

A Sufficient Condition for Polynomial Distribution-Dependent Learnability

Martin Anthony
Department of Mathematics
London School of Economics
Houghton Street
London WC2A 2AE, UK
`m.anthony@lse.ac.uk`

John Shawe-Taylor
Department of Computer Science
Royal Holloway, University of London
Egham Hill
Egham
Surrey TW20 0EX, UK
`john@dcs.rhbnc.ac.uk`

Abstract

We investigate upper bounds on the sample-size sufficient for ‘solid’ learnability with respect to a probability distribution. We obtain a sufficient condition for feasible (polynomially bounded) sample-size bounds for distribution-specific (solid) learnability.

1 Introduction

There have been extensive studies of probabilistic models of machine learning; see the books [3, 11, 12], for example. In the standard ‘PAC’ model of learning, the definition of successful learning is ‘distribution-free’. A number of researchers have examined learning where the probability distribution generating the examples is known; see [6, 5], for example. In this paper we seek conditions under which such distribution-specific learning can be achieved with a feasible (polynomial) number of training examples.

2 The PAC learning framework

In this section, we describe a probabilistic model of learning, introduced by Valiant [15] and developed by many researchers (see for example [8]). It has come to be known as the *probably approximately correct* learning model [1].

Throughout, we have an *example space* X , which is either countable or is the Euclidean space \mathbf{R}^n for some n . We have a probability space (X, Σ, μ) defined on X , where we assume that when X is countable, Σ is the set of all subsets of X and that when X is \mathbf{R}^n , Σ is the Borel σ -algebra. A *hypothesis* is a Σ -measurable $\{0, 1\}$ -valued function on X . The *hypothesis space* H is a set of hypotheses, and the *target*, c , is one particular concept from H . A *labelled example* of c is an ordered pair $(x, c(x))$. If $c(x) = 1$, we say x is a *positive example* of c , while if $c(x) = 0$, we say x is a *negative example* of c . A *sample* \mathbf{y} of c of length (or size) m is a sequence of m labelled examples of c . When the target concept is clear, we will denote the sample simply by the vector $\mathbf{x} \in X^m$, so that if $\mathbf{x} = (x_1, \dots, x_m)$ then the corresponding sample of c is $((x_1, a_1), \dots, (x_m, a_m))$, where $a_i = c(x_i)$. The learning problem is to find a good approximation to c from H , this approximation being based solely on a sample of c , each example in the sample being chosen independently and at random, according to the distribution μ .

Fix a particular target $c \in H$. For any hypothesis h of H , the *error* of h (with respect to c) is $\text{er}_\mu(h) = \mu(h\Delta c)$, where $h\Delta c$ is the set $\{x : h(x) \neq c(x)\}$, the symmetric difference of h and c . We say that a hypothesis h is ϵ -close to c if $\text{er}_\mu(h) \leq \epsilon$. For any set F of measurable subsets of X , we define the *haziness* of F (with respect to c) as

$$\text{haz}_\mu(F) = \sup\{\text{er}_\mu(h) : h \in F\}.$$

The set $H[\mathbf{x}, c]$ of hypotheses *consistent* with c on \mathbf{x} is

$$H[\mathbf{x}, c] = \{h \in H : h(x_i) = c(x_i) \ (1 \leq i \leq m)\},$$

which we shall usually denote by $H[\mathbf{x}]$ when c is understood. Now we can define what is meant by solid learnability. (This terminology comes from [5].)

Definition 2.1 *The hypothesis space H is solidly learnable if, for any $\epsilon, \delta \in (0, 1)$, there is $m_0 = m_0(\epsilon, \delta)$ such that given any $c \in H$, for all probability measures μ on X ,*

$$m > m_0 \implies \mu^m\{\mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) < \epsilon\} > 1 - \delta.$$

Here, μ^m is the product measure on X .

In words, H is solidly learnable if for a given accuracy parameter ϵ and a given certainty parameter δ , there is a sample size, independent of the distribution and the target concept, such that *any* hypothesis consistent with that many random examples will “probably” be “approximately” correct. (In this case, a learning algorithm which returns a consistent hypothesis will perform well.) From now on, ‘learnability’ shall mean ‘solid learnability’.

We assume throughout that the spaces satisfy certain measurability requirements—namely, that they are *universally separable*, so that the probabilities in the definitions and proofs are indeed defined. See [13, 8] for details.

3 Distribution-independent sample sizes

The *Vapnik-Chervonenkis dimension* (or *VC dimension*) [16] has been widely used in order to obtain some measure of the degree of expressibility of a hypothesis space, and hence to obtain learnability results [9, 8, 4]. Given a hypothesis space H , define, for each $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, a function $\mathbf{x}^* : H \rightarrow \{0, 1\}^m$ by

$$\mathbf{x}^*(h) = (h(x_1), \dots, h(x_m)).$$

The *growth function*, Π_H from the set of integers to itself is defined by

$$\Pi_H(m) = \max\{|\{\mathbf{x}^*(h) : h \in H\}| : \mathbf{x} \in X^m\} \leq 2^m.$$

If $|\{\mathbf{x}^*(h) : h \in H\}| = 2^m$ then we say that \mathbf{x} is *shattered* by H . If $\Pi_H(m) = 2^m$ for all m then the Vapnik-Chervonenkis dimension of H is infinite. Otherwise, the Vapnik-Chervonenkis dimension is the largest positive integer m for which $\Pi_H(m) = 2^m$; that is, the largest integer m such that some sample \mathbf{x} of length m is shattered. We remark that any finite hypothesis space certainly has finite VC dimension.

It can be shown that if $\text{VCdim}(H)=d$, and $m \geq d \geq 1$ then $\Pi_H(m) \leq (em/d)^d$ [14]. This is useful in obtaining bounds on the sufficient sample size $m_0(\epsilon, \delta)$. Following [10], it can be proved [8] that if the hypothesis space H has finite VC dimension d , then H is learnable. Further, if H is learnable then H must have finite VC dimension [8]. Specifically, the sufficiency result of Blumer *et al.* follows from the following, which is a refinement of a result from [16].

Theorem 3.1 (Blumer *et al.* [8]) *For any distribution μ ,*

$$\mu^m \{\mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon\} < 2\Pi_H(2m) 2^{-\epsilon m/2}.$$

This bound has been tightened [4], resulting in the following bound on sufficient sample-size.

Theorem 3.2 ([4]) *The hypothesis space H is learnable if H has finite VC dimension. If $d = \text{VCdim}(H) > 1$ is finite then a suitable m_0 is*

$$m_0 = m_0(\epsilon, \delta) = \left\lceil \frac{1}{\epsilon(1 - \sqrt{\epsilon})} \left(\ln \left(\frac{d/(d-1)}{\delta} \right) + 2d \ln \left(\frac{6}{\epsilon} \right) \right) \right\rceil,$$

where \ln denotes natural logarithm.

4 Distribution-dependent learning

Recall the definition of learnability of a hypothesis space H . H is learnable if for any accuracy parameter ϵ , any confidence parameter δ , any target concept $c \in H$ and any probability measure μ on X , there is a sample-size m_0 , which is a function of ϵ and δ alone, such that the following holds: With probability at least $1 - \delta$, if some hypothesis h is consistent with c on at least m_0 inputs chosen randomly according to the distribution μ , then h has actual error less than ϵ . As emphasised earlier, the value of m_0 must depend on neither the target concept c nor the distribution

(probability measure) μ . In many realistic learning problems, the distribution on the input space is fixed but unknown. This is the primary reason for proving learnability results and finding sufficient sample-sizes which are independent of the distribution; results that are independent of the distribution certainly hold for any particular distribution. If something *is* known of the distribution or if the distribution is of a special type, it may be possible to say more, obtaining positive results even when the hypothesis space has infinite VC dimension.

In order to introduce distribution-dependent learnability, we may define learnability of a *particular* concept c from a hypothesis space H , with respect to a *particular* probability measure μ on the input space X . We say that c is μ -*learnable* in H if given any $\epsilon, \delta \in (0, 1)$, there is an integer $m_0 = m_0(\epsilon, \delta, c, \mu)$ such that for all $m \geq m_0$, $\mu^m \{ \mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}, c]) > \epsilon \} < \delta$. In addition, we say that H itself is μ -learnable if every $c \in H$ is μ -learnable and if there is a sufficient sample-size m_0 which *is independent* of the hypothesis c . If H is μ -learnable for every distribution μ on X , then we say that H is *distribution-dependent learnable*, abbreviated as *dd-learnable*.

If one examines closely the proof in [8] of Theorem 2.1 then it is clear that the term $\Pi_H(2m)$ in the bound can be replaced by the expectation over X^{2m} of the function Π_H , where $\Pi_H(\mathbf{x}) = |\{ \mathbf{x}^*(h) : h \in H \}|$. (This will be a random variable if we assume that H is universally separable; see [2]). Thus, for distribution-dependent analysis, we can use $\mathbf{E}_{2m}(\Pi_H(\mathbf{x}))$ in place of $\Pi_H(m)$, where $\mathbf{E}_{2m}(\cdot)$ denotes expected value with respect to μ^{2m} and over X^{2m} . This yields

$$\mu^m \{ \mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon \} < 2 \mathbf{E}_{2m}(\Pi_H(\mathbf{x})) 2^{-\epsilon m/2},$$

for $m \geq 8/\epsilon$.

A function f is said to be *subexponential* if, for all $\epsilon > 0$, as x tends to infinity, $f(x) \exp(-\epsilon x)$ tends to zero. With this definition, we have the following.

Theorem 4.1 *Let μ be any probability measure on X . If $\mathbf{E}_n(\Pi_H(\mathbf{x}))$, the expected value of $\Pi_H(\mathbf{x})$ over X^n (with respect to μ^n), is a subexponential function of n , then H is μ -learnable.*

Proof: For $m \geq 8/\epsilon$, $\mu^m \{ \mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon \} < 2 \mathbf{E}_{2m} \Pi_H(\mathbf{x}) 2^{-\epsilon m/2}$. If

$$\mathbf{E}_{2m} \Pi_H(\mathbf{x}) 2^{-\epsilon m/2} \rightarrow 0 \text{ as } m \rightarrow \infty$$

for all $\epsilon > 0$, which is the case if $\mathbf{E}_n(\Pi_H(\mathbf{x}))$ is a subexponential function of n , then the quantity on the right-hand side can be made less than any $\delta > 0$ by choosing

$m \geq m_0$, where m_0 depends only on μ and not on the hypothesis c . The result follows. \square

It's fairly easy to see that demanding that $\mathbf{E}_n(\Pi_H(\mathbf{x}))$ be sub-exponential is equivalent to demanding that $n^{-1} \log \mathbf{E}_n(\Pi_H(\mathbf{x})) \rightarrow 0$ as $n \rightarrow \infty$. In fact, results of Vapnik and Chervonenkis [16] show that the weaker condition $n^{-1} \mathbf{E}_n(\log \Pi_H(\mathbf{x})) \rightarrow 0$ as $n \rightarrow \infty$ is sufficient.

We give two examples of this theorem — one discrete and the other continuous.

Example 1: Let $\{B_i\}_{i \geq 1}$ be any sequence of disjoint sets such that $|B_i| = i$, ($i \geq 1$) and take as example space the countably infinite set $X = \bigcup_{i=1}^{\infty} B_i$. Let the probability measure μ be defined on the σ -algebra of all subsets of X by

$$\mu(\{x\}) = \frac{1}{i} \frac{1}{2^i} \quad (x \in B_i).$$

Let the hypothesis space H be the set of functions $H = \bigcup_{i=1}^{\infty} \{I_C : C \subseteq B_i\}$, where $I_C : X \rightarrow \{0, 1\}$ is the characteristic function of the subset C . Then it is easy to see that H has infinite VC dimension and thus is not learnable. However, we can use Theorem 3.1 to prove that H is μ -learnable. For $\mathbf{x} \in X^n$, let $I(\mathbf{x})$ be the set of entries of \mathbf{x} . That is, $I(\mathbf{x}) = \{x_i : 1 \leq i \leq n\}$. Then it is not difficult to see that

$$\Pi_H(\mathbf{x}) = \sum 2^{|I(\mathbf{x}) \cap B_i|},$$

where the sum is over all i such that $I(\mathbf{x}) \cap B_i \neq \emptyset$. Therefore,

$$I(\mathbf{x}) \subseteq S_k = \bigcup_{i=1}^k B_i \implies \Pi_H(\mathbf{x}) \leq 2 + 2^2 + \dots + 2^k < 2^{k+1}.$$

Further, $\Pi_H(\mathbf{x}) \leq 2^n$ for all $\mathbf{x} \in X^n$.

Let η_k be the probability that $I(\mathbf{x}) \subseteq S_k$; that is, $\eta_k = \mu^n(S_k^n)$. Then,

$$\eta_k = (\mu(S_k))^n = \left(1 - \frac{1}{2^k}\right)^n.$$

For any $0 < x < 1$, $(1 - x)^n \geq 1 - nx$ and so, for $k \geq 2$,

$$\eta_k - \eta_{k-1} \leq 1 - \left(1 - \frac{1}{2^{k-1}}\right)^n \leq \frac{n}{2^{k-1}}.$$

Since the sets S_k^n cover X^n , we therefore have

$$\mathbf{E}_n(\Pi_H(\mathbf{x})) < 2\eta_1 + \sum_{k=2}^{n-1} (\eta_k - \eta_{k-1}) 2^{k+1} + 2^n (1 - \mu^n(S_{n-1}^n))$$

$$\begin{aligned} &\leq 1 + \sum_{k=2}^{n-1} \frac{n}{2^{k-1}} 2^{k+1} + 2^n \left(\frac{n}{2^{n-1}} \right) \\ &= 1 + 4n(n-2) + 2n < 4n^2. \end{aligned}$$

It follows that the expected value of $\Pi_H(\mathbf{x})$ is polynomial and therefore H is μ -learnable.

Example 2: Let X be the set of non-negative reals and let the distribution have probability density function $p(x) = e^{-x}$, so that $\mu([0, y]) = 1 - e^{-y}$. Let the hypothesis space H consist of all (characteristic functions of) finite unions of closed intervals, at most k of which intersect the interval $[0, k^2]$ for each positive integer k . Thus, for example, $[1, 2] \cup [3, 5]$ is in H , but $[0, 1] \cup [2, 3] \cup [3, 5] \cup [7, 9] \cup [17, 18]$ is not, since four of the intervals in this union intersect the interval $[0, 3^2]$. Let us denote the interval $[0, k^2]$ by S_k . Then $\mu(S_k) = 1 - e^{-k^2}$ and (see [8]) $H|_{S_k}$ has VC dimension $2k$. If $\mathbf{x} \in S_k^n$ then $\Pi_H(\mathbf{x}) \leq n^{2k+1}$, by a crude form of Sauer's result. In any case, $\Pi_H(\mathbf{x}) \leq 2^n$ and it follows that

$$\begin{aligned} \mathbf{E}_n(\Pi_H(\mathbf{x})) &\leq \sum_{k=1}^n n^{2k+1} (\mu^n(S_k) - \mu^n(S_{k-1})) + 2^n (1 - \mu^n(S_n)) \\ &< \sum_{k=1}^n n^{2k+1} \left(1 - (1 - e^{-(k-1)^2})^n \right) + 2^n \left(1 - (1 - e^{-n^2})^n \right) < \sum_{k=1}^n n^{2k+2} e^{-(k-1)^2} + 2^n n e^{-n^2}. \end{aligned}$$

The second quantity tends to 0. Further, $n^{2x+2} e^{-(x-1)^2} \leq n^4 \exp((\ln n)^2)$, as can easily be checked by calculus, so that

$$\sum_{k=1}^n n^{2k+2} e^{-(k-1)^2} \leq n^5 \exp((\ln n)^2),$$

which is sub-exponential. It follows that H is μ -learnable.

5 Polynomial learnability

Suppose that H is μ -learnable. For learning to be efficient in any sense, we certainly need a sample-size bound which, as well as being independent of c , does not increase too dramatically as ϵ and δ decrease (and the learning task becomes, consequently, more difficult). It is appropriate to demand that, for efficiency, the sample-size (and hence running time of any efficient learning algorithm) be polynomial in $1/\epsilon$. Furthermore, since if one doubles the size of a sample, then one would expect to square the probability that a bad hypothesis is consistent with the sample, we require

the sample-size to vary polynomially in $\ln(1/\delta)$. We therefore make the following definition:

Definition 5.1 *Hypothesis space H is polynomially μ -learnable if for any ϵ, δ in $(0, 1)$, there is $m_0 = m_0(\epsilon, \delta)$, polynomial in $1/\epsilon$ and $\ln(1/\delta)$, such that, given any $c \in H$,*

$$m \geq m_0 \implies \mu^m \{ \mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) < \epsilon \} > 1 - \delta.$$

We have observed that if the expectation of $\Pi_H(\mathbf{x})$ is subexponential then H is μ -learnable. We have the following result.

Theorem 5.2 *Suppose H is a hypothesis space on X and μ is a distribution on X . If there is $0 < \alpha < 1$ such that (for large n), $\log \mathbf{E}_n(\Pi_H(\mathbf{x})) < n^{1-\alpha}$ then H is polynomially μ -learnable.*

Proof: Let $n = 2^{1-\alpha}(4/\epsilon)^{1/\alpha} \log(2/\delta)$, where \log denotes binary logarithm, and suppose that $\epsilon < 1/4$. Then $n \geq (4/\epsilon) \log(2/\delta)$ and so $\epsilon n/4 \geq \log(2/\delta)$. But, also, $n \geq 2^{1-\alpha}(4/\epsilon)^{1/\alpha}$ and hence $\epsilon n/4 \geq (2n)^{1-\alpha}$. It follows that

$$\frac{\epsilon n}{2} \geq \log\left(\frac{2}{\delta}\right) + (2n)^{1-\alpha} > \log\left(\frac{2}{\delta}\right) + \log \mathbf{E}_{2n}(\Pi_H(\mathbf{x})),$$

and so

$$2 \mathbf{E}_{2n}(\Pi_H(\mathbf{x})) 2^{-\epsilon n/2} < \delta.$$

The value of n is polynomial in $1/\epsilon$ and $\ln(1/\delta)$, so H is polynomially μ -learnable. \square

The above result is essentially the best that can be obtained by using the bound

$$\mu^n \{ \mathbf{x} \in X^n : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon \} < 2 \mathbf{E}_{2n} \Pi_H(\mathbf{x}) 2^{-\epsilon n/2},$$

since if the condition of the theorem is not satisfied (for example, if the expectation is of order $2^{n/\log n}$), the resulting sample-size bound will be exponential.

Bertoni *et al.* [7] studied the question of polynomial sample complexity for distribution-dependent learning. For $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, let $C_m(\mathbf{x})$ be the size of the largest subset of $\{x_1, \dots, x_m\}$ shattered by H . The, following on from the work of Vapnik

and Chervonenkis, Bertoni *et al.* showed that if there is a positive constant β such that

$$\mathbf{E}_{\mu^n} \left(\frac{C_m(\mathbf{x})}{m} \right) = O(m^{-\beta}),$$

then H is polynomially μ -learnable.

We now take a different approach, extending work of Ben-David *et al.* [5] to determine a sufficient condition for H to be polynomially μ -learnable. In [5], the following definition was made.

Definition 5.3 *A hypothesis space H over an input space X is said to have $X\sigma$ -finite dimension if $X = \bigcup_{i=1}^{\infty} B_i$ where the restriction $H|_{B_i}$ of H to domain B_i has finite VC dimension, for each i .*

Ben-David *et al.* [5] proved that if a hypothesis space H has $X\sigma$ -finite dimension then H is dd-learnable. The spaces in the examples of the previous section are easily seen to have $X\sigma$ -finite dimension and hence are dd-learnable; that is, they are μ -learnable for all probability distributions μ (and not just for the particular distributions discussed). (Indeed, if X is countable then any hypothesis space on X has $X\sigma$ -finite dimension, and the first example is a special case of this.) It follows also that the notion of dd-learnability is not a vacuous one, since these same hypothesis spaces are dd-learnable but, being of infinite VC dimension, are not learnable.

It is straightforward to give an example of a hypothesis space H over a (necessarily) uncountable input space X such that H does not have $X\sigma$ -finite dimension. Take X to be the closed interval $X = [0, 1]$, and let H be the space of all (characteristic functions of) finite unions of closed subintervals of X . Now, for any $Y \subseteq X$, $\text{VCdim}(H|_Y) \leq k$ if and only if $|Y| \leq k$. It follows that if X were the countable union $X = \bigcup_{i=1}^{\infty} B_i$ of sets B_i such that H had finite VC dimension on B_i then, in particular, each B_i would be finite and X , as the countable union of finite sets, would be countable. However, X is uncountable and we therefore deduce that H does not have $X\sigma$ -finite VC dimension.

The result of Ben-David *et al.* provides a positive distribution-dependent learnability result. However, it does not address the size of sample required for learnability to given degrees of accuracy and confidence. A closer analysis of the proof of this result in [5] shows that the resulting sufficient sample-size will not be polynomial in $1/\epsilon$ and $\log(1/\delta)$ for many distributions. To introduce the approach taken here, we first

have the following result, in which to say that a sequence $\{S_k\}_{k=1}^\infty$ of subsets of X is *increasing* means that $S_1 \subseteq S_2 \subseteq S_3 \subseteq \dots$.

Proposition 5.4 *H has $X\sigma$ -finite dimension if and only if there exists an increasing sequence $\{S_k\}_{k=1}^\infty$ of subsets of X such that $\bigcup_{k=1}^\infty S_k = X$ and $\text{VCdim}(H|S_k) \leq k$.*

Proof: Suppose that H has $X\sigma$ -finite dimension, and let the sets B_i be as in the definition. Let $x_0 \in B_1$ and set $B_0 = \{x_0\}$. For $k \geq 1$ let $S_k = \bigcup_{i=0}^{m(k)} B_i$, where $m(k)$ is the maximum integer m such that the restriction of H to $\bigcup_{i=0}^m B_i$ has VC dimension at most k . Given any $x \in X$, there is an m such that $x \in \bigcup_{i=0}^m B_i$. Suppose that H restricted to $\bigcup_{i=0}^m B_i$ has VC dimension k . Then $m(k) \geq m$, so $x \in S_k$. Conversely, if such sets S_i exist, take $B_i = S_i$. Then $\text{VCdim}(H|B_i)$ is finite, and $\bigcup_{i=1}^\infty B_i = X$. \square

If H “nearly” has finite VC dimension, in some sense, we might hope to get polynomially bounded sample-sizes. Motivated by the above result, we make the following definition.

Definition 5.5 *Hypothesis space H has polynomial $X\sigma$ -finite dimension with respect to μ if $X = \bigcup_{k=1}^\infty S_k$ where $\{S_k\}_{k=1}^\infty$ is increasing, $\text{VCdim}(H|S_k) \leq k$, and*

$$1 - \mu(S_k) = O\left(\frac{1}{k^c}\right)$$

for some constant $c > 0$.

Benedek and Itai [6] have gone some way towards investigating sufficient sample-sizes for distribution-dependent learnability in the case of discrete distributions (that is, distributions nonzero on only countably many elements of the example space). With the definition of polynomial $X\sigma$ -finite dimension, we can develop a theory for both continuous and discrete distributions. We have the following result, which we prove by a method similar to that used in [5].

Theorem 5.6 *Let H be a hypothesis space over X , and μ a probability measure defined on X . If H has polynomial $X\sigma$ -finite dimension with respect to μ , then H is *dd-learnable* and *polynomially μ -learnable*.*

Proof: Suppose that H has polynomial $X\sigma$ -finite dimension with respect to μ . Suppose that $0 < \epsilon < 1/4$ and $S \subseteq X$ is such that $\mu(S) \geq 1 - \epsilon/2$. The probability (with respect to μ^m) that a sample of length $m = 2l$, chosen according to μ , has at least half of its members in S is at least $1 - \sum_{k=0}^l \binom{2l}{k} \left(\frac{\epsilon}{2}\right)^{2l-k} \left(1 - \frac{\epsilon}{2}\right)^k$. Now,

$$\sum_{k=0}^l \binom{2l}{k} \left(\frac{\epsilon}{2}\right)^{2l-k} \left(1 - \frac{\epsilon}{2}\right)^k \leq \sum_{k=0}^l \binom{2l}{k} \left(\frac{\epsilon}{2}\right)^{2l-k} \leq \epsilon^l 2^{-l} \sum_{k=0}^l \binom{2l}{k} = \epsilon^l 2^{l-1}.$$

Therefore, this probability is at least $1 - \epsilon^l 2^{l-1}$. If $l \geq l_0 = \log(1/\delta)$ (where \log denotes logarithm to base 2) then

$$l(\log \epsilon + 1) \leq \log\left(\frac{1}{\delta}\right) (\log \epsilon + 1) = \log \delta \left(\log\left(\frac{1}{\epsilon}\right) - 1\right) < \log \delta$$

and this implies that the above probability is greater than $1 - \delta/2$. (Note that we have used the fact that, since $\epsilon < 1/4$, $\log \epsilon + 1$ is negative.)

Let $k(\epsilon) = \min\{k : \mu(S_k) \geq 1 - \epsilon/2\}$. The above shows that, with probability at least $1 - \delta/2$, a random sample of length $m \geq 2l_0$ has at least half of its members in $S = S_{k(\epsilon)}$. Let

$$m_* = 2 \left\lceil \frac{2\sqrt{2}}{\epsilon(\sqrt{2} - \sqrt{\epsilon})} \left(\ln\left(\frac{2d/(d-1)}{\delta}\right) + 2k(\epsilon) \ln\left(\frac{12}{\epsilon}\right) \right) \right\rceil.$$

Suppose $c \in H$ is the target concept. Since $H|S$ has VC dimension at most $k(\epsilon)$, m_* is, by Theorem 2.2, twice a sufficient sample size for the learnability of $H|S$ with accuracy $\epsilon/2$ and confidence $1 - \delta/2$. Let $m \geq m_*$, and let $l = \lfloor m/2 \rfloor \geq l_0$. If $\mathbf{x} \in X^m$ is such that \mathbf{x} has at least l of its entries from $S = S_{k(\epsilon)}$, then we shall denote by \mathbf{x}_S the unique vector of length l whose entries are precisely the first l entries of \mathbf{x} from S , appearing in the same order as in \mathbf{x} . Let μ_1 be the probability measure induced on S by μ . Thus, for any measurable subset A of X ,

$$\mu_1(A \cap S) = \frac{\mu(A)}{\mu(S)}.$$

Observe that if $h \in H[\mathbf{x}]$ and $\text{er}_\mu(h) > \epsilon$ then, since $\mu(S) \geq 1 - \epsilon/2$, the function $h|S$ (h restricted to S) is such that $h|S \in (H|S)[\mathbf{x}_S]$ and

$$\text{er}_{\mu_1}(h|S) = \frac{1}{\mu(S)} \mu(\{x \in X : h(x) \neq c(x)\} \cap S) > \frac{1}{\mu(S)} \left(\epsilon - \frac{\epsilon}{2}\right) > \frac{\epsilon}{2}.$$

Therefore, denoting the number of entries of a vector \mathbf{x} which lie in S by $s(\mathbf{x})$, we have

$$\mu^m \{\mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon\}$$

$$= \mu^m \{ \mathbf{x} : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon, s(\mathbf{x}) \geq l \} + \mu^m \{ \mathbf{x} : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon, s(\mathbf{x}) < l \}.$$

The second measure here is at most $\delta/2$ since with probability at least $1 - \delta/2$, $s(\mathbf{x})$ is at least l . Further,

$$\begin{aligned} & \mu^m \{ \mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon \text{ and } s(\mathbf{x}) \geq l \} \\ = & \mu^m \{ \mathbf{x} \in X^m : \text{haz}_\mu(H[\mathbf{x}]) > \epsilon | s(\mathbf{x}) \geq l \} \mu^m \{ \mathbf{x} \in X^m : s(\mathbf{x}) \geq l \} \\ \leq & \mu^m \{ \mathbf{x} \in X^m : \exists h \in H[\mathbf{x}] \text{ with } \text{er}_\mu(h) > \epsilon | s(\mathbf{x}) \geq l \} \\ \leq & \mu^m \{ \mathbf{x} \in X^m : \exists f \in (H|S)[\mathbf{x}_S] \text{ with } \text{er}_{\mu_1}(f) > \epsilon/2 \}, \end{aligned}$$

where, for any events A and B , $\mu^m(A|B)$ is the conditional probability (with respect to μ^m) of A given B . Now, if $s(\mathbf{x}) \geq l$ and \mathbf{x} is μ -randomly chosen, then \mathbf{x}_S is a μ_1 -randomly chosen sample of length l . Therefore this last measure is at most $\delta/2$, since l is a sufficient sample-size for the learnability of $H|S$ to accuracy $\epsilon/2$ with confidence $\delta/2$.

Note that the preceding analysis, since it holds true for any distribution μ , shows that H is dd-learnable. Now, since H has polynomial $X\sigma$ -finite dimension with respect to μ , there are $c, R > 0$ such that $1 - \mu(S_k) \leq R/k^c$, so that

$$k(\epsilon) \leq \left\lceil \left(\frac{2R}{\epsilon} \right)^{1/c} \right\rceil,$$

which is polynomial in $1/\epsilon$. Therefore m_* is a sufficient sample-size which is polynomial in $1/\epsilon$ and in $\ln(1/\delta)$, and hence H is polynomially μ -learnable. \square

To illustrate the idea of polynomial $X\sigma$ -finite dimension, consider again the examples of the previous section. For the first example, we see that the space has polynomial $X\sigma$ -finite dimension by taking S_k to be the union of the sets B_1 through to B_k . The sequence $\{S_k\}_{k=1}^\infty$ is increasing and $\bigcup_{k=1}^\infty S_k = X$. Further, if $\mathbf{x} \in S_k^m$ is shattered, the entries of \mathbf{x} must lie entirely within one of the B_i ($1 \leq i \leq k$) and hence

$$\text{VCdim}(H|S_k) = \max \{ \text{VCdim}(H|B_j) : j \leq k \} = \text{VCdim}(H|B_k) = k.$$

Now, $1 - \mu(S_k) = 1/2^k$, so H has polynomial $X\sigma$ -finite dimension with respect to μ and H is polynomially μ -learnable.

For the second example, let $S_k = [0, k^2]$. Then $\{S_k\}_{k=1}^\infty$ is an increasing sequence with union X and $\text{VCdim}(H|S_k) = 2k$. (Clearly, the factor 2 here is of no consequence.) Further, $1 - \mu(S_k) = e^{-k^2}$ and so H has polynomial $X\sigma$ -finite dimension with respect to μ .

It remains to give an example of a hypothesis space H over an input space X , together with a probability distribution μ on X , such that H has $X\sigma$ -finite dimension

but does *not* have polynomial $X\sigma$ -finite dimension with respect to μ . To this end, let X be the set of all positive integers and H the set of all (characteristic functions of) subsets of X . The input space is countable, and therefore H has $X\sigma$ -finite dimension. Define the probability measure μ on X by

$$\mu(\{x\}) = \frac{1}{\log(x+1)} - \frac{1}{\log(x+2)}.$$

Suppose that the sequence of sets $\{S_k\}_{k=1}^{\infty}$ is such that

$$X = \bigcup_{k=1}^{\infty} S_k \text{ and } \text{VCdim}(H|S_k) \leq k.$$

Clearly, $\text{VCdim}(H|S_k) = |S_k|$. But H restricted to S_k is supposed to have VC dimension at most k . Therefore, for each integer k , S_k has cardinality at most k . It follows that

$$\mu(S_k) \leq \mu(\{1, 2, \dots, k\}) = 1 - \frac{1}{\log(k+2)},$$

and $1 - \mu(S_k) \geq 1/\log(k+2)$. Thus, H does not have polynomial $X\sigma$ -finite dimension with respect to μ . In fact, one can show directly that H is *not* polynomially μ -learnable. For suppose that the target is the identically-0 function and that a sample \mathbf{x} of size m is given. There is a hypothesis consistent with the target on \mathbf{x} and with error at least ϵ unless $\mu(\{x_i : 1 \leq i \leq m\}) > 1 - \epsilon$. We therefore need to have

$$1 - \epsilon < \mu(\{x_i : 1 \leq i \leq m\}) \leq 1 - \frac{1}{\log(m+2)},$$

so that $m \geq e^{1/\epsilon} - 2$, which is exponential in $1/\epsilon$.

References

- [1] Dana Angluin, Queries and concept learning, *Machine Learning*, 2(4), 1988: 319–342.
- [2] Martin Anthony, *Uniform Convergence and Learnability*, PhD thesis, University of London 1991.
- [3] Martin Anthony and Norman Biggs, *Computational Learning Theory: An Introduction*, Cambridge University Press: Cambridge, UK, 1992.
- [4] Martin Anthony, Norman Biggs and John Shawe-Taylor, The learnability of formal concepts, *Proceedings of the Third Workshop on Computational Learning Theory*, Morgan Kaufman, San Mateo, CA, 1990.

- [5] Shai Ben-David, Gyora M. Benedek and Yishay Mansour, A parameterization scheme for classifying models of learnability, *Proceedings of the Second Workshop on Computational Learning Theory*, Morgan Kaufman, San Mateo, CA, 1989.
- [6] Gyora M. Benedek and Alon Itai, Learnability with respect to fixed distributions, to appear, *Theoretical Computer Science*.
- [7] A. Bertoni, P. Campadelli, A. Morpurgo, and S. Panizza, Polynomial uniform convergence and polynomial-sample learnability, In *Proceedings 5th Annual Workshop on Computational Learning Theory*, pages 265–271. ACM Press, New York, NY, 1992.
- [8] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler and Manfred Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *Journal of the ACM*, 36(4), 1989: 929–965.
- [9] David Haussler, Quantifying inductive bias: AI learning algorithms and Valiant’s learning framework, *Artificial Intelligence*, 36, 1988: 177-221.
- [10] David Haussler and Emo Welzl, ϵ -nets and simplex range queries, *Discrete Comp. Geom.*, 2, 1987: 127-151.
- [11] Michael J. Kearns and Umesh Vazirani (1995). *Introduction to Computational Learning Theory*, MIT Press 1995.
- [12] Balas K. Natarajan, *Machine Learning: A Theoretical Approach*, Morgan Kaufmann, San Mateo, California, 1991.
- [13] David Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, 1984.
- [14] N. Sauer, On the density of families of sets, *J. Comb. Theory (A)*, 13, 1972: 145–147.
- [15] Leslie G. Valiant, A theory of the learnable. *Communications of the ACM*, 27(11), 1984: 1134–1142.
- [16] 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*, 16(2), 1971: 264-280.