Prediction, Learning, Uniform Convergence, and Scale-sensitive Dimensions

Peter L. Bartlett
Department of Systems Engineering
RSISE, Australian National University
Canberra, 0200 Australia
Peter.Bartlett@anu.edu.au

Philip M. Long
ISCS Department
National University of Singapore
Singapore 119260, Republic of Singapore
plong@iscs.nus.edu.sg

Abstract

We present a new general-purpose algorithm for learning classes of $[0,1]$-valued functions in a generalization of the prediction model, and prove a general upper bound on the expected absolute error of this algorithm in terms of a scale-sensitive generalization of the Vapnik dimension proposed by Alon, Ben-David, Cesa-Bianchi and Haussler. We give lower bounds implying that our upper bounds cannot be improved by more than a constant factor in general. We apply this result, together with techniques due to Haussler and to Benedek and Itai, to obtain new upper bounds on packing numbers in terms of this scale-sensitive notion of dimension. Using a different technique, we obtain new bounds on packing numbers in terms of Kearns and Schapire's fat-shattering function. We show how to apply both packing bounds to obtain improved general bounds on the sample complexity of agnostic learning. For each $\epsilon > 0$, we establish weaker sufficient and stronger necessary conditions for a class of $[0,1]$-valued functions to be agnostically learnable to within ϵ, and to be an ϵ-uniform Glivenko-Cantelli class.

1 Introduction

In the prediction model studied in this paper, a $[0,1]$-valued function f chosen from some known class F is hidden from the learner, the learner is given examples of f evaluated at $m-1$ elements of the domain of f that were chosen independently at random according to an arbitrary, unknown distribution, another random point x is chosen, and the learner is required to output a prediction $\hat{y}$ of $f(x)$. The learner is penalized by $|\hat{y} - f(x)|$. This can be viewed as a model of on-line learning, and is the straightforward generalization of the prediction model of Haussler, Littlestone and Warmuth [13] to real-valued functions.

In this paper, we begin by introducing a new general-purpose prediction strategy that uses a binary search to divide the problem of real-valued prediction into a number of binary-valued prediction problems. We give bounds on the expected error of this strategy in terms of fatV, the scale-sensitive generalization of the Vapnik dimension introduced by Alon, Ben-David, Cesa-Bianchi and Haussler [1] (which is similar to a notion introduced by Kearns and Schapire [14]), and show that no algorithm can improve on these bounds in general by more than a constant factor.

A packing number for a class of functions measures, in a certain sense, the largest number of significantly different behaviors functions in the class can have on a set of points of a given size. We apply the above prediction bound, together with ideas due to Haussler [11] and Benedek and Itai [5], to obtain new bounds on packing numbers in terms of fatV.

In agnostic learning [10, 15], a distribution on $X \times [0,1]$ is unknown, and the learner, given examples drawn according to this distribution, tries to find a function h from X to $[0,1]$ so that, with probability at least $1-\delta$, the expectation of $|h(x) - y|$ is at most ϵ larger than the minimum of this expectation over functions in some touchstone class F. We combine our new packing bound with the techniques of another paper of Haussler [10] to prove an upper bound[1] of

$$O\left(\frac{1}{\epsilon^2}\left(\frac{\mathrm{fatV}_F(\epsilon/5)}{\epsilon}\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

on the sample complexity of (ϵ,δ)-agnostic learning F. This improves on the bound of

$$O\left(\frac{1}{\epsilon^2}\left(\frac{\mathrm{fatV}_F(\epsilon/384)}{\epsilon}\log^2\frac{\mathrm{fatV}_F(\epsilon/384)}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

that is a straightforward consequence of the results of [1] (see [4]).

Next, using a different technique, we obtain a new packing bound in terms of Kearns and Schapire's fat-shattering function. This leads to a bound[2] of

$$O\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_F(\epsilon/5)\log^2\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

on the sample complexity of (ϵ,δ)-agnostic learning F. This improves on the dependence on fat_F of the bound

$$O\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_F(\epsilon/192)\log^2\frac{\mathrm{fat}_F(\epsilon/192)}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

that follows from the packing bound of [1] (see [4]).

In previously derived bounds on the sample complexity of agnostic learning in terms of scale-sensitive notions of dimension, the scale at which the dimension was measured was a large constant

[1] The 5 in this bound can be replaced with any constant greater than 4.

[2] This 5 can also be replaced with any constant greater than 4.

factor finer than the relative accuracy to which the learner was learning. In this paper, we investigate the question of at what scale the dimension needs to be finite for a class of functions to be agnostically learnable to within relative accuracy ϵ (also to be an ϵ-uniform Glivenko-Cantelli class). Our results narrow the range between necessary and sufficient "scales" to a factor of 2. Our weaker sufficient conditions are proved using a new general-purpose prediction strategy that directly makes use of a cover of the function class. For ϵ-agnostic learning with respect to a class F, this strategy takes a sequence of labelled examples and a single unlabelled example, and constructs an $(\epsilon - \alpha)$-cover of the restriction of the function class F to the examples. (Here α can be made as close to zero as desired.) Then the strategy divides the sample into two parts, and selects the function in the cover that has minimal error on the first half of the sample. We show that if this function is used to predict the labels of the examples in the last half of the sample, the expected error is within (approximately) ϵ of the minimal error. A standard technique converts this to an ϵ-agnostic learning algorithm.

2 Definitions

2.1 Definitions for the prediction model

For a set X, a *prediction strategy* is a mapping from $(\cup_m (X \times [0,1])^m) \times X$ to $[0,1]$. Let $\mathcal{P}_X$ be the set of all prediction strategies, and let $\mathcal{D}_X$ be the set of all probability distributions on X. For each set F of functions from X to $[0,1]$, and each positive integer m, define[3] $\mathcal{L}(F,m)$ as

$$\mathcal{L}(F,m) = \inf_{A \in \mathcal{P}_X} \sup \int_{X^m} |A(((x_1, f(x_1)), ..., (x_{m-1}, f(x_{m-1}))), x_m) - f(x_m)| \, dD^m(x_1, \ldots, x_m),$$

where the supremum is over all D in $\mathcal{D}_X$ and f in F. That is, $\mathcal{L}(F,m)$ is the worst-case expected error of the best prediction strategy. This is a generalization of the $\{0,1\}$ prediction model of [13] to $[0,1]$-valued functions.

2.2 Definitions for the agnostic learning model

Define a *learner* for a set X to be a mapping from $\cup_{n \in \mathbf{N}} (X \times [0,1])^n$ to $[0,1]^X$, i.e. to take a sequence of labelled examples, and output a hypothesis. If h is a $[0,1]$-valued function defined on X, and P is a probability distribution over $X \times [0,1]$, define the *error of h with respect to P* as

$$\mathbf{er}_P(h) = \int_{X \times [0,1]} |h(x) - y| \, dP(x,y).$$

Suppose F is a class of $[0,1]$-valued functions defined on X, $0 < \epsilon, \delta < 1$ and $m \in \mathbf{N}$. We say a learner A *(ϵ, δ)-learns in the agnostic sense with respect to F from m examples* if, for all distributions P on $X \times [0,1]$,

$$P^m \left\{ w \in (X \times [0,1])^m : \mathbf{er}_P(A(w)) \geq \inf_{f \in F} \mathbf{er}_P(f) + \epsilon \right\} < \delta.$$

For $\epsilon > 0$, the function class F is *ϵ-agnostically learnable* if there is a function $m_0 : (0,1) \to \mathbf{N}$ such that, for all $0 < \delta < 1$, there is a learner A which (ϵ, δ)-learns in the agnostic sense with respect to F from $m_0(\delta)$ examples.

[3] Throughout, we ignore issues of measurability. The reader may assume that X is countable, but significantly weaker assumptions, like those of Pollard's [17] Appendix C, suffice.

2.3 Definition of ϵ-uniform GC-classes

For $\epsilon, \delta > 0$, a set X and a set F of functions from X to $[0,1]$, if $\mathcal{D}_X$ is the set of all probability distributions over X, define

$$m_{\mathrm{GC},F}(\epsilon,\delta) = \min\left\{ n : \forall m \geq n, \forall D \in \mathcal{D}_X, \right.$$
$$\left. D^m\left\{ (x_1,...,x_m) : \exists f \in F, \left| \frac{1}{m}\left(\sum_{i=1}^m f(x_i)\right) - \int_X f(u)\, dD(u) \right| > \epsilon \right\} \leq \delta \right\}.$$

If the minimum doesn't exist, then $m_{\mathrm{GC},F}(\epsilon,\delta) = \infty$. If, for all $\delta > 0$, $m_{\mathrm{GC},F}(\epsilon,\delta)$ is finite, then F is said to be an ϵ-uniform GC- (Glivenko-Cantelli) class.

2.4 Packing and Covering

For each $n \in \mathbf{N}$, define $\ell_1 : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ by

$$\ell_1(v,w) = \frac{1}{n}\sum_{i=1}^n |v_i - w_i|.$$

For $S \subseteq \mathbf{R}^n$, define $\mathcal{N}(\epsilon, S)$ to be the size of the smallest set $T \subseteq \mathbf{R}^n$ such that for all $v \in S$, there is a $w \in T$ such that $\ell_1(v,w) \leq \epsilon$. Call such a T an *ϵ-cover* of S. Define $\mathcal{M}(\epsilon, S)$ to be the size of the largest subset T of S such that for any two elements v, w of T, $\ell_1(v,w) > \epsilon$. We will make use of the following well-known bounds that hold for all n, $S \subseteq \mathbf{R}^n$.

$$\mathcal{M}(2\epsilon, S) \leq \mathcal{N}(\epsilon, S) \leq \mathcal{M}(\epsilon, S). \tag{1}$$

2.5 Quantizing

For $\alpha > 0$ and $u \in \mathbf{R}$, let $Q_\alpha(u)$ denote the quantized version of u, with quantization width α. That is, define $Q_\alpha(u) = \alpha\lfloor u/\alpha \rfloor$. Let $Q_\alpha([0,1]) = \{Q_\alpha(u) : u \in [0,1]\}$. For $v \in \mathbf{R}^n$, define $Q_\alpha(v) = (Q_\alpha(v_1),...,Q_\alpha(v_n))$, and similarly, for a function f from some set X to $\mathbf{R}$, define $Q_\alpha(f) : X \to \mathbf{R}$ by $(Q_\alpha(f))(x) = Q_\alpha(f(x))$. Finally, for a set F of such functions, define $Q_\alpha(F) = \{Q_\alpha(f) : f \in F\}$.

2.6 Definitions relating to fat

For $m \in \mathbf{N}$, $S \subseteq [0,1]^m$, and $\gamma > 0$, we say S γ-fatly shatters a sequence $(i_1, r_1), ..., (i_d, r_d)$ of elements of $\{1,\ldots,m\} \times [0,1]$ if for all $(b_1,...,b_d) \in \{0,1\}^d$ there is a $v \in S$ such that for all $j \in \{1,...,d\}$,

$$v_{i_j} \geq r_j + \gamma \quad \text{if } b_j = 1$$
$$v_{i_j} \leq r_j - \gamma \quad \text{if } b_j = 0.$$

We then define $\mathrm{fat}_S(\gamma)$ to be the length of the longest sequence γ-fatly shattered by S. For a set F of functions from X to $[0,1]$, and a finite sequence $\xi = (x_1,...,x_n)$ of elements of X, define the restriction of F to ξ to be

$$F_{|\xi} = \{(f(x_1),...,f(x_n)) : f \in F\}.$$

We define $\mathrm{fat}_F(\gamma)$ to be the maximum, over all finite sequences ξ of elements of X, of $\mathrm{fat}_{F_{|\xi}}(\gamma)$. (This was called the fat-shattering function in [4], and was defined by Kearns and Schapire [14].)

2.7 Definitions relating to fatV

For each $r \in [0,1]$ and $\epsilon > 0$, define $\psi_{r,\epsilon} : [0,1] \to \{0, \star, 1\}$ by

$$\psi_{r,\epsilon}(y) = \begin{cases} 1 & \text{if } y \geq r + \epsilon \\ \star & \text{if } |y - r| < \epsilon \\ 0 & \text{if } y \leq r - \epsilon. \end{cases}$$

For a function f from X to $[0,1]$, define $\psi_{r,\epsilon}(f) : X \to \{0, \star, 1\}$ by

$$(\psi_{r,\epsilon}(f))(x) = \psi_{r,\epsilon}(f(x)),$$

and for a set F of such functions, define

$$\psi_{r,\epsilon}(F) = \{\psi_{r,\epsilon}(f) : f \in F\}.$$

We say $x_1, \ldots, x_d$ in X are γ-fatly Vapnik-shattered by F if there is an $r \in [0,1]$ such that

$$\{0,1\}^d \subseteq \{(\psi_{r,\gamma}(f(x_1)), ..., \psi_{r,\gamma}(f(x_d))) : f \in F\}.$$

Define $\text{fatV}_F(\gamma)$ to be the length of the longest sequence γ-fatly Vapnik-shattered by F. (This dimension was first studied in [1].)

Notice that $\text{fat}_F(\gamma)$ and $\text{fatV}_F(\gamma)$ are both non-increasing functions of γ.

3 Prediction of $[0,1]$-valued functions and fatV

This section describes our general-purpose prediction strategy and shows that it is nearly optimal. The first theorem of the paper gives the bound for the worst-case expected error incurred by this strategy.

Theorem 1 *Choose a set F of functions from X to $[0,1]$, $\gamma > 0$, and a positive integer m. Then*

$$\mathcal{L}(F, m) \leq \frac{2\text{fatV}_F(\gamma)}{m} + \gamma.$$

Fix a set X. Theorem 1 is proved by considering an algorithm that generates its prediction using binary search (details are given below). It uses subalgorithms to predict whether $f(x_m)$ is above or below $1/2$, above or below $1/4$ and $3/4$, and so on. To analyze these subalgorithms, we would like to show that, for example, the set of possible "above-below $1/2$" behaviors is not very rich. But a bound on $\text{fatV}_F(\gamma)$ only provides information about the richness of behaviors at least γ-above and γ-below $1/2$. On the other hand, if γ is small, the binary search algorithm can tolerate incorrect guesses if the truth is within γ of $1/2$, so, in a sense, we don't care about the correctness of predictions in such cases.

Therefore, we will consider a model of learning which might be called concept-with-don't-care's learning. Here, what is learned is a function from X to $\{0, \star, 1\}$. The $\star$ is interpreted as a "don't care" value, in that an incorrect prediction of the value of the function does not count against the learning algorithm if that value is $\star$. Also, when we generalize the VC-dimension, a notion of the richness of a class of $\{0,1\}$-valued functions, loosely speaking, the $\star$'s will not contribute toward a certain class being considered rich.

Formally, define a concept-with-don't-care's (CWDC) strategy to be a mapping from

$$\left(\bigcup_m (X \times \{0, \star, 1\})^m\right) \times X$$

to $\{0,1\}$. Let $\mathcal{B}_X$ be the set of all CWDC strategies. For each set G of functions from X to $\{0,\star,1\}$, define $M(G,m)$ as the worst case mistake probability of the best CWDC prediction strategy in G,

$$\begin{aligned} M(G,m) \quad = \quad & \inf_{B \in \mathcal{B}_X} \sup D^m \{(x_1,...,x_m) : g(x_m) \neq \star \\ & \qquad \text{and } B(((x_1,g(x_1)),...,(x_{m-1},g(x_{m-1}))),x_m) \neq g(x_m)\}, \end{aligned}$$

where the supremum is over all D in $\mathcal{D}_X$ and g in G. When $g(x_m) \neq \star$ and

$$B(((x_1,g(x_1)),...,(x_{m-1},g(x_{m-1}))),x_m) \neq g(x_m),$$

we say that B *makes a mistake.*

Extend the definition of VC-dimension to say that the VC-dimension $\mathrm{VCdim}(G)$ of a set G of functions from X to $\{0,\star,1\}$ is

$$\max\{d : \exists x_1,...,x_d \in X,\ \{0,1\}^d \subseteq \{(g(x_1),...,g(x_d)) : g \in G\}\}.$$

First, we will make use of the following well-known lemma, whose application is usually referred to as the "permutation trick" (see [13]). It formalizes the idea that, when m points are chosen independently at random, then any permutation of a certain sequence of points is equally likely to have been chosen.

Lemma 2 *Choose $m \in \mathbf{N}$, a distribution D on X, and a random variable ϕ defined on X^m. Let U be the uniform distribution on the permutations of $\{1,...,m\}$. Then*

$$\int \phi(x) D^m(x) \leq \sup_{(x_1,...,x_m) \in X^m} \int \phi(x_{\sigma(1)},...,x_{\sigma(m)}) U(\sigma).$$

We will make use of the following result of Haussler, Littlestone and Warmuth.

Lemma 3 ([13]) *For any $F \subseteq \{0,1\}^X$ (note the absence of "$\star$"), there is a CWDC strategy $A^{\mathrm{one-inc}}$ such that for any points $x_1,...,x_m$, if U is the uniform distribution over permutations of $\{1,...,m\}$, for any $f \in F$, the probability under U of a permutation σ for which*

$$A^{\mathrm{one-inc}}(((x_{\sigma(1)}, f(x_{\sigma(1)})), \ldots, (x_{\sigma(m-1)}, f(x_{\sigma(m-1)}))), x_{\sigma(m)}) \neq f(x_{\sigma(m)})$$

is no more than $\mathrm{VCdim}(F)/m$.

In the next lemma, we apply a generalization of this result to give a general upper bound for the CWDC model.

Lemma 4 *Choose $G \subseteq \{0,\star,1\}^X$ and $m \in \mathbf{N}$. Then*

$$M(G,m) \leq \mathrm{VCdim}(G)/m.$$

Proof: Define a strategy B as follows. Suppose B is given

$$z = (((x_1, y_1), ..., (x_{m-1}, y_{m-1})), x_m)$$

as input. Let $I_z = \{i \leq m-1 : y_i \neq \star\}$. Let $G_z = \{g \in G : \forall i \in (I_z \cup \{m\}), g(x_i) \neq \star\}$. Note that the restrictions of the functions of G_z to $\{x_i : i \in I_z \cup \{m\}\}$ form a set of $\{0,1\}$-valued functions of VC-dimension at most $\mathrm{VCdim}(G)$. If G_z is empty, B predicts arbitrarily and doesn't make a mistake, since this implies that it is certain that $g(x_m) = \star$, where g is the function being learned. If G_z is non-empty, and $i_1, ..., i_k$ are the elements of I_z in increasing order, then B applies the strategy from Lemma 3 (the one-inclusion graph algorithm) for learning G_z, using $(x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k}), x_m$ as an input.

Applying Lemma 2, we have that for any $g \in G$, the probability, with respect to m independent random draws from some fixed distribution, that B makes a mistake on the mth prediction, is at most the maximum, over $x_1, ..., x_m$, of the same probability with respect to a uniformly randomly chosen permutation of $x_1, ..., x_m$.

Fix an arbitrary target function g and sequence $x_1, ..., x_m$ of elements of X. We wish to bound the probability, over a uniformly randomly chosen permutation σ of $x_1, ..., x_m$, that B makes a mistake on the last element, given examples for the first $m-1$. Let $J = \{i \leq m : g(x_i) \neq \star\}$ (note that the mth element is included in the definition of J, but wasn't in I_z above). Let $k = |J|$.

Let E be the event that the permutation σ moves one of $\{i : g(x_i) \neq \star\}$ to be the last, i.e., has $\sigma^{-1}(m) \in J$. Conditioned on E, any order of the elements of J is equally likely, and furthermore, clearly for any pair of inputs z_1, z_2 generated from one of these permutations in the obvious way, G_{z_1} and G_{z_2} are the same. Therefore, by the definition of B and Lemma 3, the probability that B makes a mistake given that the permutation moves one of $\{i : g(x_i) \neq \star\}$ to be the last is at most $\mathrm{VCdim}(G)/k$. Next, clearly the probability that B makes a mistake given that the permutation doesn't send an element of J to be last is 0, since this means a $\star$ is last. Therefore if "Pr" is with respect to a random permutation, and "mistake" is the event that B makes a mistake on the mth element of X when given examples of the first $m-1$, then

$$\begin{aligned} \Pr(\text{mistake}) &= \Pr(\text{mistake}|E)\Pr(E) \\ &\leq (\mathrm{VCdim}(G)/k)(k/m) \\ &= \mathrm{VCdim}(G)/m. \end{aligned}$$

This completes the proof. □

Proof (of Theorem 1): Let $d = \mathrm{fatV}_F(\gamma)$. Consider the strategy A defined as follows. For each $r \in [0,1]$, define B_r to be the strategy for learning $\psi_{r,\gamma}(F)$ described in Lemma 4. Given input

$$z = (((x_1, y_1), ..., (x_{m-1}, y_{m-1})), x_m),$$

define z_r to be

$$(((x_1, \psi_{r,\gamma}(y_1)), ..., (x_{m-1}, \psi_{r,\gamma}(y_{m-1}))), x_m).$$

Strategy A performs binary search as described in the following recurrence. First, $l_1 = 0$ and $u_1 = 1$. For each $i \in \mathbf{N}$,

- if $B_{(l_i+u_i)/2}(z_{(l_i+u_i)/2}) = 1$, then $b_i = 1$, $l_{i+1} = (l_i + u_i)/2$, and $u_{i+1} = u_i$, and
- if $B_{(l_i+u_i)/2}(z_{(l_i+u_i)/2}) = 0$, then $b_i = 0$, $l_{i+1} = l_i$, and $u_{i+1} = (l_i + u_i)/2$.

The output of strategy A is then $\sum_{i=1}^{\infty} b_i 2^{-i}$, i.e. $0.b_1 b_2 ...$ in binary.

First, by a trivial induction, at any given time during the binary search, the final prediction of A is contained in $[l_i, u_i]$. By an equally trivial induction, if for $j = 1, \ldots, i-1$ either

$$B_{(l_j+u_j)/2}(z_{(l_j+u_j)/2}) = \psi_{(l_j+u_j)/2,\gamma}(f(x_m))$$

or $\psi_{(l_j+u_j)/2,\gamma}(f(x_m)) = \star$, then $f(x_m) \in [l_i - \gamma, u_i + \gamma]$.

For each positive integer i, let E_i be the event that i is the smallest number for which

$$B_{(l_i+u_i)/2}(z_{(l_i+u_i)/2}) \neq \psi_{(l_i+u_i)/2,\gamma}(f(x_m)) \text{ and } \psi_{(l_i+u_i)/2,\gamma}(f(x_m)) \neq \star.$$

Let E_∞ be the event that there is no such number. Then

$$\mathbf{E}(|A(z) - f(x_m)|) = \mathbf{E}(|A(z) - f(x_m)| \mid E_\infty) \Pr(E_\infty) + \sum_{j \in \mathbf{N}} \mathbf{E}(|A(z) - f(x_m)| \mid E_j) \Pr(E_j).$$

(Here we use the convention that, for each j, if $\Pr(E_j) = 0$, then $\mathbf{E}(|A(z) - f(x_m)| \mid E_j) \Pr(E_j)$ is taken to be 0.) Therefore since for any $\hat{y} \in [l, u]$ and $y \in [l - \gamma, u + \gamma]$, it is the case that $|\hat{y} - y| \leq |l - u| + \gamma$, we have

$$\mathbf{E}(|A(z) - f(x_m)|) \leq \gamma \Pr(E_\infty) + \sum_{j \in \mathbf{N}} (1/2^{j-1} + \gamma) \Pr(E_j).$$

Applying Lemma 4, and the fact that $\mathrm{VCdim}(\psi_{r,\gamma}(F)) \leq \mathrm{fatV}_F(\gamma)$ for all $r \in [0,1]$, we get

$$\begin{aligned} \mathbf{E}(|A(z) - f(x_m)|) &\leq \gamma + \sum_{j \in \mathbf{N}} (1/2^{j-1})(d/m) \\ &\leq \gamma + 2d/m. \end{aligned}$$

This completes the proof. □

The following theorem shows that Theorem 1 cannot be improved in general by more than a constant factor, and that the constant on the γ term is best possible. The proof uses techniques due to Ehrenfeucht, Haussler, Kearns and Valiant [8], Haussler, Littlestone and Warmuth [13], and Simon [18].

Theorem 5 *There exists c such that for all sufficiently small $\gamma \geq 0$, and all sufficiently large $d, m \in \mathbf{N}$, there is an X, and $F \subseteq [0,1]^X$ such that $\mathrm{fatV}_F(\gamma) = d$ and*

$$\mathcal{L}(F, m) \geq \max\left\{ c\frac{\mathrm{fatV}_F(\gamma)}{m}, \gamma \right\}.$$

Proof: Consider the class F of all functions f from $\mathbf{N}$ to $[0,1]$ such that $|f^{-1}([2\gamma, 1])| \leq d$. Clearly, $\mathrm{fatV}_F(\gamma) = d$.

We begin by proving the first term. Consider the distribution D on $\{1, ..., d\}$ where $D(1) = 1 - (d-1)/m$ and $D(2) = ... = D(d) = 1/m$. Clearly, F contains all functions from $\{1, ..., d\}$ to $\{0, 1\}$. For the remainder of the proof of the first term, let us assume that $\{1, ..., d\}$ is the entire domain, and that F consists exactly of those functions from $\{1, ..., d\}$ to $\{0, 1\}$. For each $b \in \{0,1\}^d$ define $f_b \in F$ by $f_b(i) = b_i$.

Fix $u_1, ..., u_m \in \{1, ..., d\}$. Notice that

$$\hat{y}_m = A((u_1, f_b(u_1)), ...(u_{m-1}, f_b(u_{m-1})), u_m)$$

is a function only of those components b_i of b for which $i \in \{u_1, ..., u_{m-1}\}$. Suppose we choose b according to the uniform distribution over $\{0,1\}^d$. Choose $i \notin \{u_1, ..., u_{m-1}\}$, and $(c_1, ..., c_{m-1}) \in \{0,1\}^{m-1}$. Then, by the independence of the choice of b_i from that of the other components, the expectation of $|\hat{y}_m - f_b(i)|$, given that $b_{u_1} = c_1, ..., b_{u_m} = c_m$, is

$$(1 - \hat{y}_m)/2 + \hat{y}_m/2,$$

which, for any value of $\hat{y}_m$, is $1/2$. Since this is true independent of $c_1, ..., c_{m-1}$, for any $i \notin \{u_1, ..., u_{m-1}\}$, the expected value of the error of A on i is at least $1/2$.

Now, suppose $u_1, ..., u_m$ are chosen independently at random according to D as well. Then the expectation of $|\hat{y}_m - f_b(u_m)|$ is at least $1/2$ times the probability that $u_m \notin \{u_1, ..., u_{m-1}\}$. This probability has been shown to be $\Omega(d/m)$ [13], and therefore, the expectation of $|\hat{y}_m - f_b(u_m)|$ over the random choice of the u_i's and b is $\Omega(d/m)$, which implies there exists b such that for that fixed b, the expectation of $|\hat{y}_m - f_b(u_m)|$ only over the random choice of the u_i's is $\Omega(d/m)$, which completes the proof of the first term.

The proof of the second term is similar. Choose $m \in \mathbf{N}$. Choose a small $\kappa > 0$, and a large $d \in \mathbf{N}$. Suppose the elements of the domain are chosen according to the uniform distribution on $\{1, ..., d\}$, and suppose the function to be learned is chosen uniformly from the set of functions from $\{1, ..., d\}$ to $\{0, 2(\gamma - \kappa)\}$. By arguing as above, we can see that the expectation of

$$|f(x_m) - A((x_1, f(x_1)), ...(x_{m-1}, f(x_{m-1})), x_m)|,$$

given that $x_m \notin \{x_1, ..., x_{m-1}\}$, is at least $(1/2)(2(\gamma - \kappa)) = \gamma - \kappa$. Furthermore, the probability that $x_m \notin \{x_1, ..., x_{m-1}\}$ is at least $1 - (m-1)/d$. Therefore, the expected error is at least $(\gamma - \kappa)(1 - (m-1)/d)$, and since κ can be made arbitrarily small, and d can be made arbitrarily large, this completes the proof. □

The following corollary shows that finiteness of fatV_F at a scale just below the desired prediction error is sufficient, and that no larger scale will suffice in general.

Corollary 6 *Suppose $\epsilon > 0$.*

For a set F of functions from X to $[0,1]$, if there is an $\alpha > 0$ with $\mathrm{fatV}_F(\epsilon - \alpha) < \infty$, then for sufficiently large m, $\mathcal{L}(F, m) < \epsilon$.

Moreover, there is a set F such that $\mathrm{fatV}_F(\epsilon) = \infty$ and, for all $\alpha > 0$, $\mathrm{fatV}_F(\epsilon + \alpha) = 0$, but $\mathcal{L}(F, m) \geq \epsilon$ for all m.

The proof of the sufficient condition follows on substituting $\gamma = \epsilon - \alpha$ in Theorem 1. The converse result is exhibited by the class F of all functions from $\mathbf{N}$ to $\{0, 2\epsilon\}$, using similar techniques to the proof of Theorem 5.

In later sections, we investigate the scale at which the dimensions fat and fatV need to be finite for agnostic learnability. The following result shows that precise bounds on this scale are important, since a constant factor gap in the scale can lead to an arbitrarily large gap in the sample complexity bounds.

Proposition 7 *For any non-increasing function ϕ from $(0, 1/2]$ to $\mathbf{N} \cup \{0, \infty\}$, there is a function class $F_\phi : \mathbf{N} \to [0,1]$ that satisfies $\mathrm{fatV}_F(\gamma) = \phi(\gamma)$ for all γ in $(0, 1/2]$.*

Proof: Let $\{A_{d,n} : d \in \mathbf{N} \cup \{\infty\}, n \in \mathbf{N}\}$ be a partition of $\mathbf{N}$, with $|A_{d,n}| = d$ for $d, n \in \mathbf{N}$, and $A_{\infty,n}$ countably infinite for all $n \in \mathbf{N}$.

Fix $d \in \mathbf{N} \cup \{\infty\}$, and consider the set $S_d = \phi^{-1}(d) \subset [0, 1/2]$. If S_d is empty, let $F_{d,n} = \emptyset$ for $n \in \mathbf{N}$. Otherwise, since ϕ is non-increasing, S_d is an interval. Suppose $r = \sup S_d$. There are two cases:

Case 1: r is in S_d.
Let $F_{d,1}$ be the set of all functions f satisfying $f(x) \in \{1/2 - r, 1/2 + r\}$ if $x \in A_{d,1}$, and $f(x) = 0$ otherwise. Let $F_{d,n} = \emptyset$ for $n > 1$.

Case 2: r is not in S_d.
For $n \in \mathbf{N}$, let $F_{d,n}$ be the set of all functions f satisfying $f(x) \in \{1/2 - r(1-1/n), 1/2 + r(1-1/n)\}$ if $x \in A_{d,n}$ and $f(x) = 0$ otherwise.

Let

$$F = \bigcup \{F_{d,n} : n \in \mathbf{N}, d \in \mathbf{N} \cup \{\infty\}\}.$$

For any d in $\mathbf{N} \cup \{\infty\}$, the sets $F_{d,n}$ ensure that for all $\gamma \in \phi^{-1}(d)$, $\text{fatV}_F(\gamma) \geq d$. Clearly, any set of points in $\mathbf{N}$ that has nonempty intersection with two distinct $A_{d,n}$'s cannot be γ-shattered for any $\gamma > 0$, which implies that the reverse inequality is also true. The case $d = 0$ is trivial, and hence $\text{fatV}_F(\gamma) = \phi(\gamma)$ for all $\gamma \in \phi^{-1}(\mathbf{N} \cup \{0, \infty\})$. □

4 Packing number bounds

In this section, we prove two new bounds on $\mathcal{M}(\epsilon, S)$. One uses fatV, and is proved using Theorem 1, together with techniques from [11, 5]. The second bound uses fat, and is proved through a refinement of a proof in [1].

For a set X, and $F \subseteq [0,1]^X$, define

$$m_{\mathcal{L}}(\epsilon, F) = \min\{m \in \mathbf{N} : \mathcal{L}(F, m) \leq \epsilon\}.$$

The following bound on $m_{\mathcal{L}}(\epsilon, F)$ follows immediately from Theorem 1.

Lemma 8 *Choose X, $F \subseteq [0,1]^X$, and $\alpha, \epsilon > 0$. Assume $\text{fatV}_F(\epsilon - \alpha) \geq 1$. Then*

$$m_{\mathcal{L}}(\epsilon, F) \leq 2\text{fatV}_F(\epsilon - \alpha)/\alpha.$$

For $m \in \mathbf{N}, x = (x_1, ..., x_m) \in X^m$, and $f : X \to [0,1]$, define

$$\text{sam}(x, f) = ((x_1, f(x_1)), ..., (x_m, f(x_m))).$$

We will also make use of the following, which is implicit in the work of Haussler, et al [13].

Lemma 9 *Choose X, $F \subseteq [0,1]^X$. There is a learner A such that for all $f \in F$, for any distribution D on X, for all $m \in \mathbf{N}$,*

$$\int \left(\int |(A(\text{sam}(x, f)))(u) - f(u)| \, dD(u) \right) \, dD^{m-1}(x)$$

is no more than $\mathcal{L}(F, m)$.

We apply these in the following. In addition to Theorem 1, the proof uses ideas due to Haussler [11] and Benedek and Itai [5].

Lemma 10 *Choose $0 < \epsilon < 1$, $b \in \mathbf{N}$ and $0 < \alpha < \epsilon/4$. Let $B = Q_{1/b}([0,1])$. Choose $m \in \mathbf{N}$, and let $S \subseteq B^m$. Set $d = \mathrm{fatV}_S(\epsilon/2 - \alpha)$. Then if $d \geq 1$,*

$$\mathcal{M}(\epsilon, S) \leq \frac{\epsilon}{2\alpha}(b+1)^{4d/\alpha}.$$

Proof: For each $v \in S$, define $f_v : \{1, ..., m\} \to [0,1]$ by $f_v(i) = v_i$, and define $F = \{f_v : v \in S\}$. Let D be the uniform distribution on $\{1, ..., m\}$. Then for all $v, w \in S$,

$$\ell_1(v, w) = \int |f_v(u) - f_w(u)| \, dD(u). \tag{2}$$

Define $\ell_1(f, g) = \int |f(u) - g(u)| \, dD(u)$. Let $m_0 = m_{\mathcal{L}}(\epsilon/2 - \alpha, F)$. Then, by Lemma 9, there is a learner A such that for all $f \in F$,

$$\int \ell_1(A(\mathrm{sam}(x, f)), f) \, dD^{m_0}(x) \leq \epsilon/2 - \alpha. \tag{3}$$

Choose an ϵ-separated subset T of $\mathcal{M}(\epsilon, S)$ elements of S. Then by (3), we have

$$\sum_{v \in T} \int \ell_1(A(\mathrm{sam}(x, f_v)), f_v) \, dD^{m_0}(x) \leq (\epsilon/2 - \alpha)|T|,$$

and hence

$$\int \sum_{v \in T} \ell_1(A(\mathrm{sam}(x, f_v)), f_v) \, dD^{m_0}(x) \leq (\epsilon/2 - \alpha)|T|. \tag{4}$$

Fix $x \in \{1, \ldots, m\}^{m_0}$. For any set $T' \subseteq T$ such that for all f_1 and f_2 in T' it is the case that $\mathrm{sam}(x, f_1) = \mathrm{sam}(x, f_2)$, since T' is ϵ-separated, the triangle inequality implies that

$$\ell_1(A(\mathrm{sam}(x, f)), f) < \epsilon/2$$

for no more than one f in T'. Therefore, if we let $f_v(x) = (f_v(x_1), \ldots, f_v(x_{m_0}))$, we have

$$\begin{aligned}
&\textstyle\sum_{v \in T} \ell_1(A(\mathrm{sam}(x, f_v)), f_v) \\
&\quad = \textstyle\sum_{l \in B^{m_0}} \sum_{v \in T, f_v(x) = l} \ell_1(A(\mathrm{sam}(x, f_v)), f_v) \\
&\quad \geq \textstyle\sum_{l \in B^{m_0}} (\epsilon/2)(|\{v \in T : l = f_v(x)\}| - 1) \\
&\quad \geq (\epsilon/2)(|T| - (b+1)^{m_0}).
\end{aligned}$$

This inequality, together with (4), implies

$$(\epsilon/2)(|T| - (b+1)^{m_0}) \leq (\epsilon/2 - \alpha)|T|.$$

Solving for $|T|$, and recalling that $|T| = \mathcal{M}(\epsilon, S)$ and $m_0 = m_{\mathcal{L}}(\epsilon/2 - \alpha, F)$, gives

$$\mathcal{M}(\epsilon, S) \leq \frac{\epsilon}{2\alpha}(b+1)^{m_{\mathcal{L}}(\epsilon/2 - \alpha, F)}.$$

Applying Lemma 8 completes the proof. □

Next, we give a new bound on $\mathcal{M}(\epsilon, S)$ in terms of fat_S. Its proof is based on that of a corresponding lemma in [1] which dealt with the ℓ_∞ norm.

Lemma 11 *Choose $\epsilon > 0$. Choose $b \in \mathbf{N}$, $b > 4/\epsilon$. Let $B = Q_{1/b}([0,1])$. Choose $m \in \mathbf{N}$, and let $S \subseteq B^m$. Set $d = \mathrm{fat}_S(\epsilon/2 - 2/b)$, and*

$$y = \sum_{i=0}^{d} \binom{m}{i} (1+b)^i.$$

Then if $m \geq d$,

$$\mathcal{M}(\epsilon, S) \leq 2b^{3(\lfloor \log_2 y \rfloor + 1)} \leq 2b^{6d \log_2(2bem/d)}.$$

Proof: Fix $m \in \mathbf{N}$. For each $h \in \mathbf{N}$ let $t(h)$ be the minimum, over all $S \subseteq B^m$ with $|S| \geq h$ that are pairwise ϵ-separated in the ℓ_1 norm, of the number of finite sets $\{(i_1, r_1), ..., (i_k, r_k)\}$ of elements of $\{1, ..., m\} \times B$ which are $(\epsilon/2 - 2/b)$-fatly shattered by S. (Here we say a set is γ-fatly shattered if any corresponding sequence is γ-fatly shattered.) Obviously, $t(2) \geq 1$.

Choose an even h, and let S be a pairwise ϵ-separated subset of B^m. Split S arbitrarily into $h/2$ pairs. For each pair v and w, if

$$l = |\{i : |v_i - w_i| \geq \epsilon - 1/b\}|,$$

then $\ell_1(v, w) = \frac{1}{m}\sum_{i=1}^m |v_i - w_i| < \frac{1}{m}(l + (\epsilon - 1/b)m) = l/m + \epsilon - 1/b$. But $\ell_1(v, w) \geq \epsilon$, thus $l/m + \epsilon - 1/b > \epsilon$, which implies $l \geq m/b$. Thus each pair (v, w) has at least m/b indices i such that $|v_i - w_i| \geq \epsilon - 1/b$. Applying the pigeonhole principle, there is some index i_0 such that $h/(2b)$ pairs v and w have $|v_{i_0} - w_{i_0}| \geq \epsilon - 1/b$. Again, by the pigeonhole principle, there are at least $h/(b^2(b-1)) \geq h/b^3$ such pairs for which the pair $\{v_{i_0}, w_{i_0}\}$ is the same.

This implies that there are two subsets S_1 and S_2 of S having at least h/b^3 elements each, and $y_1, y_2 \in B$ with $|y_1 - y_2| \geq \epsilon - 1/b$, such that, for each $v \in S_1$, $v_{i_0} = y_1$, and each $v \in S_2$, $v_{i_0} = y_2$. Obviously, any two points in S_1, respectively S_2 are ϵ-separated, thus S_1 $(\epsilon/2 - 2/b)$-fatly shatters at least $t(\lceil h/b^3 \rceil)$ sets, as does S_2. If the same set $\{(i_1, r_1), ..., (i_k, r_k)\}$ is shattered by both, then so is

$$\left\{ \left(i_0, Q_{1/b}\left(\frac{v_{i_0} + w_{i_0}}{2} \right) \right), (i_1, r_1), ..., (i_k, r_k) \right\}.$$

Thus, $t(h) \geq 2t(\lceil h/b^3 \rceil)$. Since $t(2) \geq 1$, by induction, for all k, $t(2b^{3k}) \geq 2^k$, and therefore

$$t\left(2b^{3(\lfloor \log_2 y \rfloor + 1)}\right) > y.$$

However, as argued in [1], there are only y sets of at most d elements of $\{1, \ldots, m\} \times B$. But, by the definition of t, the fact that

$$t(2b^{3(\lfloor \log_2 y \rfloor + 1)}) > y$$

implies that any ϵ-separated subset of $2b^{3(\lfloor \log_2 y \rfloor + 1)}$ elements must $(\epsilon/2 - 2/b)$-fatly shatter more than y sets, and therefore a set of length at least $d+1$. Thus, no such subset can have $\mathrm{fat}_S(\epsilon/2 - 2/b)$ at most d. Taking the contrapositive completes the proof of the first inequality in the lemma.

The second inequality is obtained by bounding y using Sauer's lemma (see, for example, [6]). □

5 Sample complexity bounds

In this section, we apply the bounds of the previous section to upper bound the sample size necessary for agnostic learnability, and for uniformly good estimates of the expectations of a set of random variables. We start with the latter.

Theorem 12 *Choose X, and a set F of functions from X to $[0, 1]$.*

If there is a $\kappa > 0$ such that for all $\epsilon > 0$, $\mathrm{fat}_F((1/4 - \kappa)\epsilon)$ is finite, then

$$m_{GC,F}(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2} \left(\mathrm{fat}_F((1/4 - \kappa)\epsilon) \log^2 \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \right). \tag{5}$$

If there is a $\kappa > 0$ such that for all $\epsilon > 0$, $\mathrm{fatV}_F((1/4 - \kappa)\epsilon)$ is finite, then

$$m_{GC,F}(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2} \left(\frac{1}{\epsilon} \mathrm{fatV}_F((1/4 - \kappa)\epsilon) \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \right). \tag{6}$$

Before sketching the proof of Theorem 12, we establish some lemmas. The first is Hoeffding's inequality (see [17], Appendix B).

Lemma 13 *Choose $a < b$, X. Let D be a probability distribution on X, and let $f_1, ..., f_m$ be independent random variables taking values in $[a, b]$. Then the probability under D^m of a sequence $(x_1, ..., x_m)$ for which*

$$\left| \left(\frac{1}{m} \sum_{i=1}^{m} f_i(x_i) \right) - \left(\frac{1}{m} \sum_{i=1}^{m} \int f_i(x) D(x) \right) \right| > \epsilon$$

is no more than $2e^{-2\epsilon^2 m/(b-a)^2}$.

The following is a restatement of Theorem 8 of Chapter II of [17].

Lemma 14 ([17]) *Suppose X and U are sets, D is a probability distribution on X, and $\Phi : X \times U \to [0, 1]$ and $\Psi : X \times U \to [0, 1]$ are functions for which $\Phi(\cdot, u_1)$ and $\Psi(\cdot, u_2)$ are independent random variables for all u_1 and u_2 in U. Suppose there exist constants $\beta, \alpha > 0$ such that for all $u \in U$, $D\{x : |\Psi(x, u)| \leq \alpha\} \geq \beta$. Then for all $\epsilon > 0$,*

$$D\{x : \sup_u \Phi(x, u) > \epsilon\} \leq \frac{1}{\beta} D\{x : \sup_u |\Phi(x, u) - \Psi(x, u)| > \epsilon - \alpha\}.$$

These are applied in the following.

Lemma 15 *Choose a set X, and a set $F \subseteq [0, 1]^X$. Choose $\epsilon > 0$, $0 < \alpha < \epsilon$ and $m \in \mathbf{N}$. Then*

$$\begin{aligned} &D^m\{(x_1, ..., x_m) : \exists f \in F \left| \left(\tfrac{1}{m} \textstyle\sum_{i=1}^m f(x_i) \right) - \int f(x)\, dD(x) \right| > \epsilon\} \\ &\leq \tfrac{1}{1-2e^{-2\alpha^2 m}} D^{2m}\{(x_1, ..., x_m, y_1, ..., y_m) : \exists f \in F \left| \left(\tfrac{1}{m} \textstyle\sum_{i=1}^m f(x_i) \right) - \left(\tfrac{1}{m} \textstyle\sum_{i=1}^m f(y_i) \right) \right| > \epsilon - \alpha\}. \end{aligned}$$

Proof: In Lemma 14, set X to be this lemma's X^{2m}, U to be F, Φ to be defined by

$$\Phi((x_1, ..., x_m, y_1, ..., y_m), f) = \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \int f(x)\ D(x)$$

and Ψ by

$$\Psi((x_1, ..., x_m, y_1, ..., y_m), f) = \frac{1}{m} \sum_{i=1}^{m} f(y_i) - \int f(x)\ D(x).$$

Applying the standard Hoeffding bound (Lemma 13), we get for all $f \in F$

$$D^{2m}\{(x_1, ..., x_m, y_1, ..., y_m) : |\Psi((x_1, ..., x_m, y_1, ..., y_m), f)| \leq \alpha\} \geq 1 - 2e^{-2\alpha^2 m}.$$

Applying Lemma 14 completes the proof. □

Lemma 16 *Choose X, $F \subseteq [0, 1]^X$. Let D be a probability distribution over X. Choose $0 < \alpha, \epsilon < 1$ with $\alpha < \epsilon/2$. Then*

$$\begin{aligned} &D^{2m}\{(x_1, ..., x_m, y_1, ..., y_m) : \exists f \in F \left| \left(\tfrac{1}{m} \textstyle\sum_{i=1}^m f(x_i) \right) - \left(\tfrac{1}{m} \textstyle\sum_{i=1}^m f(y_i) \right) \right| > \epsilon\} \\ &\leq 2 \left(\sup_{\xi \in X^{2m}} \mathcal{N}(\epsilon/2 - \alpha, F_{|\xi}) \right) e^{-2\alpha^2 m}. \end{aligned}$$

Proof: Let U be the uniform distribution over $\{-1,1\}$. Then by symmetry,

$$\begin{aligned}
&D^{2m}\{(x_1,...,x_m,y_1,...,y_m) : \exists f \in F \left|\left(\tfrac{1}{m}\textstyle\sum_{i=1}^m f(x_i)\right) - \left(\tfrac{1}{m}\textstyle\sum_{i=1}^m f(y_i)\right)\right| > \epsilon\} \\
&= (D^{2m} \times U^m)\{(x_1,...,x_m,y_1,...,y_m),(u_1,...,u_m) : \exists f \in F \left|\tfrac{1}{m}\textstyle\sum_{i=1}^m u_i(f(x_i)-f(y_i))\right| > \epsilon\} \qquad (7) \\
&\leq \sup_{(x_1,...,x_m,y_1,...,y_m)} U^m\{(u_1,...,u_m) : \exists f \in F \left|\left(\tfrac{1}{m}\textstyle\sum_{i=1}^m u_i(f(x_i)-f(y_i))\right)\right| > \epsilon\}.
\end{aligned}$$

Fix $\sigma = (x_1,...,x_m,y_1,...,y_m)$.
Choose $f \in F$. Suppose $g : X \to [0,1]$ had

$$\frac{1}{2m}\sum_{i=1}^m (|g(x_i)-f(x_i)| + |g(y_i)-f(y_i)|) \leq \epsilon/2 - \alpha,$$

and that

$$\left|\frac{1}{m}\sum_{i=1}^m u_i(f(x_i)-f(y_i))\right| > \epsilon.$$

Then

$$\begin{aligned}
\left|\frac{1}{m}\sum_{i=1}^m u_i(g(x_i)-g(y_i))\right| &= \left|\frac{1}{m}\sum_{i=1}^m u_i(g(x_i)-f(x_i)+f(x_i)-g(y_i)+f(y_i)-f(y_i))\right| \\
&= \left|\frac{1}{m}\sum_{i=1}^m u_i(f(x_i)-f(y_i)) + \frac{1}{m}\sum_{i=1}^m u_i(g(x_i)-f(x_i)-g(y_i)+f(y_i))\right| \\
&\geq \left|\frac{1}{m}\sum_{i=1}^m u_i(f(x_i)-f(y_i))\right| - \left|\frac{1}{m}\sum_{i=1}^m (|g(x_i)-f(x_i)| + |g(y_i)-f(y_i)|)\right| \\
&> \epsilon - (\epsilon - 2\alpha) \\
&= 2\alpha.
\end{aligned}$$

So if T is a (minimum-sized) $(\epsilon/2-\alpha)$-cover of $F_{|\sigma}$, then

$$\begin{aligned}
&U^m\{(u_1,...,u_m) : \exists f \in F \left|\left(\tfrac{1}{m}\textstyle\sum_{i=1}^m u_i(f(x_i)-f(y_i))\right)\right| > \epsilon\} \\
&\leq \textstyle\sum_{v\in T} U^m\{(u_1,...,u_m) : \left|\left(\tfrac{1}{m}\textstyle\sum_{i=1}^m u_i(v_i - v_{m+i})\right)\right| > 2\alpha\}.
\end{aligned} \qquad (8)$$

Fix $v \in X^{2m}$. Then $u_1(v_1 - v_{m+1}),...,u_m(v_m - v_{2m})$ form a sequence of independent $[-1,1]$ random variables with zero mean. Applying Hoeffding's inequality, we get

$$U^m\left\{(u_1,...,u_m) : \left|\left(\frac{1}{m}\sum_{i=1}^m u_i(v_i - v_{m+i})\right)\right| > 2\alpha\right\} \leq 2e^{-2\alpha^2 m}.$$

Combining this with

$$|T| \leq \sup_{\sigma \in X^{2m}} \mathcal{N}(\epsilon/2-\alpha, F_{|\sigma}),$$

(8), and (7) completes the proof. □

Lemma 17 *Choose $S \subseteq [0,1]^m$, $\epsilon > 0$, $\alpha < \epsilon/2$. Then*

$$\mathcal{N}(\epsilon, S) \leq \mathcal{N}(\epsilon - \alpha, Q_\alpha(S)).$$

Proof: Choose $v, w \in [0,1]^m$.

$$\begin{aligned}\ell_1(v, Q_\alpha(w)) &= \frac{1}{m}\sum_{i=1}^{m} |v_i - \alpha\lfloor w_i/\alpha \rfloor| \\ &= \frac{1}{m}\sum_{i=1}^{m} |(v_i - w_i) + \alpha(w_i/\alpha - \lfloor w_i/\alpha \rfloor)| \\ &\geq \ell_1(v,w) - \alpha.\end{aligned}$$

Thus $\ell_1(v,w) \leq \ell_1(v, Q_\alpha(w)) + \alpha$. Therefore, if some T is an $(\epsilon - \alpha)$-cover of $Q_\alpha(S)$, then T is an ϵ-cover of S, completing the proof. □

Next, we write down a lemma calculating a useful inverse. The lemma is proved using the by now standard technique from [3].

Lemma 18 *For any $y_1, y_2, y_4, \delta > 0$ and $y_3 \geq 1$, if*

$$m \geq \frac{2}{y_4}\left(y_2 \ln\left(\frac{2y_2y_3}{y_4}\right) + \ln\frac{y_1}{\delta}\right),$$

then

$$y_1 \exp(y_2 \ln(y_3 m) - y_4 m) \leq \delta.$$

Proof: If $\gamma = y_4/(2y_2y_3)$, then

$$m \geq \frac{2}{y_4}\left(y_2 \ln\left(\frac{2y_2y_3}{y_4}\right) + \ln\frac{y_1}{\delta}\right)$$

implies

$$\left(1 - \frac{\gamma y_2 y_3}{y_4}\right) m \geq \frac{1}{y_4}\left(y_2 \ln\left(\frac{1}{\gamma}\right) + \ln\frac{y_1}{\delta}\right).$$

Solving for m, we get

$$m \geq \frac{1}{y_4}\left(y_2\left(\gamma y_3 m + \ln\left(\frac{1}{\gamma}\right)\right) + \ln\frac{y_1}{\delta}\right).$$

Applying the fact [3] that for all $x, \gamma > 0$, $\ln x + \ln\gamma \leq \gamma x$ with $x = y_3 m$, we get

$$m \geq \frac{1}{y_4}\left(y_2 \ln(y_3 m) + \ln\frac{y_1}{\delta}\right).$$

Solving for δ completes the proof. □

Proof (of Theorem 12): Choose a probability distribution D on X. Let $\alpha = 1/\lceil 1/(\kappa\epsilon)\rceil$. Since fat_F is nonincreasing, $\mathrm{fat}_F(\epsilon/4 - \alpha) \leq \mathrm{fat}_F((1/4 - \kappa)\epsilon)$. Let $d = \mathrm{fat}_F(\epsilon/4 - \alpha)$. By Lemma 15,

$$\begin{aligned}&D^m\{(x_1,...,x_m) : \exists f \in F \left| \left(\tfrac{1}{m}\textstyle\sum_{i=1}^m f(x_i)\right) - \int f(x)\, dD(x)\right| > \epsilon\} \\ &\leq \tfrac{1}{1-2e^{-2\alpha^2 m}} D^{2m}\{(x_1,...,x_m,y_1,...,y_m) : \exists f \in F \left| \left(\tfrac{1}{m}\textstyle\sum_{i=1}^m f(x_i)\right) - \left(\tfrac{1}{m}\textstyle\sum_{i=1}^m f(y_i)\right)\right| > \epsilon - \alpha\}.\end{aligned}$$

Applying Lemmas 16 and 17 yields

$$\begin{aligned}&D^m\left\{(x_1,...,x_m) : \exists f \in F \left| \left(\frac{1}{m}\sum_{i=1}^m f(x_i)\right) - \int f(x)\, dD(x)\right| > \epsilon\right\} \\ &\leq 2(\sup_{\sigma \in X^{2m}} \mathcal{N}(\epsilon/2 - \alpha), F_{|\sigma}) \frac{e^{-\alpha^2 m/2}}{1 - 2e^{-2\alpha^2 m}} \\ &\leq 2(\sup_{\sigma \in X^{2m}} \mathcal{N}(\epsilon/2 - 8\alpha/7), Q_{\alpha/7}(F_{|\sigma})) \frac{e^{-\alpha^2 m/2}}{1 - 2e^{-2\alpha^2 m}}.\end{aligned}$$

It is immediate from the definition of fat_F that for all $0 < \beta < \gamma$, $\mathrm{fat}_{Q_\beta(F)}(\gamma) \leq \mathrm{fat}_F(\gamma - \beta)$, so $\mathrm{fat}_{Q_{\alpha/7}(F)}(\epsilon/4 - 6\alpha/7) \leq \mathrm{fat}_F(\epsilon/4 - \alpha)$. Lemma 11 implies that

$$\mathcal{N}\left(\epsilon/2 - 8\alpha/7, Q_{\alpha/7}(F_{|_\sigma})\right) \leq 2\left(\frac{7}{\alpha}\right)^{6d\log_2\left(\frac{14em}{\alpha d}\right)}.$$

If $m > \frac{1}{2\alpha^2}\ln 4$, $1/(1-2e^{-2\alpha^2 m}) < 2$. In that case, the probability above is less than

$$8\exp\left(\frac{6d}{\ln 2}\ln\frac{14em}{\alpha d}\ln\frac{7}{\alpha} - \frac{\alpha^2 m}{2}\right),$$

which, by Lemma 18, is no more than δ if

$$\begin{aligned} m &\geq \frac{4}{\alpha^2}\left(\frac{6d}{\ln 2}\ln\frac{7}{\alpha}\ln\left(\frac{336e}{\alpha\ln 2}\ln\frac{7}{\alpha}\right) + \ln\frac{8}{\delta}\right) \\ &= O\left(\frac{1}{\alpha^2}\left(d\log^2\frac{1}{\alpha} + \log\frac{1}{\delta}\right)\right), \end{aligned}$$

which completes the proof of (5).

A similar argument gives (6). In this case, let $d = \mathrm{fatV}_F(\epsilon/4 - \alpha)$. By Lemmas 15, 16, 17, and 10, and the fact that $\mathrm{fatV}_{Q_{\alpha/5}(F)}(\epsilon/4 - 4\alpha/5) \leq \mathrm{fatV}_F(\epsilon/4 - \alpha)$, we have

$$\begin{aligned} D^m\left\{(x_1,...,x_m) : \exists f \in F \left| \left(\frac{1}{m}\sum_{i=1}^m f(x_i)\right) - \int f(x)\, dD(x)\right| > \epsilon\right\} \\ \leq 2\left(\sup_{\sigma\in X^{2m}} \mathcal{N}(\epsilon/2 - 2\alpha), F_{|_\sigma}\right)\frac{e^{-2\alpha^2 m}}{1-2e^{-8\alpha^2 m}} \\ \leq 2(\sup_{\sigma\in X^{2m}} \mathcal{N}(\epsilon/2 - 11\alpha/5), Q_{\alpha/5}(F_{|_\sigma}))\frac{e^{-2\alpha^2 m}}{1-2e^{-8\alpha^2 m}} \\ \leq \frac{4\epsilon}{\alpha}\left(\frac{5}{\alpha}+1\right)^{16d/\alpha}\frac{e^{-2\alpha^2 m}}{1-2e^{-8\alpha^2 m}}, \end{aligned}$$

and this quantity is less than δ when

$$\begin{aligned} m &\geq \frac{8d}{\alpha^3}\ln\frac{6}{\alpha} + \frac{1}{2\alpha^2}\ln\frac{8\epsilon}{\delta\alpha} \\ &= O\left(\frac{1}{\alpha^2}\left(\frac{1}{\alpha}\mathrm{fatV}_F(\epsilon/4-\alpha)\log\frac{1}{\alpha} + \log\frac{1}{\delta}\right)\right). \end{aligned}$$

□

To use Theorem 12 to give sample size bounds for agnostic learnability, we will consider an algorithm that approximately minimizes empirical loss. In this case, we need to show that a class of associated loss functions is an ϵ-uniform GC class, and to do this we relate covering numbers of the loss function class to covering numbers of the function class. This lemma is implicit in the analysis of Natarajan [16].[4]

Lemma 19 *Suppose that X is a set, F is a class of functions that map from X to $[0,1]$, $x = (x_1,\ldots,x_m) \in X^m$, and $z = ((x_1,y_1),\ldots,(x_m,y_m)) \in (X\times[0,1])^m$. Define the loss function class*

$$L_F = \{(x,y) \mapsto |f(x) - y| : f \in F\}.$$

Then for any $\epsilon > 0$, $\mathcal{N}\left(\epsilon, L_{F_{|_z}}\right) \leq \mathcal{N}\left(\epsilon, F_{|_x}\right)$.

[4] It is also possible to relate the fat-shattering functions of these classes directly using Sauer's lemma (see [9]), but the proof is not as simple and the result slightly weaker.

Theorem 20 *Choose a set X, a set F of functions from X to $[0,1]$, and $\epsilon, \delta > 0$.*

If there exists $\kappa > 0$ such that for all $\epsilon > 0$, $\mathrm{fat}_F((1/4-\kappa)\epsilon)$ is finite, then there is a learner A that (ϵ,δ)-learns in the agnostic sense with respect to F from

$$O\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_F((1/4-\kappa)\epsilon)\log^2\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right) \tag{9}$$

examples.

If there exists $\kappa > 0$ such that for all $\epsilon > 0$, $\mathrm{fatV}_F((1/4-\kappa)\epsilon)$ is finite, then there is a learner A that (ϵ,δ)-learns in the agnostic sense with respect to F from

$$O\left(\frac{1}{\epsilon^2}\left(\frac{1}{\epsilon}\mathrm{fatV}_F((1/4-\kappa)\epsilon)\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right) \tag{10}$$

examples.

Proof: Fix $\beta > 0$, a small positive constant. The algorithm we will consider takes a sample

$$((x_1,y_1),\ldots,(x_m,y_m)) \in (X\times[0,1])^m$$

and chooses a function $f' \in F$ that has

$$\frac{1}{m}\sum_{i=1}^m |f'(x_i)-y_i| < \inf_{f\in F}\frac{1}{m}\sum_{i=1}^m |f(x_i)-y_i| + \beta.$$

Fix any distribution P on $X\times[0,1]$. Let $f^* \in F$ satisfy $\mathbf{er}_P(f^*) \le \inf_{g\in F}\mathbf{er}_P(g)+\beta$. From Hoeffding's inequality, with probability at least $1-2e^{-2\beta^2 m}$ over the sample,

$$\frac{1}{m}\sum_{i=1}^m |f^*(x_i)-y_i| \le \int |f^*(x)-y|dP(x,y)+\beta.$$

If $m \ge \frac{1}{2\beta^2}\log\frac{2}{\delta}$, this probability is at least $1-\delta/2$. Applying Theorem 12 to the class L_F and using Lemma 19, if m satisfies (9) and (10) above, every function f in F has

$$\frac{1}{m}\sum_{i=1}^m |f(x_i)-y_i| - \int |f(x)-y|dP(x,y) \le \epsilon - 3\beta$$

with probability at least $1-\delta/2$. It follows that, with probability at least $1-\delta$,

$$\begin{aligned}\mathbf{er}_P(f') &\le \inf_{f\in F}\frac{1}{m}\sum_{i=1}^m |f(x_i)-y_i| + \epsilon - 2\beta\\ &\le \mathbf{er}_P(f^*)+\epsilon-\beta\\ &\le \inf_{g\in F}\mathbf{er}_P(g)+\epsilon.\end{aligned}$$

□

6 Better bounds in terms of the scale

In this section, we describe a more direct approach to bounding the sample complexity of ϵ-agnostic learning, which saves a factor of two in the scale at which the dimension must be finite over that described in the previous section, sometimes at the expense of a small increase in the sample complexity.

Theorem 21 *Choose X, a set F of functions from X to $[0,1]$, and $\epsilon, \delta > 0$.*

If there is a $\kappa > 0$ such that for all $\epsilon > 0$, $\mathrm{fat}_F((1/2-\kappa)\epsilon)$ is finite, then there is a learner A that (ϵ,δ)-learns in the agnostic sense with respect to F from

$$O\left(\frac{1}{\epsilon^2}\mathrm{fat}_F((1/2-\kappa)\epsilon)\left(\log^2\frac{1}{\epsilon}\right)\left(\log\frac{1}{\delta}\right)\right) \tag{11}$$

examples.

If there is a $\kappa > 0$ such that for all $\epsilon > 0$, $\mathrm{fatV}_F((1/2-\kappa)\epsilon)$ is finite, then there is a learner A that (ϵ,δ)-learns in the agnostic sense with respect to F from

$$O\left(\frac{1}{\epsilon^3}\mathrm{fatV}_F((1/2-\kappa)\epsilon)\left(\log\frac{1}{\epsilon}\right)\left(\log\frac{1}{\delta}\right)\right) \tag{12}$$

examples.

Proof: Fix $k \in \mathbf{N}$, let $\alpha = 1/\lceil 1/(\epsilon\kappa)\rceil$, and let $\gamma = \alpha/13$. Consider a mapping Q from $(X\times[0,1])^k \times X^k$ to $[0,1]$, defined as follows. Fix a function ϕ that maps from X^{2k} to the set of finite subsets of $[0,1]^{2k}$ such that, for any $x \in X^{2k}$, $\phi(x)$ is a minimal $(\epsilon - 9\gamma)$-cover of $F_{|x}$, and $\phi(x)$ is invariant under permutations of the components of x. Then let $x = (x_1,\ldots,x_{2k}) \in X^{2k}$, and for $(y_1,\ldots,y_k) \in [0,1]^k$ let $Q((x_1,y_1),\ldots,(x_k,y_k),x_{k+1},\ldots,x_{2k}) = t'_{2k}$, where $t' = (t'_1,\ldots,t'_{2k}) \in \phi(x)$ satisfies

$$\frac{1}{k}\sum_{i=1}^{k}|t'_i - y_i| = \min_{s\in\phi(x)}\frac{1}{k}\sum_{i=1}^{k}|s_i - y_i|.$$

We will first show that, for any distribution on $X\times[0,1]$, Q predicts the value y_{2k} associated with x_{2k} almost as well as the best function in F, taking expectations over random sequences. We use this property to construct a learner that returns a hypothesis that has error within ϵ of the best in F, with high probability.

Fix a distribution P on $X\times[0,1]$. Suppose

$$((x_1,y_1),\ldots,(x_{2k},y_{2k})) \in (X\times[0,1])^{2k}$$

is a random sequence chosen according to P. Let $x = (x_1,\ldots,x_{2k})$ and $y = (y_1,\ldots,y_{2k})$. Choose $f^* \in F$ that satisfies $\mathbf{er}_P(f^*) \le \inf_{f\in F}\mathbf{er}_P(f) + \gamma$. In comparing functions defined on X, such as f^*, we will sometimes refer to the function as a vector, with the obvious interpretation that $f^*_i = f^*(x_i)$. Since $\phi(x)$ is an $(\epsilon - 9\gamma)$-cover of $F_{|x}$, there is a $t^* \in \phi(x)$ with $\ell_1(t^*,f^*) \le \epsilon - 9\gamma$. It follows that $\ell_1(t^*,y) \le \epsilon - 9\gamma + \ell_1(f^*,y)$.

Applying the Hoeffding bound,

$$P^{2k}\left\{((x_1,y_1),\ldots,(x_{2k},y_{2k})) : \ell_1(f^*,y) > \mathbf{er}_P(f^*) + \gamma\right\} \le 2e^{-4\gamma^2 k}.$$

If $k > \frac{1}{4\gamma^2}\log\frac{2}{\gamma}$, this probability is less than γ. In that case, with probability at least $1-\gamma$, $\ell_1(f^*,y) \le \mathbf{er}_P(f^*) + \gamma$, which implies $\ell_1(t^*,y) \le \epsilon - 8\gamma + \mathbf{er}_P(f^*)$.

For two vectors $a, b \in [0,1]^{2k}$, define

$$\ell_1^{\mathrm{first}}(a,b) = \frac{1}{k}\sum_{i=1}^{k}|a_i - b_i|,$$

$$\ell_1^{\mathrm{last}}(a,b) = \frac{1}{k}\sum_{i=k+1}^{2k}|a_i - b_i|.$$

Now, as in the proof of Lemma 16, let U be the uniform distribution over $\{-1,1\}$. Then, since ϕ is invariant under permutations,

$$\begin{aligned}
&P^{2k}\left\{((x_1,y_1),\ldots,(x_{2k},y_{2k})) : \exists t \in \phi(x)\ |\ell_1^{\text{first}}(t,y) - \ell_1^{\text{last}}(t,y)| > 2\gamma\right\} \\
&\leq \sup_{(x,y)} U^k\left\{(u_1,\ldots,u_k) : \exists t \in \phi(x)\ \left|\tfrac{1}{k}\textstyle\sum_{i=1}^k u_i\left(|t_i - y_i| - |t_{i+k} - y_{i+k}|\right)\right| > 2\gamma\right\}
\end{aligned}$$

For any fixed $t \in \phi(x)$, Hoeffding's inequality implies

$$U^k\left\{(u_1,\ldots,u_k) : \left|\frac{1}{k}\sum_{i=1}^k u_i\left(|t_i - y_i| - |t_{i+k} - y_{i+k}|\right)\right| > 2\gamma\right\} \leq 2e^{-2\gamma^2 k}.$$

So with probability at least $1 - |\phi(x)|2e^{-2\gamma^2 k}$, for all t in $\phi(x)$,

$$|\ell_1^{\text{first}}(t,y) - \ell_1^{\text{last}}(t,y)| \leq 2\gamma. \tag{13}$$

This implies

$$|\ell_1^{\text{first}}(t,y) - \ell_1(t,y)| \leq \gamma$$

and

$$|\ell_1^{\text{last}}(t,y) - \ell_1(t,y)| \leq \gamma.$$

The probability that this does not happen is no more than γ if

$$\sup_{x \in X^{2k}} \mathcal{N}\left(\epsilon - 9\gamma, F_{|x}\right) 2e^{-2\gamma^2 k} \leq \gamma.$$

Now, Lemmas 17 and 11, together with (1), imply that

$$\begin{aligned}
\mathcal{N}\left(\epsilon - 9\gamma, F_{|x}\right) &\leq \mathcal{N}\left(\epsilon - 10\gamma, Q_\gamma\left(F_{|x}\right)\right) \\
&\leq 2\left(\frac{1}{\gamma}\right)^{6d\log_2(142k/(d\gamma))},
\end{aligned}$$

where $d = \text{fat}_F(\epsilon - 13\gamma)$, since $\text{fat}_{Q_\gamma(F)}(\epsilon - 12\gamma) \leq \text{fat}_F(\epsilon - 13\gamma)$. So Inequality (13) will hold for all t in $\phi(x)$ with probability at least $1 - \gamma$, provided

$$4\exp\left(\frac{6d}{\ln 2}\left(\ln\frac{142k}{d\gamma}\right)\left(\ln\frac{1}{\gamma}\right) - 2\gamma^2 k\right) \leq \gamma.$$

Applying Lemma 18, we can see that there is a constant c such that

$$k \geq \frac{c}{\gamma^2}\left(d\ln^2\frac{1}{\gamma} + \ln\frac{1}{\gamma}\right)$$

will suffice. In that case, with probability at least $1 - 2\gamma$, the $t' \in \phi(x)$ with minimal $\ell_1^{\text{first}}(t',y)$ satisfies

$$\begin{aligned}
\ell_1(t',y) &\leq \ell_1^{\text{first}}(t',y) + \gamma \\
&\leq \ell_1^{\text{first}}(t^*,y) + \gamma \\
&\leq \ell_1(t^*,y) + 2\gamma \\
&\leq \epsilon - 6\gamma + \mathbf{er}_P(f^*),
\end{aligned}$$

and hence

$$\begin{aligned}
\ell_1^{\text{last}}(t',y) &\leq \epsilon - 5\gamma + \mathbf{er}_P(f^*) \\
&\leq \epsilon - 4\gamma + \inf_{f \in F} \mathbf{er}_P(f).
\end{aligned}$$

That is,

$$P^{2k}\{((x_1,y_1),\ldots,(x_{2k},y_{2k})) : \ell_1^{\text{last}}(t',y) > \epsilon - 4\gamma + \inf_{f\in F} \mathbf{er}_P(f)\} < 2\gamma,$$

which implies

$$\int \ell_1^{\text{last}}(t',y) dP^{2k}((x_1,y_1),\ldots,(x_{2k},y_{2k})) < \epsilon - 2\gamma + \inf_{f\in F} \mathbf{er}_P(f),$$

and hence

$$\int (|Q((x_1,y_1),\ldots,(x_k,y_k),x_{k+1},\ldots,x_{2k}) - y_{2k}| dP^{2k}((x_1,y_1),\ldots,(x_{2k},y_{2k})) - \inf_{f\in F} \mathbf{er}_P(f) < \epsilon - 2\gamma.$$

If we define the hypothesis h of Q as

$$h(\beta) = Q((x_1,y_1),\ldots,(x_k,y_k),x_{k+1},\ldots,x_{2k-1},\beta),$$

we have

$$P^{2k-1}\left\{((x_1,y_1),\ldots,(x_{2k-1},y_{2k-1})) : \mathbf{er}_P(h) - \inf_{f\in F} \mathbf{er}_P(f) > \epsilon - \gamma\right\} < 1 - \gamma/\epsilon.$$

To complete the proof, we use a technique from [12] to convert this prediction strategy to an agnostic learning algorithm. Consider the algorithm which takes as input $N_1(2k-1)+N_2$ labelled examples, uses the first $N_1(2k-1)$ examples and the mapping Q to compute N_1 hypotheses, and outputs the hypothesis from this set that has minimum error over the remaining N_2 examples. With probability at least $1-\delta/2$, at least one of the N_1 hypotheses has error no more than $\epsilon-\gamma$, provided that $(1-\gamma/\epsilon)^{N_1} < \delta/2$; setting $N_1 = \frac{\epsilon}{\gamma}\ln\frac{2}{\delta}$ will suffice for this. For each of these N_1 hypotheses h, the algorithm calculates the empirical error $\frac{1}{N_2}\sum_{i=1}^{N_2}|h(x_i)-y_i|$, and chooses the hypothesis with the minimum empirical error. Hoeffding's inequality implies that the probability that some hypothesis has empirical error more than $\gamma/2$ from $\mathbf{er}_P(h)$ is no more than $2N_1e^{-\gamma^2 N_2/2}$. This probability is less than $\delta/2$ when $N_2 > \frac{2}{\gamma^2}\log\frac{4N_1}{\delta}$. This implies that, with probability at least $1-\delta$ over the $N_1(2k-1)+N_2$ examples, the hypothesis returned by the algorithm has error less than ϵ. Clearly, the algorithm needs to see

$$O\left(\frac{\epsilon d}{\alpha^3}\log\frac{1}{\delta}\log^2\frac{1}{\alpha} + \frac{1}{\alpha^2}\log\frac{\epsilon}{\delta\alpha}\right)$$

examples, completing the proof of (11). The bound (12) can be proved analogously using Lemma 10 in place of Lemma 11. □

Buescher and Kumar proposed a related algorithm in [7]. Their algorithm (the "canonical estimator") splits a sequence of labelled examples into two parts. Let ξ be the sequence of points from X in the first part of the sample. The algorithm chooses a finite subset T of F such that $T_{|_\xi}$ is a cover of $F_{|_\xi}$. It then returns the function in T that has minimal error on the remaining part of the sample. Interestingly, this algorithm also discards the labels of part of the training sample.

7 Necessary conditions

In this section, we collect necessary conditions for some of the properties considered in this paper. Coupled with the positive results of the previous sections, these results considerably narrow the constant factor gap between the scales at which the finiteness of the scale-sensitive dimensions is

necessary and sufficient for learning and the GC property. We also provide examples showing that these necessary conditions are not sufficient conditions, and that they cannot be improved.

First, we prove the necessity condition for ϵ-uniform GC classes. The proof is based on that of the analogous result for fatV which was proved in [1] and follows from this new result since $\mathrm{fatV}_F \leq \mathrm{fat}_F$ for all F. It improves on the result in [1] by a factor of 2 the scale at which fat_F's finiteness is necessary for F to be an ϵ-uniform GC class.

Theorem 22 *Choose X, $F \subseteq [0,1]^X$, and $0 < \epsilon < 1$. Then if there exists $\alpha > 0$ such that $\mathrm{fat}_F(\epsilon/2 + \alpha) = \infty$, then F is not an ϵ-uniform GC class.*

Proof: Choose $0 < \epsilon < 1$. Assume for contradiction that there exist X, $F \subseteq [0,1]^X$, and $\alpha > 0$ such that $\mathrm{fat}_F(\epsilon/2+\alpha) = \infty$ but that F is an ϵ-uniform GC class. Let $m = m_{\mathrm{GC},F}(\epsilon, 1/2)$. Choose $d \in \mathbf{N}$ such that

$$d \geq \frac{m}{\alpha}(1 + \epsilon/2 + \alpha). \tag{14}$$

Let $(x_1, r_1), ..., (x_d, r_d)$ be $(\epsilon/2 + \alpha)$-fatly shattered by F, and let D be the uniform distribution over $x_1, ..., x_d$. Let $r = (1/d)\sum_{i=1}^{d} r_i$.

We claim that for *any* sequence $u_1, ..., u_m$ of elements of $\{x_1, ..., x_d\}$, there is an $f \in F$ such that

$$\left| \left(\frac{1}{m} \sum_{i=1}^{m} f(u_m) \right) - \int f(z)\, dD(z) \right| > \epsilon.$$

Choose such a $u_1, ..., u_m$, and for each $j \in \{1, ..., m\}$ let i_j be such that $u_j = x_{i_j}$.

Assume as a first case that

$$\frac{1}{m} \sum_{j=1}^{m} r_{i_j} \leq r. \tag{15}$$

Choose $f \in F$ such that

$$f(x_i) \begin{cases} \leq r_i - (\epsilon/2 + \alpha) & \text{if } i \in \{i_1, ..., i_m\} \\ \geq r_i + (\epsilon/2 + \alpha) & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} \frac{1}{m} \sum_{j=1}^{m} f(x_{i_j}) &\leq \frac{1}{m} \sum_{j=1}^{m} (r_{i_j} - (\epsilon/2 + \alpha)) \\ &\leq r - (\epsilon/2 + \alpha) \end{aligned}$$

by (15). However,

$$\begin{aligned} \int f(z)\, dD(z) &= \frac{1}{d} \sum_{i=1}^{d} f(x_i) \\ &\geq \frac{1}{d} \sum_{i \notin \{i_1, ..., i_m\}} f(x_i) \\ &\geq \frac{1}{d} \sum_{i \notin \{i_1, ..., i_m\}} r_i + (\epsilon/2 + \alpha) \\ &= \frac{d-m}{d} \left(\frac{1}{d-m} \sum_{i \notin \{i_1, ..., i_m\}} r_i + (\epsilon/2 + \alpha) \right) \\ &\geq \frac{d-m}{d} (r + (\epsilon/2 + \alpha)) \end{aligned}$$

by (15) together with the definition of r. Thus,

$$\begin{aligned}\int f(z)\, dD(z) - \frac{1}{m}\sum_{j=1}^{m} f(u_j) &\geq (1 - m/d)(r + \epsilon/2 + \alpha) - (r - (\epsilon/2 + \alpha)) \\ &= \epsilon + 2\alpha - \frac{m}{d}(r + \epsilon/2 + \alpha) \\ &\geq \epsilon + \alpha,\end{aligned}$$

from (14), completing the proof in this case. The case in which $\frac{1}{m}\sum_{j=1}^{m} r_{i_j} > r$ can be handled similarly.

Therefore, we have that for samples of size m, there is a function in F whose expectation is estimated with accuracy worse than ϵ, a contradiction, completing the proof. □

The next result shows that this condition is not sufficient for F to be an ϵ-uniform GC class.

Theorem 23 *For each $0 < \epsilon < 1/2$, there is a function class F that is not an ϵ-uniform GC class, but for all $\alpha > 0$, $\mathrm{fat}_F(\epsilon/2 + \alpha)$ is finite.*

Proof: Fix $0 < \epsilon < 1/2$ and let F be the class of all functions f from $\mathbf{N}$ to $[0, 1]$ satisfying

$$f(i) \in \{1/2 + (\epsilon/2 + 1/(i+3)), 1/2 - (\epsilon/2 + 1/(i+3))\}.$$

Clearly, for all $\alpha > 0$, $\mathrm{fat}_F(\epsilon/2 + \alpha)$ is finite. For sample size m, consider the distribution D that is uniform on $Z = \{1, \ldots, m^2 e^{\epsilon m}\}$. Then for any sequence $x_1, \ldots, x_m$, there is a function f in F with

$$\left| \frac{1}{m}\sum_{i=1}^{m} f(x_i) - \int_X f(x) dD(x) \right| > \epsilon.$$

To see this, fix a sequence $x_1, \ldots, x_m$, let $d = (m+3)^2 e^{\epsilon m}$, and consider the function f in F satisfying

$$f(n) = \begin{cases} 1/2 - (\epsilon/2 + 1/(n+3)) & \text{if } n = x_i \text{ for some } i \\ 1/2 + \epsilon/2 + 1/(n+3) & \text{otherwise.} \end{cases}$$

If we define $S_x = \{x_1, \ldots, x_m\}$, then

$$\begin{aligned}&\textstyle\int_X f(x) dD(x) - \frac{1}{m}\sum_{i=1}^{m} f(x_i) \\ &> \textstyle\frac{1}{d}\left(\sum_{i=1}^{d} f(i)\right) - 1/2 + \epsilon/2 \\ &= \textstyle\frac{1}{d}\left(\sum_{i=1}^{d} f(i) - 1/2 + \epsilon/2\right) \\ &= \textstyle\frac{1}{d}\left(\sum_{i\in S_x}(f(i) - 1/2 + \epsilon/2) + \sum_{i\in Z - S_x}(f(i) - 1/2 + \epsilon/2)\right) \\ &= \textstyle\frac{1}{d}\left(\sum_{i\in S_x}(-1/(i+3)) + \sum_{i\in Z - S_x}(\epsilon + 1/(i+3))\right).\end{aligned}$$

Both sums are clearly minimized when $S_x = \{1, \ldots, m\}$. Using the fact that $\ln(n+1) < \sum_{i=1}^{n} 1/i < \ln n + 1$, we have

$$\int_X f(x) dD(x) - \frac{1}{m}\sum_{i=1}^{m} f(x_i) > \epsilon + \left(\ln(d+4) - \epsilon m - 2\ln(m+3) + 4\right)/d,$$

but the definition of d implies that this quantity is at least ϵ. □

Note that for all function classes F and $\epsilon > 0$, $\mathrm{fat}_F(\epsilon) \geq \mathrm{fatV}_F(\epsilon)$, so that Theorem 22 implies the same thing about fatV. The following observation shows that there is no better necessary condition in terms of fat or fatV.

Proposition 24 *There is a function class F that is an ϵ-uniform GC class, but has* $\mathrm{fatV}_F(\epsilon/2)$ *infinite.*

Proof: Suppose F is the set of all functions from the natural numbers to $\{1/2 - \epsilon/2, 1/2 + \epsilon/2\}$. Clearly, $\mathrm{fatV}_F(\epsilon/2) = \infty$. However, for any sample, the estimate of the expectation of any member f of F must be in $[1/2 - \epsilon/2, 1/2 + \epsilon/2]$, as must be the true expectation of f with respect to any distribution. □

Next, we turn to proving a necessary condition for ϵ-agnostic learnability. The following variant of fat_F, due to Simon [18], will be useful. For X, $F \subseteq [0,1]^X$, and $\gamma > 0$, we say F *strongly γ-fatly shatters* a sequence $(x_1, l_1, u_1), ..., (x_d, l_d, u_d)$ of elements of $X \times [0,1]^2$ if $u_i \geq l_i + 2\gamma$ for $i = 1, \ldots, d$ and, for all $(b_1, ..., b_d) \in \{0,1\}^d$, there is an $f \in F$ such that

$$f(x_i) = u_i \Leftrightarrow b_j = 1$$
$$f(x_i) = l_i \Leftrightarrow b_j = 0$$

for $i = 1, \ldots, d$. We then define $\mathrm{sfat}_F(\gamma)$ to be the length of the longest sequence that is strongly γ-fatly shattered by F, or ∞ is there is no longest sequence.

The following lemma, whose proof closely follows parts of that of a related result in [18], as well as Theorem 5, will be useful.

Lemma 25 *Choose X, $F \subseteq [0,1]^X$, $\epsilon > 0$. Then if there exists $\alpha > 0$ such that* $\mathrm{sfat}_F(\epsilon + \alpha)$ *is infinite, then F is not ϵ-agnostically learnable.*

Proof: Assume for contradiction that F is ϵ-agnostically learnable, but that there exists $\alpha > 0$ such that $\mathrm{sfat}_F(\epsilon + \alpha)$ is infinite. Fix such an $\alpha > 0$. Let $m \in \mathbf{N}$, and a learner A be such that for all distributions P on $X \times [0,1]$,

$$P^m \left\{ z : \int |(A(z))(x) - y| \, dP(x,y) \geq (\inf_{f \in F} |f(x) - y|) + \epsilon \right\} \leq 1/2.$$

Choose $d \in \mathbf{N}$ such that

$$d > \frac{m(\epsilon + \alpha)}{\alpha}. \tag{16}$$

Choose a sequence $(x_1, l_1, u_1), \ldots, (x_d, l_d, u_d)$ from among those $(\epsilon + \alpha)$-fatly shattered by F. For each $b \in \{0,1\}^d$, choose $f_b \in F$ so that

$$f_b(x_i) = u_i \Leftrightarrow b_i = 1$$
$$f_b(x_i) = l_i \Leftrightarrow b_i = 0,$$

and let $G = \left\{ f_b : b \in \{0,1\}^d \right\}$. For each $b \in \{0,1\}^d$, let P_b be a distribution over $X \times [0,1]$ obtained by choosing the first component uniformly from $x_1, ..., x_d$, and evaluating f_b at the first component to get the second. Note that for each such P_b, $\inf_{f \in F} \int |f(x) - y| \, dP_b(x,y) = 0$.

Choose $v_1, ..., v_m \in \{x_1, ..., x_d\}$, and let $i_1, ..., i_m$ be such that for each $1 \leq j \leq m$, $v_j = x_{i_j}$. Notice that $h_b = A((v_1, f_b(v_1)), ..., (v_m, f_b(v_m)))$ is a function only of those components b_i for which $i \in \{i_1, ..., i_m\}$. Suppose we choose b uniformly according to the uniform distribution over $\{0,1\}^d$. Choose $i \notin \{i_1, ..., i_m\}$, and $(c_1, ..., c_m) \in \{-1, 1\}^m$. Then, by the independence of the choice of b_i from that of the other components, in particular those determining h_b, the expectation of $|h_b(x_i) - f_b(x_i)|$, given that $b_{i_1} = c_1, ..., b_{i_m} = c_m$, is

$$1/2|h_b(x_i) - u_i| + 1/2|h_b(x_i) - l_i| \geq 1/2|u_i - l_i| \geq \epsilon + \alpha.$$

Since this is true independent of $c_1, ..., c_m$, for any $i \notin \{i_1, ..., i_m\}$, the expected value of the error of A on x_i is at least $\epsilon + \alpha$. Therefore, the overall expected error of A's hypothesis, over the random choice of b, is at least

$$(1 - m/d)(\epsilon + \alpha).$$

This implies there exists b such that if

$$z = ((v_1, f_b(v_1)), ..., (v_m, f_b(v_m))),$$

$\int |(A(z))(x) - y| P_b(x, y)$ is at least

$$(1 - m/d)(\epsilon + \alpha).$$

Since $v_1, ..., v_m$ was chosen arbitrarily, by (16), this contradicts the fact that A $(\epsilon, 1/2)$-agnostically learns, completing the proof. □

Theorem 26 *Choose X, $F \subseteq [0,1]^X$, and $\epsilon > 0$. Then if there exists $\alpha > 0$ such that $\mathrm{fat}_F(\epsilon + \alpha)$ is infinite, then F is not ϵ-agnostically learnable.*

Proof: Fix $\alpha > 0$ such that $\mathrm{fat}_F(\epsilon + \alpha)$ is infinite. Then $\mathrm{fat}_{Q_{\alpha/3}(F)}(\epsilon + 2\alpha/3)$ is infinite. By Lemma 9 of [2], this implies $\mathrm{sfat}_{Q_{\alpha/3}(F)}(\epsilon + 2\alpha/3)$ is infinite, and then Lemma 25 implies $Q_{\alpha/3}(F)$ is not $(\epsilon + \alpha/3)$-agnostically learnable. But then F is not ϵ-agnostically learnable, since for every $f \in F$ and distribution P on $X \times [0,1]$, $\mathbf{er}_P(f) \leq \mathbf{er}_P(Q_{\alpha/3}(f)) + \alpha/3$, so a learner that ϵ-agnostically learns F can $(\epsilon + \alpha/3)$-agnostically learn $Q_{\alpha/3}(F)$. □

Next, we show that the converse of Theorem 26 is not true.

Theorem 27 *For each $0 < \epsilon < 1/4$, there is a function class F that is not ϵ-agnostically learnable, but for all $\alpha > 0$, $\mathrm{fat}_F(\epsilon + \alpha)$ is finite.*

Proof: As in the proof of Theorem 23, fix $0 < \epsilon < 1/4$ and let F be the class of all functions f from $\mathbf{N}$ to $[0,1]$ satisfying

$$f(i) \in \{1/2 + (\epsilon + 1/(i+3)), 1/2 - (\epsilon + 1/(i+3))\}.$$

Clearly, for all $\alpha > 0$, $\mathrm{fat}_F(\epsilon + \alpha)$ is finite.

Choose $d \in \mathbf{N}$. For $b \in \{-1, 1\}^d$, choose f_b such that for each $i \in \{1, .., d\}$, $f(i) = 1/2 + b_i(\epsilon + 1/(i+3))$. Define a distribution P_b over $\{1, ..., d\} \times [0,1]$ by choosing the first component uniformly from $\{1, ..., d\}$, and evaluating f_b at the first component to get the second.

Arguing as in Lemma 25, we can see that for any algorithm A and any $x_1, ..., x_m$, if for all b,

$$h_b = A((x_1, f_b(x_1)), ..., (x_m, f_b(x_m))),$$

if i is not in the sample and b is chosen uniformly at random, then the expectation of A's error is at least $\epsilon + 1/(i+3)$.

Arguing as in Theorem 23, if d is large enough, this expected error is greater than ϵ, whatever the value of $x_1, ..., x_m$. Therefore, there exists a b for which this is true, completing the proof. □

Next, we observe that none of Theorem 26 and its corollaries with regard to fatV or agnostic learning can be improved.

Proposition 28 *There is a function class F that is ϵ-agnostically learnable, but has $\mathrm{fatV}_F(\epsilon)$ infinite.*

Proof: Suppose F is the set of all functions from the natural numbers to $\{1/2 - \epsilon, 1/2, 1/2 + \epsilon\}$. Clearly, $\text{fatV}_F(\epsilon) = \infty$. However, the hypothesis of a constant $1/2$ is always ϵ-close to any $f \in F$, and therefore an algorithm that simply outputs this hypothesis ϵ-agnostically learns F. □

Our results about the relationship between the finiteness of fatV and fat, and the ϵ-uniform GC property and ϵ-agnostic learnability, are summarized in Figure 1.

Acknowledgements

We thank Pankaj Agarwal, David Haussler, Wee Sun Lee, and T.M. Murali for their help, and two anonymous referees for their comments. Peter Bartlett was supported by the Australian Telecommunications and Electronics Research Board. This work was done while Phil Long was at Duke University supported by US Office of Naval Research grant N00014–94–1–0938 and US Air Force Office of Scientific Research grant F49620–92–J0515.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Proceedings of the 1993 IEEE Symposium on Foundations of Computer Science*. IEEE Press, 1993.

[2] M. Anthony and P. Bartlett. Function learning from interpolation. In *Computational Learning Theory: EUROCOLT'95*, 1995.

[3] M. Anthony, N. Biggs, and J. Shawe-Taylor. The learnability of formal concepts. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 246–257, 1990.

[4] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

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

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

[7] K.L. Buescher and P.R. Kumar. Learning stochastic functions by smooth simultaneous estimation. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 272–279. ACM Press, 1992.

[8] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.

[9] L. Gurvits and P. Koiran. Approximation and learning of convex superpositions. In *Computational Learning Theory: EUROCOLT'95*, 1995.

[10] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

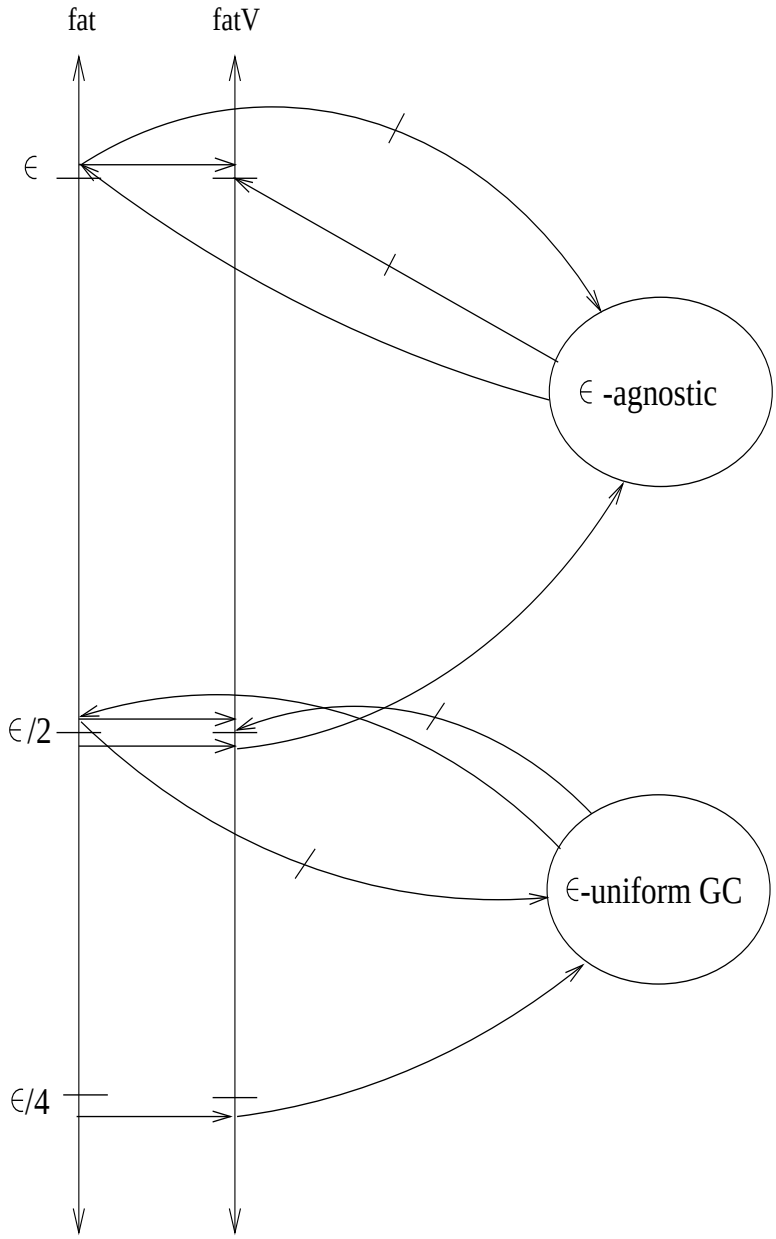

Figure 1: This figure represents the state of our knowledge with regard to the relationship between the finiteness of fat and fatV at certain scales and learnability and uniform convergence. A point on one of the number lines corresponding to fat or fatV at position γ on the line represents the statement "$\mathrm{fat}_F(\gamma)$ (respectively $\mathrm{fatV}_F(\gamma)$) is finite". The ellipses on the right have the obvious interpretation. An arrow indicates an implication, a crossed out arrow indicates that no such implication exists.

[11] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[12] D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth. Equivalence of models for polynomial learnability. *Inform. Comput.*, 95(2):129–161, December 1991.

[13] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.

[14] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[15] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

[16] B.K. Natarajan. Occam's razor for functions. In *Proceedings of the 1993 ACM Conference on Computational Learning Theory*, pages 370–376. ACM Press, 1993.

[17] D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

[18] H. U. Simon. Bounds on the number of examples needed for learning functions. In *Computational Learning Theory: EUROCOLT'93*. Oxford University Press, 1994.