Combinatorics, Probability and Computing (2000) 9, 213–225. Printed in the United Kingdom

Function Learning from Interpolation

MARTIN ANTHONY[1] and PETER L. BARTLETT[2]

[1] Department of Mathematics,
The London School of Economics and Political Science,
Houghton Street, London WC2A 2AE, England
(e-mail: m.anthony@lse.ac.uk)

[2] Department of Systems Engineering,
Research School of Information Sciences and Engineering,
The Australian National University, Canberra, 0200 Australia
(e-mail: Peter.Bartlett@anu.edu.au)

Received 6 March 1996; revised 26 September 1999

In this paper, we study a statistical property of classes of real-valued functions that we call approximation from interpolated examples. We derive a characterization of function classes that have this property, in terms of their 'fat-shattering function', a notion that has proved useful in computational learning theory. The property is central to a problem of learning real-valued functions from random examples in which we require satisfactory performance from every algorithm that returns a function which approximately interpolates the training examples.

1. Introduction

In the problem of learning a real-valued function from examples, a learner sees a sequence of values of an unknown function at a number of randomly chosen points. On the basis of these examples, the learner chooses a function – called a hypothesis – from some class $\mathscr{H}$ of hypotheses, with the aim that the learner's hypothesis is close to the target function on future random examples. In this paper we require that, for most training samples, with high probability the absolute difference between the values of the learner's hypothesis and the target function on a random point is small.

A natural learning algorithm to consider is one that chooses a function in $\mathscr{H}$ that is close to the target function on the training examples. This poses the following statistical problem: For what function classes $\mathscr{H}$ will any function in $\mathscr{H}$ that approximately interpolates the target function on the training examples probably have small absolute

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

error? More precisely, we have the following definition of approximation from interpolated examples.

Definition. Let $\mathscr{C}, \mathscr{H}$ be sets of functions that map from a set X to $\mathbb{R}$. We say that $\mathscr{H}$ *approximates* $\mathscr{C}$ *from interpolated examples* if, for all $\eta, \gamma, \epsilon, \delta \in (0,1)$, there is an $m_0(\eta, \gamma, \epsilon, \delta)$ such that, for every $t \in \mathscr{C}$ and for every probability measure† P on X, if $m \geqslant m_0(\eta, \gamma, \epsilon, \delta)$, then with P^m-probability at least $1-\delta$, $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in X^m$ has the property that, if $h \in \mathscr{H}$ and $|h(x_i) - t(x_i)| < \eta$ for $1 \leqslant i \leqslant m$, then

$$P\left(\{x \in X : |h(x) - t(x)| \geqslant \eta + \gamma\}\right) < \epsilon.$$

We say that $m_0(\eta, \gamma, \epsilon, \delta)$ is a *sufficient sample length function* for $\mathscr{H}$ to approximate $\mathscr{C}$ from interpolated examples.

Two cases of particular interest are those in which $\mathscr{C} = \mathscr{H}$ and $\mathscr{C} = \mathbb{R}^X$, the set of all functions from X to $\mathbb{R}$. If $\mathscr{H}$ approximates $\mathscr{H}$ from interpolated examples, we simply say that $\mathscr{H}$ *approximates from interpolated examples*. The main aim of this paper is to find characterizations of classes which approximate from interpolated examples and which approximate $\mathbb{R}^X$ from interpolated examples.

This problem can be interpreted as a learning problem in which we require satisfactory performance from every algorithm that returns a function that approximately interpolates the training examples. If, instead of requiring that all algorithms in this class be suitable, we require only the existence of a suitable algorithm, no necessary and sufficient conditions on the function class $\mathscr{H}$ are known. Because an arbitrary amount of information can be conveyed in a single real value, it is possible to construct complicated function classes in which the identity of a function is encoded in its value at every point, and an algorithm can take advantage of this (see [3]). We can avoid this unnatural 'conspiracy' between algorithms and function classes in two ways: by requiring that the algorithm be robust in the presence of random observation noise, as was considered in [3], or, contrastingly, by requiring satisfactory performance of every algorithm in a class of reasonable algorithms, as we consider here. Another reason for studying the problem of this paper is that it has implications for learning in the presence of *malicious noise*, in which the labels on the training sample can be any real numbers within η of the true value of the target. This will be discussed later in the paper, but for the moment simply observe that, if h is β-close to a training sample where the labels have been corrupted to a level of at most β, then h is certainly 2β-close to the target on the sample. If $\mathscr{H}$ approximates from interpolated examples, we can then deduce that if the sample is large enough then (with high probability) h is within $2\beta + \gamma$ of the target on 'most' of X.

Alon, Ben-David, Cesa-Bianchi and Haussler [1] have analysed a model of learning in which the error of a hypothesis is taken to be the expected value of $(h(x) - t(x))^2$. Their

† More formally, one has a fixed σ-algebra on X: when X is countable this is 2^X, and when $X \subseteq \mathbb{R}^n$, it is the Borel σ-algebra. Then, by 'any probability measure on X', we mean 'for any probability measure on Σ', where Σ, the fixed σ-algebra, is understood. The class $\mathscr{H}$ must have some fairly benign measurability properties; we refer to [12, 8] for details.

results can be used to provide guarantees of small *expected* absolute error. However, the results of this paper provide conditions under which we can (with high probability) have small 'pointwise' absolute error *almost everywhere* on the domain, and these results do not follow from those of Alon and co-workers.

In the next section, we define a measure of the complexity of a class $\mathscr{H}$ of functions (the fat-shattering function), and we state the main result: that the fat-shattering function is the key quantity in this problem. In Sections 3 and 4 we give upper and lower bounds on the number of examples necessary for approximation from interpolated examples. Section 5 describes the implications for learning with malicious noise.

2. Definitions and the main result

A number of ways of measuring the 'expressive power' of a class $\mathscr{H}$ of functions have been proposed. This power is quantified by associating a 'dimension' to the class. Sometimes this is simply one number depending on $\mathscr{H}$. Sometimes – in what is known as a *scale-sensitive dimension* – it is a function depending on $\mathscr{H}$.

An important example of the first type of dimension is the *pseudo-dimension* [8, 12]. We say that a finite subset $S = \{x_1, x_2, \ldots, x_d\}$ of X is *shattered* if there is an $\mathbf{r} = (r_1, r_2, \ldots, r_d) \in \mathbb{R}^d$ such that, for every $\mathbf{b} = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $h_\mathbf{b} \in \mathscr{H}$ with $h_\mathbf{b}(x_i) > r_i$ if $b_i = 1$ and $h_\mathbf{b}(x_i) < r_i$ if $b_i = 0$. The *pseudo-dimension* of $\mathscr{H}$, denoted $\mathrm{Pdim}(\mathscr{H})$, is the largest cardinality of a shattered set, or infinity if there is no bound on the cardinalities of the shattered sets.

Perhaps the most important scale-sensitive dimension that has been used to date in the development of the theory of learning real-valued functions is the *fat-shattering function*. This is a scale-sensitive version of the pseudo-dimension and was introduced by Kearns and Schapire [9]. Suppose that $\mathscr{H}$ is a set of functions from X to $[0,1]$ and that $\gamma \in (0,1)$. We say that a finite subset $S = \{x_1, x_2, \ldots, x_d\}$ of X is *γ-shattered* if there is an $\mathbf{r} = (r_1, r_2, \ldots, r_d) \in \mathbb{R}^d$ such that, for every $\mathbf{b} = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $h_\mathbf{b} \in \mathscr{H}$ with $h_\mathbf{b}(x_i) \geqslant r_i + \gamma$ if $b_i = 1$ and $h_\mathbf{b}(x_i) \leqslant r_i - \gamma$ if $b_i = 0$. Thus, S is γ-shattered if it is shattered with a 'width of shattering' of at least γ. We define the *fat-shattering function*, $\mathrm{fat}_\mathscr{H} : \mathbb{R}^+ \to \mathbb{N}_0 \cup \{\infty\}$, as

$$\mathrm{fat}_\mathscr{H}(\gamma) = \max\{|S| : S \subseteq X \text{ is } \gamma\text{-shattered by } \mathscr{H}\},$$

or $\mathrm{fat}_\mathscr{H}(\gamma) = \infty$ if the maximum does not exist. (Here, $\mathbb{N}_0$ denotes the set of nonnegative integers.) It is easy to see that $\mathrm{Pdim}(\mathscr{H}) = \lim_{\gamma \to 0} \mathrm{fat}_\mathscr{H}(\gamma)$. It should be noted, however, that it is possible for the pseudo-dimension to be infinite, even when $\mathrm{fat}_\mathscr{H}(\gamma)$ is finite for all γ. We shall say that $\mathscr{H}$ has *finite fat-shattering function* whenever it is the case that, for all $\gamma \in (0,1)$, $\mathrm{fat}_\mathscr{H}(\gamma)$ is finite.

The fat-shattering function plays an important role in the learning theory of real-valued functions. Kearns and Schapire [9] proved that if a class of probabilistic concepts is learnable, then the class has finite fat-shattering function. (A probabilistic concept f is a $[0,1]$-valued function. In this model, the learner sees examples (x_i, y_i), where $\Pr(y_i = 1) = f(x_i)$.) Alon, Ben-David, Cesa-Bianchi and Haussler [1] proved, conversely, that, if a class of probabilistic concepts has finite fat-shattering function, then it is

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

learnable. The main result in [3] is that finiteness of the fat-shattering function of a class of $[0,1]$-valued functions is a necessary and sufficient condition for learning with random observation noise.

Our main result is the following.

Theorem 2.1. *Suppose that $\mathscr{H}$ is a set of functions from a set X to $[0,1]$. Then the following propositions are equivalent.*

(1) *$\mathscr{H}$ approximates from interpolated examples.*
(2) *$\mathscr{H}$ approximates $\mathbb{R}^X$ from interpolated examples.*
(3) *$\mathscr{H}$ has finite fat-shattering function.*

3. The upper bound

In this section, we prove that finite fat-shattering function is a sufficient condition for approximation from interpolated examples and we provide a suitable sample length function $m_0(\eta, \gamma, \epsilon, \delta)$.

We first need the notion of *covering numbers* $\mathscr{N}_A(\alpha, d)$, as used extensively in [8, 1, 6], for instance. Suppose that (A, d) is a pseudo-metric space and $\alpha > 0$. Then, a subset N of A is said to be an α-*cover* for a subset B of A if, for every $x \in B$, there is an $\hat{x} \in N$ such that $d(x, \hat{x}) \leqslant \alpha$. The metric space is *totally bounded* if there is a *finite* α-cover for A, for all $\alpha > 0$. When (A, d) is totally bounded, we shall denote the minimal cardinality of an α-cover for A by $\mathscr{N}_A(\alpha, d)$ for $\alpha > 0$. A subset M of A is said to be α-*separated* if, for all distinct $x, y \in M$, $d(x, y) \geqslant \alpha$. We shall denote the maximal cardinality of an α-separated subset of A by $\mathscr{M}_A(\alpha, d)$. It is easy to show that

$$\mathscr{M}_A(2\alpha, d) \leqslant \mathscr{N}_A(\alpha, d) \leqslant \mathscr{M}_A(\alpha, d)$$

(see [10]), so $\mathscr{M}_A(\alpha, d)$ is always defined if (A, d) is totally bounded. Suppose now that $\mathscr{H}$ is a set of functions from a set X to $[0,1]$ and that $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in X^m$, where m is a positive integer. We may define a pseudo-metric $l_{\mathbf{x}}^{\infty}$ on $\mathscr{H}$ as follows: for $g, h \in \mathscr{H}$,

$$l_{\mathbf{x}}^{\infty}(g, h) = \max_{1 \leqslant i \leqslant m} |g(x_i) - h(x_i)|.$$

(This metric has been used in [6, 1], for example.) Alon, Ben-David, Cesa-Bianchi and Haussler [1] obtained (essentially) the following result bounding the $l_{\mathbf{x}}^{\infty}$-covering number of $\mathscr{H}$ in terms of the fat-shattering function of $\mathscr{H}$.

Lemma 3.1. *Suppose that $\mathscr{H}$ is a set of functions from X to $[0,1]$ and that $\mathscr{H}$ has finite fat-shattering function. Let $m \in \mathbb{N}$, and $\mathbf{x} \in X^m$. Then the pseudo-metric space $(\mathscr{H}, l_{\mathbf{x}}^{\infty})$ is totally bounded. Suppose $\alpha > 0$. Let $d = \mathrm{fat}_{\mathscr{H}}(\alpha/4)$ and*

$$y = \sum_{i=1}^{d} \binom{m}{i} \left(\left\lceil \frac{2}{\alpha} \right\rceil \right)^i.$$

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

Then, provided $m \geqslant \log y + 1$,

$$\mathcal{N}_{\mathscr{H}}(\alpha, l_{\mathbf{x}}^{\infty}) < 2\left(m\left\lceil\frac{2}{\alpha}\right\rceil^2\right)^{\log y}.$$

Here, as elsewhere in the paper, log denotes logarithm to base 2. We then have the following result.

Theorem 3.2. *Suppose that* $\mathscr{H}$ *is a class of functions mapping from a domain* X *to the real interval* $[0,1]$ *and that* $\mathscr{H}$ *has finite fat-shattering function. Let* t *be any function from* X *to* $\mathbb{R}$ *and let* $\gamma, \eta, \epsilon > 0$. *Let* P *be any probability distribution on* X *and define* B *to be the set of functions* $h \in \mathscr{H}$ *for which* $P\left(\{x \in X : |h(x) - t(x)| \geqslant \eta + \gamma\}\right) \geqslant \epsilon$. *Let* $d = \mathrm{fat}_{\mathscr{H}}(\gamma/8)$ *and let*

$$y = \sum_{i=1}^{d} \binom{2m}{i} \left(\left\lceil\frac{4}{\gamma}\right\rceil\right)^i.$$

Then, for $m \geqslant \max\left(8/\epsilon, \log y + 1\right)$, *the probability that some* h *in* B *has* $|h(x_i) - t(x_i)| < \eta$ *for* $1 \leqslant i \leqslant m$ *is at most*

$$4\left(2m\left\lceil\frac{4}{\gamma}\right\rceil^2\right)^{\log y} 2^{-\epsilon m/2}.$$

Proof. The proof is based on a technique analogous to that used in [13, 5, 8], where we 'symmetrize' and then 'combinatorially bound'. The first step – symmetrization – relates the desired probability to a 'sample-based' one. Fix t, P, m, the parameters γ, η, ϵ, and hence the set B. It is easy to show using standard techniques that

$$P^m\{\mathbf{x} \in X^m : \exists h \in B, |h(x_i) - t(x_i)| < \eta \ (1 \leqslant i \leqslant m)\} \leqslant P^{2m}(R),$$

where

$$R = \{\mathbf{xy} \in X^{2m} : \exists h \in B, |h(x_i) - t(x_i)| < \eta \ (1 \leqslant i \leqslant m) \text{ and } |\{i : |h(y_i) - t(y_i)| \geqslant \eta + \gamma\}| > \epsilon m/2\},$$

and $\mathbf{xy} \in X^{2m}$ denotes the concatenation of $\mathbf{x}, \mathbf{y} \in X^m$.

The next step is to bound the probability of R using combinatorial techniques. For this, let Γ be the 'swapping group' [12] of permutations on the set $\{1, 2, \ldots, 2m\}$. This is the group generated by the transpositions $(i, m+i)$ for $1 \leqslant i \leqslant m$. The group Γ acts in a natural way on vectors in X^{2m}: for $\sigma \in \Gamma$ and $\mathbf{z} \in X^{2m}$, we define $\sigma\mathbf{z}$ to be

$$(z_{\sigma(1)}, z_{\sigma(2)}, \ldots, z_{\sigma(2m)}).$$

Let $\Gamma(R, \mathbf{z}) = |\{\sigma \in \Gamma : \sigma\mathbf{z} \in R\}|$ be the number of permutations in Γ taking $\mathbf{z}$ into R. It is well known that, since P^{2m} is a product distribution, we have

$$P^{2m}(R) \leqslant \frac{1}{2^m} \max_{\mathbf{z} \in X^{2m}} \Gamma(R, \mathbf{z}).$$

Now, let us fix $\mathbf{z} \in X^{2m}$ and consider the pseudo-metric space $(B, l_{\mathbf{z}}^{\infty})$. Since $\mathscr{H}$ has finite

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

fat-shattering function, so does B and Lemma 3.1 implies that this pseudo-metric space is totally bounded. Let $N = \{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n\}$ be a minimal $\gamma/2$-cover for B. From Lemma 3.1,

$$n < 2\left(2m\left\lceil\frac{4}{\gamma}\right\rceil^2\right)^{\log y}.$$

Since N is a $\gamma/2$-cover, given $h \in B$, there is an $\hat{h} \in N$ such that $l_{\mathbf{z}}^{\infty}(\hat{h}, h) < \gamma/2$, which means that, for $1 \leqslant i \leqslant 2m$, $|\hat{h}(z_i) - h(z_i)| < \gamma/2$. Suppose that $\sigma\mathbf{z} = \mathbf{xy} \in R$. Then, by the definition of R, there is some $h \in B$ such that $|h(x_i) - t(x_i)| < \eta$ for $1 \leqslant i \leqslant m$ and such that, for more than $\epsilon m/2$ of the y_i, $|h(y_i) - t(y_i)| \geqslant \eta + \gamma$. But (taking $\hat{h}$ to be, as described above, a function in the cover $\gamma/2$-close to h) this implies that there is an $\hat{h} \in N$ such that, for $1 \leqslant i \leqslant m$, $|\hat{h}(x_i) - t(x_i)| < \eta + \gamma/2$, and such that, for more than $\epsilon m/2$ of the y_i, $|\hat{h}(y_i) - t(y_i)| \geqslant \eta + \gamma/2$. It follows from this that, if $\sigma\mathbf{z} \in R$, then, for some l between 1 and n, $\sigma\mathbf{z}$ belongs to the set $\hat{R}_l$, defined by

$$\hat{R}_l = \{\mathbf{xy} \in X^{2m} : |\hat{h}_l(x_i) - t(x_i)| < \eta + \gamma/2\,(1 \leqslant i \leqslant m), \text{ and } |\{i : |\hat{h}_l(y_i) - t(y_i)| \geqslant \eta + \gamma/2\}| > \epsilon m/2\}.$$

Let $\Gamma(\hat{R}_l, \mathbf{z})$ be the number of σ in Γ for which $\sigma\mathbf{z} \in \hat{R}_l$. Since $\sigma\mathbf{z} \in R$ implies $\sigma\mathbf{z} \in \hat{R}_l$ for some l, we have

$$\Gamma(R, \mathbf{z}) \leqslant \sum_{l=1}^{n} \Gamma(\hat{R}_l, \mathbf{z}).$$

Consider a particular l between 1 and n and suppose that $\Gamma(\hat{R}_l, \mathbf{z}) \neq 0$. Let k be the number of indices i between 1 and $2m$ such that $|\hat{h}_l(z_i) - t(z_i)| \geqslant \eta + \gamma/2$. Then $\epsilon m/2 < k \leqslant m$. The number of permutations σ in Γ for which $\sigma\mathbf{z}$ belongs to $\hat{R}_l$ is then equal to 2^{m-k}, which is less than $2^{m(1-\epsilon m/2)}$. (The z_i which can be 'swapped' are precisely those $m-k$ satisfying $|\hat{h}_l(z_{m+i}) - t(z_{m+i})| < \eta + \gamma/2$.) It follows that

$$P^{2m}(R) < \frac{1}{2^m}\sum_{i=1}^{n} 2^{m(1-\epsilon/2)} \leqslant n2^{-\epsilon m/2} \leqslant 2\left(2m\left\lceil\frac{4}{\gamma}\right\rceil^2\right)^{\log y} 2^{-\epsilon m/2}.$$

The statement of the theorem now follows. □

We thus obtain the following corollary, which shows that finiteness of $\mathrm{fat}_{\mathscr{H}}$ implies that $\mathscr{H}$ approximates $\mathbb{R}^X$ from interpolated examples, and hence $\mathscr{H}$ approximates $\mathscr{H}$ from interpolated examples. The proof is an easy calculation.

Corollary 3.3. *Suppose that $\mathscr{H}$ is a set of functions from X to $[0,1]$ and that $\mathscr{H}$ has finite fat-shattering function. Then $\mathscr{H}$ approximates $\mathbb{R}^X$ from interpolated examples. Furthermore, there is a positive constant K such that a sufficient sample length function is*

$$m_0(\gamma, \eta, \epsilon, \delta) = \frac{K}{\epsilon}\left(\log\left(\frac{1}{\delta}\right) + d\log^2\left(\frac{d}{\gamma\epsilon}\right)\right),$$

where $d = \mathrm{fat}_{\mathscr{H}}(\gamma/8)$. □

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

4. The lower bound

In this section, we give lower bounds on the number of examples necessary for $\mathscr{H}$ to approximate $\mathbb{R}^X$ from interpolated examples and for $\mathscr{H}$ to approximate $\mathscr{H}$ from interpolated examples. The bounds are in terms of $\text{fat}_{\mathscr{H}}$, the fat-shattering function of $\mathscr{H}$. To prove them, we consider a discretized version of $\mathscr{H}$. We then consider a number of notions of dimension for classes that map to these discrete sets, and show that a large family of these dimensions consists of closely related members. This family includes a version of the Natarajan dimension – see [11] – for which it is easy to prove lower bounds. Since the fat-shattering function is also a member of this family of closely related dimension, we obtain the lower bound. This broad outline is similar to the approach adopted by Ben-David, Cesa-Bianchi, Haussler and Long [4], who consider learning $[n]$-valued functions.

We first define the discretization we shall use. For $a \in [0,1]$, let $D_\gamma(a) = \lceil a/\gamma \rceil$. For a function $f : X \to [0,1]$, let $D_\gamma(f) : X \to \{0,1,\ldots,\lceil 1/\gamma \rceil\}$ be defined as the composition of D_γ and f. Let $D_\gamma(\mathscr{H})$ denote $\{D_\gamma(f) : f \in \mathscr{H}\}$. Functions in $D_\gamma(\mathscr{H})$ map to $\{0,1,\ldots,n\}$, where $n = \lceil 1/\gamma \rceil$. Let $[n]$ denote $\{0,1,\ldots,n\}$.

From the definition of the fat-shattering function,

$$\text{fat}_{\mathscr{H}}(\alpha) \leqslant \text{fat}_{D_\gamma(\mathscr{H})}\left(\frac{1}{2}\left\lfloor \frac{2\alpha}{\gamma} \right\rfloor\right)$$

for $\alpha, \gamma \in \mathbb{R}^+$.

We consider the following notions of dimension, defined using classes of $\{0,1,*\}$-valued functions on $[n]$.

Definition. If $\mathscr{F}$ is a class of $[n]$-valued functions defined on X and Ψ is a class of $\{0,1,*\}$-valued functions defined on $[n]$, we say that $\mathscr{F}$ Ψ-shatters $\mathbf{x} = (x_1,\ldots,x_d) \in X^d$ if there is a sequence $(\phi_1,\ldots,\phi_d) \in \Psi^d$ such that

$$\{0,1\}^d \subseteq \{(\phi_1(f(x_1)),\ldots,\phi_d(f(x_d))) : f \in \mathscr{F}\}.$$

The Ψ-dimension of $\mathscr{F}$, denoted $\Psi\text{-dim}(\mathscr{F})$, is the size of the largest Ψ-shattered sequence, or infinity if there is no largest sequence.

We can express the fat-shattering function $\text{fat}_{\mathscr{F}}(k)$ as a dimension of this type, for $k \geqslant 2$. Define $\Psi_{\text{fat}}(k) = \{\psi_i : i \in \{0,\ldots,n-k\}\}$, with

$$\psi_i(z) = \begin{cases} 1, & z \geqslant i+k, \\ *, & i < z < i+k, \\ 0, & z \leqslant i, \end{cases}$$

for $z \in [n]$. Then $\text{fat}_{\mathscr{F}}(\alpha) = \Psi_{\text{fat}}(\lfloor 2\alpha \rfloor)\text{-dim}(\mathscr{F})$ for all classes $\mathscr{F}$ of functions from X to $[n]$.

The Ψ_{gnat}-dimension (a 'gapped' version of the Natarajan dimension) will be useful, since it is easy to prove lower bounds using this dimension.

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

Definition. Let $\Psi_{\text{gnat}}(k)$ be the following set of $\{0,1,*\}$-valued functions defined on $[n]$, where $k \in \{2,3,\ldots\}$:

$$\Psi_{\text{gnat}}(k) = \{\psi_{i,j} : i,j \in [n], |i-j| \geqslant k\},$$

with

$$\psi_{i,j}(\alpha) = \begin{cases} 1, & \alpha = i, \\ 0, & \alpha = j, \\ *, & \text{otherwise.} \end{cases}$$

The following result and its proof (which we omit) are similar to the key result in [1], Lemma 15, which bounds covering numbers of $\mathscr{F}$ in terms of $\text{fat}_{\mathscr{F}}(1)$. We will see later that it generalizes that lemma, since the $\Psi_{\text{gnat}}(2)$-dimension is the smallest of a family of dimensions that includes $\text{fat}_{\mathscr{F}}(1)$.

Lemma 4.1. *Let $k \geqslant 2$ and $n \geqslant 1$ be integers. Suppose that $\mathscr{F}$ is a class of $[n]$-valued functions defined on X satisfying $\Psi_{\text{gnat}}(k)\text{-dim}(\mathscr{F}) \leqslant d$, and that $m \geqslant \log y + 1$, where*

$$y = \sum_{i=0}^{d} \binom{m}{i} n^{2i}.$$

Then

$$\max_{\mathbf{x} \in X^m} \mathscr{M}_{\mathscr{F}}(k, l_{\mathbf{x}}^{\infty}) < 2(2mn^2)^{\log y}.$$

The 'k-gapped distinguishers' correspond to a family of dimensions that includes the $\Psi_{\text{gnat}}(k)$-dimension and the fat-shattering function at a certain scale.

Definition. Let $k \geqslant 2$. A set Ψ of functions from $[n]$ to $\{0,1,*\}$ is a *k-gapped distinguisher* if it satisfies:

(1) for all $i \in \{0,1,\ldots,n-k\}$ and $j \in \{i+k,\ldots,n\}$, there is a function $\psi \in \Psi$ and a bit $b \in \{0,1\}$ such that $\psi(i) = b$ and $\psi(j) = 1-b$;
(2) $\min\{|i-j| : i,j \in [n], \exists \psi \in \Psi, \psi(i) = 0, \psi(j) = 1\} = k$.

In addition to the set $\Psi_{\text{gnat}}(k)$, another important example of a k-gapped distinguisher is the class

$$\Psi_{\text{g}}(k) = \left\{\psi \in \{0,1,*\}^{[n]} : \min\{|i-j| : i,j \in [n], \psi(i) = 0, \psi(j) = 1\} = k\right\}.$$

In fact $\Psi_{\text{g}}(k)$ is the largest k-gapped distinguisher, in the sense that it contains any other k-gapped distinguisher.

Lemma 4.2. *Suppose $\mathscr{F}$ is a class of $[n]$-valued functions defined on X, Ψ is a class of $\{0,1,*\}$-valued functions defined on $[n]$, and $k \geqslant 2$. If Ψ is a k-gapped distinguisher then $\Psi_{\text{gnat}}(k)\text{-dim}(\mathscr{F}) \leqslant \Psi\text{-dim}(\mathscr{F}) \leqslant \Psi_{\text{g}}(k)\text{-dim}(\mathscr{F})$.*

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

Proof. Take a $\Psi_{\text{gnat}}(k)$-shattered sequence $\mathbf{x} \in X^d$. Since Ψ is a k-gapped distinguisher, for all $\psi_{i,j} \in \Psi_{\text{gnat}}(k)$ there is a $\phi \in \Psi$ and a $b \in \{0,1\}$ for which $\phi(i) = b$ and $\phi(j) = 1-b$. It follows that $\mathbf{x}$ is Ψ-shattered, which gives the first inequality. The second inequality follows from the fact that $\Psi \subseteq \Psi_{\text{g}}(k)$. □

It follows from Lemma 4.2 that Lemma 4.1 generalizes Alon, Ben-David, Cesa-Bianchi and Haussler's Lemma 15 [1], which gave a similar result for the $\Psi_{\text{fat}}(2)$-dimension. The following result shows that the $\Psi_{\text{gnat}}(k)$-dimension, the $\Psi_{\text{g}}(k)$-dimension, and the Ψ-dimension (for any k-gapped distinguisher Ψ) are all closely related.

Lemma 4.3. *Let $k \geqslant 2$. Let $\mathscr{F}$ be a class of functions that map from X to $[n]$, satisfying $\Psi_{\text{g}}(k)\text{-dim}(\mathscr{F}) \geqslant d \geqslant 2$. Then*

$$\Psi_{\text{gnat}}(k)\text{-dim}(\mathscr{F}) > \frac{d}{3\log^2(2dn^2)}.$$

Proof. Suppose $\mathbf{x} = (x_1, \ldots, x_d) \in X^d$ is $\Psi_{\text{g}}(k)$-shattered by $\mathscr{F}$. The definition of $\Psi_{\text{g}}(k)$ implies that any (minimal) subset of $\mathscr{F}$ that $\Psi_{\text{g}}(k)$-shatters $\mathbf{x}$ is k-separated, so $\mathscr{M}_{\mathscr{F}}(k, l_{\mathbf{x}}^{\infty}) \geqslant 2^d$.

Suppose $\Psi_{\text{g}}(k)\text{-dim}(\mathscr{F}) = d_N$, and let

$$y = \sum_{i=0}^{d_N} \binom{d}{i} n^{2i}.$$

If $d > \log y$ then, by Lemma 4.1,

$$2^d \leqslant \mathscr{M}_{\mathscr{F}}(k, l_{\mathbf{x}}^{\infty}) < 2(2dn^2)^{\log y},$$

so

$$d < 1 + \log y \log(2dn^2). \tag{4.1}$$

Alternatively, if $d \leqslant \log y$, then (4.1) is obviously true. Clearly, $d_N = 0$ only if $d = 0$, so assume $d_n \geqslant 1$. Then $y \leqslant 2d_N d^{d_N} n^{2d_N}$, so we have

$$\begin{aligned} d &< 1 + \log y \log(2dn^2) \\ &\leqslant 1 + (\log(2d_N) + d_N \log(dn^2)) \log(2dn^2) \\ &\leqslant 3d_N \log^2(2dn^2). \end{aligned}$$

□

The following result follows easily from [7, Theorem 1], which gives a lower bound on the number of examples necessary for learning $\{0,1\}^{[d]}$ in the probably approximately correct model (see also [5]).

Lemma 4.4. *Let $0 < \epsilon \leqslant 1/8$, $0 < \delta < 1/100$, and $d \geqslant 1$. If*

$$m < \max\left(\frac{d}{32\epsilon}, \frac{1-\epsilon}{\epsilon} \ln \frac{1}{\delta}\right),$$

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

then there is a distribution P on $[d]$ and a function $t \in \{0,1\}^{[d]}$ such that

$$P^m\left\{\mathbf{x} \in X^m : \exists f \in \{0,1\}^{[d]} \text{ such that } f(x_i) = t(x_i), i = 1,\ldots,m \text{ and } P\{y : f(y) \neq t(y)\} \geqslant \epsilon\right\} \geqslant \delta.$$

We use Lemma 4.4 to prove the following lower bound on the sample length function for approximating $\mathbb{R}^X$ from interpolated examples.

Theorem 4.5. *Suppose $\mathscr{H}$ is a class of $[0,1]$-valued functions defined on a set X, $0 < \gamma < \eta < 1$, and $\epsilon, \delta \in (0,1)$. Then if $\mathrm{fat}_{\mathscr{H}}(\gamma) \geqslant d \geqslant 1$ and $\gamma^2 \geqslant 4d2^{-\sqrt{d/6}}$, any sample length function m_0 for $\mathscr{H}$ to approximate $\mathbb{R}^X$ from interpolated examples satisfies*

$$m_0(\eta,\gamma,\epsilon,\delta) \geqslant \max\left(\frac{1}{32\epsilon}\left(\frac{d}{3\log^2(4d/\gamma^2)} - 1\right), \frac{1}{\epsilon}\log\frac{1}{\delta}\right).$$

Proof. Fix $0 < \gamma < \eta < 1$, define $n = \lceil 1/\gamma \rceil$, and suppose $\mathrm{fat}_{\mathscr{H}}(\gamma) \geqslant d$. Let $\mathscr{F} = D_\gamma(\mathscr{H})$. Then $\mathrm{fat}_{\mathscr{F}}(1) \geqslant d$, so $\Psi_{\mathrm{gnat}}(2)\text{-dim}(\mathscr{F}) > k$, where $k = d/(3\log^2(2dn^2))$. Consider a sequence $(x_1,\ldots,x_k) \in X^k$ that is $\Psi_{\mathrm{gnat}}(2)$-shattered by $\mathscr{F}$. Clearly, there is a subset $\mathscr{H}_0 \subseteq \mathscr{H}$ with $|\mathscr{H}_0| = 2^k$ and a sequence $\left(\psi_{a_1,b_1},\ldots,\psi_{a_k,b_k}\right) \in \Psi_{\mathrm{gnat}}(2)^k$ such that

$$\left\{\left(\psi_{a_1,b_1}(f(x_1)),\ldots,\psi_{a_k,b_k}(f(x_k))\right) : f \in D_\gamma(\mathscr{H}_0)\right\} = \{0,1\}^k.$$

Without loss, we can assume that $a_j > b_j$ for $j = 1,\ldots,k$.

Now, if $m < \max((k-1)/(32\epsilon), \left((1-\epsilon)/\epsilon\right)\ln(1/\delta))$ and $k \geqslant 2$ (for which $\gamma^2 \geqslant 4d2^{-\sqrt{d/6}}$ suffices), Lemma 4.4 implies that there is a distribution P on $\{1,\ldots,k\}$ and a function $p : \{1,\ldots,k\} \to \{0,1\}$ such that

$$P^m\{l \in \{1,\ldots,k\}^m : \exists p' : \{1,\ldots,k\} \to \{0,1\} \text{ such that } p(l_i) = p'(l_i) \text{ for } i = 1,\ldots,m \text{ and } P\{y \in \{1,\ldots,k\} : p(y) \neq p'(y)\} \geqslant \epsilon\} \geqslant \delta.$$

Choose a function $t : X \to \mathbb{R}$ satisfying

$$t(x_j) = \begin{cases} (a_j - 1)\gamma + \eta, & p(j) = 1, \\ b_j\gamma - \eta + \Delta, & p(j) = 0, \end{cases}$$

for $j = 1,\ldots,k$, where

$$\Delta = \tfrac{1}{2}\min\{h(x_j) - (a_j - 1)\gamma : h \in \mathscr{H}_0 \text{ and } h(x_j) > (a_j - 1)\gamma, j = 1,\ldots,k\}.$$

For each function $h \in \mathscr{H}_0$ define $f_h = D_\gamma(h)$. Let $p_h : \{1,\ldots,k\} \to \{0,1\}$ be defined by

$$p_h(j) = \begin{cases} 1, & f_h(x_j) = a_j, \\ 0, & f_h(x_j) = b_j. \end{cases}$$

Clearly, if $|h(x_j) - t(x_j)| < \eta$ for some $h \in \mathscr{H}_0$ and some $j \in \{1,\ldots,k\}$, then $h(x_j) \in ((a_j - 1)\gamma, a_j\gamma] \cup ((b_j - 1)\gamma, b_j\gamma]$ so $p_h(j) = p(j)$. Also, if $p_h(j) \neq p(j)$ for some $h \in \mathscr{H}$, then $|h(x_j) - t(x_j)| > \eta + \gamma$. It follows that $P\{y \in \{1,\ldots,k\} : p_h(y) \neq p(y)\} \geqslant \epsilon$ implies $Q\{y \in X : |h(y) - t(y)| \geqslant \eta + \gamma\} \geqslant \epsilon$, where Q is the discrete probability distribution on X

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

satisfying $Q(x_j) = P(j)$ for $j = 1, \ldots, k$. So

$$\begin{aligned} Q^m\{\mathbf{y} \in X^m : \exists h \in \mathscr{H} \text{ such that } |h(y_i) - t(y_i)| < \eta \text{ for } i = 1, \ldots, m \text{ and} \\ Q\{\mathbf{y} \in X : |h(y) - t(y)| \geqslant \eta + \gamma\} \geqslant \epsilon\} \geqslant \\ Q^m\{\mathbf{y} \in X^m : \exists h \in \mathscr{H}_0 \text{ such that } |h(y_i) - t(y_i)| < \eta \text{ for } i = 1, \ldots, m \text{ and} \\ Q\{y \in X : |h(y) - t(y)| \geqslant \eta + \gamma\} \geqslant \epsilon\} \geqslant \\ P^m\{l \in \{1, \ldots, k\}^m : \exists p' : \{1, \ldots, k\} \to \{0, 1\} \text{ such that} \\ p(l_i) = p'(l_i) \text{ for } i = 1, \ldots, m \text{ and} \\ P\{y \in \{1, \ldots, k\} : p(y) \neq p'(y)\} \geqslant \epsilon\} \geqslant \delta. \end{aligned}$$ □

We also have the following result which bounds from below the sample length function for $\mathscr{H}$ to approximate $\mathscr{H}$ from interpolated examples.

Theorem 4.6. *Suppose $\mathscr{H}$ is a class of $[0,1]$-valued functions defined on a set X, $0 < \gamma < 1$, $3\gamma/2 \leqslant \eta < 1$, and $\epsilon, \delta \in (0,1)$. If d satisfies $\mathrm{fat}_{\mathscr{H}}(\eta + \gamma) \geqslant d \geqslant 1$ and $\gamma^2 \geqslant 4d2^{-\sqrt{d/6}}$, then any sample length function m_0 for $\mathscr{H}$ to approximate $\mathscr{H}$ from interpolated examples satisfies*

$$m_0(\eta, \gamma, \epsilon, \delta) \geqslant \max\left(\frac{1}{32\epsilon}\left(\frac{d}{3\log^2(4d/\gamma^2)} - 1\right), \frac{1}{\epsilon}\log\frac{1}{\delta}\right).$$

Proof. Fix $0 < \gamma < 1$ and $3\gamma/2 \leqslant \eta < 1$, define $n = \lceil 1/\gamma \rceil$, and suppose $d \leqslant \mathrm{fat}_{\mathscr{H}}(\eta + \gamma)$. Let $\mathscr{F} = D_\gamma(\mathscr{H})$. Then

$$\mathrm{fat}_{\mathscr{F}}\left(\frac{1}{2}\left\lfloor\frac{2(\eta+\gamma)}{\gamma}\right\rfloor\right) \geqslant d,$$

so $\Psi_{\mathrm{fat}}(\lfloor 2\eta/\gamma \rfloor + 1)\text{-dim}(\mathscr{F}) \geqslant d$, hence $\Psi_{\mathrm{gnat}}(\lfloor 2\eta/\gamma \rfloor + 1)\text{-dim}(\mathscr{F}) > k$, where $k = d/(3\log^2(2dn^2))$. Consider a sequence $(x_1, \ldots, x_k) \in X^k$ that is $\Psi_{\mathrm{gnat}}(\lfloor 2\eta/\gamma \rfloor + 1)$-shattered by $\mathscr{F}$. Clearly, there is a subset $\mathscr{H}_0 \subseteq \mathscr{H}$ with $|\mathscr{H}_0| = 2^k$ and a sequence $\left(\psi_{a_1,b_1}, \ldots, \psi_{a_k,b_k}\right) \in \Psi_{\mathrm{gnat}}(\lfloor 2\eta/\gamma \rfloor + 1)^k$ such that

$$\left\{\left(\psi_{a_1,b_1}(f(x_1)), \ldots, \psi_{a_k,b_k}(f(x_k))\right) : f \in D_\gamma(\mathscr{H}_0)\right\} = \{0,1\}^k.$$

Fix a function $t \in \mathscr{H}_0$. Any function $h \in \mathscr{H}$ that has

$$\psi_{a_i,b_i}(D_\gamma(h)(x_i)) = \psi_{a_i,b_i}(D_\gamma(t)(x_i))$$

satisfies $|h(x_i) - t(x_i)| < \gamma < \eta$. Any function h in $\mathscr{H}$ that has

$$\psi_{a_i,b_i}(D_\gamma(h)(x_i)) \neq \psi_{a_i,b_i}(D_\gamma(t)(x_i))$$

satisfies

$$\begin{aligned} |h(x_i) - t(x_i)| &\geqslant \left\lfloor\frac{2\eta}{\gamma}\right\rfloor\gamma \\ &= 2\gamma + \left\lfloor\frac{2(\eta-\gamma)}{\gamma}\right\rfloor\gamma \\ &\geqslant 2\gamma + \eta - \gamma = \eta + \gamma, \end{aligned}$$

since $(\eta - \gamma)/\gamma \geqslant 1/2$ and $\lfloor 2\alpha \rfloor \geqslant \alpha$ for $\alpha \geqslant 1/2$.

https://doi.org/10.1017/S0963548300004247 Published online by Cambridge University Press

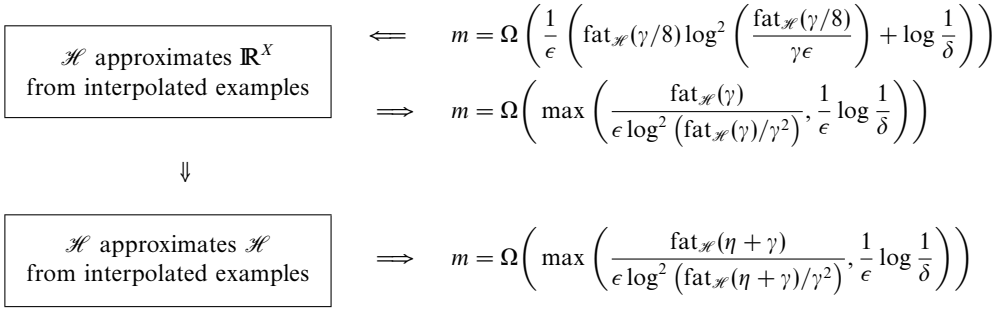

Figure 1 Sample complexity bounds

Using the same argument as in the proof of Theorem 4.5, there is a distribution P on X such that if m is too small then, with P^m-probability at least δ, some $h \in \mathscr{H}$ is within η of t on a random sample, but $P(|h-t| \geqslant \eta+\gamma) \geqslant \epsilon$. □

5. Discussion

Figure 1 shows the sample complexity bounds for approximation from interpolated examples. (The diagram omits the requirement in the lower bounds that γ is not too small as a function of $\mathrm{fat}_{\mathscr{H}}(\gamma)$ and $\mathrm{fat}_{\mathscr{H}}(\eta+\gamma)$.) These bounds imply Theorem 2.1.

Notice that the upper and lower bounds on the sample length for $\mathscr{H}$ to approximate $\mathbb{R}^X$ from interpolated examples are within log factors of each other. These sample complexity bounds are also relevant to the problem of learning real-valued functions in the presence of malicious noise. Suppose a learner sees a sequence of training examples that correspond to the values of a target function corrupted with arbitrary bounded additive noise. That is, each example is of the form $(x_i, t(x_i)+n_i)$, where $t \in \mathscr{H}$ and $|n_i| < \eta$. Clearly, any function $h \in \mathscr{H}$ that is η-close to the training sample will satisfy

$$\Pr\left(|h-t| \geqslant 2\eta+\gamma\right) < \epsilon,$$

provided that $\mathscr{H}$ approximates from interpolated examples and the training sample is sufficiently large. In addition, if there is an algorithm that can learn in the presence of malicious noise (in this sense), then it can certainly learn in the presence of uniformly distributed random noise (as defined in [3]), which implies $\mathrm{fat}_{\mathscr{H}}$ is finite ([3, Theorem 3]). That is, a function class $\mathscr{H}$ is learnable with malicious noise if and only if $\mathrm{fat}_{\mathscr{H}}$ is finite.

Acknowledgements

This research was supported in part by the Australian Telecommunications and Electronics Research Board. The work of Martin Anthony is supported in part by the European Union through the 'Neurocolt' ESPRIT Working Group. The research reported here was conducted while Martin Anthony was visiting the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University.

References

[1] Alon, N., Ben-David, S., Cesa-Bianchi, N. and Haussler, D. (1997) Scale-sensitive dimensions, uniform convergence, and learnability. *J. Assoc. Comput. Mach.* **44** 615–631.

[2] Anthony, M. and Biggs, N. (1992) *Computational Learning Theory: An Introduction*, Cambridge University Press.

[3] Bartlett, P. L., Long, P. M. and Williamson, R. C. (1994) Fat-shattering and the learnability of real-valued functions. In *Proc. Seventh Annual ACM Conference on Computational Learning Theory*, ACM Press, New York.

[4] Ben-David, S., Cesa-Bianchi, N., Haussler, D. and Long, P. (1995) Characterizations of learnability for classes of $\{0,\ldots,n\}$-valued functions. *J. Comput. System Sci.* **50** 74–86. (An earlier version appeared in *Proc. Fifth Annual ACM Workshop on Computational Learning Theory*, ACM Press, New York.)

[5] Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K. (1989) Learnability and the Vapnik–Chervonenkis dimension. *J. Assoc. Comput. Mach.* **36** 929–965.

[6] Dudley, R. M., Giné, E. and Zinn, J. (1991) Uniform and universal Glivenko–Cantelli classes. *J. Theoret. Probab.* **4** 485–510.

[7] Ehrenfeucht, A., Haussler, D., Kearns, M. and Valiant, L. (1989) A general lower bound on the number of examples needed for learning. *Inform. Comput.* **82** 247–261.

[8] Haussler, D. (1992) Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comput.* **100** 78–150.

[9] Kearns, M. J. and Schapire, R. E. (1994) Efficient distribution-free learning of probabilistic concepts. *J. Comput. System Sci.* **48** 464–497.

[10] Kolmogorov, A. N. and Tihomirov, V. M. (1961) ϵ-entropy and ϵ-capacity of sets in functional spaces. *AMS Translations Ser. 2* **17** 277–364.

[11] Natarajan, B. K. (1993) Occam's razor for functions. In *Proc. Sixth Annual Workshop on Computational Learning Theory*, ACM Press, New York, pp. 370–376.

[12] Pollard, D. (1984) *Convergence of Stochastic Processes*, Springer.

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