JOURNAL OF COMPUTER AND SYSTEM SCIENCES 50, 74–86 (1995)

Characterizations of Learnability for Classes of $\{0, ..., n\}$-Valued Functions

SHAI BEN-DAVID*

Technion, Haifa 32000, Israel

NICOLÒ CESA-BIANCHI†

Università di Milano, Via Comelico 39, 20135 Milano, Italy

DAVID HAUSSLER‡

University of California at Santa Cruz, Santa Cruz, California 95064

AND

PHILIP M. LONG§

Duke University P.O. Box 90129 Durham, North Carolina 27708

Received August 10, 1993; revised April 7, 1994

We investigate the PAC learnability of classes of $\{0, ..., n\}$-valued functions ($n < \infty$). For $n = 1$ it is known that the finiteness of the Vapnik–Chervonenkis dimension is necessary and sufficient for learning. For $n > 1$ several generalizations of the VC-dimension, each yielding a distinct characterization of learnability, have been proposed by a number of researchers. In this paper we present a general scheme for extending the VC-dimension to the case $n > 1$. Our scheme defines a wide variety of notions of dimension in which all these variants of the VC-dimension, previously introduced in the context of learning, appear as special cases. Our main result is a simple condition characterizing the set of notions of dimension whose finiteness is necessary and sufficient for learning. This provides a variety of new tools for determining the learnability of a class of multi-valued functions. Our characterization is also shown to hold in the "robust" variant of PAC model and for any "reasonable" loss function.

1. INTRODUCTION

A central task in computational learning theory is to provide simple mathematical characterizations of what is learnable under natural formal models of learning. An example along these lines is the characterization of those classes of binary functions that are learnable in Valiant's PAC model in terms of the Vapnik–Chervonenkis dimension[1] proved by Blumer, Ehrenfeucht, Haussler, and Warmuth [4]. A natural way to extend the PAC model is to consider the learning of general multi-valued (instead of just binary) functions. Natarajan [10] introduced a generalization of the Vapnik–Chervonenkis dimension and showed that his notion of dimension characterizes the learnability of classes of $\{0, ..., n\}$-valued functions for all fixed $n > 1$.

Intuitively, we can reduce the problem of learning any multi-valued function f to the problem of learning a related set of binary functions by providing a binary encoding of f's range. For example, a function $f: X \to \{0, ..., n\}$ in some class $\mathcal{F}$ can be learnt by learning the $\lceil \log(n+1) \rceil$ binary functions $f_i \in \mathcal{F}_i$, where $f_i(x)$ is the i th bit of $f(x)$. However, checking the learnability of each $\mathcal{F}_i$ might result in a much harder task than directly inspecting the class $\mathcal{F}$. In this paper we offer several simple combinatorial properties of the class $\mathcal{F}$ itself each characterizing the learnability of $\mathcal{F}$. More precisely, we present a general scheme for extending the VC-dimension to classes of $\{0, ..., n\}$-valued functions for any positive integer n. Our scheme defines a wide family of notions of dimension including as special cases the Natarajan dimension [10], the graph dimension [5, 10], Pollard's pseudo-dimension [12, 13, 7], and a generalization proposed by Vapnik (see, e.g., [18]).

In extending Valiant's PAC model, we assume (see also [10]) that a "target" function is chosen from a given class

* shai@cs.technion.ac.il.
† cesabian@dsi.unimi.it.
‡ haussler@cse.ucsc.edu.
§ plong@cs.duke.edu.

[1] Defined by Vapnik and Chervonenkis [19].

0022-0000/95 $6.00

of multi-valued functions and the learner is to select from the same class a function that yields a good approximation of the target. The class is said to be learnable if for any target function and for any probability distribution on the domain an arbitrarily accurate approximation can be obtained with high probability with respect to a random sample of finite size.

Our main result is a simple combinatorial condition characterizing the set of notions of dimension (from among those generated by our scheme) whose finiteness is necessary and sufficient for learning (Theorem 16). This provides a variety of new tools for determining the learnability of a given class $\mathcal{F}$ of multi-valued functions and, furthermore, establishes the equivalence between the learnability of $\mathcal{F}$ and the learnability of the classes (of binary functions) generated by any reasonable binary encoding of $\mathcal{F}$'s range. As a side effect we establish the equivalence between PAC-learnability and the property of uniform convergence of frequencies to probabilities over an associated class of binary "loss functions." Another interesting side effect (Corollary 6 and Theorem 10) is that the ratio of (1) the dimension of a set $\mathcal{F}$ of $\{0, ..., n\}$-valued functions as measured by any of the previously studied notions of dimensions listed above, and (2) $\mathcal{F}$'s dimension as measured by any other of those notions of dimension, is at most $4.67 \log_2(n+1)$. In fact, this relationship can be seen to hold for any pair of the notions of dimension in our scheme, each of who's finiteness provides a characterization of learnability. Thus, one may use whichever of these notions of dimension is most convenient for analyzing a given class of functions and have a good estimate for all of them.

A further extension to Valiant's learning framework can be obtained within the more general pattern recognition model studied by Vapnik [17]. This framework, often called the "robust" or "agnostic" PAC model, was discussed in an appendix of [4] and studied in a more general setting in [7]. In the robust PAC model the learner's task is to generate (with high probability) a nearly optimal deterministic approximation of a stochastic relationship using hypotheses chosen from a given class of functions. The finiteness of each of the dimensions defined by our condition is shown to be sufficient for robust learning and necessary for the weaker nonrobust learning. Therefore they characterize learnability in both models. Moreover, this double characterization is also shown not to depend on which particular bounded nonnegative function is used to measure the loss.

In Section 2 we prove some combinatorial properties of our family of generalizations. The applications of these properties to learning are described in Section 3. In Section 4 we show how our results can be extended to robust learning models and to more general loss functions. We also show how sample size bounds can be computed using the results of Section 2. Finally, Section 5 is devoted to open problems and conclusions.

2. GENERALIZATIONS OF THE VC-DIMENSION

We begin by introducing a general scheme for extending the VC-dimension. The results presented in this section can be stated more easily if we define the VC-dimension and its generalizations as dimensions of subsets of $\{0, ..., n\}^m$ for any positive integer m. In Section 3 we will extend these definitions and demonstrate their relevance to the learnability of classes of $\{0, ..., n\}$-valued functions on arbitrary domains.

Let N be the positive integers. Choose $m, n \in N$ and let $S \subseteq \{0, ..., n\}^m$. For each k-tuple $\bar{\imath} = (i_1, ..., i_k)$ of indices from the set $\{1, ..., m\}$, define the $\bar{\imath}$-projection of S in $\{0, ..., n\}^k$ by

$$S|_{\bar{\imath}} = \{(s_{i_1}, ..., s_{i_k}) : (s_1, ..., s_m) \in S\}.$$

Suppose for a moment that $n = 1$. In such a case we say that $S \subseteq \{0, 1\}^m$ VC-shatters a k-tuple $\bar{\imath} = (i_1, ..., i_k)$ of indices if and only if

$$S|_{\bar{\imath}} = \{0, 1\}^k.$$

The VC-dimension of S is the length of the longest sequence of indices VC-shattered by S.[2]

Now let us return to the more general case in which $n \geq 1$. A natural way to extend the above definition of shattering is to say that S shatters the k-tuple $\bar{\imath} = (i_1, ..., i_k)$ if and only if

$$S|_{\bar{\imath}} = \{0, ..., n\}^k$$

and define a notion of dimension as we did with the VC-dimension. This definition of shattering was investigated in [1, 9, 15, 2]. Unfortunately, using this extension, if $n > 1$, the set $\{0, 1\}^m$, which has 2^m elements, has dimension 0. This fact prevents the finiteness of such a notion of dimension from being a characterization of learnability in the model studied here [4].

To define a generalization that yields bounds on $|S|$ for sets S of a given dimension that are polynomial in m and therefore sufficiently strong for learning in our model, we look for a "translation" of multi-valued vectors into binary vectors. This is done by considering mappings ψ from the set $\{0, ..., n\}$ to $\{0, 1, *\}$ ($*$ will be thought of as a null element).

[2] Note that this definition is equivalent to that we would obtain if we insisted that the shattered indices satisfy $1 \leq i_1 < i_2 < \cdots < i_k \leq m$, which is perhaps the easiest way to think of this and the following definitions of shattering.

More formally, let Ψ be a family of mappings ψ from $\{0, ..., n\}$ to $\{0, 1, *\}$. For $\bar{u} \in \{0, ..., n\}^m$ and $\bar{\psi} = (\psi_1, ..., \psi_m) \in \Psi^m$, denote $(\psi_1(u_1), ..., \psi_m(u_m))$ by $\bar{\psi}(\bar{u})$. For a set $U \subseteq \{0, ..., n\}^m$, define $\bar{\psi}(U) = \{\bar{\psi}(\bar{u}) : \bar{u} \in U\}$.

We say $\bar{\imath} = (i_1, ..., i_k)$ is *Ψ-shattered* by S if there exists $\bar{\psi} \in \Psi^k$ such that

$$\{0, 1\}^k \subseteq \bar{\psi}(S|_{\bar{\imath}}).$$

In the case in which there exists such a $\bar{\psi}$, which in addition has $\psi_1 = \psi_2 = \cdots = \psi_k$, we say that $\bar{\imath}$ is *uniformly Ψ-shattered* by S.

Let the *Ψ-dimension* of S (denoted by Ψ-dim(S)) be the maximum d for which there exists a d-tuple $\bar{\imath} \in \{1, ..., m\}^d$ of indices Ψ-shattered by S and let the *uniform Ψ-dimension* of S (denoted by Ψ-dim$_U(S)$) be the corresponding definition for uniform shattering.

By choosing different subsets Ψ of the set of all functions from $\{0, ..., n\}$ to $\{0, 1, *\}$, we obtain a whole family of notions of dimension. In Section 2.2 we will investigate some properties of this family that will prove useful for showing results about learnability.

2.1. *Previously Known Examples*

Several previously defined notions of dimension correspond to particular choices of the set Ψ of mappings. Some of them are listed below.

- Pollard's pseudo-dimension [13, 7] is the Ψ_P-dimension, where $\Psi_P = \{\psi_{P,k} : 0 < k \leqslant n\}$ and $\psi_{P,k}$ is defined by

$$\psi_{P,k}(a) = \begin{cases} 1 & \text{if } a \geqslant k \\ 0 & \text{otherwise.} \end{cases}$$

- Vapnik's dimension [18] is the uniform Ψ_P-dimension, where Ψ_P is defined above.

- The graph dimension [5, 10] is the Ψ_G-dimension, where $\Psi_G = \{\psi_{G,k} : k \in \{0, ..., n\}\}$ and $\psi_{G,k}$ is defined by

$$\psi_{G,k}(a) = \begin{cases} 1 & \text{if } a = k \\ 0 & \text{otherwise.} \end{cases}$$

- The Natarajan dimension [10] is the Ψ_N-dimension, where $\Psi_N = \{\psi_{N,k,l} : k, l \in \{0, ..., n\},\ k \neq l\}$ and $\psi_{N,k,l}$ is defined by

$$\psi_{N,k,l}(a) = \begin{cases} 1, & \text{if } a = k \\ 0, & \text{if } a = l \\ *, & \text{otherwise.} \end{cases}$$

Let Ψ_B be the set of all mappings from $\{0, ..., n\}$ to $\{0, 1\}$ and define the Ψ_B-dimension accordingly.

Note that the graph dimension, the Natarajan dimension, and the Ψ_B-dimension do not make use of the natural ordering on $\{0, ..., n\}$ and could just as easily be defined for abstract finite sets.

2.2. *Elementary Properties*

We may define an order $\sqsubseteq$ on the set of all subsets of $\{0, 1, *\}^{\{0, ..., n\}}$ in the following way: $\Psi \sqsubseteq \Phi$ iff for all $m \in \mathbf{N}$ and $S \subseteq \{0, ..., n\}^m$, S Ψ-shatters $\bar{\imath} \subseteq \{1, ..., m\}$ implies that S Φ-shatters $\bar{\imath} \subseteq \{1, ..., m\}$.

It can be shown that for $m \geqslant 2$, $n \geqslant 3$, this order is partial.

THEOREM 1. *If $m = 2$, $n = 3$, there exists $S \subseteq \{0, ..., n\}^m$ that Ψ_G-shatters $(1, 2)$ and does not Ψ_P-shatter it, and there exists $T \subseteq \{0, ..., n\}^m$ that Ψ_P-shatters $(1, 2)$ and does not Ψ_G-shatter it. Hence Ψ_P and Ψ_G are incomparable.*

Proof. In Appendix A. ∎

LEMMA 2. *Let Ψ, Φ be classes of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ such that $\Psi \sqsubseteq \Phi$. Then for all $S \subseteq \{0, ..., n\}^m$*

$$\Psi\text{-dim}(S) \leqslant \Phi\text{-dim}(S)$$

$$\Psi\text{-dim}_U(S) \leqslant \Phi\text{-dim}_U(S).$$

Proof. Follows directly from the definitions. ∎

The next lemma gives a sufficient condition for $\Psi \sqsubseteq \Phi$ for families Ψ and Φ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$.

LEMMA 3. *Let Ψ, Φ be classes of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ such that for all $\psi \in \Psi$ there exists $\phi \in \Phi$ such that $\psi^{-1}(0) \subseteq \phi^{-1}(b)$ and $\psi^{-1}(1) \subseteq \phi^{-1}(1 - b)$ holds for b either 0 or 1. Then $\Psi \sqsubseteq \Phi$.*

Proof. Assume that for all $\psi \in \Psi$ there is a $\phi \in \Phi$ such that $\psi^{-1}(0) \subseteq \phi^{-1}(0)$ and $\psi^{-1}(1) \subseteq \phi^{-1}(1)$. (The case in which for all $\psi \in \Psi$ there is a $\phi \in \Phi$ such that $\psi^{-1}(0) \subseteq \phi^{-1}(1)$ and $\psi^{-1}(1) \subseteq \phi^{-1}(0)$ can be handled analogously.) Choose $S \subseteq \{0, ..., n\}^m$ and $\bar{\imath} \in \{1, ..., m\}^k$ such that S Ψ-shatters $\bar{\imath}$. Choose $\bar{\psi} \in \Psi^k$ such that

$$\{0, 1\}^k \subseteq \bar{\psi}(S|_{\bar{\imath}}).$$

For each j, $1 \leqslant j \leqslant k$, let ϕ_j be such that $\psi_j^{-1}(0) \subseteq \phi_j^{-1}(0)$ and $\psi_j^{-1}(1) \subseteq \phi_j^{-1}(1)$. Let $\bar{\phi} = (\phi_1, ..., \phi_k)$.

We claim that $\{0, 1\}^k \subseteq \bar{\phi}(S|_{\bar{\imath}})$. Choose $\bar{b} = (b_1, ..., b_k) \in \{0, 1\}^k$. Let $\bar{r} \in S|_{\bar{\imath}}$ be such that $\bar{\psi}(\bar{r}) = \bar{b}$. Choose $j \in \{1, ..., k\}$. Since $\psi_j^{-1}(0) \subseteq \phi_j^{-1}(0)$, $\psi_j^{-1}(1) \subseteq \phi_j^{-1}(1)$, and $b_j \in \{0, 1\}$, $\phi_j(r_j) = \psi_j(r_j)$. Since j was chosen arbitrarily, $\bar{\phi}(\bar{r}) = \bar{\psi}(\bar{r}) = \bar{b}$. Therefore, since $\bar{b}$ was chosen arbitrarily,

$$\{0, 1\}^k \subseteq \bar{\phi}(S|_{\bar{\imath}}).$$

Thus S Φ-shatters $\bar{\imath}$. The uniform case follows analogously. ∎

Finally, we have the following simple observation.

LEMMA 4. *For any set Ψ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ and any $S \subseteq \{0, ..., n\}^m$,*

$$\Psi\text{-dim}_U(S) \leqslant \Psi\text{-dim}(S).$$

2.3. *Distinguishers*

Let Ψ be a family of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$. We say that a pair a, b of distinct elements in $\{0, ..., n\}$ is *Ψ-distinguishable* if there exists $\psi \in \Psi$ such that $\psi(a) = 0$ and $\psi(b) = 1$ or vice versa. We call Ψ a *distinguisher* if each pair a, b of distinct elements in $\{0, ..., n\}$ is Ψ-distinguishable.

All the examples of notions of dimension given in Section 2.1 are easily seen to correspond to distinguishers. It is also immediate to see that if $n = 1$, for any distinguisher Ψ the definitions of the Ψ-dimension and the uniform Ψ-dimension are equivalent to the definition of the VC-dimension.

THEOREM 5. *For any distinguisher Ψ,*

$$\Psi_N \sqsubseteq \Psi \sqsubseteq \Psi_B.$$

Proof. Follows immediately from Lemma 3 and the definition of a distinguisher. ∎

Theorem 5 trivially yields the following corollary about the Ψ-dimension and the uniform Ψ-dimension for various Ψ's.

COROLLARY 6. *Choose a distinguisher Ψ and $S \subseteq \{0, ..., n\}^m$:*

$$\Psi_N\text{-dim}(S) \leqslant \Psi\text{-dim}(S) \leqslant \Psi_B\text{-dim}(S)$$

$$\Psi_N\text{-dim}_U(S) \leqslant \Psi\text{-dim}_U(S) \leqslant \Psi_B\text{-dim}_U(S).$$

We now turn to the proof of some combinatorial bounds about distinguishers that will be used in the next section. First, we establish the following bound on the uniform Ψ-dimension of S in terms of its (nonuniform) Ψ-dimension for any Ψ.

THEOREM 7. *Choose a set Ψ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ and choose a subset S of $\{0, ..., n\}^m$. Then*

$$\Psi\text{-dim}(S) \leqslant |\Psi|\,(\Psi\text{-dim}_U(S)).$$

Proof. Let $d = \Psi\text{-dim}_U(S)$. Suppose that the Ψ-dimension d' of S is greater than $d|\Psi|$. Let $\bar{\imath} = (i_1, ..., i_{d'})$ be a sequence shattered by S and let $\bar{\psi} = (\psi_1, ..., \psi_{d'})$ be such that

$$\{0, 1\}^{d'} \subseteq \bar{\psi}(S|_{\bar{\imath}}).$$

By the pigeonhole principle, since $d' > d|\Psi|$, there exists a subsequence $(i_{j_1}, ..., i_{j_{d+1}})$ of $\bar{\imath}$ such that for all $1 \leqslant k, l \leqslant d+1$, $\psi_{j_k} = \psi_{j_l}$. Therefore, S uniformly Ψ-shatters $(i_{j_1}, ..., i_{j_{d+1}})$, contradicting the assumption that $\Psi\text{-dim}_U(S) = d$. ∎

We next show that this bound is the best possible in terms of d and $|\Psi|$.

THEOREM 8. *Choose positive integers d and r. Then if m and n are large enough, there is a family Ψ of functions from $\{0, ..., n\}$ to $\{0, 1, *\}$, and $S \subseteq \{0, ..., n\}^m$ for which*

1. $|\Psi| = r$,
2. $\Psi\text{-dim}_U(S) = d$,
3. $\Psi\text{-dim}(S) = dr$.

Proof. In Appendix B. ∎

We will make use of the following result which bounds from above the cardinality of a set $S \subseteq \{0, ..., n\}^m$ in terms of its Natarajan dimension.[3]

THEOREM 9 [8]. *Choose $S \in \{0, ..., n\}^m$. Then if $\Psi_N\text{-dim}(S) \leqslant d$,*

$$|S| \leqslant \sum_{i=0}^{d} \binom{m}{i}\binom{n+1}{2}^i \leqslant \left(\frac{me(n+1)^2}{2d}\right)^d.$$

We apply this theorem to obtain an upper bound on the Ψ_B-dimension of a given class in terms of its Natarajan (Ψ_N) dimension. Recall that Corollary 6 established that the Natarajan dimension of a class is at most its Ψ_B-dimension.

THEOREM 10. *Let $S \subseteq \{0, ..., n\}^m$. Let $d_N = \Psi_N\text{-dim}(S)$ and $d_B = \Psi_B\text{-dim}(S)$. Then*

$$d_B \leqslant 4.67 d_N \log_2(n+1).$$

Proof. Let $\bar{\imath} = (i_1, ..., i_{d_B})$ be a sequence of indices Ψ_B-shattered by S. Let $T = S|_{\bar{\imath}}$. Since there exists $\bar{\psi} \in \Psi_B$ such that

$$\{0, 1\}^{d_B} \subseteq \bar{\psi}(T),$$

we have that $|T| \geqslant 2^{d_B}$. From Theorem 9, we may conclude that $|T| \leqslant (d_B e(n+1)^2/2d_N)^{d_N}$. Thus,

$$\left(\frac{d_B e(n+1)^2}{2d_N}\right)^{d_N} \geqslant 2^{d_B}.$$

[3] A looser bound was proved by Natarajan [10]. We apparently need the stronger bound to prove Theorem 10.

Using the approximation $\ln x \leq xy - \ln(ey)$ (see [14]), which holds for any pair x, y of real positive numbers, we derive the following chain of implications for all $y < \ln 2$,

$$2^{d_B} \leq \left(\frac{d_B e(n+1)^2}{2d_N}\right)^{d_N}$$

$$\Leftrightarrow \quad d_B \ln 2 \leq d_N \left[\ln \frac{d_B}{d_N} + \ln \frac{e(n+1)^2}{2}\right]$$

$$\Rightarrow \quad d_B \ln 2 \leq d_N \left[\frac{d_B}{d_N} y - \ln(ey) + \ln \frac{e(n+1)^2}{2}\right]$$

$$\Leftrightarrow \quad d_B \ln 2 \leq d_B y + d_N \ln \frac{(n+1)^2}{2y}$$

$$\Leftrightarrow \quad d_B \leq \frac{d_N}{\ln 2 - y} \ln \frac{(n+1)^2}{2y}.$$

If we assume further that $y \leq \frac{1}{2}$, we obtain

$$d_B \leq \frac{2d_N \ln(n+1) - d_N \ln(2y)}{\ln 2 - y}$$

$$= \frac{(2 \ln 2) d_N \log_2(n+1) - d_N \ln(2y)}{\ln 2 - y}$$

$$\leq \left(\frac{(2 \ln 2) - \ln(2y)}{\ln 2 - y}\right) d_N \log_2(n+1)$$

$$\left(\text{since } n \geq 1,\ y \leq \frac{1}{2}\right)$$

$$= \left(\frac{\ln 2 - \ln y}{\ln 2 - y}\right) d_N \log_2(n+1).$$

Choosing $y = \frac{1}{5}$ and verifying that $\ln 10/(\ln 2 - \frac{1}{5}) < 4.67$ completes the proof. ∎

Finally, we show that the bound of Theorem 10 is within a constant factor of the best possible in terms of d_N and n.

THEOREM 11. *Choose a positive integer d. Then if $m = d\lceil \log_2(n+1) \rceil$, there exists $S \subseteq \{0, ..., n\}^m$ for which*

- $\Psi_N\text{-dim}(S) \leq d$.
- $\Psi_B\text{-dim}(S) = d\lfloor \log_2(n+1) \rfloor$.

Proof. In Appendix C. ∎

3. APPLICATIONS TO LEARNING

We move on to apply the results of Sections 2.2 and 2.3 to a natural extension of the PAC learning model. We describe a number of characterizations of learnability for classes of $\{0, ..., n\}$-valued functions proving, in particular, that for any distinguisher Ψ, a class is learnable if and only if its Ψ-dimension is finite. After Vapnik [18], we will adopt a naive attitude toward measurability, assuming that every set encountered in our proofs is measurable. If one prefers, one may assume that the domain of any probability space we describe is countable, although considerably weaker assumptions, similar to those used in [4, 7], suffice. If X is a set, P is a probability distribution over X, and f maps X to $\boldsymbol{R}$, let $E_{x \in P}[f(x)]$ denote the expectation of f with respect to P.

Choose a set X, a positive integer n, and a family $\mathcal{F}$ of $\{0, ..., n\}$-valued functions defined on X. For a probability measure D over X and a function $f \in \mathcal{F}$ we define the error of a function h with respect to D and f, denoted by $\text{error}_{D,f}(h)$, to be

$$D\{x \in X : f(x) \neq h(x)\}.$$

A *learning strategy* for $\mathcal{F}$ is a mapping from finite sequences of elements of $X \times \{0, ..., n\}$ to $\mathcal{F}$.

Intuitively, our definition of learnability requires the existence of a learning strategy able to yield an arbitrarily good approximation of any target function in the class with high probability with respect to a random sample of finite size.

More formally, we say that $\mathcal{F}$ is *learnable* if there exists a learning strategy A (not necessarily computable) and an integer-valued function $m = m(\varepsilon, \delta)$ such that for any $\varepsilon, \delta > 0$, for any probability measure D over X and for any $f \in \mathcal{F}$ the event

$$\text{error}_{D,f}(A(\bar{v})) > \varepsilon$$

occurs with probability at most δ for random sequences $\bar{v} = ((x_1, f(x_1)), ..., (x_m, f(x_m)))$, where $(x_1, ..., x_m) \in X^m$ is drawn according to D^m.

This definition of learnability is essentially that studied in [10], which in turn was based on Valiant's PAC model [16]. As we discuss in Section 4, the characterizations of learnability we present here also hold for more general and perhaps more realistic "loss functions." For instance, the "loss" on x can be allowed to depend on the extent the target's value $f(x)$ differs from the value $h(x)$ of the learner's hypothesis.

In order to apply the results of the previous section, we now extend the notion of the dimension of a set of vectors to sets of functions.

For a finite sequence $\bar{x} = (x_1, ..., x_k)$ of elements of X define the $\bar{x}$-restriction of $\mathcal{F}$ by

$$\mathcal{F}|_{\bar{x}} = \{(f(x_1), ..., f(x_k)) : f \in \mathcal{F}\}.$$

For a class Ψ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ define the Ψ-dimension of $\mathcal{F}$ (denoted by $\Psi\text{-dim}(\mathcal{F})$) to be the maximum over all positive integers k and all $\bar{x} \in X^k$ of the Ψ-dimension of its $\bar{x}$-restriction, if such a maximum exists,

and infinity otherwise. Define the uniform Ψ-dimension of $\mathcal{F}$ analogously.

As a first step, we mention the following result showing that the finiteness of the Natarajan dimension is necessary for learning.

THEOREM 12 [6, 10]. *If Ψ_N-dim$(\mathcal{F}) = \infty$ then $\mathcal{F}$ is not learnable.*

The next theorem follows relatively straightforwardly from the results obtained in the previous section.

THEOREM 13. *Choose distinguishers Ψ and Φ. Then the following are equivalent:*

1. Ψ-dim$(\mathcal{F}) = \infty$.
2. Φ-dim$(\mathcal{F}) = \infty$.
3. Ψ-dim$_U(\mathcal{F}) = \infty$.
4. Φ-dim$_U(\mathcal{F}) = \infty$.

Proof. $(1 \Rightarrow 2)$ Assume for contradiction that Ψ-dim$(\mathcal{F}) = \infty$ and Φ-dim$(\mathcal{F})$ is finite. Let $d = \Psi$-dim$(\mathcal{F})$. Let m_1 and $\bar{x} = (x_1, ..., x_{m_1})$ be such that that

$$\Phi\text{-dim}(\mathcal{F}|_{\bar{x}}) = d.$$

Let m_2 and $\bar{y} = (y_1, ..., y_{m_2})$ be such that

$$\Psi\text{-dim}(\mathcal{F}|_{\bar{y}}) = \lceil 1 + 4.67d \log_2(n+1) \rceil.$$

Let $\bar{z} = (x_1, ..., x_{m_1}, y_1, ..., y_{m_2})$. Let $S = \mathcal{F}|_{\bar{z}}$. Trivially, Φ-dim$(S) = d$ and

$$\Psi\text{-dim}(S) > 4.67d \log_2(n+1).$$

Applying Corollary 6 we have that

$$\Psi_B\text{-dim}(S) > 4.67 \Psi_N\text{-dim}(S) \log_2(n+1),$$

but by Theorem 10 this is a contradiction.

$(2 \Rightarrow 1)$ This follows from $(1 \Rightarrow 2)$ by symmetry.

$(1 \Rightarrow 3)$ Assume for contradiction that Ψ-dim$(\mathcal{F}) = \infty$ and Ψ-dim$_U(\mathcal{F})$ is finite. Let $d = \Psi$-dim$_U(\mathcal{F})$. Let m_1 and $\bar{x} = (x_1, ..., x_{m_1})$ be such that

$$\Psi\text{-dim}_U(\mathcal{F}|_{\bar{x}}) = d.$$

Let m_2 and $\bar{y} = (y_1, ..., y_{m_2})$ be such that

$$\Psi\text{-dim}(\mathcal{F}|_{\bar{y}}) > |\Psi|\, d.$$

Let $\bar{z} = (x_1, ..., x_{m_1}, y_1, ..., y_{m_2})$. Let $S = \mathcal{F}|_{\bar{z}}$. Again, trivially, Ψ-dim$_U(S) = d$ and Ψ-dim$(S) > |\Psi|\, d$, which contradicts Theorem 7.

$(2 \Rightarrow 4)$ This follows immediately from $(1 \Rightarrow 3)$.

Finally, $(3 \Rightarrow 1)$ and $(4 \Rightarrow 2)$ follow immediately from Lemma 4. This completes the proof. ∎

Combining Theorem 12 with Theorem 13, we can show that the finiteness of any dimension defined by a distinguisher is necessary for learning.

COROLLARY 14. *Let Ψ be a distinguisher. If Ψ-dim$(\mathcal{F}) = \infty$, then $\mathcal{F}$ is not learnable.*

We now turn to the definition of a class of $\{0, 1\}$-valued "loss functions" for $\mathcal{F}$, whose combinatorial and statistical properties are related to the learnability of $\mathcal{F}$ in a way that will be shown later.

Define the $0-1$ *loss function* l_{0-1} from $\{0, ..., n\}^2$ to $\{0, 1\}$ by

$$l_{0-1}(a, b) = \begin{cases} 1, & \text{if } a \neq b \\ 0, & \text{otherwise.} \end{cases}$$

For any function f from X to $\{0, ..., n\}$ define the function $l_{0-1,f}$ from $X \times \{0, ..., n\}$ to $\{0, 1\}$ by $l_{0-1,f}(x, a) = l_{0-1}(f(x), a)$. Finally, define the class $l_{0-1,\mathcal{F}} = \{l_{0-1,f} : f \in \mathcal{F}\}$ of $0-1$ *loss functions induced by* $\mathcal{F}$.

LEMMA 15. *The VC-dimension of $l_{0-1,\mathcal{F}}$ equals the graph dimension of $\mathcal{F}$.*

Proof. Suppose that the sequence $x_1, ..., x_k$ of elements of X are Ψ_G-shattered by $\mathcal{F}$. Then there exist $\psi_1, ..., \psi_k \in \Psi_G$ such that

$$\{(\psi_1(f(x_1)), ..., (\psi_k(f(x_k)))\} = \{0, 1\}^k.$$

Let $a_1, ..., a_k \in \{0, ..., n\}$ be such that for all j, $1 \leq j \leq k$, ψ_j is defined by

$$\psi_j(b) = \begin{cases} 1, & \text{if } b = a_j \\ 0, & \text{otherwise.} \end{cases}$$

Such a sequence $a_1, ..., a_k$ exists due to the definition of Ψ_G-shattering. We claim that the sequence $(x_1, a_1), ..., (x_k, a_k)$ of elements of $X \times \{0, ..., n\}$ is VC-shattered by $l_{0-1,\mathcal{F}}$. Choose $\bar{b} \in \{0, 1\}^k$. Let $f \in \mathcal{F}$ be such that

$$\bar{b} = (1 - \psi_1(f(x_1)), ..., 1 - \psi_k(f(x_k))).$$

Since, by definition, for all j, $1 \leq j \leq k$, $l_{0-1,f}(x_j, a_j) = 1 - \psi_j(f(x_j))$, we have

$$\bar{b} = (l_{0-1,f}(x_1, a_1), ..., l_{0-1,f}(x_k, a_k)).$$

Since $\bar{b}$ was chosen arbitrarily, $l_{0-1,\mathcal{F}}$ shatters $(x_1, a_1), ..., (x_k, a_k)$. Thus the VC-dimension of $l_{0-1,\mathcal{F}}$ is at least the graph dimension of $\mathcal{F}$.

Now assume that a sequence $(x_1, a_1), ..., (x_k, a_k)$ of elements of $X \times \{0, ..., n\}$ is shattered by $l_{0-1,\mathcal{F}}$. We claim that $x_1, ..., x_k$ is Ψ_G-shattered by $\mathcal{F}$. Define $\psi_1, ..., \psi_k \in \Psi_G$, by

$$\psi_j(b) = \begin{cases} 1, & \text{if } b = a_j \\ 0, & \text{otherwise.} \end{cases}$$

Applying the fact that for all j, $1 \leq j \leq k$, $l_{0-1,f}(x_j, a_j) = 1 - \psi_j(f(x_j))$, in a similar manner to the above verifies that $x_1, ..., x_k$ is Ψ_G-shattered by $\mathcal{F}$, and therefore the graph dimension of $\mathcal{F}$ is at least the VC-dimension of $l_{0-1,\mathcal{F}}$. This completes the proof. ∎

Next we define a statistical property of $l_{0-1,\mathcal{F}}$ that in the next theorem will be shown equivalent to learning. We say that $l_{0-1,\mathcal{F}}$ is *uniformly convergent* if there exists an integer-valued function $m = m(\varepsilon)$ such that for all $\varepsilon > 0$ and for all probability measures P over $X \times \{0, ..., n\}$, if $((x_1, a_1), ..., (x_m, a_m))$ is chosen according to P^m, the event

$$\exists f \in \mathcal{F} : \left| \frac{1}{m} \sum_{j=1}^{m} l_{0-1,f}(x_j, a_j) - \boldsymbol{E}_{(x,a)\in P}[l_{0-1,f}(x, a)] \right| \geq \varepsilon$$

occurs with probability at most ε. Since we require that the same m be sufficient for all distributions P, this is sometimes called *distribution-free* uniform confergence.

Now we are ready for our main result which shows a variety of ways in which learnability can be characterized.

THEOREM 16. *For any distinguisher Ψ the following are equivalent:*

1. Ψ-dim$(\mathcal{F})$ *is finite.*
2. Ψ-dim$_U(\mathcal{F})$ *is finite.*
3. $l_{0-1,\mathcal{F}}$ *is uniformly convergent.*
4. *The VC-dimension of* $l_{0-1,\mathcal{F}}$ *is finite.*
5. $\mathcal{F}$ *is learnable.*

Proof. Theorem 13 implies that $(1 \Rightarrow 2)$. Corollary 14 implies that $(5 \Rightarrow 1)$. Lemma 15 and Theorem 13 imply that $(1 \Leftrightarrow 4)$. The implication $(4 \Rightarrow 3)$ is an immediate consequence of the results in [19] and the implication $(3 \Rightarrow 5)$ is a special case of [7, Lemma 1]. This completes the proof. ∎

We remark that a result essentially equivalent to implication $(4 \Rightarrow 5)$ was also shown in [11, Theorem 5.4, p. 114].

The concept of distinguisher is a kind of meta-characterization, as it characterizes those Ψ which in turn characterize learnability both through the finiteness of the Ψ-dimension and through the finiteness of the uniform Ψ-dimension. To see this, all that remains is to show that for any family Ψ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ which is not a distinguisher, neither the Ψ-dimension nor the uniform Ψ-dimension characterizes learnability.

LEMMA 17. *If Ψ is a family of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ which is not a distinguisher and if X is infinite, then there is a family $\mathcal{F}$ of functions from X to $\{0, ..., n\}$ which has Ψ-dimension 0 and has uniform Ψ-dimension 0, but which is not learnable.*

Proof. Suppose Ψ fails to distinguish $a_1, a_2 \in \{0, ..., n\}$. Then the set of all functions from X to $\{a_1, a_2\}$ trivially has Ψ-dimension and uniform Ψ-dimension 0. However, this class is isomorphic to the set of all $\{0, 1\}$-valued functions defined on X, which was shown in [4] to not be PAC-learnable if X is infinite. ∎

Say that a family Ψ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ provides a characterization of learnability if and only if for any family $\mathcal{F}$ of $\{0, ..., n\}$-valued functions the learnability of $\mathcal{F}$ is equivalent to the finiteness of either its Ψ-dimension or its uniform Ψ-dimension. Then Theorem 16 and Lemma 17 yield the following result.

THEOREM 18. *A family Ψ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$ provides a characterization of learnability if and only if Ψ is a distinguisher.*

3.1. *Reductions between Learning Problems*

We now show a different way in which distinguishers can be used to characterize learnability. A natural approach to learning many-valued functions is to represent them by sets of $\{0, 1\}$-valued functions. For example, a function $f : X \to \{0, ..., n\}$ can be represented by $\lceil \log(n+1) \rceil$ binary functions f_i, where $f_i(x)$ is the i th bit of $f(x)$. The problem of learning f can then be reduced to the problem of learning each function f_i for $i = 1, ..., \lceil \log(n+1) \rceil$. More generally speaking, reductions between learning problems can be built by representing sets of multi-valued functions through sets of $\{0, 1, *\}$-valued functions. This is done using some set Ψ of mappings from $\{0, ..., n\}$ to $\{0, 1, *\}$, whereby we represent each f in a class $\mathcal{F}$ of $\{0, ..., n\}$-valued functions through a set of $\{0, 1, *\}$-valued functions. In this section we show that whenever a set Ψ of such mappings is a distinguisher, then its representation of $\mathcal{F}$ preserves learnability. More precisely, the learnability of a class $\mathcal{F}$ is equivalent to the learnability of each set of $\{0, 1, *\}$-valued functions in the collection representing $\mathcal{F}$ through the distinguisher Ψ.

We begin by introducing a few preliminary definitions. For all $f : X \to \{0, ..., n\}$ and all $\psi : \{0, ..., n\} \to \{0, 1, *\}$ let ψ_f be the function from X to $\{0, 1, *\}$ defined by $\psi_f(x) = \psi(f(x))$. Moreover, for any class $\mathcal{F}$ of such functions f let $\psi_{\mathcal{F}} = \{\psi_f : f \in \mathcal{F}\}$.

The definition of VC-shattering (and therefore the associated notion of VC-dimension) can be trivially generalized to classes of $\{0, 1, *\}$-valued functions by insisting that a sequence $\bar{x}$ be VC-shattered if and only if

$\mathcal{F}|_{\bar{x}} = \{0,1\}^{|\bar{x}|}$ (as usual, *'s do not contribute to shattering). In the rest of the section we will use the above slightly extended definition of VC-dimension.

The next lemma relates the notions of uniform Ψ-dimension and VC-dimension.

LEMMA 19. *For all sets Ψ of mappings from $\{0, ..., n\}$ to $\{0,1,*\}$,*

$$\Psi\text{-dim}_U(\mathcal{F}) = \max\{\text{VC-dim}(\psi_{\mathcal{F}}) : \psi \in \Psi\}.$$

Proof. Immediate from the definitions. ∎

We then obtain the following characterization.

THEOREM 20. *For all distinguishers Ψ, $\mathcal{F}$ is learnable if and only if $\psi_{\mathcal{F}}$ is learnable for all $\psi \in \Psi$.*

Proof. Choose a distinguisher Ψ. By Theorem 16, $\mathcal{F}$ is learnable if and only if the uniform Ψ-dimension of $\mathcal{F}$ is finite. By Lemma 19, this is equivalent to the finiteness of the VC-dimension of $\psi_{\mathcal{F}}$ for all $\psi \in \Psi$. By the results of [4], for each $\psi \in \Psi$, $\psi_{\mathcal{F}}$ is learnable exactly when it has finite VC-dimension. This completes the proof. ∎

As a final remark to this section we point out that Theorem 16, as most of the results presented in this paper, holds also if we use a more general definition of distinguisher where each mapping ψ depends on both the domain and the range of the functions in $\mathcal{F}$. More formally, in this framework a distinguisher is a set Ψ of mappings ψ from $X \times \{0, ..., n\}$ to $\{0, 1, *\}$. A sequence $\bar{x}$ on X (say of length d) is Ψ-shattered by $\mathcal{F}$ whenever there exists a sequence $(\psi_1, ..., \psi_d) \in \Psi^d$ such that for any $\bar{b} \in \{0,1\}^d$ there exists a $f \in \mathcal{F}$ for which

$$(\psi_1(x_1, f(x_1)), ..., \psi_d(x_d, f(x_d))) = \bar{b}.$$

Accordingly, the Ψ-dimension of $\mathcal{F}$ is the length of the longest sequence $\bar{x}$ Ψ-shattered by $\mathcal{F}$.

4. FURTHER APPLICATIONS

In this section we describe how our results can be applied to more general and perhaps more realistic learning problems. We also apply the bounds of Section 2 to the derivation of sample complexities, that is, sample sizes sufficient for learning in our framework.

4.1. *Robust Learning*

For practical learning tasks one might want to relax the hypothesis that the pairs $(x, a) \in X \times \{0, ..., n\}$ in the sample are generated according to a function belonging to some known class, or even that there exists a functional relationship between the x's and a's. A more realistic assumption might be to require the existence of a probability distribution on $X \times \{0, ..., n\}$ from which all pairs in the sample are independently drawn. In such learning settings, which are usually called "robust," the learner's goal might be to find a good approximation, according to a proper criterion, of the unknown distribution within the class $\mathcal{F}$ of hypotheses. In this section we prove that learnability and robust learnability are in fact equivalent properties of classes of $\{0, ..., n\}$-valued functions.

In the robust variant of our learning model, the error of a function $h \in \mathcal{F}$ is defined with respect to a distribution P over the set $X \times \{0, ..., n\}$ by

$$\text{error}_P(h) = P\{(x, a) : f(x) \neq a\}.$$

We say that a class $\mathcal{F}$ of $\{0, ..., n\}$-valued functions is *robustly learnable* if there exists a learning strategy A (again not necessarily computable) and an integer-valued function $m = m(\varepsilon, \delta)$ such that for any $\varepsilon, \delta > 0$ and for any probability measure P over $X \times \{0, ..., n\}$, the event

$$\text{error}_P(A(\bar{v})) > \inf_{f \in \mathcal{F}} \text{error}_P(f) + \varepsilon$$

occurs with probability at most δ over all samples $\bar{v} \in (X \times \{0, ..., n\})^m$ drawn according to P^m (the m-fold product measure derived from P).

This definition of learnability is a restriction of that studied in [17, 7] and we refer the interested reader to these sources for additional motivation. Using the results of previous sections we can quickly prove the following theorem.

THEOREM 21. *$\mathcal{F}$ is learnable if and only if $\mathcal{F}$ is robustly learnable.*

Proof. The implication $(3 \Rightarrow 5)$ in Theorem 16 holds also in the robust learning model. For the other direction just observe that learnability is clearly implied by robust learnability. ∎

Note that the equivalence between learnability and robust learnability could have been more directly demonstrated by combining Natarajan's [10] and Haussler's [7] results.

4.2. *General Loss Functions*

A more general error model than that of Section 3 can be considered. A natural choice could be a model in which certain errors are more serious than others. Call any function l from $\{0, ..., n\}^2$ to the nonnegative reals such that for any $a \in \{0, ..., n\}$, $l(a, a) = 0$ a *loss function*. Let the error of a function h with respect to a loss function l, a distribution D over X, and a function f be

$$\text{error}_{l,D,f}(h) = E_{x \in D}[l(f(x), h(x))].$$

We then say that $\mathcal{F}$ is *learnable w.r.t. a loss function* l if $\mathcal{F}$ is learnable according to the definition of learnability given in Section 3 with $\text{error}_{D,f}$ replaced by $\text{error}_{l,D,f}$. Note that the learnability of $\mathcal{F}$ is equivalent to the learnability of $\mathcal{F}$ w.r.t. the loss function l_{0-1} defined in Section 3.

THEOREM 22. *$\mathcal{F}$ is learnable if and only if $\mathcal{F}$ is learnable w.r.t. all loss functions.*

Proof. Since $\mathcal{F}$ is learnable if and only if $\mathcal{F}$ is learnable w.r.t. l_{0-1}, the "if" direction is trivial. For the other direction assume $\mathcal{F}$ is learnable w.r.t. l_{0-1} and choose a loss function l. Let M be the maximum value taken by l on its (finite) domain. Then for any $f, g \in \mathcal{F}$

$$\text{error}_{l,D,f}(g) \leqslant M \cdot \text{error}_{D,f}(g).$$

Therefore, for any $\varepsilon > 0$, distribution D, and functions $f, g \in \mathcal{F}$, $\text{error}_{D,f}(g) \leqslant \varepsilon/M$ implies that $\text{error}_{l,D,f}(g) \leqslant \varepsilon$. The theorem is proven. ∎

The proof of the "only if" part in Theorem 22 is analogous to the proof of [11, Theorem 5.6, p. 121], where a similar statement is proven for classes of real-valued functions using the notion of a metric instead of the more general notion of loss.

In Section 3 we introduced the class $l_{0-1,\mathcal{F}}$ of $0-1$ loss functions induced by $\mathcal{F}$. This notation can be extended to any loss function l as follows. Let l_f be the function on $X \times \{0, ..., n\}$ defined by

$$l_f(x, a) = l(f(x), a)$$

and let $l_{\mathcal{F}} = \{l_f : f \in \mathcal{F}\}$. Note that $l_{\mathcal{F}}$ is a class of functions with finite range. Note further that the definition of the Ψ_P dimension given in Section 2.1 can be extended to classes of real-valued functions, the union of whose ranges contains nonintegral values as follows.

For each real κ define $\psi_{P,\kappa} : \boldsymbol{R} \to \{0, 1\}$ by

$$\psi_{P,\kappa}(a) = \begin{cases} 1, & \text{if } a \geqslant \kappa \\ 0, & \text{otherwise.} \end{cases}$$

We then say that the sequence $(x_1, ..., x_k) \in X^k$ is Ψ_P-shattered by $\mathcal{F}$ iff there esists $\kappa_1, ..., \kappa_k$ such that

$$\{0, 1\}^k \subseteq \{(\psi_{P,\kappa_1}(f(x_1)), ..., \psi_{P,\kappa_k}(f(x_k))) : f \in \mathcal{F}\}$$

and define the Ψ_P-dimension of $\mathcal{F}$ to be the length of the longest sequence Ψ_P-shattered by $\mathcal{F}$.

We now extend Theorem 22 to robust learning. A preliminary lemma is needed.

LEMMA 23. *For all classes $\mathcal{F}$ and loss functions l the Ψ_P-dimension of $l_{\mathcal{F}}$ is at most the Ψ_B-dimension of $\mathcal{F}$.*

Proof. To prove the lemma is sufficient to show that for all positive integers d and all $\bar{z} \in (X \times \{0, ..., n\})^d$, if $l_{\mathcal{F}}$ Ψ_P-shatters $\bar{z} = (x_1, a_1), ..., (x_d, a_d)$, then $\mathcal{F}$ Ψ_B-shatters $\bar{x} = (x_1, ..., x_d)$. Assuming that $\bar{z}$ is Ψ_P-shattered by $l_{\mathcal{F}}$ amounts to saying that there is a sequence $\bar{r}$ of d positive reals such that for all $\bar{b} \in \{0, 1\}^d$ there is some $f \in \mathcal{F}$ satisfying

$$l(f(x_i), a_i) \geqslant r_i \Leftrightarrow b_i = 1 \quad \text{for} \quad i = 1, ..., d. \tag{1}$$

For each $i = 1, ..., d$, define the mapping ψ_i with domain $\{0, ..., n\}$ by

$$\psi_i(c) = \begin{cases} 1, & \text{if } l(c, a_i) \geqslant r_i \\ 0, & \text{otherwise.} \end{cases}$$

Since for $1 \leqslant i \leqslant d$, ψ_i maps $\{0, ..., n\}$ to $\{0, 1\}$, it belongs to the distinguisher Ψ_B, since Ψ_B contains all such mappings. Therefore, condition (1) implies that

$$\psi_i(f(x_i)) = 1 \Leftrightarrow l(f(x_i), a_i) \geqslant r_i \quad \text{for} \quad i = 1, ..., d.$$

Since such a set of ψ_i can be found in Ψ_B for all choices of $\bar{z}$ and $\bar{r}$, the lemma is proven. ∎

We will also use the following result.

THEOREM 24 [7]. *If the Ψ_P dimension of $l_{\mathcal{F}}$ is finite then $\mathcal{F}$ is robustly learnable.*

Now we are ready for the main result of this section.

THEOREM 25. *$\mathcal{F}$ is robustly learnable if and only if $\mathcal{F}$ is robustly learnable w.r.t. all loss functions.*

Proof. It is easily verified that the robust learnability of $\mathcal{F}$ is equivalent to the robust learnability of $\mathcal{F}$ w.r.t. l_{0-1}, and this proves the "if" part. For the other direction if $\mathcal{F}$ is robustly learnable, then $\mathcal{F}$ is learnable as well. Since Ψ_B is a distinguisher, by Theorem 16 the Ψ_B-dimension of $\mathcal{F}$ is finite. Choose a loss function l. By Lemma 23, the Ψ_P-dimension of $l_{\mathcal{F}}$ is finite. By Theorem 24, this implies the robust learnability of $\mathcal{F}$ w.r.t. l. ∎

4.3. *Sample Size Bounds*

The definition of learnability for a class $\mathcal{F}$ of functions calls for the existence of both a learning strategy and an integer-valued function $m = m(\varepsilon, \delta)$ satisfying the given learning criterion. The term *PAC learning sample complexity* is often used to denote the slowest growing function m for which such a learning strategy exists. By generalizing results from [4], we now prove upper bounds on the PAC learning sample complexity for the multi-valued case.

THEOREM 26. *The PAC learning sample complexity of a class $\mathcal{F}$ of $\{0, ..., n\}$-valued functions is at most*

$$O\left(\frac{1}{\varepsilon}\left(d_G \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right), \tag{2}$$

where d_G is the graph dimension of $\mathcal{F}$.

Proof. Choose a class $\mathcal{F}$ of functions from X to $\{0, ..., n\}$ and let d_G be its graph dimension. Choose a distribution D on X and a target function $f \in \mathcal{F}$. Let P be the distribution induced on $X \times \{0, ..., n\}$ by D and f. By Lemma 15, the VC-dimension of the class $l_{0-1,\mathcal{F}}$ of $0-1$ loss functions on $X \times \{0, ..., n\}$ induced by $\mathcal{F}$ equals the graph dimension of $\mathcal{F}$. Now observe that the problem of learning the target function $f \in \mathcal{F}$ to within accuracy $\varepsilon > 0$ reduces to the problem of identifying any $f \in \mathcal{F}$ for which $\Pr(l_{0-1,f} = 1) \leqslant \varepsilon$. By the results of [4], a sample size of order as specified in formula (2) is sufficient to ensure that, with probability at least $1 - \delta$, any hypothesis h for which $l_{0-1,h}(x, a) = 0$ for all pairs (x, a) in the sample achieves this goal. ∎

We can generalize Theorem 26 and obtain a similar bound in terms of the Ψ-dimension and the uniform Ψ-dimension of $\mathcal{F}$ for each distinguisher Ψ.

THEOREM 27. *Choose a class $\mathcal{F}$ of $\{0, ..., n\}$-valued functions and a distinguisher Ψ. Then the PAC learning sample complexity of $\mathcal{F}$ is at most*

$$O\left(\frac{1}{\varepsilon}\left(d(\log n)\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right), \tag{3}$$

where d is the Ψ-dimension of $\mathcal{F}$ and at most

$$O\left(\frac{1}{\varepsilon}\left(u\,|\Psi|\,(\log n)\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right), \tag{4}$$

where u is the uniform Ψ-dimension of $\mathcal{F}$.

Proof. Corollary 6 and Theorem 10 imply that for any class $\mathcal{F}$ of Ψ-dimension d and Φ-dimension d', where Ψ and Φ are two distinguishers, $d' = O(d \log n)$ holds. Therefore, by Theorem 26 the upper bound (3) holds whenever $\mathcal{F}$ has Ψ-dimension d with respect to some distinguisher Ψ. Finally, the bound (4) is an immediate consequence of Theorem 7. ∎

Theorems 26 and 27 can both be easily extended to arbitrary loss functions by replacing each $1/\varepsilon$ with $(\max_{a,b} l(a, b))/\varepsilon$.

Regarding the second part of the above theorem, observe that the size of a distinguisher can have very different rates of growth with respect to n. For instance, $|\Psi_B|$ is exponential in n, whereas $|\Psi_P|$ is linear. Also, there are distinguishers whose size is logarithmic in n: Let Ψ_L be the set defined by

$$\{\psi_{L,k} : k \leqslant \lceil \log_2(n+1) \rceil\}, \tag{5}$$

where $\psi_{L,k}(j)$ is the kth least significant bit in a binary encoding of j. Then Ψ_L has size $\lceil \log_2(n+1) \rceil$ and is a distinguisher, as one can easily verify.

5. CONCLUSIONS AND OPEN PROBLEMS

In this work we gave a general scheme for extending the VC-dimension to classes of multi-valued functions and we proved a combinatorial condition characterizing those generalizations of the VC-dimension whose finiteness is necessary and sufficient for learning in a natural extension of the PAC framework. We also provided further characterizations of learnability for classes of multi-valued functions in terms of the VC-dimension and in terms of the uniform convergence property of a class of induced loss functions. We then proved equivalence between learning and robust learning and independence of our notion of learnability on the choice of the loss function. We finally showed applications of these results to the problem of estimating the PAC learning sample complexity sufficient for learning with consistent hypotheses.

A possible direction for future work is the investigation of the relationships of the results proven here to the real-valued case. In fact certain notions of dimension, such as the Pollard's Ψ_P-dimension, are naturally extended to classes of real-valued functions taking values in a bounded real interval (say $[0, 1]$ for simplicity). A real-valued learning problem can be reduced to the multi-valued case through a discretization of the range $[0, 1]$ into a set $\{0, ..., n\}$ (see [8]). The number of discrete elements into which the continuous range is broken is proportional to $1/\varepsilon$, where ε is the required bound on the error of the hypothesis. Also, the discretization does not increase the Ψ_P-dimension of the original class, so the finiteness of either the Ψ_P-dimension or Vapnik's uniform Ψ_P-dimension are sufficient for robust learning in the real-valued case. On the other hand, some properties true in the discrete case are lost in the continuous one. For instance, while in the discrete case, the finiteness of the Ψ_P-dimension is equivalent to the finiteness of the uniform Ψ_P dimension, the class of monotone increasing functions on $[0, 1]$ has infinite Ψ_P-dimension but uniform Ψ_P-dimension equal to 1.

Recent results [3] show that the equivalence between learning and robust learning does not carry on to the real-valued case. Namely, there are classes of $[0, 1]$-valued functions which are learnable but not in a robust way.

APPENDIX A: PROOF OF THEOREM 1

Suppose that

$$S = \{(0, 0), (0, 1), (1, 1), (1, 2)\}$$
$$T = \{(0, 1), (1, 2), (2, 3), (3, 0)\}.$$

If $\bar{\psi} = (\psi_{G,1}, \psi_{G,1})$, then

$$\bar{\psi}(S) = \{(0, 0), (0, 1), (1, 1), (1, 0)\} = \{0, 1\}^2,$$

and $(1, 2)$ is Ψ_G-shattered by S.

Assume for contradiction that $(1, 2)$ is Ψ_P-shattered by S. Then there exists $\bar{\psi} = (\psi_{P,k}, \psi_{P,l})$, where $k, l \in \{1, 2, 3\}$ such that $\{0, 1\}^2 \subseteq \bar{\psi}(S)$. Assume as a first case that $k \geq l$. Since there is $(u, v) \in S$ such that $(1, 0) \in \bar{\psi}(S)$, $u \geq k \geq l > v$, contradicting the easily observed fact that $u \leq v$ for all $(u, v) \in S$. Assume as a second case that $k < l$. Since there is $(u, v) \in S$ such that $(0, 1) \in \bar{\psi}(S)$, $v \geq l > k > u$, contradicting another easily observed fact: that $u \geq v - 1$ for all $(u, v) \in S$. Therefore $(1, 2)$ is not Ψ_P-shattered by S.

If $\bar{\psi} = (\psi_{P,2}, \psi_{P,2})$, then

$$\bar{\psi}(T) = \{(0, 0), (0, 1), (1, 1), (1, 0)\} = \{0, 1\}^2$$

and $(1, 2)$ is Ψ_P-shattered by T.

Choose $\bar{\psi} = (\psi_{G,k}, \psi_{G,l})$, where $(k, l) \in T$. Then for all $\bar{t} \in T$ for which $\bar{t} \neq (k, l)$, $\bar{\psi}(\bar{t}) = (0, 0)$. If $(k, l) \notin T$, and $\bar{\psi} = (\psi_{G,k}, \psi_{G,l})$, then for no $\bar{t} \in T$ is $\bar{\psi}(\bar{t}) = (1, 1)$. Thus, $(1, 2)$ is not Ψ_G-shattered by T. This completes the proof. ∎

This result may trivially be extended to arbitrary $m > 1$, $n > 2$.

APPENDIX B: PROOF OF THEOREM 8

Let $n = r$, $m = dn$, and $\Psi = \Psi_P$. Let

$$J_0 = \{1, ..., n\},$$
$$J_1 = \{n + 1, ..., 2n\}, ..., J_{d-1} = \{(d - 1)n + 1, ..., dn\}.$$

Let S be the set of all $\bar{s} \in \{0, ..., n\}^m$ for which for all $j \in \{0, ..., d - 1\}$, for all $u, v \in J_j$, if $u < v$, then $s_u \leq s_v$.

Note that part 1 of Theorem 8 is satisfied by the assumption $\Psi = \Psi_P$. We now split the proof of parts 2 and 3 into a number of lemmas.

LEMMA 28. *The uniform Ψ_P-dimension of S is at least d.*

Proof. We claim that $(n, 2n, ..., dn)$ is uniformly Ψ_P-shattered by S. Choose $\bar{b} \in \{0, 1\}^d$. Define $\bar{s} \in \{0, ..., n\}^m$ by

$$s_i = \begin{cases} 0, & \text{if } n \text{ does not divide } i, \text{ or if } b_{\lfloor i/n \rfloor} = 0 \\ 1, & \text{otherwise.} \end{cases}$$

Clearly $\bar{s} \in S$, and

$$(\Psi_{P,1}(s_n), \Psi_{P,1}(s_{2n}), ..., \Psi_{P,1}(s_{dn})) = \bar{b}.$$

Since $\bar{b}$ was chosen arbitrarily

$$\{0, 1\}^d \subseteq \{(\Psi_{P,1}(s_n), \Psi_{P,1}(s_{2n}), ..., \Psi_{P,1}(s_{dn})) : \bar{s} \in S\}.$$

Thus $(n, 2n, ..., dn)$ is uniformly Ψ_P-shattered by S, completing the proof. ∎

LEMMA 29. *Choose a positive integer q. If S uniformly Ψ_P-shatters $\bar{i} \in \{1, ..., m\}^q$, then for each $1 \leq j \leq d$,*

$$|\{i_l : 1 \leq l \leq q\} \cap J_j| \leq 1.$$

Proof. Assume the negation of the lemma for contradiction. Let $1 \leq k \leq n$ be such that

$$\{0, 1\}^q \subseteq \{(\Psi_{P,k}(s_{i_1}), \Psi_{P,k}(s_{i_2}), ..., \Psi_{P,k}(s_{i_q})) : \bar{s} \in S\}$$

and let j be such that

$$|\{i_{j'} : 1 \leq j' \leq q\} \cap J_j| > 1.$$

Let u and v be distinct elements of J_j for which $u < v$. We have

$$\{0, 1\}^2 \subseteq \{(\Psi_{P,k}(s_u), \Psi_{P,k}(s_v)) : \bar{s} \in S\}.$$

In particular,

$$(1, 0) \in \{(\Psi_{P,k}(s_u), \Psi_{P,k}(s_v)) : \bar{s} \in S\}.$$

Thus,

$$s_u \geq k \quad \text{and} \quad s_v < k,$$

which implies

$$s_u > s_v,$$

which, by the definition of S, is a contradiction since $u, v \in J_j$ and $u < v$. This completes the proof. ∎

THEOREM 30 (Requirement (2) from Theorem 8). *The uniform Ψ_P-dimension of S is d.*

Proof. Follows immediately from Lemmas 28 and 29. ∎

COROLLARY 31. *The Ψ_P-dimension of S is at most dn.*

Proof. Follows immediately from Theorem 30 and Theorem 7. ∎

LEMMA 32. *The Ψ_P-dimension of S is at least dn.*

Proof. We claim that $(1, ..., dn)$ is shattered. Define $\bar{\psi}$ by letting

$$\psi_i = \psi_{P,(i-n\lfloor (i-1)/n \rfloor)},$$

for $i = 1, 2, ..., dn$. Thus

$$\bar{\psi} = (\psi_{P,1}, ..., \psi_{P,n}, \psi_{P,1}, ..., \psi_{P,n}, ..., \psi_{P,1}, ..., \psi_{P,n}).$$

Choose $\bar{b} \in \{0, 1\}^{dn}$. Define $\bar{s} \in \{0, ..., n\}^{dn}$ by

$$s_i = \begin{cases} (i-1) - n\lfloor (i-1)/n \rfloor, & \text{if } b_i = 0 \\ i - n\lfloor (i-1)/n \rfloor, & \text{otherwise.} \end{cases}$$

Since $s_i \leqslant i - n\lfloor (i-1) \rfloor$ and $s_{i+1} \geqslant i - n\lfloor i/n \rfloor$, if i and $i+1$ are both in J_j for some j, $\lfloor (i-1)/n \rfloor = \lfloor i/n \rfloor$, and therefore $s_i \leqslant s_{i+1}$. Therefore $\bar{s} \in S$. Choose $i \in \{1, ..., dn\}$. Then, by definition,

$$s_i \geqslant i - n\lfloor (i-1)/n \rfloor \quad \text{iff } b_i = 1,$$

and, therefore,

$$\psi_i(s_i) = \psi_{P,(i-n\lfloor (i-1)/n \rfloor)}(s_i) = b_i.$$

Since $\bar{b}$ was chosen arbitrarily,

$$\{0, 1\}^{dn} \subseteq \bar{\psi}(S).$$

This completes the proof. ∎

THEOREM 33 (Requirement (3) from Theorem 8). *The Ψ_P dimension of S is dn.*

Proof. Follows from Corollary 31 and Lemma 32. ∎

APPENDIX C: PROOF OF THEOREM 11

Let Ψ_L be the set defined by

$$\{\psi_{L,k} : 1 \leqslant k \leqslant \lceil \log_2(n+1) \rceil\}, \tag{6}$$

where $\psi_{L,k}(r)$ is the kth least significant bit in a binary encoding of r.

Define $\gamma : \{0, ..., n\}^d \to \{0, ..., n\}^m$ as follows (recall that $m = d\lceil \log_2(n+1) \rceil$):

Suppose that $\gamma(\bar{z}) = \bar{s}$. Informally, $\bar{s}$ is formed by first writing the binary representation of z_1 using the alphabet $\{0, z_1\}$, then writing the binary representation of z_2 using the alphabet $\{0, z_2\}$, and so on, and then "concatenating" the results. (If some $z_j = 0$, we just put $\lceil \log_2(n+1) \rceil$ zero's in its place.) Formally, for each $1 \leqslant i \leqslant m$, if $j = \lfloor (i-1)/\lceil \log_2(n+1) \rceil \rfloor$, then

$$s_i = z_{j+1} \psi_{L,t}(z_{j+1}),$$

where $t = \lceil \log_2(n+1) \rceil - ((i+1) \bmod \lceil \log_2(n+1) \rceil)$. Note that γ is one-to-one. Let

$$S = \gamma(\{0, ..., n\}^d)$$

be the range of γ. For example, if $n = 3$, $d = 2$,

$$\begin{aligned} S = \{&(0,0,0,0), (0,0,0,1), (0,0,2,0), (0,0,3,3), \\ &(0,1,0,0), (0,1,0,1), (0,1,2,0), (0,1,3,3), \\ &(2,0,0,0), (2,0,0,1), (2,0,2,0), (2,0,3,3), \\ &(3,3,0,0), (3,3,0,1), (3,3,2,0), (3,3,3,3)\}. \end{aligned}$$

LEMMA 34. *Choose a positive integer q. If S Ψ_N-shatters $\bar{\imath} \in \{1, ..., m\}^q$, then for each $0 \leqslant j \leqslant d-1$,*

$$|\{j' : \lceil (i_{j'} - 1)/\lceil \log_2(n+1) \rceil \rfloor = j\}| \leqslant 1.$$

Proof. Assume the negation of the lemma for contradiction. Let $\bar{\psi} \in \Psi_N^q$ be such that

$$\{0, 1\}^q \subseteq \bar{\psi}(S|_{\bar{\imath}})$$

and let j be such that

$$|\{i_{j'} : \lfloor (i_{j'} - 1)/\lceil \log_2(n+1) \rceil \rfloor = j\}| > 1.$$

Let i_u and i_v be distinct elements of

$$\{i_{j'} : \lfloor (i_{j'} - 1)/\lceil \log_2(n+1) \rceil \rfloor = j\}.$$

Let k_u, l_u, k_v, l_v satisfy

$$\psi_u = \psi_{N,k_u,l_u}, \quad \psi_v = \psi_{N,k_v,l_v}.$$

Assume without loss of generality that $k_u < l_u$, $k_v < l_v$. Let $\bar{s} \in S$ be such that

$$(\psi_u(s_{i_u}), \psi_v(s_{i_v})) = (1, 1).$$

Let $\bar{z} = \gamma^{-1}(\bar{s})$. Since

$$\begin{aligned} s_{i_u} &= z_{j+1} \psi_{L,t_u}(z_{j+1}) \\ s_{i_v} &= z_{j+1} \psi_{L,t_v}(z_{j+1}), \end{aligned}$$

where

$$t_u = \lceil \log_2(n+1) \rceil - ((i_u + 1) \bmod \lceil \log_2(n+1) \rceil)$$

and

$$t_v = \lceil \log_2(n+1) \rceil - ((i_v + 1) \bmod \lceil \log_2(n+1) \rceil).$$

Also $l_u, l_v > 0$ (since $l_u > k_u$, $l_v > k_v$), we have

$$s_{i_u} = l_u = z_{j+1}$$
$$s_{i_v} = l_v = z_{j+1}$$

and $l_u = l_v = z_{j+1}$.

Let $\vec{s}' \in S$ be such that

$$(\psi_u(s'_{i_u}), \psi_v(s'_{i_v})) = (0, 1).$$

Let $\vec{z}' = \gamma^{-1}(\vec{s}')$. Since $\psi_u(s'_u) \neq \psi_u(s_u)$, the binary representation of z_{j+1} is not the same as that of z'_{j+1}, and therefore $z_{j+1} \neq z'_{j+1}$. But since $\psi_v(s'_v) = 1$, $s'_v > 0$, and therefore $s'_v = z'_{j+1}$ and $l_v = z'_{j+1}$. Thus, we have

$$l_v = z'_{j+1} \neq z_{j+1} = l_v,$$

a contradiction. This completes the proof. ∎

COROLLARY 35. *The Ψ_N-dimension of S is at most d.*

THEOREM 36. *The uniform Ψ_P-dimension of S is at least $d\lfloor \log(n+1) \rfloor$.*

Proof. Assume first that $(n+1)$ is a power of 2. In this case

$$\lfloor \log_2(n+1) \rfloor = \lceil \log_2(n+1) \rceil = \log_2(n+1).$$

Choose $\bar{b} \in \{0,1\}^m$. Let z_1 be the number represented in binary by the first $\log_2(n+1)$ bits of $\bar{b}$, let z_2 be the number represented by the next $\log_2(n+1)$ bits, and so on. Let $\bar{s} = \gamma(\bar{z})$. Trivially,

$$(\psi_{P,1}(s_1), \ldots, \psi_{P,1}(s_m)) = \bar{b}.$$

Since $\bar{b}$ was chosen arbitrarily,

$$\{0,1\}^m \subseteq \{(\psi_{P,1}(s_1), \ldots, \psi_{P,1}(s_m)) : s \in S\},$$

completing the proof in this case.

If $(n+1)$ is not a power of two, then

$$\lfloor \log_2(n+1) \rfloor = \lceil \log_2(n+1) \rceil - 1.$$

One may easily show in an analogous way that

$$(2, \ldots, \lceil \log_2(n+1) \rceil, \lceil \log_2(n+1) \rceil + 2, \ldots, 2\lceil \log_2(n+1) \rceil, \ldots, (d-1)\lceil \log_2(n+1) \rceil + 2, \ldots, d\lceil \log_2(n+1) \rceil)$$

is uniformly Ψ_P-shattered by S, completing the proof. ∎

COROLLARY 37. *The Ψ_B-dimension of S is at least $d\lfloor \log(n+1) \rfloor$.*

Proof. Follows directly from Lemma 4, Corollary 6, and Theorem 36. ∎

ACKNOWLEDGMENTS

Part of this research was done while Nicolò Cesa-Bianchi was visiting U.C. Santa Cruz partially supported by the "Progetto finalizzato sistemi informatici e calcolo parallelo" of CNR under Grant 91.00884.69.115.09672. David Haussler was supported by ONR Grant NO0014-91-J-1162 and NSF Grant IRI-9123692. Phil Long was supported first by a UCSC Chancellor's dissertation-year fellowship and later by a Lise Meitner postdoctoral fellowship from the Fonds zur Förderung der wissenschaftlichen Forschung (Austria).

REFERENCES

1. N. Alon, On the density of sets of vectors, *Discrete Math.* **24** (1983), 177–184.
2. R. P. Anstee, A forbidden configuration theorem of Alon, *J. Combin. Theory Ser. A* **47** (1988), 16–27.
3. P. L. Barlett, P. M. Long, and R. C. Williamson, Fat-shattering and the learnability of real-valued functions, manuscript, 1993.
4. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, *J. Assoc. Comput. Mach.* **36**, No. 4 (1989), 929–965.
5. R. M. Dudley, Universal Donsker classes and metric entropy, *Ann. Probab.* **15**, No. 4 (1987), 1306–1326.
6. A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant, A general lower bound on the number of examples needed for learning, *Inform. and Comput.* **82**, No. 3 (1989), 247–251.
7. D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. and Comput.* **100** (1992), 78–150.
8. D. Haussler and P. M. Long, "A Generalization of Sauer's Lemma," Technical Report UCSC-CRL-90-15, U.C. Santa Cruz, 1990.
9. M. G. Karpovsky and V. D. Milman, Coordinate density of sets of vectors, *Discrete Math.* **24** (1978), 177–184.
10. B. K. Natarajan, On learning sets and functions, *Mach. Learning* **4** (1989), 67–97.
11. B. K. Natarajan, "Machine Learning: A Theoretical Approach," Morgan Kaufmann, San Mateo, CA, 1991.
12. D. Pollard, "Convergence of Stochastic Processes," Springer-Verlag, New York/Berlin, 1984.
13. D. Pollard, "Empirical Processes: Theory and Applications, of NSF–CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Institute of Math. Statist. and Am. Stat. Assoc., Alexandria, VA, 1990.
14. J. Shawe-Taylor, M. Anthony, and R. L. Biggs, "Bounding Sample Size with the Vapnik–Chervonenkis Dimension," Technical Report CSD-TR-618, University of London, Royal Halloway and Bedford New College, 1989.
15. J. M. Steele, Existence of submatrices with all possible columns, *J. Combin. Theory Ser. A* **24** (1978), 84–88.
16. L. Valiant, A theory of the learnable, *Comm. ACM* **27**, No. 11 (1984), 1134–1142.
17. V. N. Vapnik, "Estimation of Dependences Based on Empirical Data," Springer Verlag, Berlin/New York, 1982.
18. V. N. Vapnik, Inductive principles of the search for empirical dependences, *in* "Proceedings, 2nd Annual Workshop on Computational Learning Theory, 1989."
19. V. N. Vapnik and A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab. Appl.* **16**, No. 2 (1971), 264–280.