On the sample complexity of PAC learning halfspaces against the uniform distribution

Philip M. Long[*]
Research Triangle Institute
3040 Cornwallis Road
P.O. Box 12194
Research Triangle Park, North Carolina 27709 USA

November 2, 1995

Abstract

We prove an

$$\Omega\left(\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$$

lower bound on the PAC learning sample complexity of learning halfspaces against the uniform distribution on the unit ball in $\mathbf{R}^d$.

[*] This work was done while the author was at Duke University supported by Air Force Office of Scientific Research grant F49620–92–J–0515.

1 Introduction

In the PAC ("probably approximately correct") learning model [Val84], the learner wishes to approximately learn an unknown function f from some domain X to $\{0,1\}$, chosen from a known class F of such functions. The learner is given examples $(x_1, f(x_1)), ..., (x_m, f(x_m))$ of the behavior of f at elements $x_1, ..., x_m$ of the domain X chosen independently at random according to some distribution D. In the original formulation of the model, distribution-free PAC learning, the distribution D is arbitrary and unknown. Benedek and Itai [BI91] studied learning with respect to particular, known D. In each case, after receiving $(x_1, f(x_1)), ..., (x_m, f(x_m))$, the learner outputs a hypothesis h for f. The accuracy of h is measured by the probability that h would be different from f on another point chosen randomly according to D. The sample complexity indicates, informally, what is the least number m of examples so that there is a learner which, for any $f \in F$ (and in the distribution-free case, for any D), with probability (with respect to the random choice of $x_1, ..., x_m$) at least $1 - \delta$, outputs a hypothesis of accuracy at least as good as ϵ. (See Section 2 for a precise definition.)

Ehrenfeucht, Haussler, Kearns and Valiant [EHKV89] proved a general lower bound of

$$\Omega\left(\frac{\mathrm{VCdim}(F)}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right) \tag{1}$$

on the sample complexity of distribution-free PAC learning, where $\mathrm{VCdim}(F)$ is the VC-dimension [VC71] of F, a measure of its complexity (see Section 2 for a precise definition). For each class of finite VC-dimension, this matched earlier upper bounds on the sample-complexity of distribution-free PAC learning [BEHW89, HLW90] to within log factors. Since the VC-dimension of homogeneous halfspaces in $\mathbf{R}^d$ is d, (1) implies a lower bound of

$$\Omega\left(\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right) \tag{2}$$

on the sample complexity of distribution-free PAC learning for this class. The distributions used in the lower bound proof of [EHKV89] are concentrated on $\mathrm{VCdim}(F)$ elements of the domain, and almost all the weight is concentrated on one point. It has been argued that since such distributions are inaccurate models of real-world learning situations, the lower bound of [EHKV89], while interesting theoretically, does not provide even qualitative guidance concerning the difficulty of real-world learning. In this note, we prove that a lower bound of (2) on the sample complexity of PAC learning d-dimensional homogeneous halfspaces still holds even if the distribution is uniform over the unit ball, which is among the best-behaved distributions one can imagine.

The $(1/\epsilon)\log(1/\delta)$ term is easy, but the d/ϵ term requires a trick. For each distribution D, one may place a metric ρ_D on the set of functions from X to $\{0,1\}$ by letting $\rho_D(f,g) = \mathbf{Pr}_{x \in D}(f(x) \neq g(x))$. One can then define $\mathcal{M}_D(F, \epsilon)$ to be the largest subset of elements of F, each pair of which is at least ϵ far apart as measured by ρ_D.

Benedek and Itai [BI91] proved a general lower bound of

$$\log_2((1-\delta)\mathcal{M}_D(F, 2\epsilon)) \tag{3}$$

on the sample complexity of learning F with respect to D. However, as Haussler [Hau91] has proved that in general

$$\mathcal{M}_D(F, \epsilon) \leq \left(\frac{10}{\epsilon}\right)^{\mathrm{VCdim}(F)}, \tag{4}$$

the best one could *ever* get by directly applying (3) is a lower bound on the sample complexity of

$$\Omega\left(\mathrm{VCdim}(F)\log\frac{1}{\epsilon}+\log(1-\delta)\right).$$

However, we adapt the proof of [BI91] to prove a lower bound on the sample complexity for the class HALF_d of homogeneous halfspaces in d dimensions, when $\delta = 1/2$, for all distributions D, of

$$\frac{d-1}{e}\left(\mathcal{M}_D(\mathrm{HALF}_d, 2\epsilon)/4\right)^{1/(d-1)}. \tag{5}$$

One can easily see how to use the same technique to prove a general lower bound of

$$\frac{\mathrm{VCdim}(F)}{e}\left(\mathcal{M}_D(F, 2\epsilon)/2\right)^{1/\mathrm{VCdim}(F)}.$$

The latter bound shows that for all class/distribution pairs for which one can show (4) is tight to within a constant factor in the base of the exponent, an $\Omega(d/\epsilon)$ lower bound holds. In the case of halfspaces, however, we needed the stronger bound of (5).

The lower bound for halfspaces is then proved by establishing a tight enough lower bound on $\mathcal{M}_D(\mathrm{HALF}_d, 2\epsilon)$ and applying (5). For this, we use a probabilistic method trick from [ASE92]. Usually, to use the probabilistic method to show that a large set with a given property exists, one randomly picks a suitably large set, and shows that the probability that the randomly chosen set has the desired property is nonzero. This methodology appears to be too crude for this application. Instead, we use the "removing blemishes" trick from [ASE92], where one randomly picks a large set, shows that the expected number of elements which interfere with the large set having the desired property is not too big, and therefore that there exists a large set without too many such "blemishes", then removes those offending elements to get the required set.

The most closely related previous work that we are aware of is that of Opper and Haussler [OH91]. Using techniques from statistical mechanics, they calculated the learning curve (in the large d limit) for the Bayes optimal algorithm for learning homogeneous halfspaces in $\mathbf{R}^d$ according to the uniform distribution on the unit ball, when the normal vector to the halfspace to be learned was chosen according to the uniform distribution on the unit ball. Since their results yield lower bounds on the performance of the Bayes optimal algorithm, they lower bound any algorithm. Further, obviously lower bounds for a randomly chosen target yield the same lower bounds for a worst-case target. However, their bound is for a model in which an algorithm is measured by the expected value of the accuracy of its hypothesis. It is not clear how to modify their argument to obtain the bounds of this paper which are for the PAC model, where the goal of the algorithm is to get a hypothesis of a given accuracy with a given probability.[1] Moreover, their results depend on the unproven replica hypothesis.

2 Definitions

Denote the reals by $\mathbf{R}$ and the positive integers by $\mathbf{N}$.

[1] The distribution constructed in the proof of [EHKV89] had the property that any reasonable algorithm *always* had an error of at most a constant times ϵ, which enabled them to argue that producing a hypothesis that was good on average was effectively equivalent to producing a hypothesis that is good with a certain probability (see [EHKV89] for further details). Since the uniform distribution does not have this property, this trick cannot be used for our problem.

These definitions are a commonly studied version of Valiant's PAC model [Val84]; study of the distribution specific version was initiated by Benedek and Itai [BI91]. Choose a set X, a probability distribution D over X, and a set F of functions from X to $\{0,1\}$. Define $\rho_D : \{0,1\}^X \times \{0,1\}^X \rightarrow [0,1]$ by $\rho_D(f,g) = \mathbf{Pr}_{u \in D}(f(u) \neq g(u))$. For each $\epsilon > 0$, define

$$\mathcal{M}_D(F,\epsilon) = \max\{n \in N : \exists f_1,...,f_n \in F, \forall 1 \leq i,j, \leq n, \rho_D(f_i,f_j) \geq \epsilon\}.$$

For a finite sequence $\vec{x} = (x_1,...,x_m)$ of elements of X and $f \in F$, define $\mathrm{sample}(\vec{x},f) \in (X \times \{0,1\})^m$ by

$$\mathrm{sample}(\vec{x},f) = ((x_1,f(x_1)),...,(x_m,f(x_m))).$$

A *learning strategy* for X is a mapping from $\cup_{m \in \mathbf{N}}(X \times \{0,1\})^m$ to $\{0,1\}^X$. That is, given a finite sample, the learner outputs a hypothesis. Note that the learner is not required to output hypotheses from any restricted class. If $\mathcal{A}$ is the set of learning strategies for X, for a probability distribution D on X, $\epsilon > 0$, and $\delta > 0$, define

$$m(F,D,\epsilon,\delta) = \min\{r \in \mathbf{N} : \exists A \in \mathcal{A}, \forall f \in F, \mathbf{Pr}_{\vec{x} \in D^r}(\rho_D(A(\mathrm{sample}(\vec{x},f)),f) \geq \epsilon) \leq \delta\}.$$

The VC-dimension [VC71] of F is defined to be

$$\max\{d : \exists x_1,...,x_d \in X, \{(f(x_1),...,f(x_d)) : f \in F\} = \{0,1\}^d\}.$$

For $d \in N$, define

- HALF_d to be the set of homogeneous halfspaces in $\mathbf{R}^d$, i.e., the set of all $f : \mathbf{R}^d \rightarrow \{0,1\}$ for which there exists $\vec{w} \in \mathbf{R}^d$ such that for all $\vec{x} \in \mathbf{R}^d$, $f(\vec{x}) = 1 \Leftrightarrow \sum_{i=1}^d w_i x_i \geq 0$, and
- UBALL_d to be the uniform distribution on the surface of the unit ball in $\mathbf{R}^d$.

3 The lower bound

The following is the main result of this note.

Theorem 1

$$m(\mathrm{HALF}_d, \mathrm{UBALL}_d, \epsilon, \delta) = \Omega\left(\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right).$$

3.1 The d/ϵ term

We will make use of the following bound on the "growth function" of halfspaces, due to Cover [Cov65] (see [HKP91, page 113]).

Lemma 2 ([Cov65]) *Choose $d \in \mathbf{N}$. For any $x_1,...,x_m \in \mathbf{R}^d$,*

$$|\{(f(x_1),...,f(x_m)) : f \in \mathrm{HALF}_d\}| \leq 2(em/(d-1))^{d-1}.$$

The following is the key lemma used in proving the d/ϵ term. Its proof is based on [BI91, Lemma 4.8].

Lemma 3 *Choose $d \in \mathbf{N}, d \geq 2$, a distribution D over $\mathbf{R}^d$, and $\epsilon > 0$. Then*

$$m(\mathrm{HALF}_d, D, \epsilon, 1/2) \geq \frac{d-1}{e}\left(\frac{\mathcal{M}_D(\mathrm{HALF}_d, 2\epsilon)}{4}\right)^{1/(d-1)}.$$

Proof: Fix d,D, and ϵ, and let $m = m(\mathrm{HALF}_d, D, \epsilon, 1/2)$. Let A be a learning strategy for X such that for all $f \in \mathrm{HALF}_d$,

$$\mathbf{Pr}_{\vec{x} \in D^m}(\rho_D(A(\mathrm{sample}(\vec{x}, f)), f) \geq \epsilon) \leq 1/2. \tag{6}$$

Let $G \subseteq \mathrm{HALF}_d$ be a set of $\mathcal{M}_D(\mathrm{HALF}_d, 2\epsilon)$ elements of HALF_d such that for all $g_1, g_2 \in G$, $\rho_D(g_1, g_2) \geq 2\epsilon$.

For each $\vec{x} \in X^m, f \in \mathrm{HALF}_d$, define

$$\varphi(\vec{x}, f) = \begin{cases} 1 & \text{if } \rho_D(A(\mathrm{sample}(\vec{x}, f)), f) < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

and define

$$S = \sum_{g \in G} \mathbf{E}_{\vec{x} \in D^m}(\varphi(\vec{x}, g)).$$

By (6),

$$S \geq |G|/2 = \mathcal{M}_D(\mathrm{HALF}_d, 2\epsilon)/2. \tag{7}$$

Moving the sum inside the expectation, we have

$$S = \mathbf{E}_{\vec{x} \in D^m}(\sum_{g \in G} \varphi(\vec{x}, g)). \tag{8}$$

For any $g_1, g_2 \in G, \vec{x} \in X^m$ for which $\mathrm{sample}(\vec{x}, g_1) = \mathrm{sample}(\vec{x}, g_2)$, since $\rho_D(g_1, g_2) \geq 2\epsilon$ and ρ_D is a metric, it cannot be the case that both $\varphi(\vec{x}, g_1) = 1$ and $\varphi(\vec{x}, g_2) = 1$. Thus, (8) implies

$$S \leq \mathbf{E}_{\vec{x} \in D^m}(|\{\mathrm{sample}(\vec{x}, g) : g \in G\}|).$$

Applying Lemma 2 yields

$$S \leq \mathbf{E}_{\vec{x} \in D^m}(2(em/(d-1))^{d-1}) = 2(em/(d-1))^{d-1}.$$

Combining with (7) and solving for m completes the proof. □

We will make use of the following easily verified technical lemmas.

Lemma 4 *For all $x \in [0, \pi/4]$, $\tan x \leq 2x$.*

Lemma 5 (see [Bau90]) *For all $d \in \mathbf{N}$, if V_d is the volume of the unit ball in $\mathbf{R}^d$, then for all $d \geq 2$,*

$$V_{d-1}/V_d \leq \sqrt{d}.$$

Now we are ready to lower bound $\mathcal{M}_{\mathrm{UBALL}_d}(\mathrm{HALF}_d, \epsilon)$. For this, we use the "removing blemishes" probabilistic method trick from [ASE92].

Lemma 6 *For all* $0 < \epsilon < 1/2$, $d \in \mathbf{N}$,

$$\mathcal{M}_{\mathrm{UBALL}_d}(\mathrm{HALF}_d, \epsilon) \geq \frac{\sqrt{d}}{2}\left(\frac{1}{2\pi\epsilon}\right)^{d-1} - 1.$$

Proof: For each unit-length $\vec{w} \in \mathbf{R}^d$, define $f_{\vec{w}}$ to be the homogeneous halfspace whose normal vector is $\vec{w}$, i.e. the function $f_{\vec{w}} : \mathbf{R}^d \to \{0,1\}$ defined by

$$f_{\vec{w}}(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{x} \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

It is well known, and very intuitive, that $\rho_{\mathrm{UBALL}_d}(f_{\vec{v}}, f_{\vec{w}})$ is equal to the angle between $\vec{v}$ and $\vec{w}$ (in radians) divided by π. Further, for all $\alpha > 0$, $f_{\alpha\vec{w}} = f_{\vec{w}}$, so, by symmetry, choosing an element of HALF_d randomly by sampling its normal vector from the unit ball is equivalent to choosing an element of HALF_d by sampling its normal vector uniformly from the interior of the unit ball. Thus, for fixed $\vec{v}$, the probability that a uniformly randomly chosen $\vec{w}$ has $\rho_{\mathrm{UBALL}_d}(f_{\vec{v}}, f_{\vec{w}}) \leq \epsilon$ is equal to the volume of those vectors $\vec{w}$ in the interior of the unit ball whose angle with $\vec{v}$ is at most $\epsilon\pi$, divided by the volume of the unit ball. We upper bound the first volume by calculating the volume of the smallest cone containing it centered at $\vec{v}$ with its base on the hyperplane tangent to the unit ball at $\vec{v}$. In this manner, we get that if $V_d(r)$ is the volume of a ball of radius r in $\mathbf{R}^d$,

$$\mathbf{Pr}_{\vec{w} \in \mathrm{UBALL}_d}(\rho_{\mathrm{UBALL}_d}(f_{\vec{v}}, f_{\vec{w}}) \leq \epsilon) \leq \frac{1}{V_d(1)} \int_0^1 V_{d-1}(x \tan(\epsilon\pi))\, dx.$$

Using the fact that $V_{d-1}(r) = r^{d-1} V_{d-1}(1)$, we get

$$\mathbf{Pr}_{\vec{w} \in \mathrm{UBALL}_d}(\rho_{\mathrm{UBALL}_d}(f_{\vec{v}}, f_{\vec{w}}) \leq \epsilon) \leq \frac{V_{d-1}(1)}{V_d(1)} (\tan(\epsilon\pi))^{d-1} \int_0^1 x^{d-1}\, dx$$

which implies

$$\mathbf{Pr}_{\vec{w} \in \mathrm{UBALL}_d}(\rho_{\mathrm{UBALL}_d}(f_{\vec{v}}, f_{\vec{w}}) \leq \epsilon) \leq \frac{V_{d-1}(1)}{d V_d(1)} (\tan(\epsilon\pi))^{d-1}$$

and applying Lemmas 4 and 5 yields

$$\mathbf{Pr}_{\vec{w} \in \mathrm{UBALL}_d}(\rho_{\mathrm{UBALL}_d}(f_{\vec{v}}, f_{\vec{w}}) \leq \epsilon) \leq \frac{1}{\sqrt{d}} (2\epsilon\pi)^{d-1}.$$

By symmetry, if $\vec{v}$ is chosen randomly,

$$\mathbf{Pr}_{(\vec{v},\vec{w}) \in (\mathrm{UBALL}_d)^2}(\rho_{\mathrm{UBALL}_d}(f_{\vec{v}}, f_{\vec{w}}) \leq \epsilon) \leq \frac{1}{\sqrt{d}} (2\epsilon\pi)^{d-1}.$$

Thus, for $s \in \mathbf{N}$, if we pick s normal vectors uniformly at random from the unit ball, the expected number of pairs of resulting halfspaces that are ϵ-close according to ρ_{UBALL_d} is at most

$$\frac{s^2}{2\sqrt{d}} (2\epsilon\pi)^{d-1}.$$

Hence, for each $s \in \mathbf{N}$, there exists a set S of s homogeneous halfspaces such that at most $\frac{s^2}{2\sqrt{d}}(2\epsilon\pi)^{d-1}$ pairs of elements of S are ϵ-close according to ρ_{UBALL_d}. Removing one element of each such pair from S yields a set of

$$s - \frac{s^2}{2\sqrt{d}} (2\epsilon\pi)^{d-1}$$

halfspaces, each pair of which is of distance at least ϵ according to ρ_{UBALL_d}. Setting

$$s = \left\lfloor \frac{\sqrt{d}}{(2\epsilon\pi)^{d-1}} \right\rfloor$$

and simplifying completes the proof. □

3.2 The $(1/\epsilon)\log(1/\delta)$ term

Definition 7 *Choose X, a probability distribution D over X, and $F \subseteq \{0,1\}^X$. We say that F* has continuous hard pairs with respect to *D if for each $0 < \epsilon < 1$, there exist $f, g \in F$ such that $\mathbf{Pr}_{x \in D}(f(x) \neq g(x)) = \epsilon$.*

This definition is stronger than we need. Obviously, the set of halfspaces has continuous hard pairs with respect to the uniform distribution on the unit ball.

The following theorem is a fairly straightforward extension of the corresponding result in [BEHW89].

Theorem 8 *Choose X, a probability distribution D over X, and $F \subseteq \{0,1\}^X$. If F has continuous hard pairs with respect to D then*

$$m(F, D, \epsilon, \delta) = \Omega\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right).$$

Proof: Choose $0 < \epsilon < 1/2$. Recall that $\rho_D(f, g) = \mathbf{Pr}_{x \in D}(f(x) \neq g(x))$. Let $f, g \in F$ be such that $\rho_D(f, g) = 2\epsilon$. Choose a learning algorithm A for F.

Choose $m \in \mathbf{N}$. Define

$$\mathrm{MISS} = \{(x_1, ..., x_m) : \forall j, f(x_j) = g(x_j)\}.$$

Note that

$$\begin{aligned} &\mathbf{Pr}_{(x_1,...,x_m) \in D^m}(\rho_D(A((x_1, f(x_1)), ..., (x_m, f(x_m))), f) \geq \epsilon | \mathrm{MISS}) \\ &= \mathbf{Pr}_{(x_1,...,x_m) \in D^m}(\rho_D(A((x_1, g(x_1)), ..., (x_m, g(x_m))), f) \geq \epsilon | \mathrm{MISS}). \end{aligned}$$

Call the above quantity b_f (for "bad"). Define

$$\begin{aligned} b_g &= \mathbf{Pr}_{(x_1,...,x_m) \in D^m}(\rho_D(A((x_1, f(x_1)), ..., (x_m, f(x_m))), g) \geq \epsilon | \mathrm{MISS}) \\ &= \mathbf{Pr}_{(x_1,...,x_m) \in D^m}(\rho_D(A((x_1, g(x_1)), ..., (x_m, g(x_m))), g) \geq \epsilon | \mathrm{MISS}). \end{aligned}$$

By the triangle inequality, a hypothesis cannot be ϵ-close to both f and g, and therefore $b_f + b_g \geq 1$.

Assume without loss of generality that $b_f \geq 1/2$. Then, if f is the target concept,

$$\begin{aligned} &\mathbf{Pr}_{(x_1,...,x_m) \in D^m}(\rho_D(A((x_1, f(x_1)), ..., (x_m, f(x_m))), f) \geq \epsilon) \\ &\geq \mathbf{Pr}_{(x_1,...,x_m) \in D^m}(\rho_D(A((x_1, f(x_1)), ..., (x_m, f(x_m))), f) \geq \epsilon | \mathrm{MISS}) \mathbf{Pr}_{(x_1,...,x_m) \in D^m}(\mathrm{MISS}) \\ &\geq (1/2)(1 - 2\epsilon)^m. \end{aligned}$$

Using the fact that for small ϵ, $1 - 2\epsilon \approx e^{-2\epsilon}$ and solving for m completes the proof. □

Putting together Lemmas 3 and 6, together with Theorem 8, proves Theorem 1. The proof can be trivially modified to establish the same lower bound for any distribution whose density is within a constant factor of the uniform distribution. It would be nice to determine a wider class of distributions for which the lower bound holds.

4 Acknowledgements

We are very grateful to Wolfgang Maass for suggesting this problem. We are also very grateful to two anonymous referees for their comments, particularly the referee who pointed out a mistake in an earlier version of this paper, and suggested how it might be fixed.

References

[ASE92] N. Alon, J.H. Spencer, and P. Erdös. *The Probabilistic Method.* Wiley, 1992.

[Bau90] E.B. Baum. The perceptron algorithm is fast for nonmalicious distributions. *Neural Computation*, 2:248–260, 1990.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

[BI91] G. Benedek and A. Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

[Cov65] Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. on Electronic Computers*, EC-14:326–334, 1965.

[EHKV89] A. Ehrenfeucht, D. Haussler, M. Kearns, and L.G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

[Hau91] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. Technical Report UCSC-CRL-91-41, University of California at Santa Cruz, 1991.

[HKP91] J. A. Hertz, A. Krogh, and R. Palmer. *Introduction to the theory of neural computation.* Addison-Wesley, 1991.

[HLW90] D. Haussler, N. Littlestone, and M.K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. Technical Report UCSC-CRL-90-54, University of California Santa Cruz, Computer Research Laboratory, December 1990. To appear in Information and Computation.

[OH91] M. Opper and D. Haussler. Calculation of the learning curve of Bayes optimal classification algorithm for learning a perceptron with noise. In *Computational Learning Theory: Proceedings of the Fourth Annual Workshop*, pages 75–87. Morgan Kaufmann, 1991.

[Val84] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[VC71] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.