learning algorithm for C as a subroutine. Subsequently, Freund [7, 8] has given a different technique for converting a weak learning algorithm into a strong learning algorithm. Combining this result with the lower bound provided by Blumer et al., one obtains an initial lower bound on weak learning sample complexity. This bound does not give an unconditional lower bound on the sample size required by any weak learning algorithm, but instead describes a tradeoff between the advantage obtained and the number of examples required.

Theorem 2 *Let C be a parameterized concept class, let $p(n)$ be a polynomial, and let $d(n) = \mathrm{VCD}(C_n)$. Then any weak learning algorithm that outputs deterministic hypotheses of accuracy $\frac{1}{2} + \frac{1}{p(n)}$ must use*

$$\Omega\left(\frac{d(n)}{p(n)^2(\log d(n))^2}\right)$$

examples whenever $\delta \leq 1/2$.

Proof: We prove this result by showing that a weak learning algorithm that violates this lower bound can be used to *compress* data beyond what is information-theoretically possible.

Fix $\delta = 1/2$ and let A be a weak learning algorithm that outputs $(\frac{1}{2} - \frac{1}{p(n)})$-good hypotheses with probability $1/2$, and that requires $m(n)$ examples. (By running A repeatedly and hypothesis testing, the probability δ of failing to weakly learn can be made arbitrarily small. This technique is described by Haussler et al. [9].) Note that the output hypothesis can be encoded by the $m(n)$ examples on which A was successfully trained. Under this encoding, the size $s(n)$ of the output hypothesis in bits is $m(n)$ times the number of bits needed to encode each example.

Schapire [14] and Freund [7, 8] describe techniques for converting this weak learning algorithm into a strong learning algorithm A' outputting hypotheses of size

$$O\left(s(n) \cdot (p(n))^\alpha \cdot (\log(1/\epsilon))^\beta\right) \tag{1}$$

for some constants α and β. If A' is run against a uniform distribution over a shattered set of size $d(n)$ with $\epsilon < 1/d(n)$, then the output hypothesis is consistent with the sample with high probability. Since each example in the shattered set can be encoded by $O(\log d(n))$ bits it follows from the above that $s(n) = O\left(m(n) \log d(n)\right)$. Substituting this bound as well as the bound $1/\epsilon = O(d(n))$ into Equation (1) we see that the size of the hypothesis output by A' is $O\left(m(n) \cdot (p(n))^\alpha \cdot (\log d(n))^{\beta+1}\right)$. Finally, since all $2^{d(n)}$ labelings of the instances in the shattered set are possible, it is clear that at least $d(n)$ bits are needed to encode these labelings, and thus $d(n)$ lower bounds the size of the hypothesis output by A'. Thus,

$$d(n) \leq O\left(m(n) \cdot (p(n))^\alpha \cdot (\log d(n))^{\beta+1}\right).$$

Since, in Freund's construction, $\alpha = 2$ and $\beta = 1$, the stated lower bound on $m(n)$ follows. ■

We now demonstrate that for concept classes with polynomial Vapnik-Chervonenkis dimension, the lower bound of Blumer et al. [2] breaks down in the weak learning model. If $\mathrm{VCD}(C_n)$ is polynomial in n and the target distribution is over a shattered set, then $O(\log(1/\delta))$ examples suffice for weak learning.

Theorem 3 *Let C be a parameterized concept class and let $p(n)$ be a polynomial. Then there exists an algorithm outputting a randomized hypothesis with accuracy $\frac{1}{2} + \frac{1}{p(n)}$ on any target distribution with a support set of cardinality $d(n)$; the number of examples required is*

$$O\left(\frac{d(n)}{p(n)} + \log(1/\delta)\right).$$

Proof: The algorithm draws enough points so the weight of the points in the sample cover at least a fraction β of the distribution. The output hypothesis h correctly classifies the seen points, and flips a fair coin elsewhere. Thus the error of h is at most $(1-\beta)/2$. To insure that the error is at most $\frac{1}{2} - \frac{1}{p(n)}$, it suffices to select $\beta = 2/p(n)$.

We use Hoeffding's inequality to prove that a sample of size $O(\beta d(n) + \log(1/\delta))$ covers a fraction β of the distribution with probability at least $1-\delta$. Since we have shown that $\beta = 2/p(n)$ suffices, without loss of generality assume that $\beta \leq 1/3$. If at least 1/3 of the distribution is covered, then we are done. Suppose instead that fewer than 1/3 of the distribution has been covered. Thus when drawing a new example x from D:

1. $\Pr[x \text{ is already covered}] \leq 1/3$
2. $\Pr[x \text{ is new point with weight} \leq 1/3d(n)] \leq 1/3$
3. $\Pr[x \text{ is new point with weight} \geq 1/3d(n)] \geq 1/3$

We say that a trial is *successful* if x is a new point with weight at least $1/3d(n)$. Thus after $3d(n)\beta$ successful trials a fraction β of the distribution will be covered. Using Hoeffding's inequality with $p = 1/3$ and $\beta = 1/6$ it can easily be shown that a sample of size $\max\{18d(n)\beta, 18\ln 1/\delta\}$ is sufficient to ensure that with probability $1-\delta$ the number of successful trials is at least $3d(n)\beta$. Finally, substituting $2/p(n)$ for β gives the desired result. ∎

By setting $p(n) = d(n) = \text{VCD}(C_n)$ in Theorem 3, we obtain:

Corollary 4 *Let C be a parameterized concept class over $\{0,1\}^n$ for which $\text{VCD}(C_n)$ is polynomial in n. Then there exists an algorithm outputting a randomized hypothesis that weakly learns C_n on any distribution over a set of cardinality $\text{VCD}(C_n)$; the number of examples required is $O(\log(1/\delta))$.*

Thus, for fixed δ, $O(1)$ examples suffice for weak learning against target distributions over small support sets; this should be contrasted with the lower bound of $\Omega(\text{VCD}(C_n))$ for the same class of distributions in the strong learning model [2]. In Section 6 we show that for the weak learning model, randomized hypotheses are *necessary* to obtain such significant decreases in sample complexity.

5 Removing Randomness from Hypotheses

In this section we give a sampling technique for converting randomized hypotheses into deterministic hypotheses in both the strong and weak learning models. If computation time is not a concern, then in the strong learning model randomized and deterministic hypothesis classes give essentially the same power with respect to sample complexity (this follows from the results of Blumer et al.). We extend this result to hold even when considering computation time: we describe a technique to *efficiently* convert any randomized hypothesis into a deterministic hypothesis using $O((1/\epsilon)\log(1/\delta))$ additional examples.

We use the following definitions in the next two theorems. Given a randomized hypothesis h, let $h(x,r)$ be the prediction made by h on instance x with random bits r. We define the error of h on random bits r as $e_h(r) = \Pr_x[h(x,r) \neq c(x)]$ where $c(x)$ is the correct classification for x. Likewise for a deterministic hypothesis h and a sample S drawn randomly from D, let $e_h = \Pr_x[h(x) \neq c(x)]$ and let $\hat{e}_h(S)$ denote the estimated error of hypothesis h based on sample S. That is, $\hat{e}_h(S) =$ (number of misclassified examples from $S)/|S|$.

Theorem 5 *Let A be a strong learning algorithm for a parameterized class C that outputs a randomized hypothesis and requires $m(n,\epsilon,\delta)$ examples. Then there exists a strong learning algorithm A' for C that outputs a deterministic hypothesis and requires*

$$O\Big(m(n,\epsilon,\delta) + \frac{1}{\epsilon}\log(1/\delta)\Big)$$

examples.

Proof: We begin by running algorithm A (with parameters $\epsilon/4$ and $\delta/2$) once to obtain a single randomized hypothesis h, that with probability at least $1-\delta/2$, has error at most $\epsilon/4$. It is easily shown that

$$\Pr_{x,r}[h(x,r) \neq c(x)] = \mathrm{E}_r[e_h(r)] \leq \frac{\epsilon}{4}.$$

Let $q = \Pr_r[e_h(r) \geq \epsilon/2]$. Since $\mathrm{E}_r[e_h(r)] \geq \epsilon q/2$ it follows that $q \leq 1/2$ and thus

$$\Pr_r\left[e_h(r) < \frac{\epsilon}{2}\right] \geq 1/2. \qquad (2)$$

We are now ready to describe the technique for converting the randomized hypothesis into a deterministic one. We choose t random strings $r_1, \ldots, r_t$ to obtain t *deterministic hypotheses* $h_i = h(\cdot, r_i)$. It follows from Equation (2) that

$$\Pr\left[\text{all } h_i\text{'s have error} > \frac{\epsilon}{2}\right] \leq 2^{-t}.$$

So for $t = \lg(6/\delta)$, with probability $1-\delta/6$ at least one of the h_i's will have error at most $\epsilon/2$.

Next we use hypothesis testing (as described by Haussler et al. [9]) to estimate the error of each hypothesis and output the one with the lowest error. For hypothesis h_i, if $e_{h_i} \geq \epsilon$ then Chernoff bounds can be used to show that if a sample S of size $(8/\epsilon)\ln(6t/\delta)$ is drawn then $\Pr[\hat{e}_{h_i}(S) \leq \epsilon/2] \leq \delta/6t$. Since at most t such estimates are made, with probability at least $1-\delta/6$, for any hypothesis h_i with $e_{h_i} > \epsilon$, the estimated error $\hat{e}_{h_i}(S) > \epsilon/2$.

Likewise, Chernoff bounds can be used to show that if a sample S of size $(12/\epsilon)\ln(6t/\delta)$ is drawn then with probability $1-\delta/6$, for any hypothesis h_i with $e_{h_i} < \epsilon/4$, the estimated error $\hat{e}_{h_i}(S) < \epsilon/2$. Thus by drawing an additional sample of size

$$\frac{12}{\epsilon}\ln\frac{12t}{\delta} = \frac{12}{\epsilon}\left(\ln\frac{12}{\delta} + \ln\lg\frac{6}{\delta}\right) = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$$

we can ensure with probability at least $1-\delta$ that the hypothesis output by A' has error at most ϵ. ■

Thus for the case of strong learning the distinction between deterministic and randomized hypothesis spaces is not significant. Next we give a similar conversion for the weak learning model, but the increase in sample complexity is now significant. This result will be used in the next section to obtain improved sample sizes for weak learning with deterministic hypotheses.

Theorem 6 *Let C be a parameterized concept class and let $p(n)$ be a polynomial. Let A be a weak learning algorithm for C that outputs a randomized hypothesis of accuracy $\frac{1}{2}+\frac{1}{p(n)}$ and that requires $m(n,\delta)$ examples. Then there exists a weak learning algorithm A' for C that outputs a deterministic hypothesis and that requires $O(m(n,\delta)+p(n)^2\log(p(n)/\delta))$ examples.*

Proof: As in the proof of Theorem 5 we begin by running algorithm A once to obtain a single randomized hypothesis h, that with probability at least $1-\delta/2$, has error at most $\frac{1}{2}-\frac{1}{p(n)}$. It is easily shown that

$$\Pr_{x,r}[h(x,r) \neq c(x)] = \mathrm{E}_r[e_h(r)] \leq \frac{1}{2} - \frac{1}{p(n)}.$$

Let $q = \Pr_r[e_h(r) \geq \frac{1}{2} - \frac{1}{2p(n)}]$. Since $\mathrm{E}_r[e_h(r)] \geq q(\frac{1}{2}-\frac{1}{2p(n)})$ it follows that

$$\Pr_r\left[e_h(r) \leq \frac{1}{2} - \frac{1}{2p(n)}\right] \geq \frac{1}{p(n)-1}. \qquad (3)$$

As in the proof of Theorem 5 to convert the randomized hypothesis into a deterministic one, we choose t random strings $r_1,\ldots,r_t$ to obtain t deterministic hypotheses $h_i = h(\cdot, r_i)$. Using Equation (2) it is easily shown that for $t = (p(n)-1)\ln(6/\delta)$, with probability $1-\delta/6$ at least one of the h_i's will have error at most $\frac{1}{2}-\frac{1}{2p(n)}$.

Finally, we use hypothesis testing to accurately estimate the error of each hypothesis and output the one with the lowest error. We want to draw enough examples so that the following two requirements are met for all h_i's:

1. If $e_{h_i} \geq \frac{1}{2} - \frac{1}{4p(n)}$, then $\Pr[\hat{e}_{h_i}(S) \leq \frac{1}{2} - \frac{1}{2p(n)}] \leq \delta/6t$.
2. If $e_{h_i} \leq \frac{1}{2} - \frac{1}{p(n)}$, then $\Pr[\hat{e}_{h_i}(S) \geq \frac{1}{2} - \frac{1}{2p(n)}] \leq \delta/6t$.

Using Hoeffding's inequality it can be shown that drawing a sample of size $8p(n)^2 \ln 6t/\delta$ is sufficient to ensure that with probability $1 - \delta/6$, the first requirement is met for all h_i's. Likewise by drawing a sample of size $2p(n)^2 \ln 6t/\delta$ we can ensure that with probability $1 - \delta/6$ the second requirement holds for all h_i's. Thus an additional sample of size

$$8p(n)^2 \ln \frac{6t}{\delta} = 8p(n)^2 \left(\ln \frac{6}{\delta} + \ln(p(n) - 1) + \ln \ln \frac{6}{\delta}\right) = O\left(p(n)^2 \log \frac{p(n)}{\delta}\right)$$

is sufficiently large so that with probability at least $1 - \delta$ the hypothesis output by A' has error at most $\frac{1}{2} - \frac{1}{4p(n)}$. ∎

As we shall see, often when designing an algorithm with a randomized hypothesis, only a single random bit is needed. If the hypothesis output by A only requires a constant number of random bits, then only a constant number of hypotheses need to be generated. Thus in Theorem 6, $t = O(1)$ giving the following corollary.

Corollary 7 *Let C be a parameterized concept class and let $p(n)$ be a polynomial. Let A be a weak learning algorithm for C that outputs a randomized hypothesis of accuracy $\frac{1}{2} + \frac{1}{p(n)}$ and that requires $m(n, \delta)$ examples. Furthermore, suppose that h requires a constant number of random bits. Then there exists a weak learning algorithm A' for C that outputs a deterministic hypothesis and that requires $O(m(n, \delta) + p(n)^2 \log(1/\delta))$ examples.*

6 Deterministic Hypotheses for Weak Learning

In this section we consider the weak sample complexity when using deterministic hypotheses. We begin by showing that any weak learning algorithm for a parameterized concept class C using deterministic hypotheses requires $\Omega(\sqrt{\mathrm{VCD}(C_n)})$ examples.

Theorem 8 *Let C be a parameterized concept class. Then the sample size required for weakly learning C_n using deterministic hypotheses is $\Omega(\sqrt{\mathrm{VCD}(C_n)})$ for any $\delta \leq \delta_0$, where $0 < \delta_0 < 1$ is a constant.*

Proof: Let $d(n) = \mathrm{VCD}(C_n)$, and let A be a weak learning algorithm for C that outputs a deterministic hypothesis. For each $c \in C_n$, let the target distribution D be uniform over a shattered set T of size $d(n)$. Let $C'_n \subseteq C_n$ be such that C'_n shatters T and $|C'_n| = 2^{d(n)}$ (thus, there is exactly one concept in C' for each induced labeling of T).

Consider the following experiment: first the target concept c is chosen uniformly at random from C'_n. Then a sample S of $\sqrt{d(n)}$ points labeled according to c is chosen from the target distribution D and is given to A. The outcome of the experiment is the accuracy of the deterministic hypothesis output by A.

This experiment is easily seen to be equivalent to the following one: first a sample S of $\sqrt{d(n)}$ points is chosen randomly from T and is randomly labeled. Then the target c is chosen randomly among all concepts in C' consistent with the chosen labeling. Then the labeled sample is given to S, and the accuracy of the hypothesis output by A is measured. Now since the hypothesis of A is chosen independently from the random choice of c, this experiment is equivalent to the following: first a sample S of $\sqrt{d(n)}$ points is chosen randomly from T and is randomly labeled. Then the labeled sample S is given to A, and the deterministic hypothesis h of A is obtained. Then a target concept c is chosen randomly from among all concepts in C'_n consistent with S.

We assume, without loss of generality, that h makes no errors on the $\sqrt{d(n)}$ points in the sample S. It can be seen that the accuracy of h on D exceeds $1/2$ only if h is incorrect on at most $d(n)/2$ of the $d(n) - \sqrt{d(n)}$ points of $T - S$. However, we may regard the random draw of c in the third description of the experiment above as a sequence of unbiased coin flips, since each possible labeling of the points in $T - S$ is represented exactly once in C'_n. But the probability that at least $d(n)/2$ tails occur in a sequence of $d(n) - \sqrt{d(n)}$ coin flips is at least δ_0 for some constant $0 < \delta_0 < 1$ (for example, see Feller [6]). Letting tails represent points in $T - S$ on which h is incorrect, and applying an averaging argument, we see that there must exist some $c \in C'_n$ for which A has probability at least δ_0 of failing to output a hypothesis of accuracy $1/2$ on D. ■

We now show that for fixed δ, the bound of Theorem 8 is tight on the uniform distribution over a shattered set. Thus if the result of Theorem 8 is to be improved, a different distribution must be used.

Theorem 9 *Let C be a parameterized concept class, and let $d(n)$ be a polynomial. Then there exists an efficient algorithm that weakly learns C_n against the uniform distribution on any set of cardinality $d(n)$; the number of examples required is $O\left(\sqrt{d(n)}\log(1/\delta) + \log(1/\delta)\right)$.*

Proof: The algorithm is simple. First draw a large enough sample so that with probability at least $1 - \delta/2$ this sample will include $s(n)$ distinct points from the support set. Using Hoeffding's inequality (as in the proof of Theorem 3) it is easily shown that a sample of size $O(s(n) + \log(1/\delta))$ is sufficient to achieve this goal. The output hypothesis h will be constructed as follows. For each point in the sample, predict the known value. For all other points, the learning algorithm flips a fair coin to select the classification.

Thus we only need to determine how large to make $s(n)$ so that the accuracy of the hypothesis is at least $\frac{1}{2}+\frac{1}{d(n)}$. Let β denote the fraction of the $d(n)-s(n)$ unseen instances that are classified correctly by h. Then, to achieve a $1/d(n)$ advantage, we need that

$$\frac{s(n)+\beta(d(n)-s(n))}{d(n)} \geq \frac{1}{2}+\frac{1}{d(n)}.$$

Solving for β gives the requirement that

$$\beta \geq \frac{d(n)-2s(n)+2}{2(d(n)-s(n))}.$$

Finally, we use Hoeffding's inequality (with $m=d(n)-s(n)$, and $p=1/2$) to ensure that β is sufficiently large with probability at least $1-\delta/2$. This yields the following:

$$\exp\left\{\frac{-(s(n)-2)^2}{2(d(n)-s(n))}\right\} \leq \delta/2.$$

Thus choosing $s(n)=\sqrt{2d(n)\ln(1/\delta)}+2$ suffices. ■

We now wish to extend the upper bound of Theorem 9 to hold for *any* distribution on a shattered set. This is obtained by applying the conversion technique of Corollary 7 to the example-efficient algorithm of Theorem 3. The result is an efficient algorithm using deterministic hypotheses for learning any concept class of polynomial Vapnik-Chervonenkis dimension against any distribution on a set of size $\textsc{vcd}(C_n)$ using $O\Big(\textsc{vcd}(C_n)^{2/3}\log(1/\delta)\Big)$ examples:

Theorem 10 *Let C be a parameterized concept class such that $\textsc{vcd}(C_n)$ is polynomial in n. Then there exists an algorithm using deterministic hypotheses for weakly learning C_n against any distribution over a set of size $\textsc{vcd}(C_n)$; the number of examples required is $O\Big(\textsc{vcd}(C_n)^{2/3}\log(1/\delta)\Big)$.*

Proof: We apply the conversion technique of Corollary 7 to the algorithm of Theorem 3. In applying this conversion we get an interesting trade-off between hypothesis accuracy and sample complexity—the additional sample complexity needed for the conversion is reduced as the accuracy of the randomized hypothesis improves. Specifically, if $d(n)=\textsc{vcd}(C_n)$ then a sample of size $O(d(n)/p(n)+p(n)^2\log(1/\delta))$ is required to obtain a hypothesis with accuracy $\frac{1}{2}+\frac{1}{4p(n)}$. Letting $p(n)=d(n)^{1/3}$ we obtain the desired result. ■

Thus for any class C_n of polynomial Vapnik-Chervonenkis dimension, the strong learning sample complexity and the sample complexity for weak learning with deterministic hypotheses are always polynomially related; this follows from the results of Blumer et al. and Theorem 8. However, for any distribution on a set of size $\textsc{vcd}(C_n)$, the number of examples required for weakly learning C_n with a deterministic hypothesis is provably less than that required for strong learning; this follows from Blumer et al. and Theorem 10. For weak learning with randomized hypotheses, $O(\log(1/\delta))$ examples suffice for any distribution on a set of size $\mathrm{VCD}(C_n)$, a provable and significant decrease from the sample size for weak learning with deterministic hypotheses and for strong learning. For some classes of Boolean functions, such as symmetric functions, *any* distribution reduces to a distribution on a shattered set (symmetric functions are Boolean functions over $\{0,1\}^n$ whose output is invariant under all permutations of the input bits). Thus for symmetric functions we obtain a separation of the sample complexities for the various models.

Theorem 11 *Let C be the parameterized concept class of symmetric Boolean functions, and let $0 < \delta \leq 1/2$ be fixed. Then the sample size required for strongly learning C_n is $\Theta(n)$, the sample size required for weakly learning C_n with deterministic hypotheses is $\Omega(\sqrt{n})$ and $O(n^{2/3})$, and the sample size required for weakly learning C_n with randomized hypotheses is $\Theta(1)$.*

It is interesting to note that the algorithms for weak learning with randomized hypotheses all use a method of *localization* not available to a strong learning algorithm: a small set of examples is used to classify some local region of the domain. For symmetric functions, for instance, a single vector $\vec{v}$ can be used to correctly classify all those vectors with the same number of bits set to 1 as $\vec{v}$. The hypothesis output deterministically classifies this small region and flips a fair coin elsewhere. Thus, the hypothesis space used is actually considerably weaker in terms of representational power than the true target class. This should be contrasted with results showing that the *computational* complexity of learning can sometimes be reduced by using a hypothesis space that is *more* powerful than the target class (see for example Pitt and Valiant [13]).

7 Almost Every Class Has Weak Sample Complexity $\Omega(\mathrm{VCD}(C_n))$

We have seen that the power of using a randomized hypothesis may be dramatic in some cases for weak learning sample size. Our results thus far leave open the possibility that every concept class C_n over $\{0,1\}^n$ such that $\mathrm{VCD}(C_n)$ is polynomial in n can be weakly learned with only a constant number of examples (for fixed δ). The next theorem shows that this is not the case for almost every concept class of polynomial dimension.

Theorem 12 *For each $n \geq 1$, there is a parameterized class C of Boolean concepts over $\{0,1\}^n$ such that $\mathrm{VCD}(C_n) = \Theta(n)$, and the number of examples required for weak learning C_n (using either deterministic or randomized hypotheses) is $\Omega(n)$.*

Proof: The proof is a probabilistic construction showing that a randomly chosen concept class has the desired properties with overwhelming probability. From this we conclude that some fixed concept class C_n has the desired properties. Note that a weak learning algorithm A for C_n is given access to a complete description (truth table) of every concept in C_n. Thus the choice of a random target class is only for the purposes of *constructing* C_n in the proof; algorithm A is not being given examples of a "random" concept.

The class C_n we construct will consist of 2^{kn} randomly chosen Boolean concepts on $\{0,1\}^n$ for some constant $k \geq 1$; it follows that $\textsc{vcd}(C_n) \leq kn = O(n)$, and from the proof below it can be shown that $\textsc{vcd}(C_n) = \Omega(n)$.

Let S be any fixed set of n arbitrarily labeled examples from $\{0,1\}^n$. Now let $N = 2^n - n$, and let $T = \{0,1\}^n - S$. We think of S as the sample given to a learning algorithm, and T as those points not seen by the algorithm. With respect to the N points in T, any Boolean concept c is represented by characteristic vector $\vec{v}_c \in \{0,1\}^N$ on the N-dimensional Boolean hypercube and any randomized hypothesis h is represented by a vector $\vec{v}_h \in [0,1]^N$ in the N-dimensional real cube. In both cases we regard the ith components $(\vec{v}_c)_i$ and $(\vec{v}_h)_i$ as the probability that 1 is output when the input is the ith point of T. For the moment we are concerned only with behavior on the set T, and equate concepts and randomized hypotheses over T with these characteristic vectors.

We now define a distance measure between concepts and randomized hypotheses by

$$d_N(\vec{v}_c, \vec{v}_h) = \frac{\sum_{i=1}^N |(\vec{v}_c)_i - (\vec{v}_h)_i|}{N}.$$

It is easily verified that $d_N(\vec{v}_c, \vec{v}_h)$ is a metric and is in fact the probability that the concept c and the randomized hypothesis h disagree with respect to the uniform distribution on T. Thus, for any randomized hypothesis h, the ball in $\{0,1\}^N$ under the d_N metric defined by

$$b_h = \left\{ \vec{v}_c \in \{0,1\}^N : d_N(\vec{v}_c, \vec{v}_h) < \frac{1}{2} \right\}$$

is the set of all target concepts c over T such that h has accuracy more than $1/2$ for c (with respect to the uniform distribution on T). The next lemma shows that any ball b_h contains less than half of $\{0,1\}^N$; thus any randomized hypothesis h is a weak learning hypothesis for at most half of all concepts over T.

Lemma 13 *For any randomized hypothesis $\vec{v}_h \in [0,1]^N$, at most $1/2$ of the $\vec{v}_c \in \{0,1\}^N$ satisfy $\vec{v}_c \in b_h$.*

Proof: For any $\vec{v}_c \in \{0,1\}^N$, we have $d_N(\vec{v}_c, \mathrm{comp}(\vec{v}_c)) = 1$ where $\mathrm{comp}(\vec{v}_c)$ denotes the complement of $\vec{v}_c$. Thus $d_N(\vec{v}_c, \vec{v}_h) < 1/2$ implies $d_N(\mathrm{comp}(\vec{v}_c), \vec{v}_h) > 1/2$ since d_N is a metric. ■

Thus if we draw a concept over T at random, the probability that h has accuracy more than $1/2$ with respect to this concept is at most $1/2$. Using Chernoff bounds, it is easy to show that if we draw many concepts at random, the fraction of the concepts drawn for which h has accuracy more than $1/2$ rapidly approaches some value bounded above by $1/2$. We want this statement to hold simultaneously for *all* randomized hypotheses h. Since there are infinitely many such hypotheses, we need a uniform convergence result for the *class of concept classes* $B = \{b_h : h \in [0,1]^N\}$. This is exactly the approach taken in our next lemma, which shows that with overwhelming probability, *any* fixed randomized hypothesis h has accuracy more than $1/2$ (with respect to the uniform distribution over T) for at most half of all the concepts in C_n.

In the following lemma, it is assumed that C_n is generated by choosing 2^{kn} random characteristic vectors from $\{0,1\}^{2^n}$. Until now, we have implicitly restricted our attention to those concepts consistent with the fixed sample S. Now that we are drawing all 2^n labels for each concept at random, we must explicitly state this restriction. Finally, we sum the probability of failure over all choices for S.

Lemma 14 *Fix $0 < \beta \leq 1/2$. The probability (over the random choice of the class C_n) that there exists $\vec{v}_h \in [0,1]^N$ such that $d_N(\vec{v}_c, \vec{v}_h) < 1/2$ for a fraction $1/2 + \beta$ of the $\vec{v}_c \in C_n$ consistent with S is at most $c_0 e^{-2^n}$ for some constant $c_0 > 0$.*

Proof: It can be shown that if we draw 2^{kn} concepts randomly from $\{0,1\}^{2^n}$, then the probability we fail to get at least $2^{(k-1)n-1}$ concepts consistent with S is at most $e^{-2^{(k-1)n}/8}$.

Now on the uniform distribution over $\{0,1\}^N$ any $b_h \in B$ has probability weight at most $1/2$ by Lemma 13. It can be shown that $\text{VCD}(B) = c_1 N$ for some constant $c_1 > 0$. Thus, if we draw $2^{(k-1)n-1}$ concepts uniformly at random from $\{0,1\}^N$ then the probability that there is some $\vec{v}_h$ such that $d_N(\vec{v}_c, \vec{v}_h) < 1/2$ for a fraction $1/2 + \beta$ of the $\vec{v}_c$ drawn is at most

$$4(2^{(k-1)n \cdot c_1 N})e^{-\beta^2 2^{(k-1)n-1}/8}$$

by Vapnik and Chervonenkis [17]. Note that this is a generalized use of the Vapnik-Chervonenkis dimension, which is usually used to prove uniform convergence of some concept class over a domain set. Here we are actually interested in the uniform convergence of a *set* of concept classes, and each point of the "domain set" is now actually a concept itself.

Since the probability that we fail to have $2^{(k-1)n-1}$ concepts in C_n consistent with S is at most $e^{-2^{(k-1)n}/8}$, the total probability that there exists a $\vec{v}_h$ satisfying the condition of this lemma is bounded above by

$$e^{-2^{(k-1)n}/8} + 4(2^{(k-1)n \cdot c_1 N})e^{-\beta^2 2^{(k-1)n-1}/8}$$

which is at most $c_0 e^{-2^n}$ for $k \geq 5$, n large enough and a constant $c_0 > 0$. ■

To complete the proof of Theorem 12, we sum over all possible choices of the labeled sample S of size n. The number of such samples is at most 2^{n^2+n}; thus the probability (over the random choice of C_n) that there is some labeled sample S of n points such that there exists $\vec{v}_h \in [0,1]^N$ satisfying $d_N(\vec{v}_c, \vec{v}_h) < 1/2$ for a fraction $1/2+\beta$ of the concepts in $\vec{v}_c \in C_n$ consistent with S is at most $c_0 2^{n^2+n}/e^{2^n}$. From this we conclude that there must be some fixed C_n such that for any labeled sample S of n points, and any randomized hypothesis h, h has error less than $1/2$ on at most $1/2+\beta$ of the concepts in C_n consistent with S. By choosing the target $c \in C_n$ randomly from among all concepts consistent with S, the desired bound is achieved by an averaging argument. (See Lemma 15 below.) ■

Note that the above proof holds for *almost every* class C_n.

8 A Sufficient Condition for Large Weak Sample Complexity

We have now shown that there are classes C_n such that $\text{VCD}(C_n) = \Theta(n)$ and $\Omega(n)$ examples are required to weakly learn C_n (even using a randomized hypothesis space). However, since the proof of Theorem 12 is non-constructive in nature, so far we have no example of a class C_n of small (polynomial-size) circuits over $\{0,1\}^n$ with an $\Omega(\text{VCD}(C_n))$ weak learning sample size lower bound. Indeed, we do not even have non-constant lower bounds for any such class. Our goal now is twofold. First, we wish to extract a combinatorial property of concept classes from the proof of Theorem 12 that is sufficient to imply an $\Omega(\text{VCD}(C_n))$ lower bound. Second, we wish to exhibit a class of small circuits that has this property, and thus requires $\Omega(\text{VCD}(C_n))$ examples to obtain even a small advantage over random guessing.

Let C_n be a concept class over $\{0,1\}^n$. For any labeled sample S, we define $C_n(S)$ to be the set of concepts in C_n consistent with S. If h is any randomized hypothesis over $\{0,1\}^n$, and $p(n)$ is any polynomial, we denote by $C_n(S)[h, p(n)]$ all concepts $c \in C_n(S)$ such that h is a $(\frac{1}{2} - \frac{1}{p(n)})$-good hypothesis for target concept c with respect to the uniform distribution over $\{0,1\}^n$.

For any function $t(n)$, we say that the parameterized concept class C is $t(n)$-*unapproximable* if there exists some constant $\delta_0 > 0$ such that for any $c \in C_n$, for any labeled sample S of c consisting of at most $t(n)$ examples, and for any randomized hypothesis h and polynomial $p(n)$, we have

$$|C_n(S)[h, p(n)]| < (1 - \delta_0) \cdot |C_n(S)|$$

for sufficiently large n. In other words, a concept class is $t(n)$-unapproximable if for every sample S of size $t(n)$ there exists no hypothesis h that weakly approximates a fraction $1 - \delta_0$ of the concepts consistent with S. Note that the proof of Theorem 12 shows implicitly that a randomly selected concept class is $\Omega(n)$-unapproximable with high probability.

Lemma 15 *Let C be a $t(n)$-unapproximable concept class. Then $t(n)$ examples are insufficient to weakly learn C_n when $\delta \leq \delta_0$ for some constant $\delta_0 > 0$, and for n sufficiently large.*

Proof: A probabilistic argument is used to prove this lemma.

Let $\delta_0 > 0$ witness that C_n is $t(n)$-unapproximable, and suppose for contradiction that there exists an algorithm A that requires at most $t(n)$ examples to find a $(\frac{1}{2} - \frac{1}{p(n)})$-good hypothesis with probability at least $1 - \delta_0$. Assume n is sufficiently large.

Consider an experiment in which a target concept c is chosen uniformly at random from C_n, and A is trained on c under a uniform distribution on the domain. By assumption, A sees a sample S of cardinality at most $t(n)$. Let h be the hypothesis output by A. The chance that h is a $(\frac{1}{2} - \frac{1}{p(n)})$-good hypothesis is equal to the probability that c is chosen in $C_n(S)[h, p(n)]$, given that c is chosen from among the consistent concepts $C_n(S)$. Since c was selected uniformly at random, this probability is

$$\frac{|C_n(S)[h, p(n)]|}{|C_n(S)|} < 1 - \delta_0.$$

Thus, the probability (over random choices of c) that A fails to output a $(\frac{1}{2} - \frac{1}{p(n)})$-good hypothesis is greater than δ_0. Since this probability is the average failure probability of A over random choices of c, it follows that there exists some concept $c \in C_n$ for which the probability of failure exceeds δ_0. This contradicts our assumption about A. ■

We note that Lemma 15 can be proved under weaker versions of $t(n)$-unapproximability. For example, the lemma still holds even if we modify $t(n)$-unapproximability to hold only for *most* samples S of cardinality $t(n)$.

9 Small Circuits with Large Weak Sample Complexity

We turn next to the problem of finding a class of small circuits with large weak sample complexity. In particular, we show that the class of parity functions on n variables is $\Omega(n)$-unapproximable. Specifically, let P_n be the class of concepts c on domain $\{0,1\}^n$ of the form $c(x) = x_{i_1} \oplus \cdots \oplus x_{i_k}$. Thus, each concept is just the parity function computed on a subset of the n variables. It is known that P_n is learnable in polynomial time [10, 15]. It is not hard to show that $\text{VCD}(P_n) = n$. Also, note that each concept in P_n can be represented by a vector in $\{0,1\}^n$. Each vector $\vec{c} \in \{0,1\}^n$ is associated with the parity function c defined by

$$c(\vec{x}) \equiv \vec{c} \cdot \vec{x} = \bigoplus_{i=1}^{n} c_i x_i.$$

We use this vector representation throughout the following proof.

Theorem 16 *For any constant $\alpha < 1$, for sufficiently large n, and for $\delta < 1/2$, the number of examples required to weakly learn P_n (using either deterministic or randomized hypotheses) is at least αn.*

Proof: From Lemma 15, it suffices to show that P_n is αn-unapproximable. Consider a sample, $S = \{(\vec{x}_1, \ell_1), \ldots, (\vec{x}_t, \ell_t)\}$, generated by some concept in P_n where $t = \lfloor \alpha n \rfloor$. Let h be any randomized hypothesis. Then a concept (represented as a vector) $\vec{c}$ is consistent with S if and only if $\vec{c} \cdot \vec{x}_i = \ell_i$ for $i = 1, \ldots, t$. Thus, the sample S defines a system of t linear equations on n variables over the field Z_2; the solution space of that system of equations consists exactly of those concepts consistent with S.

Let M be the $t \times n$ matrix whose ith row is the vector $\vec{x}_i$. Let $r \leq t$ be the rank of M. Then, using standard linear algebra techniques, it can be shown that r of the bits of $\vec{c}$ can be solved for in terms of the remaining bits. That is, by possibly renaming variables, we may write

$$c_i = b_i \oplus \bigoplus_{j=r+1}^{n} a_{ij} c_j \tag{4}$$

for $i = 1, \ldots, r$, and for some $b_i, a_{ij} \in \{0, 1\}$ which can be determined from M using Gaussian elimination. Put another way, for every assignment to the bits $c_{r+1}, \ldots, c_n$, equation (4) gives an expression for the bits $c_1, \ldots, c_r$ with the property that the resulting concept $\vec{c}$ is consistent with S. Explicitly, this concept is defined by

$$c(\vec{x}) = \bigoplus_{i=1}^{n} c_i x_i = \bigoplus_{i=1}^{r} b_i x_i \oplus \bigoplus_{j=r+1}^{n} c_j (x_j \oplus \bigoplus_{i=1}^{r} a_{ij} x_i). \tag{5}$$

To complete the proof that P_n is αn-unapproximable, we will show that every hypothesis h has accuracy less than $\frac{1}{2} + \frac{4}{2^{(n-r)/2}} \leq \frac{1}{2} + \frac{4}{2^{(1-\alpha)n/2}}$ on over half of the consistent concepts. Consider an experiment in which one of these consistent concepts is chosen uniformly at random. Let c be the random variable representing this randomly chosen concept, and let e_c denote h's error on c. Then

$$e_c = \mathrm{E}_{\vec{x}}[|h(\vec{x}) - c(\vec{x})|]$$

where $\vec{x}$ is a vector chosen uniformly at random from $\{0, 1\}^n$, and $h(\vec{x})$ denotes the probability that h outputs 1 on input $\vec{x}$. For vector $\vec{x}$, and $r + 1 \leq j \leq n$, we write $\tilde{x}_j$ to denote $x_j \oplus \bigoplus_{i=1}^{r} a_{ij} x_i$. Using this notation,

$$c(\vec{x}) = \bigoplus_{i=1}^{r} b_i x_i \oplus \bigoplus_{j=r+1}^{n} c_j \tilde{x}_j. \tag{6}$$

We say that $\vec{x}$ is *known* if $\tilde{x}_j = 0$ for all j $(r + 1 \leq j \leq n)$ since $c(\vec{x})$ can be determined in this case using Equation (6).

Lemma 17 $\mathrm{E}_c[e_c] \geq \frac{1}{2} - \frac{1}{2^{n-r+1}}$.

Proof: We have that $\mathrm{E}_c[e_c] = \mathrm{E}_{\vec{x}}[d_{\vec{x}}]$ where

$$d_{\vec{x}} = \mathrm{E}_c[|h(\vec{x}) - c(\vec{x})|].$$

Clearly, $d_{\vec{x}} \geq 0$ for all $\vec{x}$.

If $\vec{x}$ is not known, then $\tilde{x}_j = 1$ for some j and so equation (6) implies that, for random c, $c(\vec{x}) = 1$ with probability 1/2. Thus, $d_{\vec{x}} = 1/2$ in this case.

Since the probability that a randomly chosen vector $\vec{x}$ is known is exactly $2^{-(n-r)}$, it follows that

$$\mathrm{E}_c[e_c] = \mathrm{E}_x[d_x] = \frac{1}{2}(1 - 2^{-(n-r)})$$

as claimed. ■

Lemma 18 $\mathrm{E}_c[e_c^2] \leq \frac{1}{4} + \frac{3}{2^{n-r}}$.

Proof: We have that

$$\begin{aligned}
\mathrm{E}_c[e_c^2] &= \mathrm{E}_c[(E_{\vec{x}}[|h(\vec{x}) - c(\vec{x})|])^2] \\
&= \mathrm{E}_c[\mathrm{E}_{\vec{x},\vec{y}}[|h(\vec{x}) - c(\vec{x})| \cdot |h(\vec{y}) - c(\vec{y})|]] \\
&= \mathrm{E}_{\vec{x},\vec{y}}[s_{\vec{x}\vec{y}}]
\end{aligned}$$

where $s_{\vec{x}\vec{y}} = \mathrm{E}_c[|h(\vec{x}) - c(\vec{x})| \cdot |h(\vec{y}) - c(\vec{y})|]$. Clearly, $s_{\vec{x}\vec{y}} \leq 1$ for all $\vec{x}, \vec{y}$.

Suppose that $\vec{x}$ and $\vec{y}$ are not known, and suppose further that $\tilde{x}_j \neq \tilde{y}_j$ for some j. Without loss of generality, assume that $\tilde{x}_j = 1$ and $\tilde{y}_j = 0$. Since $\vec{y}$ is not known, $\tilde{y}_k = 1$ for some k. For $a, b \in \{0, 1\}$, we have that

$$\Pr_c[c(\vec{x}) = a \wedge c(\vec{y}) = b] = 1/4.$$

To see that this is so, fix all the bits of c except for c_j and c_k. Choosing c_k now determines the value of $c(\vec{y})$ (since $\tilde{y}_j = 0$) to be 1 with probability 1/2. Finally, $c(\vec{x})$ is determined by choosing c_j, and its value will be 1 with probability 1/2, independent of $c(\vec{y})$. Thus, it follows that $s_{\vec{x}\vec{y}} = 1/4$ in this case.

The probability that either $\vec{x}$ or $\vec{y}$ is known is at most $2 \cdot 2^{-(n-r)}$. The probability that $\tilde{x}_j = \tilde{y}_j$ for all j is $2^{-(n-r)}$.

Combining these facts gives the stated bound on $\mathrm{E}_c[e_c^2]$. ■

Thus, $\mathrm{Var}[e_c] = \mathrm{E}[e_c^2] - (\mathrm{E}[e_c])^2 \leq \frac{4}{2^{n-r}}$. Applying Chebyshev's inequality, it follows that

$$\Pr_c\left[e_c \leq \frac{1}{2} - \frac{4}{2^{(n-r)/2}}\right] < \frac{1}{2}.$$

That is, h has error at least $\frac{1}{2} - \frac{4}{2^{(n-r)/2}}$ on more than half of the remaining concepts. ■

10 Limitations on the Power of Membership Queries

An interesting question in Valiant's learning model is under what conditions the sample size required for learning can be significantly reduced by allowing the learning algorithm to make *membership queries*, in addition to receiving random examples from the target distribution. Briefly, a membership query allows the learner to choose an input x and receive in unit time the label assigned to x by the unknown target concept. In Valiant's model with membership queries, the learner is still required to output a hypothesis that is accurate (in either the strong or weak learning sense) against the target distribution, but is now allowed both random examples and membership queries during the learning process.

It can be shown that Lemma 15 holds even when both random examples and membership queries are allowed. More precisely, if C_n is $t(n)$-unapproximable, then any algorithm weakly learning C_n must see more than $t(n)$ labeled examples of the target, regardless of whether these examples are chosen randomly from the target distribution or are membership queries. This again holds even when the target distribution is known to be uniform. In fact, we can prove that the $t(n)$ lower bound still holds even when the learning algorithm is allowed to *choose the answers* to the membership queries; that is, the learning algorithm is allowed to choose an input x and its corresponding label, and is then guaranteed that the target concept will be consistent with this labeled example (provided such a concept exists). Applying these results to the class of parity functions, we have a natural and simple class of efficiently learnable Boolean circuits for which the $\Theta(n)$ random sample size required for strong learning cannot be reduced even by relaxing to weak learning, restricting the target distribution to be uniform, providing membership queries, and allowing the learner to play a significant role in the choice of the target concept.

Similar issues have recently been investigated in Euclidean domains by Eisenberg and Rivest [4].

11 Toward a Characterization of Weak Sample Complexity

As we have mentioned, it is well-known that the sample size required for strong learning is characterized by the Vapnik-Chervonenkis dimension. In Section 6, we saw that this same measure fails to characterize weak sample complexity — for instance, the weak learning sample complexity of symmetric Boolean functions is significantly smaller than the strong learning sample complexity. Perhaps the most interesting open problem suggested by the research presented here is that of finding a clean *combinatorial* characterization of weak sample complexity. We provided an initial step in this direction in Section 8 by defining the notion of $t(n)$-unapproximability and proving that this is sufficient to imply a $t(n)$ lower bound. However, the necessity of this property (or even a weakened variant of it) has not been demonstrated. A promising alternative that was suggested to us is the property that *every* set of $d(n)$ points in the domain is shattered by C_n, with the hard distribution being uniform. However, it is possible to show the existence of classes C_n such that $\mathrm{VCD}(C_n) = O(n^2)$ and every set of size n is shattered, yet there is an algorithm that successfully weakly learns C_n against the uniform distribution using *zero* examples! Thus, the combinatorial characterization of weak sample complexity remains both open and elusive.

Acknowledgements

We are grateful to David Haussler, Ron Rivest and Manfred Warmuth for helpful suggestions and discussions on this material.

References

[1] Dana Angluin and Leslie G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, April 1979.

[2] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, October 1989.

[3] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[4] Bonnie Eisenberg and Ronald L. Rivest. On the sample complexity of pac-learning using random and chosen examples. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 154–162. Morgan Kaufmann, August 1990.

[5] P. Erdös and J. Spencer. *Probabilistic Methods in Combinatorics.* Academic Press, New York, 1974.

[6] William Feller. *An Introduction to Probability and Its Applications*, volume 1. John Wiley and Sons, third edition, 1968.

[7] Yoav Freund. Boosting a weak learning algorithm by majority. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 202–216. Morgan Kaufmann, August 1990.

[8] Yoav Freund. An improved boosting algorithm and its implications on learning complexity. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 391–398, July 1992.

[9] David Haussler, Michael Kearns, Nick Littlestone, and Manfred K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 25(2):129–161, December 1991.

[10] David Helmbold, Robert Sloan, and Manfred K. Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

[11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[12] Michael Kearns and Leslie Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. In *Proceedings of the Twenty First Annual ACM Symposium on Theory of Computing*, pages 433–444, May 1989. To appear in JACM.

[13] Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *Journal of the Association for Computing Machinery*, 35(4):965–984, 1988.

[14] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[15] Robert Hal Sloan. *Computational Learning Theory: New Models and Algorithms.* PhD thesis, MIT Dept. of Electrical Engineering and Computer Science, May 1989.

[16] Leslie Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

[17] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, XVI(2):264–280, 1971.