Fat-Shattering and the Learnability of Real-Valued Functions

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

Philip M. Long
Computer Science Department
Duke University, P.O. Box 90129
Durham, NC 27708 USA
plong@cs.duke.edu

Robert C. Williamson
Department of Engineering
Australian National University
Canberra, 0200 Australia
Bob.Williamson@anu.edu.au

Abstract

We consider the problem of learning real-valued functions from random examples when the function values are corrupted with noise. With mild conditions on independent observation noise, we provide characterizations of the learnability of a real-valued function class in terms of a generalization of the Vapnik-Chervonenkis dimension, the fat-shattering function, introduced by Kearns and Schapire. We show that, given some restrictions on the noise, a function class is learnable in our model if and only if its fat-shattering function is finite. With different (also quite mild) restrictions, satisfied for example by gaussian noise, we show that a function class is learnable from polynomially many examples if and only if its fat-shattering function grows polynomially. We prove analogous results in an agnostic setting, where there is no assumption of an underlying function class.

1 INTRODUCTION

In many common definitions of learning, a learner sees a sequence of values of an unknown function at random points, and must, with high probability, choose an accurate approximation to that function. The function is assumed to be a member of some known class. Using a popular definition of the problem of learning $\{0,1\}$-valued functions (probably approximately correct learning — see [9], [22]), Blumer, Ehrenfeucht, Haussler, and Warmuth have shown [9] that the Vapnik-Chervonenkis dimension (see [23]) of a function class characterizes its learnability, in the sense that a function class is learnable if and only if its Vapnik-Chervonenkis dimension is finite. Natarajan [15] and Ben-David, Cesa-Bianchi, Haussler and Long [7] have characterized the learnability of $\{0,...,n\}$-valued functions for fixed n. Alon, Ben-David, Cesa-Bianchi, and Haussler have proved an analogous result for the problem of learning probabilistic concepts [1]. In this case, there is an unknown $[0,1]$-valued function, but the learner does not receive a sequence of values of the function at random points. Instead, with each random point it sees either 0 or 1, with the probability of a 1 given by the value of the unknown function at that point. Kearns and Schapire [12] introduced a generalization of the Vapnik-Chervonenkis dimension, which we call the fat-shattering function, and showed that a class of probabilistic concepts is learnable only if the class has a finite fat-shattering function. The main learning result of [1] is that finiteness of the fat-shattering function of a class of probabilistic concepts is also sufficient for learnability.

In this paper, we consider the learnability of $[0,1]$-valued function classes. We show that a class of $[0,1]$-valued functions is learnable from a finite training sample with observation noise satisfying some mild conditions (the distribution has bounded support and its density satisfies a smoothness constraint) if and only if the class has a finite fat-shattering function. We also consider small-sample learnability, for which the sample size is allowed to grow only polynomially with the required performance parameters. We show that a real-valued function class is learnable from a small sample with observation noise satisfying some other quite mild conditions (the distribution need not have bounded support, but it must have light tails and be symmetric about zero; gaussian noise satisfies these conditions) if and only if the fat-shattering function of the class has a polynomial rate of growth. We also consider agnostic learning [11] [13], in which there is no assumption of an underlying function generating the training examples, and the performance of the learning algorithm is measured by comparison with some function class F. We show that the fat-shattering function of F characterizes finite-sample and small-sample learnability in this case also.

The proof of the lower bound on the number of examples necessary for learning is in two steps. First, we show that the problem of learning real-valued functions in the presence of noise is not much easier than that of learning functions in a discrete-valued function class obtained by quantizing the real-valued function class. This formalizes the intuition that a noisy, real-valued measurement provides little more informa-

tion than a quantized measurement, if the quantization width is sufficiently small. Existing lower bounds on the number of examples required for learning discrete-valued function classes [7], [15] are not strong enough for our purposes. We improve these lower bounds by relating the problem of learning the quantized function class to that of learning $\{0,1\}$-valued functions. The proof of the upper bound departs from the basic outline of proofs of related upper bounds [1], [11] in one key way (see the discussion preceding Lemma 15), and might therefore contribute a useful new technique.

In addition to the aforementioned papers, other general results about learning real-valued functions have been obtained. Haussler [11] gives sufficient conditions for agnostic learnability. Anthony and Shawe-Taylor [3] provide sufficient conditions that a function that approximately interpolates the target function is a good approximation to it. Natarajan [16] considers the problem of learning a class of real-valued functions in the presence of bounded observation noise, and presents sufficient conditions for learnability. Merhav and Feder [14], and Auer, Long, Maass, and Woeginger [4] study function learning in a worst-case setting.

In the next section, we define admissible noise distribution classes and the learning problems, and present the characterizations of learnability. Sections 3 and 4 give lower and upper bounds on the number of examples necessary for learning real-valued functions. Section 5 presents the characterization of agnostic learnability. Section 6 discusses our results.

2 DEFINITIONS AND MAIN RESULT

Denote the integers by $\mathbb{Z}$, the positive integers by $\mathbb{N}$, the reals by $\mathbb{R}$ and the positive reals by $\mathbb{R}^+$. We use log to denote logarithm to base two, and ln to denote the natural logarithm. Fix an arbitrary set X. Throughout the paper, X denotes the input space on which the real-valued functions are defined. We refer to probability distributions on X without explicitly defining a σ-algebra $\mathcal{S}$. For countable X, let $\mathcal{S}$ be the set of all subsets of X. If X is a metric space, let $\mathcal{S}$ be the Borel sets of X. All functions and sets we consider are assumed to be measurable.

2.1 CLASSES OF NOISE DISTRIBUTIONS

The noise distributions we consider are absolutely continuous, and their densities have bounded variation. A function $f : \mathbb{R} \to \mathbb{R}$ is said to have **bounded variation** if there is a constant $C > 0$ such that for every ordered sequence $x_0 < \cdots < x_n$ in $\mathbb{R}$ we have

$$\sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \leq C.$$

In that case, the **total variation** of f on $\mathbb{R}$ is

$$V(f) = \sup \left\{ \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| : x_0 < \cdots < x_n \right\}.$$

Definition 1 *An* **admissible noise distribution class** $\mathcal{D}$ *is a class of distributions on* $\mathbb{R}$ *that satisfies*

1. *Each distribution in* $\mathcal{D}$ *has mean* 0 *and finite variance.*

2. *Each distribution in* $\mathcal{D}$ *is absolutely continuous and its probability density function (pdf) has bounded variation: there is a function* $v : \mathbb{R}^+ \to \mathbb{R}^+$ *such that, if* f *is the pdf of a distribution in* $\mathcal{D}$ *with variance* σ^2*, then* $V(f) \leq v(\sigma)$*. The function* v *is called the* **total variation function** *of the class* $\mathcal{D}$*.*

If $\mathcal{D}$ *also satisfies the following condition, we say it is a* **bounded admissible distribution class***.*

3. *There is a non-decreasing function* $s : \mathbb{R}^+ \to \mathbb{R}^+$ *such that, if* D *is a distribution in* $\mathcal{D}$ *with variance* σ^2*, then the support of* D *is contained in a closed interval of length* $s(\sigma)$*. The function* s *is called the* **support function** *of* $\mathcal{D}$*.*

If $\mathcal{D}$ *satisfies Conditions 1, 2, and the following condition*[1]*, we say it is an* **almost-bounded admissible distribution class***.*

3′. *Each distribution* D *in* $\mathcal{D}$ *has an even pdf* ($f(x) = f(-x)$) *and light tails: there are constants* s_0 *and* c_0 *in* $\mathbb{R}^+$ *such that, for all distributions* D *in* $\mathcal{D}$ *with variance* σ^2*, and all* $s > s_0\sigma$*,*

$$D\{\eta : |\eta| > s/2\} \leq c_0 e^{-s/\sigma}.$$

Example (Uniform noise) Let $\mathcal{U} = \{U_\sigma : \sigma > 0\}$, where U_σ is uniform on $(-\sqrt{3}\sigma, \sqrt{3}\sigma)$. Then this noise has mean 0, standard deviation σ, total variation function $v(\sigma) = 2/\sigma$, and support function $s(\sigma) = 2\sqrt{3}\sigma$, so $\mathcal{U}$ is a bounded admissible distribution class. □

Example (Gaussian noise) Let $\mathcal{G} = \{G_\sigma : \sigma > 0\}$, where G_σ is the zero mean gaussian distribution with variance σ^2. Since the density f_σ of G_σ has $f_\sigma(0) = (\sqrt{2\pi}\sigma)^{-1}$, and $f_\sigma(x)$ is monotonically decreasing for $x > 0$, the total variation function is $v(\sigma) = 2(\sqrt{2\pi}\sigma)^{-1}$. Obviously, f_σ is an even function. Standard bounds on the area under the tails of the gaussian density (see [17], p.64, Fact 3.7.3) give

$$G_\sigma\{\eta \in \mathbb{R} : |\eta| > s/2\} \leq \exp\left(-\frac{s^2}{8\sigma^2}\right), \qquad (1)$$

and if $s > 8\sigma$, $\exp(-s^2/(8\sigma^2)) < \exp(-s/\sigma)$, so the constants $c_0 = 1$ and $s_0 = 8$ will satisfy Condition 3′. So the class $\mathcal{G}$ of gaussian distributions is almost-bounded admissible. □

[1] In fact, Condition 3′ is stronger than we need. It suffices that the distributions be "close to" symmetric and have light tails in the following sense: there are constants s_0 and c_0 in $\mathbb{R}^+$ such that, for all distributions D in $\mathcal{D}$ with variance σ^2, and all $s > s_0\sigma$, if $l \in \mathbb{R}$ satisfies $\int_l^{l+s} xf(x)\,dx = 0$, then

$$\int_l^{l+s} f(x)\,dx \geq 1 - c_0 e^{-s/\sigma},$$

where f is the pdf of D.

2.2 THE LEARNING PROBLEM

Choose a set F of functions from X to $[0,1]$. For $m \in \mathbb{N}$, $f \in F$, $x \in X^m$, and $\eta \in \mathbb{R}^m$, let

$$\text{sam}(x, \eta, f) = ((x_1, f(x_1) + \eta_1), \ldots, (x_m, f(x_m) + \eta_m)).$$

(We often dispense with the parentheses in tuples of this form, to avoid cluttering the notation.) Informally, a **learning algorithm** takes a sample of the above form, and outputs a hypothesis for f. More formally, a **deterministic learning algorithm**[2] is defined to be a mapping from $\cup_m (X \times \mathbb{R})^m$ to $[0,1]^X$. A **randomized learning algorithm** L is a pair (A, P_Z), where P_Z is a distribution on a set Z, and A is a mapping from $\cup_m (X \times \mathbb{R})^m \times Z^m$ to $[0,1]^X$. Given a sample of length m, the randomized algorithm chooses a sequence $z \in Z^m$ at random from P_Z^m, and passes it to the (deterministic) mapping A as a parameter.

For a probability distribution P on X, $f \in F$ and $h : X \to [0,1]$, define

$$\mathbf{er}_{P,f}(h) = \int_X |h(x) - f(x)| dP(x).$$

The following definition of learning is based on those of [9], [15], [22].

Definition 2 *Let $\mathcal{D}$ be a class of distributions on $\mathbb{R}$. Choose $0 < \epsilon, \delta < 1$, $\sigma > 0$, and $m \in \mathbb{N}$. We say a learning algorithm $L = (A, P_Z)$ $(\epsilon, \delta, \sigma)$-**learns** F **from** m **examples with noise** $\mathcal{D}$ if for all distributions P on X, all functions f in F, and all distributions $D \in \mathcal{D}$ with variance σ^2,*

$$P^m \times D^m \times P_Z^m \{(x, \eta, z) : \mathbf{er}_{P,f}(A(\text{sam}(x, \eta, f), z)) \geq \epsilon\} < \delta.$$

*Similarly, L (ϵ, δ)-**learns** F **from** m **examples without noise** if, for all distributions P on X and all functions f in F,*

$$(P^m \times P_Z^m) \{(x, z) : \mathbf{er}_{P,f}(A(\text{sam}(x, 0, f), z)) \geq \epsilon\} < \delta.$$

*We say F is **learnable with noise** $\mathcal{D}$ if there is a learning algorithm L and a function $m_0 : (0,1) \times (0,1) \times \mathbb{R}^+ \to \mathbb{N}$ such that for all $0 < \epsilon, \delta < 1$, for all $\sigma > 0$, algorithm L $(\epsilon, \delta, \sigma)$-learns F from $m_0(\epsilon, \delta, \sigma)$ examples with noise $\mathcal{D}$. We say F is **small-sample learnable with noise** $\mathcal{D}$ if, in addition, the function m_0 is bounded above by a polynomial in $1/\epsilon$, $1/\delta$, and σ.*

The following definition comes from [12]. Choose $x_1, \ldots, x_d \in X$. We say $x_1, \ldots, x_d$ are γ-**shattered by** F if there exists $r \in [0,1]^d$ such that for each $b \in \{0,1\}^d$, there is an $f \in F$ such that for each i

$$f(x_i) \begin{cases} \geq r_i + \gamma & \text{if } b_i = 1 \\ \leq r_i - \gamma & \text{if } b_i = 0. \end{cases}$$

For each γ, let

$$\text{fat}_F(\gamma) = \max\{d \in \mathbb{N} : F \text{ } \gamma\text{-shatters some } x_1, \ldots, x_d\}$$

if such a maximum exists, and ∞ otherwise. If $\text{fat}_F(\gamma)$ is finite for all γ, we say F has a **finite fat-shattering function.**

The following is our main result.

[2]Despite the name "algorithm," there is no requirement that this mapping be computable. Throughout the paper, we ignore issues of computability.

Theorem 3 *Suppose F is a permissible[3] class of $[0,1]$-valued functions defined on X.*

If $\mathcal{D}$ is a bounded admissible distribution class, then F is learnable with observation noise $\mathcal{D}$ if and only if F has a finite fat-shattering function.

If $\mathcal{D}$ is an almost-bounded admissible distribution class, then F is small-sample learnable with observation noise $\mathcal{D}$ if and only if there is a polynomial p that satisfies $\text{fat}_F(\gamma) < p(1/\gamma)$ for all $\gamma > 0$.

3 LOWER BOUND

In this section, we give a lower bound on the number of examples necessary to learn a real-valued function class in the presence of observation noise. Lemma 5 in Section 3.1 shows that an algorithm that can learn a real-valued function class with observation noise can be used to construct an algorithm that can learn a quantized version of the function class to slightly worse accuracy and confidence with the same number of examples, provided the quantization width is sufficiently small. Lemma 10 in Section 3.2 gives a lower bound on the number of examples necessary for learning a quantized function class in terms of its fat-shattering function. In Section 3.3, we combine these results to give the lower bound for real-valued functions, Theorem 11.

3.1 LEARNABILITY WITH NOISE IMPLIES QUANTIZED LEARNABILITY

In this section, we relate the problem of learning a real-valued function class with observation noise to the problem of learning a quantized version of that class, without noise.

Definition 4 *For $\alpha \in \mathbb{R}^+$, define the quantization function*

$$Q_\alpha(y) = \alpha \left\lceil \frac{y - \alpha/2}{\alpha} \right\rceil.$$

For a set $S \subset \mathbb{R}$, let $Q_\alpha(S) = \{Q_\alpha(y) : y \in S\}$. For a function class $F \subset [0,1]^X$, let $Q_\alpha(F)$ be the set $\{Q_\alpha \circ f : f \in F\}$ of $Q_\alpha([0,1])$-valued functions defined on X.

Lemma 5 *Suppose F is a set of functions from X to $[0,1]$, $\mathcal{D}$ is an admissible noise distribution class with total variation function v, A is a learning algorithm, $0 < \epsilon, \delta < 1$, $\sigma \in \mathbb{R}^+$, $m \in \mathbb{N}$. If the quantization width $\alpha \in \mathbb{R}^+$ satisfies*

$$\alpha \leq \min\left(\frac{\delta}{v(\sigma)m}, 2\epsilon\right),$$

and A $(\epsilon, \delta, \sigma)$-learns F from m examples with noise $\mathcal{D}$, then there is a randomized learning algorithm (C, P_Z) that $(2\epsilon, 2\delta)$-learns $Q_\alpha(F)$ from m examples.

Figure 1 illustrates our approach. Suppose an algorithm A can $(\epsilon, \delta, \sigma)$-learn from m noisy examples. If we quantize the observations to accuracy α and add noise that is uniform on $(-\alpha/2, \alpha/2)$, Lemma 6(a) shows that the distribution of

[3]This is a benign measurability constraint defined in Section 4.

the observations is approximately unchanged (in the notation of Figure 1, the distributions P_1 and P_2 are close), so A learns almost as well as it did previously. If we define Algorithm B as this operation of adding uniform noise and then invoking Algorithm A, B solves a certain quantized learning problem. Lemma 6(b) shows that this problem is similar to the problem of learning the quantized function class when the observations are contaminated with independent noise whose distribution is a quantized version of the original observation noise (that is, distributions P_3 and P_4 in Figure 1 are close). It follows that Algorithm C, which adds this quantized noise to the observations and passes them to Algorithm B, learns the quantized function class without observation noise.

For distributions P and Q on $\mathbb{R}$, define the total variation distance between P and Q as

$$d_{TV}(P,Q) = 2 \sup_E |P(E) - Q(E)|$$

where the supremum is over all Borel sets. If P and Q are discrete, it is easy to show that

$$d_{TV}(P,Q) = \sum_x |P(x) - Q(x)|,$$

where the sum is over all x in the union of the supports of P and Q. Similarly, if P and Q are continuous with probability density functions p and q respectively,

$$d_{TV}(P,Q) = \int_{-\infty}^{\infty} |p(x) - q(x)|\, dx.$$

Lemma 6 *Let $\mathcal{D}$ be an admissible noise distribution class with total variation function v. Let $\sigma > 0$ and $0 < \alpha < 1$. Let D be a distribution in $\mathcal{D}$ with variance σ^2. Let η, ζ, and ν be random variables, and suppose that η and ν are distributed according to D, and ζ is distributed uniformly on $(-\alpha/2, \alpha/2)$.*
(a) *For any $y \in [0,1]$, if P_1 is the distribution of $y + \eta$ and P_2 is the distribution of $Q_\alpha(y+\eta) + \zeta$, we have*

$$d_{TV}(P_1, P_2) \le \alpha v(\sigma).$$

(b) *For any $y \in [0,1]$, if P_3 is the distribution of $Q_\alpha(y+\eta)$ and P_4 is the distribution of $Q_\alpha(y) + Q_\alpha(\nu)$, we have*

$$d_{TV}(P_3, P_4) \le \alpha v(\sigma).$$

Proof Let p be the pdf of D.
(a) The random variable $y+\eta$ has density $p_1(a) = p(a-y)$, and $Q_\alpha(y+\eta)+\zeta$ has density p_2 given by

$$p_2(a) = \frac{1}{\alpha} \int_{Q_\alpha(a)-\alpha/2}^{Q_\alpha(a)+\alpha/2} p(x-y)\, dx$$

for $a \in \mathbb{R}$. So

$$\begin{aligned} &d_{TV}(P_1, P_2) \\ &\quad = \int_{-\infty}^{\infty} \left| p(x-y) - \tfrac{1}{\alpha} \int_{Q_\alpha(x)-\alpha/2}^{Q_\alpha(x)+\alpha/2} p(\theta - y)\, d\theta \right| dx \\ &\quad = \int_{-\alpha/2}^{\alpha/2} \sum_{n=-\infty}^{\infty} \Big| p(x-y+n\alpha) \\ &\qquad\qquad - \tfrac{1}{\alpha} \int_{-\alpha/2}^{\alpha/2} p(\theta - y + n\alpha)\, d\theta \Big|\, dx. \end{aligned}$$

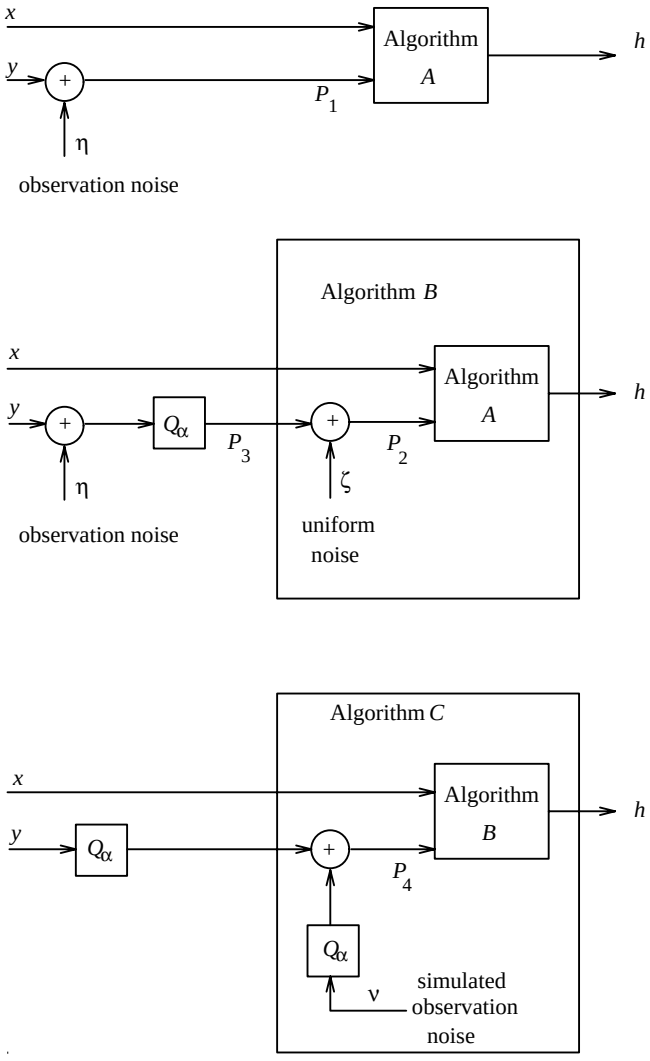

Figure 1: Lemma 5 shows that a learning algorithm for real-valued functions (Algorithm A) can be used to construct a randomized learning algorithm for quantized functions (Algorithm C).

By the mean value theorem, there are z_1 and z_2 in $[-\alpha/2, \alpha/2]$ such that

$$p(z_1 - y + n\alpha) \le \frac{1}{\alpha} \int_{-\alpha/2}^{\alpha/2} p(\theta - y + n\alpha)\, d\theta \le p(z_2 - y + n\alpha),$$

so for all $x \in [-\alpha/2, \alpha/2]$,

$$\begin{aligned} &\textstyle\sum_{n=-\infty}^{\infty} \left| p(x-y+n\alpha) - \frac{1}{\alpha} \int_{-\alpha/2}^{\alpha/2} p(\theta - y + n\alpha)\, d\theta \right| \\ &\quad \le \textstyle\sum_{n=-\infty}^{\infty} \sup_{z \in (-\alpha/2, \alpha/2)} |p(x-y+n\alpha) \\ &\qquad\qquad - p(z-y+n\alpha)| \\ &\quad \le v(\sigma), \end{aligned}$$

and therefore

$$d_{TV}(P_1, P_2) \le \alpha v(\sigma).$$

(b) The distribution P_3 of $Q_\alpha(y+\eta)$ is discrete, and satisfies

$$P_3(a) = \int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p(x-y)\, dx$$

if $a = n\alpha$ for some $n \in \mathbb{Z}$, and $P_3(a) = 0$ otherwise. Since ν has distribution D, the distribution P_4 of the random variable $Q_\alpha(y) + Q_\alpha(\nu)$ is also discrete, and satisfies

$$P_4(a) = \int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p(x)\,dx$$

if $a = n\alpha + Q_\alpha(y)$ for some $n \in \mathbb{Z}$, and $P_4(a) = 0$ otherwise. So

$$\begin{aligned}
&d_{TV}(P_3, P_4) \\
&\quad = \textstyle\sum_{n=-\infty}^{\infty} |P_3(n\alpha) - P_4(n\alpha)| \\
&\quad = \textstyle\sum_{n=-\infty}^{\infty} \Big| \int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p(x-y)\,dx \\
&\qquad\qquad - \int_{n\alpha-\alpha/2}^{n\alpha+\alpha/2} p\left(x - Q_\alpha(y)\right)\,dx \Big| \\
&\quad \le \textstyle\int_{-\alpha/2}^{\alpha/2} \sum_{n=-\infty}^{\infty} |p(x-y+n\alpha) \\
&\qquad\qquad - p\left(x - Q_\alpha(y) + n\alpha\right)|\,dx \\
&\quad \le \alpha v(\sigma).
\end{aligned}$$

□

We will use the following lemma. The proof is by induction, and is implicit in the proof of Lemma 12 in [6].

Lemma 7 *If P_i and Q_i are distributions on a set Y $(i = 1, \ldots, m)$, and E is a measurable subset of Y^m, then*

$$\left| \left(\prod_{i=1}^m P_i\right)(E) - \left(\prod_{i=1}^m Q_i\right)(E) \right| \le \tfrac{1}{2} \textstyle\sum_{i=1}^m d_{TV}(P_i, Q_i).$$

Proof (of Lemma 5) We will describe a randomized algorithm (Algorithm C) that is constructed from Algorithm A, and show that it $(2\epsilon, 2\delta)$-learns the quantized function class $Q_\alpha(F)$. Fix a noise distribution D in $\mathcal{D}$ with variance σ^2, a function $f \in F$, and a distribution P on X. Since A $(\epsilon, \delta, \sigma)$-learns F, we have

$$P^m \times D^m \left\{(x, \eta) : \mathbf{er}_{P,f}(A(\mathrm{sam}(x, \eta, f))) \ge \epsilon\right\} < \delta.$$

That is, the probability (over all $x \in X^m$ and $\eta \in \mathbb{R}^m$) that Algorithm A chooses a bad function is small. We will show that this implies that the probability that Algorithm C chooses a bad function is also small, where the probability is over all $x \in X^m$ and all values of the random variables that Algorithm C uses.

Now, fix a sequence $x = (x_1, \ldots, x_m) \in X^m$, and define the events

$$\begin{aligned}
E &= \left\{\eta \in \mathbb{R}^m : \mathbf{er}_{P,f}(A(\mathrm{sam}(x, \eta, f))) \ge \epsilon\right\}, \\
E_1 &= \left\{y \in \mathbb{R}^m : \mathbf{er}_{P,f}(A(x_1, y_1, \ldots, x_m, y_m)) \ge \epsilon\right\}.
\end{aligned}$$

That is, E is the set of noise sequences that make A choose a bad function, and E_1 is the corresponding set of y sequences. Clearly,

$$D^m(E) = \left(\prod_{i=1}^m P_{1|x_i}\right)(E_1),$$

where $P_{1|x_i}$ is the distribution of $f(x_i) + \eta$ (see Figure 1).

Let ζ be a random variable with distribution U_α, where U_α is the uniform distribution on $(-\alpha/2, \alpha/2)$. Let Algorithm B be the randomized algorithm that adds noise ζ to each y value it receives, and passes the sequence to Algorithm A. That is,

$$B(x_1, y_1, \ldots, x_m, y_m) = A(x_1, y_1 + \zeta_1, \ldots, x_m, y_m + \zeta_m).$$

Let $P_{2|x_i}$ be the distribution of $Q_\alpha(f(x_i) + \eta) + \zeta$ (see Figure 1). From Lemma 6(a), $d_{TV}(P_{1|x_i}, P_{2|x_i}) \le \alpha v(\sigma)$. Lemma 7 implies

$$\left(\prod_{i=1}^m P_{2|x_i}\right)(E_1) \le D^m(E) + \frac{m\alpha v(\sigma)}{2} \le D^m(E) + \frac{\delta}{2}$$

where the second inequality follows from the hypothesis that $\alpha \le \delta/(mv(\sigma))$.

Let $P_{3|x_i}$ be the distribution of $Q_\alpha(f(x_i) + \eta)$ (see Figure 1), and let

$$E_3 = \left\{y \in \mathbb{R}^m : \mathrm{E}\left(\mathbf{er}_{P,f}(B(x_1, y_1, \ldots, x_m, y_m))\right) \ge \epsilon\right\},$$

where the expectation is over all values of ζ, the uniform noise that B introduces. In this case, E_3 is the set of y sequences that make B choose a bad function. Clearly,

$$\left(\prod_{i=1}^m P_{3|x_i}\right)(E_3) = \left(\prod_{i=1}^m P_{2|x_i}\right)(E_1).$$

Let ν be a random variable with distribution D. Let Algorithm C be the randomized algorithm that adds noise $Q_\alpha(\nu)$ to each y value it receives, and passes the sequence to Algorithm B. That is,

$$\begin{aligned}
&C(x_1, y_1, \ldots, x_m, y_m) = \\
&\quad B\left(x_1, y_1 + Q_\alpha(\nu_1), \ldots, x_m, y_m + Q_\alpha(\nu_m)\right).
\end{aligned}$$

Let $P_{4|x_i}$ be the distribution of $Q_\alpha(f(x_i)) + Q_\alpha(\nu)$ (see Figure 1). From Lemma 6(b), $d_{TV}(P_{4|x_i}, P_{3|x_i}) \le \alpha v(\sigma)$. Lemma 7 implies

$$\begin{aligned}
\left(\prod_{i=1}^m P_{4|x_i}\right)(E_3) &\le \left(\prod_{i=1}^m P_{3|x_i}\right)(E_3) + \frac{m\alpha v(\sigma)}{2} \\
&\le D^m(E) + \delta.
\end{aligned}$$

It follows that the probability under $P^m \times U_\alpha^m \times D^m$ that x, ζ, and ν satisfy

$$\begin{aligned}
&\mathbf{er}_{P,f}\left(A(x_1, Q_\alpha(f(x_1)) + Q_\alpha(\nu_1) + \zeta_1, \right. \\
&\qquad \left. \ldots, x_m, Q_\alpha(f(x_m)) + Q_\alpha(\nu_m) + \zeta_m)\right) \\
&\quad = \mathbf{er}_{P,f}\left(C\left(x_1, Q_\alpha(f(x_1)), \ldots, x_m, Q_\alpha(f(x_m))\right)\right) \\
&\quad \ge \epsilon
\end{aligned} \tag{2}$$

is less than 2δ.

Since $\alpha \le 2\epsilon$, for all $x \in X$, $|f(x) - Q_\alpha(f(x))| \le \epsilon$, and therefore (2) implies

$$\mathbf{er}_{P,Q_\alpha(f)}\left(C\left(x_1, Q_\alpha(f(x_1)), \ldots, x_m, Q_\alpha(f(x_m))\right)\right) < 2\epsilon$$

with probability at least $1 - 2\delta$. This is true for any $Q_\alpha(f)$ in $Q_\alpha(F)$, so this algorithm $(2\epsilon, 2\delta)$-learns $Q_\alpha(F)$ from m examples. □

3.2 LOWER BOUNDS FOR QUANTIZED LEARNING

In the last section, we showed that if a class F can be $(\epsilon, \delta, \sigma)$-learned with a certain number of examples, then an associated class $Q_\alpha(F)$ of discrete-valued functions can be $(2\epsilon, 2\delta)$-learned with the same number of examples. Given this result, one would be tempted to apply techniques of Natarajan [15] or Ben-David, Cesa-Bianchi, Haussler, and Long [7] (who consider the learnability of discrete-valued functions) to bound from below the number of examples required for learning $Q_\alpha(F)$. The main results of those papers, however, were for the discrete loss function, where the learner "loses" 1 whenever its hypothesis is incorrect. When those results are applied directly to get bounds for learning with the absolute loss, the resulting bounds are not strong enough for our purposes because of the restrictions on α required to show that learning F is not much harder than learning $Q_\alpha(F)$.

In this section, we present a new technique, inspired by the techniques of [5]. We show that an algorithm for learning a class of discrete-valued functions can effectively be used as a subroutine in an algorithm for learning binary-valued functions. We then apply a lower bound result for binary-valued functions.

For each $d \in \mathbb{N}$, let POWER$_d$ be the set of all functions from $\{1, \ldots, d\}$ to $\{0, 1\}$. We will make use of the following special case of a general result about POWER$_d$ ([9], Theorem 2.1b).

Theorem 8 ([9]) *Let A be a randomized learning algorithm which always outputs $\{0,1\}$-valued hypotheses. If A is given fewer than $d/2$ examples, A fails to $(1/8, 1/8)$-learn* POWER$_d$.

Theorem 2.1b of [9] is stated for deterministic algorithms, but an almost identical proof gives the same result for randomized algorithms.

We will make use of the following lemma, which is implicit in the results of Benedek and Itai.

Theorem 9 ([8]) *Choose X, a probability distribution P on X, and $f \in \{0,1\}^X$. If* **(a)** $h_1, ..., h_r \in \{0,1\}^X$ *are such that there exists i for which* $\mathbf{er}_{P,f}(h_i) \leq 1/32$, **(b)** $m = \lceil 96 \ln(8r) \rceil$, **(c)** $x_1, ..., x_m$ *are drawn independently at random according to P, and* **(d)** $h \in \{h_1, ..., h_r\}$ *satisfies*

$$\forall 1 \leq j \leq r, |\{1 \leq i \leq m : h(x_i) \neq f(x_i)\}| \leq |\{1 \leq i \leq m : h_j(x_i) \neq f(x_i)\}|,$$

then $\Pr(\mathbf{er}_{P,f}(h) \geq 1/8) \leq 1/16$.

Lemma 10 *Choose a set F of functions from X to $Q_\alpha([0,1])$, $d \in \mathbb{N}$ and $\gamma > 0$ such that* $\mathrm{fat}_F(\gamma) \geq d$. *If a randomized learning algorithm A is given fewer than*

$$\frac{d - 400}{4 + 192 \ln \lceil 1/\alpha \rceil}$$

examples, A fails to $(\gamma/32, 1/16)$-learn F without noise.

Proof Let $x_1, \ldots, x_d \in X$ be γ-shattered by F, and let $r_1, \ldots, r_d \in [0,1]^d$ be such that for each $b \in \{0,1\}^d$, there is an $f_b \in F$ such that for all $j, 1 \leq j \leq d$,

$$f_b(x_j) \begin{cases} \geq r_j + \gamma & \text{if } b_j = 1 \\ \leq r_j - \gamma & \text{if } b_j = 0. \end{cases}$$

Choose an algorithm A for learning F. For each $q \in \mathbb{N}$, consider the algorithm $\tilde{A}_q$ (which will be used for learning POWER$_d$) which uses A as a subroutine as follows. Given $m > q$ examples, $(\kappa_1, y_1), \ldots, (\kappa_m, y_m)$ in $\{1, \ldots, d\} \times \{0,1\}$, Algorithm $\tilde{A}_q$ first, for each $v \in \{0, \ldots, \lceil 1/\alpha \rceil - 1\}^q$, sets $h_{\kappa,v} = A((x_{\kappa_1}, v_1), \ldots, (x_{\kappa_q}, v_q))$. Algorithm $\tilde{A}_q$ then uses this to define a set $\tilde{S}$ of $\{0,1\}$-valued functions defined on $\{1, \ldots, d\}$ by

$$\tilde{S} = \left\{ \tilde{h}_{\kappa,v} : v \in \{0, \ldots, \lceil 1/\alpha \rceil - 1\}^q \right\},$$

where

$$\tilde{h}_{\kappa,v}(j) = \begin{cases} 1 & \text{if } h_{\kappa,v}(x_j) \geq r_j \\ 0 & \text{otherwise,} \end{cases}$$

for all $j \in \{1, \ldots, d\}$. Finally, $\tilde{A}_q$ returns an $\tilde{h}^*$ in $\tilde{S}$ for which the number of disagreements with the last $m - q$ examples is minimized. That is,

$$|\{j \in \{q+1, \ldots, m\} : \tilde{h}^*(\kappa_j) \neq y_j\}| = \min_{\tilde{h} \in \tilde{S}} \left\{ |\{j \in \{q+1, \ldots, m\} : \tilde{h}(\kappa_j) \neq y_j\}| \right\}.$$

We claim that if A $(\gamma/32, 1/16)$-learns F from m_0 examples without noise, then $\tilde{A}_{m_0}$ $(1/8, 1/8)$-learns POWER$_d$ from

$$m_0 + \lceil 96 \left(\ln 8 + m_0 \ln \lceil 1/\alpha \rceil \right) \rceil$$

examples without noise. Assume A $(\gamma/32, 1/16)$-learns F from m_0 examples, and let $\tilde{A} = \tilde{A}_{m_0}$. Suppose $\tilde{A}$ is trying to learn $g \in$ POWER$_d$ and the distribution on the domain $\{1, \ldots, d\}$ is $\tilde{P}$. Let P be the corresponding distribution on $\{x_1, \ldots, x_d\}$, and let $b = (g(1), \ldots, g(d)) \in \{0,1\}^d$. Since A $(\gamma/32, 1/16)$-learns F, we have

$$\tilde{P}^{m_0} \{(\kappa_1, \ldots, \kappa_{m_0}) : \mathbf{er}_{P,f_b} \left(A \left(\mathrm{sam}((x_{\kappa_1}, ..., x_{\kappa_{m_0}}), 0, f_b) \right) \right) \geq \gamma/32 \} < 1/16,$$

which implies

$$\tilde{P}^{m_0} \{\kappa : \forall v \in \{0, \ldots, \lceil 1/\alpha \rceil - 1\}^{m_0}, \mathbf{er}_{P,f_b}(h_{\kappa,v}) \geq \gamma/32\} < 1/16.$$

This can be rewritten as

$$\tilde{P}^{m_0} \left\{ \kappa : \forall v, \int |h_{\kappa,v}(x_j) - f_b(x_j)| \, d\tilde{P}(j) \geq \frac{\gamma}{32} \right\} < \frac{1}{16},$$

which, applying Markov's inequality, yields

$$\tilde{P}^{m_0} \left\{ \kappa : \forall v, \tilde{P} \{j : |h_{\kappa,v}(x_j) - f_b(x_j)| \geq \gamma\} \geq \tfrac{1}{32} \right\} < \tfrac{1}{16}. \quad (3)$$

Now, for all j, $|f_b(x_j) - r_j| \geq \gamma$, so if $|\tilde{h}_{\kappa,v}(j) - b_j| = 1$ the definitions of $\tilde{h}_{\kappa,v}$ and f_b imply $|h_{\kappa,v}(x_j) - f_b(x_j)| \geq \gamma$. Therefore $\mathbf{er}_{\tilde{P},g}(\tilde{h}_{\kappa,v}) \geq 1/32$ implies

$$\tilde{P} \{j : |h_{\kappa,v}(x_j) - f_b(x_j)| \geq \gamma\} \geq 1/32,$$

so (3) implies

$$\tilde{P}^{m_0}\left\{\kappa : \forall v, \mathbf{er}_{\check{P},g}(\tilde{h}_{\kappa,v}) \geq 1/32\right\} < 1/16. \qquad (4)$$

That is, $\tilde{A}$ is unlikely to choose $\tilde{S}$ so that all elements have large error.

Let E be the event that some hypothesis in $\tilde{S}$ has error below $1/32$,

$$E = \left\{(\kappa,\lambda) \in \{1,\ldots,d\}^{m_0+u} : \exists v, \mathbf{er}_{\check{P},g}(\tilde{h}_{\kappa,v}) < \frac{1}{32}\right\}.$$

(Notice that this event is independent of the examples $\lambda \in \{1,\ldots,d\}^u$ that are used to assess the functions in $\tilde{S}$.) For $\kappa \in \{1,\ldots,d\}^{m_0}$ and $\lambda \in \{1,\ldots,d\}^u$, let $\tilde{A}_{\kappa,\lambda,g}$ denote

$$\tilde{A}\left(\kappa_1, g(\kappa_1), \ldots, \kappa_{m_0}, g(\kappa_{m_0}), \lambda_1, g(\lambda_1), \ldots, \lambda_u, g(\lambda_u)\right).$$

Then Theorem 9 implies that

$$\Pr\left(\mathbf{er}_{\check{P},g}\left(\tilde{A}_{\kappa,\lambda,g}\right) > 1/8 \middle| E\right) < 1/16, \qquad (5)$$

where the probability is taken over all values of κ and λ conditioned on $(\kappa,\lambda) \in E$. But (4), which shows that $\Pr(\text{not } E) < 1/16$, and (5) imply

$$\begin{aligned}\Pr&\left(\mathbf{er}_{\check{P},g}\left(\tilde{A}_{\kappa,\lambda,g}\right) > 1/8\right)\\ &\leq \Pr\left(\mathbf{er}_{\check{P},g}\left(\tilde{A}_{\kappa,\lambda,g}\right) > 1/8 \middle| E\right) + \Pr(\text{not } E)\\ &< 1/8.\end{aligned}$$

That is, $\tilde{A}$ $(1/8, 1/8)$-learns POWER_d using

$$m_0 + \lceil 96(\ln 8 + m_0 \ln\lceil 1/\alpha\rceil)\rceil$$

examples, as claimed. Applying Theorem 8 completes the proof. □

3.3 THE LOWER BOUND

We can combine Lemmas 5 and 10 to prove the following lower bound on the number of examples necessary for learning with observation noise. Obviously the constants have not been optimized.

Theorem 11 *Suppose F is a set of $[0,1]$-valued functions defined on X, $\mathcal{D}$ is an admissible noise distribution class with total variation function v, $0 < \gamma < 1$, $0 < \epsilon \leq \gamma/65$, $0 < \delta \leq 1/32$, $\sigma \in \mathbb{R}^+$, and $d \in \mathbb{N}$. If $\text{fat}_F(\gamma) \geq d > 800$, then any algorithm that (ϵ,δ,σ)-learns F with noise $\mathcal{D}$ requires at least m_0 examples, where*

$$m_0 > \min\left\{\frac{d}{800\ln(2+dv(\sigma)/10)}, \frac{d}{800\ln(2+d/120)}, \frac{d}{400\ln(40/\gamma)}\right\}. \qquad (6)$$

In particular, if $v(\sigma) > \max\left(1/12, 64/(d\gamma^{1/2})\right)$, then

$$m_0 > \frac{d}{800\ln(2+dv(\sigma)/10)}.$$

This theorem shows that if there is a $\gamma > 0$ such that $\text{fat}_F(\gamma)$ is infinite then we can choose ϵ, δ, and σ for which (ϵ,δ,σ)-learning is impossible from a finite sample. Similarly, if $\text{fat}_F(\gamma)$ grows faster than polynomially in $1/\gamma$, we can fix σ and Theorem 11 implies that the number of examples necessary for learning must grow faster than polynomially in $1/\epsilon$. This proves the "only if" parts of the characterization theorem (Theorem 3).

4 UPPER BOUND

In this section, we prove an upper bound on the number of examples required for learning with observation noise, finishing the proof of Theorem 3.

For $n \in \mathbb{N}, v, w \in \mathbb{R}^n$, let $d(v,w) = \frac{1}{n}\sum_{i=1}^n |v_i - w_i|$. For $U \subseteq \mathbb{R}^n$, $\epsilon > 0$, we say $C \subseteq \mathbb{R}^n$ is an **ϵ-cover** of U if and only if for all $v \in U$, there exists $w \in C$ such that $d(v,w) \leq \epsilon$, and we denote by $\mathcal{N}(\epsilon, U)$ the size of the smallest ϵ-cover of U (the **ϵ-covering number of** U).

For a function $f : X \to [0,1]$, define $\ell_f : X \times \mathbb{R} \to \mathbb{R}$ by $\ell_f(x,y) = (f(x)-y)^2$, and if $F \subseteq [0,1]^X$, let $\ell_F = \{\ell_f : f \in F\}$.

If W is a set, $f : W \to \mathbb{R}$, and $w \in W^m$, let $f_{|_w} \in \mathbb{R}^m$ denote $(f(w_1),\ldots,f(w_m))$. Finally, if F is a set of functions from W to $\mathbb{R}$, let $F_{|_w} \subseteq \mathbb{R}^m$ be defined by $F_{|_w} = \{f_{|_w} : f \in F\}$.

The following theorem is due to Haussler [11] (Theorem 3, p107); it is an improvement of a result of Pollard [18]. We say a function class is **PH-permissible** if it satisfies the mild measurability condition defined in Haussler's Section 9.2 [11]. We say a class F of real-valued functions is **permissible** if the class ℓ_F is PH-permissible. Notice that this implies that the class $\ell_F^a = \{(x,y) \mapsto |f(x)-y| : f \in F\}$ is PH-permissible, since the square root function on $\mathbb{R}^+$ is measurable.

Theorem 12 ([11]) *Let Y be a set and G a PH-permissible class of $[0,M]$-valued functions defined on $Z = X \times Y$, where $M \in \mathbb{R}^+$. For any $\alpha > 0$ and any distribution P on Z,*

$$\begin{aligned}P^m&\left\{z \in Z^m : \exists g \in G, \left|\frac{1}{m}\sum_{i=1}^m g(z_i) - \int_Z g\,dP\right| > \alpha\right\}\\ &\leq 4 \max_{z \in Z^{2m}} \left(\mathcal{N}\left(\alpha/16, G_{|_z}\right)\right) e^{-\alpha^2 m/(64M^2)}.\end{aligned}$$

Corollary 13 *Let F be a permissible class of $[0,1]$-valued functions defined on X. Let $Y = [a,b]$ with $a \leq 0$ and $b \geq 1$, and let $Z = X \times Y$. There is a mapping B from $(0,1) \times \cup_i Z^i$ to $[0,1]^X$ such that, for any $0 < \epsilon < 1$ and any distribution P on Z,*

$$\begin{aligned}P^m&\left\{z \in Z^m : \int_Z \ell_{B(\epsilon,z)}\,dP \geq \inf_{f \in F}\int_Z \ell_f\,dP + \epsilon\right\}\\ &\leq 4 \max_{z \in Z^{2m}} \left(\mathcal{N}\left(\epsilon/48, (\ell_F)_{|_z}\right)\right) e^{-\epsilon^2 m/(576(b-a)^4)}.\end{aligned}$$

The proof is similar to the proof of Haussler's Lemma 1 [11], but the mapping B selects a function f^* from F that satisfies

$$\frac{1}{m}\sum_{i=1}^m \ell_{f^*}(z_i) < \inf_{f \in F}\frac{1}{m}\sum_{i=1}^m \ell_f(z_i) + \frac{\epsilon}{3}.$$

The following result follows from Alon, Ben-David, Cesa-Bianchi and Haussler's Lemmas 14 and 15 [1].

Theorem 14 *If F is a class of $[0,1]$-valued functions defined on X, $0 < \epsilon < 1/2$ and $m \geq \text{fat}_F(\epsilon/4)/2$, then for all x in*

X^m

$$\mathcal{N}(\epsilon, F_{|_x}) \leq \exp\left(\frac{2}{\ln 2}\mathrm{fat}_F(\epsilon/4)\ln^2\frac{9m}{\epsilon^2}\right).$$

Alon *et al.* [1] showed that $\mathrm{fat}_{\tilde{\ell}_F}(\gamma) \leq \mathrm{fat}_F(\gamma/2)$ (where $\tilde{\ell}_F$ is obtained by restricting the second arguments of the functions in ℓ_F to $\{0,1\}$) and used Theorem 14 with the function class $\tilde{\ell}_F$ to get upper bounds for probabilistic concepts. We have been unable to demonstrate a similar relationship between fat_{ℓ_F} and fat_F. Instead, we push the problem down a level, bounding the covering numbers related to ℓ_F directly in terms of those for F.

Lemma 15 *Choose a set F of functions from X to $[0,1]$. Then for any $\epsilon > 0$, for any $m \in \mathbb{N}$, if $a \leq 0$ and $b \geq 1$,*

$$\max_{z \in (X\times[a,b])^m} \mathcal{N}(\epsilon, (\ell_F)_{|_z}) \leq \max_{x\in X^m} \mathcal{N}\left(\frac{\epsilon}{3(b-a)}, F_{|_x}\right).$$

Proof Choose $(x_1,y_1),\ldots,(x_m,y_m) \in X\times[a,b]$, and $f, g : X \to [0,1]$. We have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m |(g(x_i)-y_i)^2 - (f(x_i)-y_i)^2| \\
&= \frac{1}{m}\sum_{i=1}^m |(f(x_i)-g(x_i))^2 - 2(f(x_i)-g(x_i))(g(x_i)-y_i)| \\
&\leq \frac{1}{m}\sum_{i=1}^m ((f(x_i)-g(x_i))^2 + 2|f(x_i)-g(x_i)||g(x_i)-y_i|) \\
&\leq \frac{1}{m}\sum_{i=1}^m 3(b-a)|f(x_i)-g(x_i)|.
\end{aligned}$$

It follows that we can construct an ϵ-cover of $(\ell_F)_{|_z}$ from an $\epsilon/(3(b-a))$-cover of $F_{|_x}$. □

In our proof of upper bounds on the number of examples needed for learning, we will make use of the following lemma.

Lemma 16 *For any $y_1, y_2, y_4, \delta > 0$ and $y_3 \geq 1$, if*

$$m \geq \frac{2}{y_4}\left(4y_2\left(4+\ln\left(\frac{y_2y_3}{y_4}\right)\right)^2 + \ln\frac{y_1}{\delta}\right),$$

then $y_1\exp(y_2\ln^2(y_3m) - y_4m) \leq \delta$.

The proof uses the fact [20] that for all $a, b > 0, \ln a \leq ab + \ln(1/b)$ in a manner similar to [20].

We can now present the upper bound. Again, the constants have not been optimized.

Theorem 17 *For any permissible class F of functions from X to $[0,1]$, there is a learning algorithm A such that, for all bounded admissible distribution classes $\mathcal{D}$ with support function s, for all probability distributions P on X, and for all $0 < \epsilon < 1/2$, $0 < \delta < 1$, and $\sigma > 0$, if $d = \mathrm{fat}_F\left(\epsilon^2/(576(s(\sigma)+1))\right)$, then A (ϵ,δ,σ)-learns F from*

$$\frac{1152(1+s(\sigma))^4}{\epsilon^4}\left(12d\left(25+\ln\frac{d(1+s(\sigma))^6}{\epsilon^8}\right)^2 + \ln\frac{4}{\delta}\right)$$

examples with noise $\mathcal{D}$.

Proof Let B be the mapping from Corollary 13. Choose $0 < \epsilon < 1/2$, $0 < \delta < 1$, and $\sigma > 0$. Let $\epsilon_0 = \epsilon^2$. Let D be a distribution in $\mathcal{D}$ with variance σ^2 and support contained in $[c,d]$, so $d - c \leq s(\sigma)$. Choose a distribution P on X and a function $f \in F$.

For $x \in X^m$ and $\eta \in [c,d]^m$, let $B_{x,\eta} = B(\epsilon_0, \mathrm{sam}(x,\eta,f))$. Define the event BAD to be

$$\left\{(x,\eta) \in \left(X^m \times [c,d]^m\right) : \int_X\int_{[c,d]} \left[B_{x,\eta}(u) - (f(u)+\kappa)\right]^2 dD(\kappa)dP(u) \geq \sigma^2 + \epsilon_0\right\}.$$

Since D has variance σ^2 and mean 0,

$$\inf_{g\in F}\int_X\int_{[c,d]} (g(u) - (f(u)+\kappa))^2\, dD(\kappa)dP(u) = \sigma^2,$$

so BAD consists of all those (x,η) such that

$$\begin{aligned}
&\textstyle\int_X\int_{[c,d]} \left[B_{x,\eta}(u) - (f(u)+\kappa)\right]^2 dD(\kappa)dP(u) \\
&\geq \inf_{g\in F}\int_X\int_{[c,d]} \left[g(u) - (f(u)+\kappa)\right]^2 dD(\kappa)dP(u) + \epsilon_0.
\end{aligned}$$

The random variable $f(u)+\kappa$ has a distribution on $[c, 1+d]$, determined by the distributions P and D and the function f. Thus, by Corollary 13,

$$\begin{aligned}
\Pr(\mathrm{BAD}) \leq 4\left(\max_{z\in(X\times[c,1+d])^{2m}} \mathcal{N}\left(\epsilon_0/48, (\ell_F)_{|_z}\right)\right) \\
\cdot \exp\left(\frac{-\epsilon_0^2 m}{576(1+s(\sigma))^4}\right).
\end{aligned}$$

Lemma 15 implies

$$\begin{aligned}
\Pr(\mathrm{BAD}) \leq 4\left(\max_{x\in X^{2m}} \mathcal{N}\left(\frac{\epsilon_0}{144(1+s(\sigma))}, F_{|_x}\right)\right) \\
\cdot \exp\left(\frac{-\epsilon_0^2 m}{576(1+s(\sigma))^4}\right).
\end{aligned}$$

Applying Theorem 14, if

$$d = \mathrm{fat}_F\left(\frac{\epsilon_0}{576(1+s(\sigma))}\right),$$

and $m \geq d/4$, then

$$\begin{aligned}
\Pr(\mathrm{BAD}) \leq 4\exp\left(\frac{2}{\ln 2}d\ln^2\frac{373248m(1+s(\sigma))^2}{\epsilon_0^2}\right. \\
\left. - \frac{\epsilon_0^2 m}{576(1+s(\sigma))^4}\right).
\end{aligned} \tag{7}$$

For any particular $x \in X^m$, $\eta \in [c,d]^m$,

$$\begin{aligned}
&\textstyle\int_X\int_{[c,d]}(B_{x,\eta}(u) - (f(u)+\kappa))^2\, dD(\kappa)dP(u) \\
&= \left(\textstyle\int_X\int_{[c,d]}(B_{x,\eta}(u) - f(u))^2\, dD(\kappa)dP(u)\right) \\
&\quad -2\textstyle\int_X\int_{[c,d]}(B_{x,\eta}(u) - f(u))\kappa\, dD(\kappa)dP(u) + \sigma^2 \\
&= \textstyle\int_X\int_{[c,d]}(B_{x,\eta}(u) - f(u))^2\, dD(\kappa)dP(u) + \sigma^2
\end{aligned}$$

because of the independence of the noise, and the fact that it has zero mean. Thus BAD consists of those $(x,\eta) \in (X^m \times [a,b]^m)$ for which

$$\int_X (B_{x,\eta}(u) - f(u))^2 \, dP(u) \geq \epsilon_0$$

If

$$m \geq \frac{1152(1+s(\sigma))^4}{\epsilon_0^2}\left(12d\left(25 + \ln\frac{d(1+s(\sigma))^6}{\epsilon_0^4}\right)^2 + \ln\frac{4}{\delta}\right), \tag{8}$$

then applying Lemma 16, with $y_1 = 4$, $y_2 = 2d/\ln 2$, $y_3 = 373248(1+s(\sigma))^2/\epsilon_0^2$, and $y_4 = \epsilon_0^2/(576(1+s(\sigma))^4)$, we have that (7) and (8) imply

$$P^m \times D^m\left\{(x,\eta) : \int_X (B_{x,\eta}(u) - f(u))^2 \, dP(u) \geq \epsilon_0\right\} < \delta. \tag{9}$$

From Jensen's inequality,

$$\left\{(x,\eta) : \int_X |B_{x,\eta}(u) - f(u)| \, dP(u) \geq \sqrt{\epsilon_0}\right\} \subseteq \left\{(x,\eta) : \int_X \left(B_{x,\eta}(u) - f(u)\right)^2 \, dP(u) \geq \epsilon_0\right\},$$

so if $m \geq m_0(\epsilon,\delta,\sigma)$,

$$P^m \times D^m\left\{(x,\eta) : \int_X \left|(B(\epsilon^2, \mathrm{sam}(x,\eta,f)))(u) - f(u)\right| \, dP(u) \geq \epsilon\right\} < \delta,$$

where

$$m_0(\epsilon,\delta,\sigma) = \frac{1152(1+s(\sigma))^4}{\epsilon^4} \cdot \left(12d\left(25 + \ln\frac{d(1+s(\sigma))^6}{\epsilon^8}\right)^2 + \ln\frac{4}{\delta}\right),$$

and

$$d = \mathrm{fat}_F\left(\frac{\epsilon^2}{576(1+s(\sigma))}\right).$$

Now, let A be the algorithm that counts the number m of examples it receives and chooses ϵ_1 such that $m_0(\epsilon_1, 1, 0) = m$. This is always possible, since d and hence m_0 are non-increasing functions of ϵ. Algorithm A then passes ϵ_1^2 and the examples to the mapping B, and returns B's hypothesis. Since $s(\sigma)$ is a non-decreasing function of σ, m_0 is a non-decreasing function of $1/\epsilon$, $1/\delta$, and σ, so for any ϵ, δ, and σ satisfying $m_0(\epsilon,\delta,\sigma) \leq m$, we must have $\epsilon \geq \epsilon_1$. It follows that, for any ϵ, δ, and σ for which A sees at least $m_0(\epsilon,\delta,\sigma)$ examples, if P is a distribution on $X \times Y$ and $D \in \mathcal{D}$ has variance σ^2 then

$$P^m \times D^m\left\{(x,\eta) : \int_X \left|(A(\mathrm{sam}(x,\eta,f)))(u) - f(u)\right| \, dP(u) \geq \epsilon\right\} < \delta,$$

completing the proof. □

As an immediate consequence of Theorem 17, if F has a finite fat-shattering function and $\mathcal{D}$ is a bounded admissible distribution class, then F is learnable with observation noise $\mathcal{D}$. The following corollary provides the one implication in Theorem 3 we have yet to prove.

Corollary 18 *Let F be a class of functions from X to $[0,1]$. Let p be a polynomial, and suppose $\mathrm{fat}_F(\gamma) < p(1/\gamma)$ for all $0 < \gamma < 1$. Then for any almost-bounded admissible distribution class $\mathcal{D}$, F is small-sample learnable with noise $\mathcal{D}$.*

Proof We will show that Algorithm A from Theorem 17 can (ϵ,δ,σ)-learn F from a polynomial number of examples with noise $\mathcal{D}$.

Let $s : \mathbb{R}^+ \to \mathbb{R}^+$ (we will define s later). Choose $0 < \epsilon, \delta < 1$, $\sigma > 0$. Fix a distribution P on X, a function f in F, and a noise distribution D in $\mathcal{D}$ with variance σ^2.

Construct a distribution D_s from D as follows. Let f be the pdf of D. Define the pdf f_s of D_s as

$$f_s(x) = \begin{cases} \frac{f(x)}{\int_{-s(\sigma)/2}^{s(\sigma)/2} f(x)\,dx} & \text{if } -s(\sigma)/2 < x < s(\sigma)/2 \\ 0 & \text{otherwise.} \end{cases}$$

Since $\mathcal{D}$ is an almost-bounded admissible class, there are universal constants $s_0, c_0 \in \mathbb{R}^+$ such that, if $s(\sigma) > s_0\sigma$,

$$\int_{-s(\sigma)/2}^{s(\sigma)/2} f(x)\,dx \geq 1 - c_0 e^{-s(\sigma)/\sigma}.$$

It is easy to show that the total variation distance between D and D_s is

$$d_{TV}(D, D_s) \leq 2c_0 e^{-s(\sigma)/\sigma}. \tag{10}$$

For some m in $\mathbb{N}$, fix $x \in X^m$ and define the event

$$E_1 = \left\{\eta \in \mathbb{R}^m : \mathbf{er}_{P,f}(A(\mathrm{sam}(x,\eta,f))) \geq \epsilon\right\}.$$

Then (10) and Lemma 7 show that

$$D^m(E_1) \leq D_s^m(E_1) + mc_0 \exp(-s(\sigma)/\sigma).$$

If we choose $s(\sigma) = \sigma(s_0 + |\log(mc_0/\delta)|)$, then (10) holds and $s(\sigma) \geq \sigma\log(2mc_0/\delta)$, so $D^m(E_1) \leq D_s^m(E_1) + \delta/2$, Since this is true for any $x \in X^m$,

$$P^m \times D^m(E_2) \leq P^m \times D_s^m(E_2) + \delta/2,$$

where

$$E_2 = \left\{(x,\eta) \in X^m \times \mathbb{R}^m : \mathbf{er}_{P,f}(A(\mathrm{sam}(x,\eta,f))) \geq \epsilon\right\}.$$

Clearly, D_s has mean 0, finite variance, and support contained in an interval of length $s(\sigma)$. From the proof of Theorem 17, there is a polynomial p_1 such that if

$$m \geq p_1(s(\sigma), d, 1/\epsilon, \ln 1/\delta)$$

then

$$P^m \times D_s^m(E_2) < \delta/2. \tag{11}$$

Now, $\mathrm{fat}_F(\gamma) < p(1/\gamma)$, so for some polynomial p_2, $m > p_2(\sigma, 1/\epsilon, \log(1/\delta), \log m)$ implies (11). Clearly, for some polynomial p_3, if $m > p_3(\sigma, 1/\epsilon, \log(1/\delta))$ then $P^m \times D^m(E_2) < \delta$. Since this is true for any P and any D in $\mathcal{D}$ with variance σ^2, Algorithm A (ϵ,δ,σ)-learns F with noise $\mathcal{D}$ from $p_3(\sigma, 1/\epsilon, \log(1/\delta))$ examples. □

5 AGNOSTIC LEARNING

In this section, we consider an agnostic learning model, a model of learning in which assumptions about the target function and observation noise are removed. In this model, we assume labelled examples (x, y) are generated by some joint distribution P on $X \times [0,1]$. The agnostic learning problem can be viewed as the problem of learning a real-valued function f with observation noise when the constraints on the noise are relaxed — in particular, we no longer have the constraint that the noise is independent of the value $f(x)$. This model has been studied in [11], [13].

If h is a $[0,1]$-valued function defined on X, define the **error of h with respect to P** as

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

We require that the learner chooses a function with error little worse than the best function in some "touchstone" function class F. Notice that the learner is not restricted to choose a function from F; the class F serves only to provide a performance measurement standard (see [13]).

Definition 19 *Suppose F is a class of $[0,1]$-valued functions defined on X, P is a probability distribution on $X \times [0,1]$, $0 < \epsilon, \delta < 1$ and $m \in \mathbb{N}$. We say a learning algorithm $L = (A, D_Z)$ (ϵ,δ)-**learns in the agnostic sense with respect to** F **from** m **examples** if, for all distributions P on $X \times [0,1]$,*

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

The function class F is **agnostically learnable** *if there is a learning algorithm L and a function $m_0 : (0,1)\times(0,1) \to \mathbb{N}$ such that, for all $0 < \epsilon, \delta < 1$, Algorithm L (ϵ,δ)-learns in the agnostic sense with respect to F from $m_0(\epsilon,\delta)$ examples. If, in addition, m_0 is bounded by a polynomial in $1/\epsilon$ and $1/\delta$, we say that F is* **small-sample agnostically learnable.**

The following result is analogous to the characterization theorem of Section 2.

Theorem 20 *Suppose F is a permissible class of $[0,1]$-valued functions defined on X. Then F is agnostically learnable if and only if its fat-shattering function is finite, and F is small-sample agnostically learnable if and only if there is a polynomial p such that $\mathrm{fat}_F(\gamma) < p(1/\gamma)$ for all $\gamma > 0$.*

The following result proves the "only if" parts of the theorem.

Theorem 21 *Let F be a class of $[0,1]$-valued functions defined on X. Suppose $0 < \gamma < 1$, $0 < \epsilon \leq \gamma/65$, $0 < \delta \leq 1/16$, and $d \in \mathbb{N}$. If $\mathrm{fat}_F(\gamma) \geq d > 800$, then any learning algorithm that (ϵ,δ)-learns in the agnostic sense with respect to F requires at least m_0 examples, where*

$$m_0 > \frac{d}{400 \log \frac{40}{\gamma}}.$$

Proof The proof is similar to, though simpler than, the argument in Section 3. We will show that the agnostic learning problem is not much harder than the problem of learning a quantized version of the function class F, and then apply Lemma 10.

Set $\epsilon = \gamma/65$ and $\delta = 1/16$. Consider the class of distributions P on $X \times [0,1]$ for which there exists an f in F such that, for all $x \in X$,

$$P(y|x) = \begin{cases} 1 & y = Q_{2\epsilon}(f) \\ 0 & \text{otherwise} \end{cases}$$

Fix a distribution P in this class. Let L be a randomized learning algorithm that can (ϵ,δ)-learn in the agnostic sense with respect to F. Then

$$\Pr\left(\mathbf{er}_P(L) \geq \inf_{f\in F}(\mathbf{er}_P(f)) + \epsilon\right) < \delta,$$

where $\mathbf{er}_P(L)$ is the error of the function that the learning algorithm chooses. But the definition of P ensures that $\inf_{f\in F} \mathbf{er}_P(f) \leq \epsilon$, so

$$\Pr\left(\mathbf{er}_P(L) \geq 2\epsilon\right) < \delta.$$

Clearly Algorithm L $(2\epsilon,\delta)$-learns the quantized function class $Q_{2\epsilon}(F)$.

By hypothesis, $\mathrm{fat}_F(\gamma) \geq d$, but then the definition of fat-shattering implies that $\mathrm{fat}_{Q_{2\epsilon}(F)}(\gamma - \epsilon) \geq d$. Since $\epsilon = \gamma/65$, $2\epsilon \leq (\gamma-\epsilon)/32$. Also, $\delta = 1/16$, so Lemma 10 implies

$$\begin{aligned} m_0 &> \frac{d - 400}{2 + 192 \log \lceil 1/(2\epsilon)\rceil} \\ &> \frac{d}{384 \log(34/\gamma)}. \end{aligned}$$

□

With minor modifications, the proof of Theorem 17 yields the following analogous result for agnostic learning.

Theorem 22 *Choose a permissible set F of functions from X to $[0,1]$. There exists an algorithm A such that, for all $0 < \epsilon < 1/2$, for all $0 < \delta < 1$, if $\mathrm{fat}_F(\epsilon/192) = d$, then A agnostically (ϵ,δ)-learns F from*

$$\frac{1152}{\epsilon^2}\left(12d\left(23 + \ln\frac{d}{\epsilon^4}\right)^2 + \ln\frac{4}{\delta}\right)$$

examples.

Proof Sketch First, the analog of Corollary 13 where the expected absolute error is used to measure the "quality" of a hypothesis in place of the expected squared error, and $b = 1$ and $a = 0$, can be proved using essentially the same argument. Second, the analog of Lemma 15 where ℓ_F is replaced with a corresponding class constructed from absolute loss in place of ℓ, where $a = 0, b = 1$, and where the $\epsilon/(3|b-a|)$ of the upper bound is replaced with ϵ, also is obtained using a simpler, but similar, proof. These results are combined with Theorem 14 and Lemma 16 in much the same way as was done for Theorem 17. □

6 DISCUSSION

All of our results can be extended easily to the case of $[L, U]$-valued functions by scaling the parameters ϵ, γ, and σ to convert the learning problem to an equivalent $[0, 1]$-valued learning problem.

It seems likely that the characterization of learnability in terms of finiteness of the fat-shattering function could be extended to the case of unbounded noise. Perhaps the techniques used in [10] to prove uniform convergence with unbounded noise could be useful here.

There are several ways in which our results could be improved. The sample complexity upper bound in Theorem 17 increases at least as $1/\epsilon^4$. It seems plausible that this rate is excessive; perhaps it is an artifact of the use of Jensen's inequality in the proof. Obviously, the constants in our bounds are large.

The lower bound on the sample complexity of real-valued learning (Theorem 11) does not increase with $1/\epsilon$ and $1/\delta$. In fact, the lower bound of that theorem is trivially true if the standard deviation of the noise is sufficiently small,[4] i.e.

$$\frac{1}{v(\sigma)} < de^{-d/800}/10.$$

However, the following example shows that a condition of this form is essential, and that when the noise variance is small there need be no dependence of the lower bound on the desired accuracy and confidence.

Example Fix $d \in \mathbb{N}$. Let the nonempty, measurable sets S_j, $j = 0, \ldots, d-1$, form a partition of X (that is, $\cup_j S_j = X$, and $S_j \cap S_k = \emptyset$ if $j \neq k$). Consider the function class

$$F_d = \left\{ f_{b_0,\ldots,b_{d-1}} : b_i \in \{0,1\}, i = 0, \ldots, d-1 \right\}$$

of functions defined by

$$f_{b_0,\ldots,b_{d-1}}(x) = \frac{3}{4} \sum_{j=0}^{d-1} 1_{S_j}(x) b_j + \frac{1}{8} \sum_{k=0}^{d-1} b_k 2^{-k},$$

where 1_{S_j} is the indicator function for S_j ($1_{S_j}(x) = 1$ iff $x \in S_j$). That is, the labels b_j determine the two most significant bits of the value of the function in S_j, and the d least significant bits of its value at any $x \in X$ encode the identity of the function. Clearly, for any $\gamma \leq 1/4$, $\mathrm{fat}_{F_d}(\gamma) = d$.

With no observation noise, one example (x, y) suffices to learn F_d exactly, because the learning algorithm can identify the function from the d least significant bits of y. (The union of these function classes, $F = \cup_{d=1}^{\infty} F_d$, has $\mathrm{fat}_F(\gamma) = \infty$ for $\gamma \leq 1/4$, but any f in F can be identified from a single example (x, y) with no observation noise[5].) One example also suffices with uniform observation noise provided the variance is sufficiently small; if $\sigma < 2^{-d-3}3^{-1/2}$, a learning algorithm that sees one example (x, y) and chooses the integral multiple of 2^{-d-2} that is closest to y will be able to identify the target function. That is, if $1/v(\sigma) < 2^{-d-2}3^{-1/2}$, then $(\epsilon, \delta, \sigma)$-learning with uniform noise is possible from a single example, for any $\epsilon, \delta \geq 0$.

Suppose the observation noise is gaussian, of variance σ^2, and $\sigma < 2^{-d-5/2}(\log 4)^{-1/2}$. Consider the following algorithm. For each example (x, y), the algorithm chooses the integral multiple of 2^{-d-2} that is closest to y, and stores the corresponding function label (the d least significant bits). After m examples, it outputs the function with the most common label. The bound on σ and Inequality (1) (the bound on the area under the tails of the gaussian density) imply that, with probability at least $3/4$ a noisy observation is closer to the value $f(x)$ than to any other integral multiple of 2^{-d-2}. From the standard Chernov bounds [2], if $m \geq 12\log(1/\delta)$ the probability that the algorithm will store the correct label for fewer than half of the examples is less than δ. So this algorithm can $(\epsilon, \delta, \sigma)$-learn from $12\log(1/\delta)$ examples, for any $\epsilon \geq 0$. □

The above example shows that a gap in the growth of the upper and lower bounds with $1/\epsilon$ and $1/\delta$ is essential. However, if the noise variance is sufficiently large, it seems likely that there is a general lower bound that grows with these quantities. Simon [21] shows that a stronger notion of shattering provides a lower bound for the problem of learning without noise. However, the finiteness of this strong-fat-shattering function is not sufficient for learnability, as the following example shows.

Example Simon says that a sequence $x_1, \ldots, x_d$ is **strongly γ-shattered** by F if there exists $r \in [0, 1]^d$ such that for each $b \in \{0, 1\}^d$, there is an $f \in F$ such that for each i

$$f(x_i) = \begin{cases} r_i + \gamma & \text{if } b_i = 1 \\ r_i - \gamma & \text{if } b_i = 0. \end{cases}$$

For each γ, let

$$\mathrm{sfat}_F(\gamma) = \max\{d \in \mathbb{N} : \exists x_1, \ldots, x_d, \\ F \text{ strongly } \gamma\text{-shatters } x_1, \ldots, x_d\}.$$

if such a maximum exists, and ∞ otherwise. If $\mathrm{sfat}_F(\gamma)$ is finite for all γ, we say F has a **finite strong-fat-shattering function.**

Suppose $X = \mathbb{N}$. For each $q : \mathbb{N} \to \{0, 1\}$, let y_q be the element of $[0, 1]$ whose binary decimal expansion is given by q, i.e., let $y_q = \sum_{i=1}^{\infty} q(i)2^{-i}$. Also, let $f_q : \mathbb{N} \to [0, 1]$ be defined by

$$f_q(j) = \begin{cases} 3/4 + y_q/4 & \text{if } q(j) = 1 \\ 1/4 - y_q/4 & \text{if } q(j) = 0. \end{cases}$$

Let

$$Q = \{q \in \{0,1\}^{\mathbb{N}} : \forall j_0 \, \exists j > j_0, q(j) = 0\}.$$

Informally, Q represents the set of all infinite binary sequences that don't end with repeating 1's. Each real number in $[0, 1)$ has a unique representation in Q [19]. Suppose $F = \{f_q : q \in Q\}$. Since X is countable, F is permissible. Trivially, $\mathrm{fat}_F(1/4) = \infty$, so F is not learnable in any sense described in this paper. However, since for any $q_1, q_2 \in Q$ for which $q_1 \neq q_2$, for any $j \in \mathbb{N}$, $f_{q_1}(j) \neq f_{q_2}(j)$, trivially,

[4] Note that as the standard deviation gets small, the total variation of the density function must get large.

[5] Thanks to David Haussler for suggesting this function class.

$\text{sfat}_F(\gamma) \leq 1$ for all $\gamma < 1$, so neither the finiteness nor the polynomial growth of sfat_F characterizes learnability in any of the senses of this paper. □

Simon provides examples in his paper that show that his general lower bounds are tight. These classes have identical strong-fat-shattering and fat-shattering functions.

ACKNOWLEDGEMENTS

This research was supported by the Australian Telecommunications and Electronics Research Board and the Australian Research Council. Phil Long was also supported by Air Force Office of Scientific Research grant F49620–92–J0515. Thanks to Wee Sun Lee and Martin Anthony for helpful comments.

REFERENCES

[1] N. ALON, S. BEN-DAVID, N. CESA-BIANCHI AND D. HAUSSLER, *Scale-sensitive dimensions, uniform convergence, and learnability*, Symposium on Foundations of Computer Science, 1993.

[2] D. ANGLUIN AND L. G. VALIANT, *Fast probabilistic algorithms for Hamiltonian circuits and matchings*, Journal of Computer and System Sciences, 18 (1979), pp. 155–193.

[3] M. ANTHONY AND J. SHAWE-TAYLOR, *Valid generalization from approximate interpolation*, Proceedings of the 1993 IMA European Conference on Computational Learning Theory, to appear.

[4] P. AUER, P. M. LONG, W. MAASS AND G. J. WOEGINGER, *On the complexity of function learning*, Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 1993.

[5] P. AUER AND P. M. LONG, *Simulating access to hidden information while learning*, Proceedings of the 26th Annual ACM Symposium on the Theory of Computation, 1994.

[6] P. L. BARTLETT, *Learning with a slowly changing distribution*, Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 1992.

[7] S. BEN-DAVID, N. CESA-BIANCHI, D. HAUSSLER AND P. LONG, *Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions*, Manuscript, an earlier version was presented at the Fifth Annual ACM Workshop on Computational Learning Theory, 1993.

[8] G. BENEDEK AND A. ITAI, *Learnability with respect to fixed distributions*, Theoretical Computer Science, 86 (1991), pp. 377–389.

[9] A. BLUMER, A. EHRENFEUCHT, D. HAUSSLER AND M. K. WARMUTH, *Learnability and the Vapnik-Chervonenkis dimension*, Journal of the Association for Computing Machinery, 36 (1989), pp. 929–965.

[10] S. VAN DE GEER, *Regression Analysis and Empirical Processes*, Centrum voor Wiskunde en Informatica, Amsterdam, 1988.

[11] D. HAUSSLER, *Decision theoretic generalizations of the PAC model for neural net and other learning applications*, Information and Computation, 100 (1992), pp. 78–150.

[12] M. J. KEARNS AND R. E. SCHAPIRE, *Efficient distribution-free learning of probabilistic concepts (extended abstract)*, Proceedings of the 31st Annual Symposium on the Foundations of Computer Science, 1990.

[13] M. J. KEARNS, R. E. SCHAPIRE AND L. M. SELLIE, *Toward efficient agnostic learning*, Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 1992.

[14] N. MERHAV AND M. FEDER, *Universal schemes for sequential decision from individual data sequences*, IEEE Transaction on Information Theory, 39 (1993), pp. 1280–1292.

[15] B. K. NATARAJAN, *On learning sets and functions*, Machine Learning, 4 (1989), pp. 67–97.

[16] B. K. NATARAJAN, *Occam's razor for functions*, Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 1993.

[17] J. K. PATEL AND C. B. READ, *The Big Book of Facts About the Normal Distribution*, Marcel Dekker, New York, 1982.

[18] D. POLLARD, *Convergence of Stochastic Processes*, Springer, New York, 1984.

[19] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1988.

[20] J. SHAWE-TAYLOR, M. ANTHONY AND N. BIGGS, *Bounding sample size with the Vapnik-Chervonenkis dimension*, Discrete Applied Mathematics, 42 (1993), pp. 65–73.

[21] H. U. SIMON, *Bounds on the number of examples needed for learning functions*, FB Informatik, LS II, Universität Dortmund, Forschungsbericht Nr. 501, 1993.

[22] L. G. VALIANT, *A theory of the learnable*, Communications of the ACM, 27 (1984), pp. 1134–1143.

[23] V. N. VAPNIK AND A. Y. CHERVONENKIS, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications, XVI (1971), pp. 264–280.