# Some Contributions to Fixed-Distribution Learning Theory

M. Vidyasagar and Sanjeev R. Kulkarni

***Abstract*—In this paper, we consider some problems in learning with respect to a fixed distribution. We introduce two new notions of learnability; these are probably uniformly approximately correct (PUAC) learnability which is a stronger requirement than the widely studied PAC learnability, and minimal empirical risk (MER) learnability, which is a stronger requirement than the previously defined notions of "solid" or "potential" learnability. It is shown that, although the motivations for defining these two notions of learnability are entirely different, these two notions are in fact equivalent to each other and, in turn, equivalent to a property introduced here, referred to as the shrinking width property. It is further shown that if the function class to be learned has the property that empirical means converge uniformly to their true values, then all of these learnability properties hold. In the course of proving conditions for these forms of learnability, we also obtain a new estimate for the VC-dimension of a collection of sets obtained by performing Boolean operations on a given collection; this result is of independent interest. We consider both the case in which there is an underlying target function, as well as the case of "model-free" (or agnostic) learning. Finally, we consider the issue of *representation* of a collection of sets by its subcollection of equivalence classes. It is shown by example that, by suitably choosing representatives of each equivalence class, it is possible to affect the property of uniform convergence of empirical probabilities.**

***Index Terms*—Fixed distribution, PAC learning, representation.**

## I. Introduction

DURING the past decade or so, there has been considerable interest among computer scientists, engineers, and probabilists in so-called probably approximately correct (PAC) learning theory. Although PAC learning theory was originally proposed as a mathematical model of machine learning and generalization, it soon became clear that an intimate connection exists between PAC learning theory and a much older branch of probability known as the theory of empirical processes. A significant difference between the two theories, however, is that, although a large part of empirical process theory deals with a *single fixed* probability, PAC learning theory is largely focused on the so-called *distribution-free* case, especially in the computer science community.

The PAC learning problem in the case of a single fixed probability, usually referred to as *fixed-distribution learning*, is by now relatively well-understood. One of the aims of the present paper is to point out that, in the fixed-distribution case, PAC learning is only one possible definition of learning, and that it is possible to define some alternate, and significantly stronger, forms of learning. The reason for introducing these stronger notions of learnability is that, if we use the results from empirical process theory to deduce learnability, then, in fact, we can conclude these much stronger forms of learnability. In particular, if a function class has the property that empirical means converge uniformly to their true values, then such a class also possesses these stronger learnability properties.

Almost all of the literature on fixed-distribution PAC learning is limited to the case in which the data driving the learning algorithm is generated by a "target" concept. Recently, however, there has been considerable interest in learning problems in which the data is not necessarily generated by a target concept. This type of learning is known by various names, such as the "loss function" approach, "agnostic" learning, etc. In this paper, this type of learning is referred to as "model-free" learning, and known results for the standard PAC learning problem are extended to the model-free setting.

A point often overlooked in fixed-distribution learning theory is the importance of *representation*. Given a class of concepts to be learned, it is natural that we would first partition the given concept class into a collection of *equivalence classes*, whereby two concepts are considered to be equivalent if they are at a distance of zero from each other. Then, we would replace the original concept class by a subset consisting of exactly one representative of each distinct equivalence class, because there is no point in trying to distinguish between two concepts at zero distance from each other. Clearly, the choice of the representative from each equivalence class is not unique in general. It is shown here that the choice of the representatives affects whether or not the resulting collection of sets has the property that empirical probabilities converge to their true values, but *does not* affect PAC learnability. Thus, in this sense, PAC learnability is a more "robust" notion than the uniform convergence of empirical means.

Until recently, very few points of contact have existed between (PAC-type) learning theory and system identification. During the past few years, however, there has been increasing interest in bringing about a rapprochement between the two areas; see, for example, [3]. Some subtle differences exist between the underlying problem formulations in the two areas, which need to be bridged to bring about such a rapprochement. Specifically, PAC-type learning theory in its "pure" form is ideally suited to the problem of identifying a memoryless nonlinearity excited by i.i.d. inputs, with possibly noisy

Manuscript received October 21, 1997; revised November 1, 1998. Recommended by Associate Editor, J. Farrell.

M. Vidyasagar is with the Centre for Artificial Intelligence and Robotics, Raj Bhavan Circle, High Grounds, Bangalore 560 001, India (e-mail: sagar@cair.res.in).

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kulkarni@ee.princeton.edu).

Publisher Item Identifier S 0018-9286(00)01063-1.

measurements of the resulting outputs. No natural notion of "time" in conventional PAC learning theory exists; indeed, introducing dynamics into PAC-type learning models is stated as an open problem in [36, Ch. 12]. If we attempt to identify Hammerstein-type nonlinear systems of the type studied in [15] and [21], the inputs to the memoryless nonlinearity are no longer independent. In [9] and [10], an attempt is made to extend the PAC-type learning formulation to wide sense stationary inputs, instead of i.i.d. inputs. Another related approach is that of [22] and [23], whereby a "universal" predictor is proposed for a large class of regular systems with fading memory. In some sense, the problem studied in [22] and [23] is close to that in PAC-type learning theory, in that the emphasis is on making a good prediction of the next output, instead of on identifying the underlying system itself. In [13], PAC-type learning theories are worked out for input–output mappings of recurrent perceptron networks (in contrast with the nonrecurrent perceptron networks more commonly studied). In [12] and [28], the inputs or outputs are assumed to be corrupted by "deterministic noise," and *worst-case* estimates are derived for the complexity of identification; not surprisingly, these estimates are terribly conservative. In [2], an attempt is made to reduce this conservatism via a two-step approach. First, complexity estimates for identification are derived *without* making any assumptions about the measurement noise, by the covering numbers of the class of systems; it is further assumed that the systems all have "fading memory," so that each system can be well approximated by another system whose output depends only on a *finite number* of past inputs. Second, by using standard inequalities in probability theory, the probability that the identification process fails is estimated. These and other papers represent the first steps in the development of a PAC approach to system identification.

In fixed distribution learning theory, a central role is played by the concept of covering numbers. Roughly speaking, the $\epsilon$-covering number of a set is the minimum number of balls of radius $\epsilon$ needed to cover the set. Kearns and Vazirani [19] provide a very large number of examples that give the $\epsilon$-covering numbers of several sets of functions. Interestingly, the computation of covering numbers is already a well-established subject in control theory, albeit in a different context. In [37] and [38], covering numbers appear in the context of trying to reduce the uncertainty about a plant to be controlled using feedback. These papers, especially [38], contain explicit upper bounds for the covering numbers of families of impulse response of linear time-invariant systems. With a little effort, these estimates can be extended to Hammerstein- and Wiener-type nonlinear systems (i.e., systems in which an LTI subsystem is preceded or followed by a memoryless nonlinearity).

The paper is organized as follows. In Section II, the standard formulation of the PAC learning problem is recalled for the convenience of the reader. Two types of learning are studied, namely, learning an unknown target function (or concept) and model-free learning. To facilitate a comparison with the literature in the theory of empirical processes, some relevant results from this theory are summarized. In Section III, some new notions of learnability are introduced, and their interrelationships to PAC learning are discussed. Two new definitions are introduced, namely, probably uniformly approximately correct (PUAC) learnability and minimal empirical risk (MER) learnability. The relationships between these new notions and existing notions of PAC learnability and solid learnability are illustrated through examples. In Section IV, it is shown that MER learnability and PUAC learnability are in fact equivalent to each other and equivalent to another property introduced here called the "shrinking width" property. In Section V, some sufficient conditions for PUAC (MER) learnability are given. In particular, it is shown that whenever the function class to be learned has the property that empirical means converge uniformly to their true values, the function class is PUAC learnable. In the course of proving these results, a new estimate is given of the VC-dimension of a collection of sets obtained by performing Boolean operations on a given collection. This result is of independent interest. In Section VI, the problem of model-free learning is studied, and previously known necessary and sufficient conditions for learning a target concept (or function) are extended to the model-free setting. In Section VII, the focus is on the issue of *representation* of a collection of sets by its subcollection of equivalence classes. It is shown by example that, by suitably choosing representatives of each equivalence class, it is possible to affect the property of uniform convergence of empirical probabilities, whereas PAC learnability is not affected by the exact choice of representatives. Finally, Section VIII contains the conclusions.

## II. Problem Statements and Summary of Known Results

In this section, the basic learning problems are formulated, and some known results are summarized for the convenience of the reader.

We begin with statements of the two types of learning problems studied here, namely learning an unknown target function (or concept) and model-free learning. The first type of learning problem is as in [1], [19], or [26], with obvious modifications to cater to function instead of concept learning. The second type of learning problem, variously referred to as the "loss function approach" or "agnostic" learning, follows [17]. Throughout, the notation and terminology follows [36], which also gives more details and background for the material of this section.

### A. Learning an Unknown Target Concept or Function

The basic ingredients of the learning problem under a fixed probability are as follows:

- set $X$;
- $\sigma$-algebra $\mathcal{S}$ of subsets of $X$;
- fixed known probability measure $P$ on the measurable space $(X, \mathcal{S})$;
- subset $\mathcal{C} \subseteq \mathcal{S}$, called the *concept class*, or a family $\mathcal{F}$ of measurable functions mapping $X$ into $[0, 1]$,[1] called the *function class*.

Let us begin with the function learning problem and then specialize to concept learning. Given two measurable functions

[1]Actually, there is nothing special about the interval $[0, 1]$, and it can be replaced by any *bounded* interval; however, some technical difficulties develop if an *a priori* bound on the functions in $\mathcal{F}$ does not exist.

$a, b: X \to [0, 1]$ and the probability measure $P$ on $(X, \mathcal{S})$, define

$$d_P(a, b) = \int_X |a(x) - b(x)| \, P(dx).$$

This equation is a pseudometric on the set of measurable functions in $[0, 1]^X$ and equals the expected value of the difference $|a(x) - b(x)|$. In function learning, a fixed but unknown target function $f \in \mathcal{F}$ exists, and the objective is to "learn" this target function using the values of the unknown function $f$ at randomly generated input values from the set $X$. Specifically, i.i.d. samples $x_1, \cdots, x_m \in X$ are drawn in accordance with the probability measure $P$, and for each index $j$, an "oracle" returns the value of $f(x_j)$. Thus, after $m$ samples have been generated, the information available to the learner consists of the "labeled multisample"

$$(x_1, f(x_1)), \cdots, (x_m, f(x_m)) \in (X \times [0, 1])^m .$$

The objective is to construct an approximation to the unknown function $f$ using this information, through an appropriate "algorithm." For this paper, an "algorithm" is an indexed family of maps $\{A_m\}$, in which

$$A_m: (X \times [0, 1])^m \to \mathcal{F}.$$

Define

$$h_m(f; \mathbf{x}) := A_m[(x_1, f(x_1)), \cdots, (x_m, f(x_m))]$$

to be the hypothesis generated by the algorithm when the target function is $f$ and the multisample is $\mathbf{x}$, and let

$$r(m, \epsilon) := \sup_{f \in \mathcal{F}} P^m \{\mathbf{x} \in X^m: d_P[f, h_m(f; \mathbf{x})] > \epsilon\}. \quad (1)$$

*Definition 1:* The algorithm $\{A_m\}$ is said to be PAC if $r(m, \epsilon) \to 0$ as $m \to \infty$ for each $\epsilon > 0$. The function class $\mathcal{F}$ is PAC *learnable with respect to* $P$ if an algorithm exists that is PAC.

The concept learning problem is a special case of the above. Suppose we are given a concept class $\mathcal{C} \subseteq \mathcal{S}$. With each set $C \in \mathcal{C}$, we can identify its indicator function, which maps $C$ into $\{0, 1\}$. An algorithm is said to be PAC for the concept class $\mathcal{C}$ if it is PAC for the corresponding class of binary-valued functions $\{I_C(\cdot): C \in \mathcal{C}\}$ in the sense of Definition 1. Similarly, a concept class $\mathcal{C}$ is said to be PAC learnable if the corresponding family $\{I_C(\cdot): C \in \mathcal{C}\}$ is PAC learnable in the sense of Definition 1.

### B. Uniform Convergence of Empirical Means

Now, we discuss a related problem in the theory of empirical processes, known as the uniform convergence of empirical means (UCEM). Suppose we are given a family of *functions* $\mathcal{F}$, in which each $f \in \mathcal{F}$ maps $X$ into $[0, 1]$ and is measurable with respect to the $\sigma$-algebra $\mathcal{S}$. For each $f \in \mathcal{F}$, we can define its *empirical mean* $\hat{E}(f; \mathbf{x})$ as follows:

$$\hat{E}(f; \mathbf{x}) := \frac{1}{m} \sum_{j=1}^{m} f(x_j).$$

Define

$$q(m, \epsilon) := P^m \left\{ \mathbf{x} \in X^m: \sup_{f \in \mathcal{F}} |\hat{E}(f; x) - E_P(f)| > \epsilon \right\}.$$

Then, $1 - q(m, \epsilon)$ is the probability that *every* empirical mean is within $\epsilon$ of its true value. We say that the family of functions $\mathcal{F}$ has the property of UCEM if $q(m, \epsilon) \to 0$ as $m \to \infty$ for each $\epsilon > 0$.

Suppose $\mathcal{A} \subseteq \mathcal{S}$ is a given collection of measurable sets. Note that for each $A \in \mathcal{A}$, the corresponding indicator function $I_A(\cdot)$ maps $X$ into $\{0, 1\}$ and that $E_P(I_A)$ is precisely $P(A)$. We say that the collection of sets $\mathcal{A}$ has the property of *uniform convergence of empirical probabilities* (UCEP) if the family of functions $\{I_A(\cdot), A \in \mathcal{A}\}$ has the UCEM property.

In [6]–[8], another notion called "simultaneous estimability" is introduced, which corresponds to the ability to estimate each true probability $P(A)$ based on the data $(x_i, I_A(x_i))$, $i = 1, \cdots, m$, *without* insisting that the estimate must in fact be the empirical probability (mean) as defined above. Simultaneous estimability is a weaker notion than the UCEM property, but it is not discussed further in this paper.

### C. Model-Free Learning

In the learning problem formulated in the preceding subsection, it is assumed that the input to the learning algorithm is a labeled multisample of the form $(x_1, f(x_1)), \cdots, (x_m, f(x_m))$, in which $f$ is a fixed but unknown element of the family $\mathcal{F}$. In the model-free learning problem, we dispense with this assumption. The ingredients of the present learning problem are as follows:

- sets $X$, $Y$, and $U$;
- $\sigma$-algebra $\overline{\mathcal{S}}$ on $X \times Y$, and a family of probability measures $\overline{\mathcal{P}}$ on $(X \times Y, \overline{\mathcal{S}})$;
- family of functions $\mathcal{H}$ mapping $X$ into $U$, called the set of *hypotheses*;
- function $\ell$ mapping $Y \times U$ into $[0, 1]$, called the "loss" function.

Learning takes place as follows. An unknown probability measure $\overline{P} \in \overline{\mathcal{P}}$ is fixed, and i.i.d. samples $(x_1, y_1), \cdots, (x_m, y_m)$ are drawn from $X \times Y$ in accordance with $\overline{P}$. In this setting, an "algorithm" is an indexed family of maps $\{A_m\}_{m \geq 1}$, in which

$$A_m: (X \times Y)^m \to \mathcal{H}.$$

As before, define

$$h_m := A_m [(x_1, y_1), \cdots, (x_m, y_m)].$$

Associated with $h_m$ and $\overline{P}$ is an error measure

$$J(h_m, \overline{P}) := \int_{X \times Y} \ell[y, h_m(x)] \, \overline{P}(dx, dy).$$

Also, associated with $\overline{P}$ alone is a number

$$J^*(\overline{P}) := \inf_{h \in \mathcal{H}} \int_{X \times Y} \ell[y, h(x)] \, \overline{P}(dx, dy).$$

We can think of $J^*(\overline{P})$ as the *best possible* performance by any hypothesis function $h \in \mathcal{H}$, when the samples are drawn in ac-

cordance with the probability measure $\overline{P}$. Similarly, $J(h_m, \overline{P})$ can be thought of as the *actual* performance of the algorithm after $m$ samples are drawn in accordance with $\overline{P}$. The quantity $J(h_m, \overline{P})$ is sometimes referred to as the *risk* associated with the hypothesis $h_m$, when the underlying probability measure is $\overline{P}$. Clearly

$$0 \leq J^*(\overline{P}) \leq J(h_m, \overline{P}) \leq 1.$$

Also, although $J^*(\overline{P})$ is a deterministic number, $J(h_m, \overline{P})$ is a random number, because it depends on the random samples $(x_1, y_1), \cdots, (x_m, y_m)$. Define

$$r_{\mathrm{mf}}(m, \epsilon) := \sup_{\overline{P} \in \overline{\mathcal{P}}} \overline{P}^m \{(\mathbf{x}, \mathbf{y}) \in X^m \times Y^m : J(h_m, \overline{P}) > J^*(\overline{P}) + \epsilon\} \tag{2}$$

where

$$\mathbf{x} := [x_1 \cdots x_m]^t \in X^m, \qquad \mathbf{y} := [y_1, \cdots, y_m]^t \in Y^m$$

and by a slight abuse of notation, we write the $m$-fold sample $[(x_1, y_1), \cdots, (x_m, y_m)]$ as an element of $X^m \times Y^m$ [instead of as an element of $(X \times Y)^m$, which it is]. Thus, for a fixed $\epsilon$, we can think of $r_{\mathrm{mf}}(m, \epsilon)$ as the measure of the set of "bad" samples, in which a sample is deemed to be "bad" if it leads to a hypothesis that performs more than $\epsilon$-worse compared with the optimum achievable performance. The quantity $r_{\mathrm{mf}}(m, \epsilon)$ is analogous to the quantity $r(m, \epsilon)$ defined in (1), except that the subscript "mf" is used to remind us that the problem under study is one of model-free learning.

*Definition 2:* The algorithm $\{A_m\}$ is PAC if $r_{\mathrm{mf}}(m, \epsilon) \to 0$ as $m \to \infty$ for each $\epsilon > 0$. The triplet $(\mathcal{H}, \overline{\mathcal{P}}, \ell)$ is model-free learnable if an algorithm exists that is PAC.

The above problem formulation follows [17]. In contrast with the fixed-distribution problem formulated in Section II-A, in the present case, a *family* of probability measures $\overline{\mathcal{P}}$ on the product space $X \times Y$ exists. In [17] and other papers on this subject, it is common to take $\overline{\mathcal{P}}$ to be the set of *all* probability measures on $X \times Y$, so that we get a model-free version of *distribution-free* learning. It is easy to see that the model-free learning problem is very simple if $\overline{\mathcal{P}}$ is a *singleton set*. So, clearly, choosing $\overline{\mathcal{P}}$ to be a singleton set is *not* the correct model-free version of fixed-distribution learning. Thus, it is reasonable to ask: what is good a model-free version of fixed distribution learning?

Suppose $\overline{P}$ is a probability measure on $X \times Y$. We can define the corresponding "marginal" probability $\overline{P}_x$ on $X$ alone as follows. Suppose $A \subseteq X$ is measurable. Then

$$\overline{P}_x(A) := \overline{\mathcal{P}}(A \times Y).$$

Now, given the family $\overline{\mathcal{P}}$ of probability measures on $X \times Y$, it is assumed that a *known fixed* probability measure $P$ on $X$ exists such that

$$\overline{P}_x = P, \ \forall \overline{P} \in \overline{\mathcal{P}}.$$

In other words, it is assumed that, although several probability measures on $X \times Y$ exist, they all have the same marginal probability on $X$. This is a natural generalization of the idea of trying to learn a family of functions $\mathcal{F}$ under a fixed probability measure. To illustrate this generalization, it is assumed for notational simplicity that $P$ has a density $p(\cdot)$. In the standard PAC learning formulation, we can define a family $\{\overline{P}_f : f \in \mathcal{F}\}$ of probability measures on $X \times Y$ with the density

$$\overline{p}_f(x, y) := p(x)\, \delta(y - f(x)).$$

If a sample $(x, y)$ is drawn from $X \times Y$ in accordance with a fixed $\overline{P}_f$, then with probability one $y$ equals $f(x)$. Thus, the multisample $((x_1, y_1), \cdots, (x_m, y_m))$ corresponds to the standard labeled multisample $(x_1, f(x_1)), \cdots, (x_m, f(x_m))$. If the loss function $\ell(y, u)$ is taken $|y - u|$, then the model-free learning problem reduces to the standard PAC learning problem. Note that, with the family $\overline{\mathcal{P}}$ defined as above, the marginal probability on $X$ of every probability measure $\overline{P}_f$ equals $P$. Thus, a reasonable model-free version of fixed distribution learning is obtained by assuming that a *known fixed* probability measure $P$ on $X$ alone exists such that every $\overline{P} \in \overline{\mathcal{P}}$ has the marginal probability $P$ over $X$.

### D. Summary of Relevant Known Results

In this subsection, we give a brief summary of the known results in the problems of uniform convergence of empirical probabilities and of PAC learning under a fixed distribution.

Using the notion of the Vapnik–Chervonenkis dimension, it is possible to state a simple necessary and sufficient condition for a collection of sets to have the UCEP property. The original result in this direction is from [33], but the refinement below is from [32].

*Definition 3:* Let $(X, \mathcal{S})$ be a given measurable space, and let $\mathcal{A} \subseteq \mathcal{S}$. A set $S = \{x_1, \cdots, x_n\} \subseteq X$ is said to be *shattered* by $\mathcal{A}$ if, for every subset $B \subseteq S$, a set $A \in \mathcal{A}$ exists such that $S \cap A = B$. The *Vapnik–Chervonenkis dimension* of $\mathcal{A}$, denoted by VC-dim$(\mathcal{A})$, equals the largest integer $n$ such that a set of cardinality $n$ exists that is shattered by $\mathcal{A}$.

In order to state the theorem, we introduce the symbol $d(\mathbf{x})$. Suppose $\mathbf{x} \in X^m$, and define $S = \{x_1, \cdots, x_m\} \subseteq X$. Because it is possible for the vector $x$ to have repeated elements, in general, the cardinality of $S$ need not equal $m$. Look at the collection $\{A \cap S : A \in \mathcal{A}\} \subseteq 2^S$. This is a family of subsets of $S$. The integer $d(\mathbf{x})$ is defined as the VC-dimension of the collection $\{A \cap S : A \in \mathcal{A}\}$. Equivalently, $d(\mathbf{x})$ can be defined as the largest integer $n$ such that some subset $\{x_{i_1}, \cdots, x_{i_n}\} \subseteq S$ exists that is shattered by $\mathcal{A}$. Clearly, $d(x) \leq m$, and in fact, $d(\mathbf{x})$ is no larger than the number of *distinct* elements among $x_1, \cdots, x_m$.

*Theorem 1:* The collection of sets $\mathcal{A}$ has the property of uniform convergence of empirical probabilities if and only if

$$\lim_{m \to \infty} \frac{E_{P^m}[d(\mathbf{x})]}{m} = 0. \tag{3}$$

For a proof of this theorem, see [29, Theorem 21, p. 22], [32, Theorem 4.2], or [36, Theorem 5.4].

It is also possible to state necessary and sufficient conditions for a *function* family to have the UCEM property in terms of the covering numbers of some sets. The original result is from [34], however, and is reproduced in [35, Appendix to Ch. 6], and in

[36, Theorem 5.3]. These results, however, are not stated here because they are not pursued in this paper.

Now, let us turn to PAC learning under a fixed distribution. The fundamental results in this problem are from [4] and based on the notion of a covering number. Suppose $Y$ is a given set and $\rho$ is a pseudometric on $Y$. We say that a collection $\{y_1, \cdots, y_n\} \subseteq Y$ is an *$\epsilon$-cover* of $Y$ (with respect to $\rho$) if, for every $y \in Y$, an index $j$ exists such that $\rho(y, y_j) \leq \epsilon$. The *$\epsilon$-covering number* of $Y$ with respect to $\rho$ is defined as the smallest integer $n$ for which an $\epsilon$-cover of cardinality $n$ exists, and is denoted by $N(\epsilon, Y, \rho)$.

In the case of *concept* learning, the following theorem gives a definitive result.

*Theorem 2 [4]:* A concept class $\mathcal{C} \subseteq \mathcal{S}$ is PAC learnable with respect to a fixed probability measure $P$ if and only if the covering number $N(\epsilon, \mathcal{C}, d_P)$ is finite for every $\epsilon > 0$.

In case a concept class $\mathcal{C}$ satisfies the above condition, it is shown in [4] that a so-called "minimal empirical risk" algorithm is PAC to accuracy $\epsilon$, for each $\epsilon$. The algorithm presented in [4] is for *concept* learning, but an entirely analogous argument goes through for *function* learning as well. In particular, we can state the following result.

*Theorem 3:* Suppose a function family $\mathcal{F}$ has the property that the covering number $N(\epsilon, \mathcal{F}, d_P)$ is finite for every $\epsilon > 0$; then $\mathcal{F}$ is PAC learnable with respect to $P$.

Note that, unlike in the case of concept learning, the condition $N(\epsilon, \mathcal{F}, d_P) < \infty \ \forall \epsilon > 0$ is *not* necessary for function learnability; (see [36, Example 6.11, p. 183]).

## III. Some New Notions of Learning

In this section, we introduce two new notions of learnability, namely, PUAC learnability and MER learnability. PUAC learnability is a stronger property than the standard notion of PAC learnability, and while MER learnability is stronger property than "solid" or "potential" learnability, which are properties introduced earlier in the literature. The reason for introducing these two new notions is that many of the standard sufficient conditions for PAC learnability actually imply these stronger forms of learnability; thus, these stronger forms of learnability "come for free," so to speak. In the next section, it is shown that actually MER learnability and PUAC learnability are equivalent properties and, in turn, equivalent to another property referred to here as the "shrinking width" property. Thus, the shrinking width property plays nearly the same role in PUAC (MER) learning as the finite metric entropy property in PAC learning. A noteworthy difference, however, is that the shrinking width property is a necessary and sufficient condition for PUAC learnability even for *function* classes; in contrast, the finite metric entropy condition is necessary and sufficient for PAC learnability only for *concept* classes—for function classes, it is known to be sufficient, but not necessary.

*Definition 4:* The algorithm $\{A_m\}$ is said to be (PUAC) if the quantity

$$s(m, \epsilon) := P^m\{\mathbf{x} \in X^m: \sup_{f \in \mathcal{F}} d_P[f, h_m(f; \mathbf{x})] > \epsilon\} \quad (4)$$

approaches zero as $m \to \infty$, for each fixed $\epsilon > 0$. The function class $\mathcal{F}$ is said to be PUAC *learnable* if an algorithm exists that is PUAC.

To contrast the definition of a PUAC algorithm with that of a PAC algorithm, recall the quantity $r(m, \epsilon)$ defined in (1), namely

$$r(m, \epsilon) := \sup_{f \in \mathcal{F}} P^m\{\mathbf{x} \in X^m: d_P[f, h_m(f; \mathbf{x})] > \epsilon\}.$$

Suppose an accuracy parameter $\epsilon > 0$ and an integer $m$ are specified. Given a target function $f \in \mathcal{F}$, refer to a multisample $\mathbf{x} \in X^m$ as "bad" if the hypothesis $h_m(f; \mathbf{x})$ produced by the algorithm differs from $f$ by more than $\epsilon$, i.e., if $d_P[f, h_m(f; \mathbf{x})] > \epsilon$. Let

$$B_m(f, \epsilon) := \{\mathbf{x} \in X^m: d_P[f, h_m(f; \mathbf{x})] > \epsilon\}$$

denote the set of bad samples for the target function $f$. Then

$$B_m(\epsilon) := \bigcup_{f \in \mathcal{F}} B_m(f, \epsilon) = \left\{\mathbf{x} \in X^m: \sup_{f \in \mathcal{F}} d_P[f, h_m(f; \mathbf{x})] > \epsilon\right\}$$

consists of the set of multisamples that are bad for *even a single* target function $f \in \mathcal{F}$. Thus, an algorithm is PAC if the measure of *each single* set $B_m(f, \epsilon)$ approaches zero as $m \to \infty$, uniformly with respect to $f \in \mathcal{F}$; an algorithm is PUAC if the measure of *the union of the sets* approaches zero as $m \to \infty$. With this interpretation, it is obvious that every PUAC algorithm is also PAC. If the function class is finite, then every PAC algorithm is also PUAC, because

$$P^m\left[\bigcup_{f \in \mathcal{F}} B_m(f, \epsilon)\right] \leq |\mathcal{F}| \sup_{f \in \mathcal{F}} P^m[B_m(f, \epsilon)].$$

Note that, for concept classes, PUAC learnability is the same as the notion of "simultaneous learnability" defined in [6].

The PUAC property can be interpreted by the convergence of a stochastic process. Let $X^\infty$ denote the countable Cartesian product of $X$, and let $\mathcal{S}^\infty$, $P^\infty$ denote the corresponding product $\sigma$-algebra and product probability measure, respectively. Use the symbol $\mathbf{x}^*$ to denote a typical element of $X^\infty$; thus, $\mathbf{x}^* = (x_1, x_2, \cdots)$, where each $x_i \in X$. Now, define the stochastic process $\{b_m(\cdot)\}$ on $X^\infty$ by

$$b_m(\mathbf{x}^*) := \sup_{f \in \mathcal{F}} d_P[f, h_m(f; \mathbf{x}^*)] \quad (5)$$

where $h_m(f; \mathbf{x}^*)$ denotes the hypothesis produced by the algorithm based on the first $m$ elements of the multisample $\mathbf{x}^*$. Thus, $b_m(\mathbf{x}^*)$ (which depends only on the first $m$ terms of the sequence $\mathbf{x}^*$) is the *worst-case error* between a target function and the corresponding hypothesis, when the multisample is $\mathbf{x}^*$. Now, it is easy to see that the algorithm is PUAC if and only if the stochastic process $\{b_m(\cdot)\}$ converges to zero *in probability*, with respect to $P^\infty$. It is possible to define an apparently stronger property called almost surely eventually correct (ASEC) algorithm, by requiring the stochastic process $\{b_m(\cdot)\}$ to converge to zero *almost surely* with respect to $P^\infty$. We do

not choose to define this other notion, because it turns out that in many situations the two properties are equivalent.

*Example 1:* The objective of this example is to present an algorithm that is PAC, but not PUAC. Let $X = [0, 1]$, $\mathcal{S}$ = the Borel $\sigma$-algebra on $X$. Let $P$ denote the uniform probability measure on $X$. Let $\mathcal{G}$ denote the collection of all *finite* subsets of $X$, and let $\tau$: $\mathcal{G} \to [0, 0.5)$ be a one-to-one (but not necessarily onto) mapping. One such mapping is constructed at the end of the example, but the exact nature of the mapping is not important for the example. If $a \in [0, 0.5)$ belongs to the range of the map $\tau$, let $\tau^{-1}(a)$ denote the unique finite set $G$ such that $\tau(G) = a$. If $a$ does not belong to the range of $\tau$, define $\tau^{-1}(a)$ to equal the empty set $\emptyset$. With this notation, let the concept class $\mathcal{C}$ consist of all unions of the form $[0, a] \cup \tau^{-1}(a)$ as $a$ varies over $[0, 0.5)$, together with $X$ itself. In symbols

$$\mathcal{C} = \{[0, a] \cup \tau^{-1}(a)\colon 0 \le a < 0.5\} \cup \{X\}.$$

The algorithm is defined next. Suppose $\mathbf{x} \in X^m$, and let $\boldsymbol{L}(\mathbf{x}) = [I_T(x_1) \cdots I_T(x_m)]^t \in \{0, 1\}^n$ denote the set of labels of the components of $x$ generated by the unknown target concept $T$. The algorithm is as follows. If the label vector $\boldsymbol{L}(\mathbf{x})$ consists of all one's, then define $H_m(T; \mathbf{x}) := X$. If the label vector $\boldsymbol{L}(\mathbf{x})$ does not equal the vector of all one's, then define $H_m(T; \mathbf{x}) := [0, h] \cup \tau^{-1}(h)$, where

$$h := \min\{0.5, \max\{x_i\colon I_T(x_i) = 1\}\}.$$

The algorithm is intuitive as follows. If each component of the multisample $\mathbf{x}$ is labeled with a one, then the algorithm declares that the unknown target concept is the entire set $X$. If at least one component of the multisample $\mathbf{x}$ *fails* to belong to the unknown target concept, then the algorithm declares that $T$ is the largest interval of the form $[0, h]$ that contains all positive examples (i.e., all $x_i$ that belong to $T$), together with its associated "tail" $\tau^{-1}(h)$. This tail is appended to the interval $[0, h]$ so as to ensure that the hypothesis produced by the algorithm indeed belongs to the concept class $\mathcal{C}$. Similarly, the "min" is introduced to ensure that $h$ is never larger than 0.5. If we were to modify the definitions of PAC and PUAC in the obvious manner by *discarding* the requirement that the hypothesis belongs to $\mathcal{C}$, then we could simply define $H_m(T; \mathbf{x}) := [0, h]$ without the tail; we could also drop the "min" and just take $h$ to be the largest positive example $x_i$. These modifications are easy and left to the reader.

It is shown first that the algorithm is PAC. Observe that $P(G) = 0$ for every finite set $G$. Hence, given an $x \in X$ selected at random according to $P$, we have

$$I_{[0, a] \cup \tau^{-1}(a)}(x) = I_{[0, a]}(x) \text{ w.p. } 1, \qquad \forall\, a \in [0, 0.5)$$

where "w.p. 1" is an abbreviation for "with probability one." Hence, if the target concept $T$ is of the form $[0, a] \cup \tau^{-1}(a)$, then the label $I_T(x)$ is the same as the indicator function $I_{[0, a]}(x)$, with probability one. Therefore, after a multisample $\mathbf{x}$ is drawn, if the target concept is of the above form, then it can be assumed w.p. 1 that none of the samples $x_i$ belongs to the "tail" $\tau^{-1}(a)$, and that

$$h = \max\{x_i\colon x_i \le a\}.$$

Now, it is easy to see that, w.p. 1, the interval $[0, h]$ in the hypothesis $H_m(T; \mathbf{x})$ is a subinterval of $[0, a]$, so that $d_P(T, H_m) = a - h$. Suppose $\epsilon > 0$ is specified. Then, $d_P(T, H_m) > \epsilon$ only if every one of the samples $x_i$ *fails* to belong to the interval $[a - \epsilon, a]$. For a fixed $i$, the probability of this happening is $1 - \epsilon$, when it follows that the probability of this happening $m$ times in a row is $(1 - \epsilon)^m$. In other words, it has been shown that

$$P^m\{\mathbf{x} \in X^m\colon d_P(T, H_m) > \epsilon\} \le (1 - \epsilon)^m.$$

The above analysis applies whenever the target concept $T$ is of the form $[0, a] \cup \tau^{-1}(a)$. If, on the other hand, the target concept $T$ equals $X$, then the labels $I_T(x)$ will all equal one; in which case, the algorithm will output $H_m = X$, which happens to be correct; thus, $d_P(T, H_m) = 0$ for all $x$ and all $m$ in this case. Combining these steps, we conclude that the quantity $r(m, \epsilon)$ defined in (1) is bounded by

$$r(m, \epsilon) \le (1 - \epsilon)^m.$$

Because the right side of this inequality approaches zero as $m \to \infty$ for every fixed $\epsilon$, we conclude that the algorithm is PAC.

It is shown next that the algorithm is *not* PUAC. To establish this claim, let $m$ and $\mathbf{x} \in X^m$ be arbitrary, and define $G(\mathbf{x}) := \{x_1, \cdots, x_m\}$ after deleting repeated components if any. Suppose the target concept $T$ is $[0, \tau(G(\mathbf{x}))] \cup G(\mathbf{x})$. Then, the label vector $I_T(x_i)$ equals the vector of all one's; as a result, the algorithm returns the hypothesis $H_m = X$. Because the measure of $T$ equals $\tau(G(\mathbf{x})) < 0.5$, it follows that, *for this particular choice of target concept,*

$$d_P[T; H_m(T; \mathbf{x})] > 0.5.$$

This reasoning can be applied to *every* multisample $\mathbf{x}$. Hence, we conclude that

$$\sup_{T \in \mathcal{C}} d_P[T; H_m(T; \mathbf{x})] > 0.5, \qquad \forall\, \mathbf{x} \in X^m.$$

In other words, whenever $\epsilon \le 0.5$, we have

$$\left\{\mathbf{x} \in X^m\colon \sup_{T \in \mathcal{C}} d_P[T; H_m(T; \mathbf{x})] > \epsilon\right\} = X^m.$$

Hence, the quantity $s(m, \epsilon)$ defined in (4) is given by

$$s(m, \epsilon) = P^m(X^m) = 1, \ \forall\, m.$$

This equation shows that the algorithm is *not* PUAC.

Though the details of the above example are a little messy, the idea is simple. For *each fixed* target concept, the set of multisamples $\mathbf{x}$ that lead to a poor hypothesis has small measure. *Every* multisample, however, is "bad" for *some* target concept, so that the union of the multisamples that are "bad" for at least one target concept is in fact the entire space $X^m$.

The example is completed by demonstrating the existence of a mapping $\tau$ with the desired properties. As stated above, the exact nature of the mapping is not important for the example.

Let $I_0, I_1, I_2, \cdots$ be a partition of the set $\{2, 3, 4, \cdots\}$ such that each set $I_i$ is infinite. For example, let $p_i$ denote the $i$th

prime number, and for $i \geq 1$, let $I_i$ consist of all powers of $p_i$, that is, $I_i = \{p_i, p_i^2, p_i^3, \cdots\}$. Finally, let $I_0$ be the complement of the union $\cup_{i \geq 1} I_i$. Thus, $I_0$ consists of all numbers that have at least two distinct prime divisors.[2] Given a finite set $G = \{x_1, \cdots, x_n\}$, arrange the $x_i$'s such that $x_1 > x_2 > \cdots > x_n$. Now, the number $\tau(G)$ is defined by its binary expansion

$$\tau(G) = \sum_{i=1}^{\infty} b_i 2^{-i}.$$

Set $b_1 = 0$ always, which ensures that $\tau(G) \leq 0.5$. Next, set the first $n$ bits in $I_0$ equal to one, and the rest to zero which encodes the value of the integer $n$, i.e., the cardinality of the set $G$. Finally, for each $i \leq n$, set the bits in $I_i$ equal to the bits in the binary representation of $x_i$; and for $i > n$, set all bits in $I_i$ equal to zero. Then, it is easy to see that $\tau$ is one-to-one. Moreover, because only a finite number of bits in $I_0$ are nonzero, it follows that $\tau(G) < 0.5$. ■

Next, define the notion of "MER" learnability. Suppose $f \in \mathcal{F}$, $\mathbf{x} \in X^m$, and let $h_m(f; \mathbf{x})$ denote, as before, the hypothesis generated by the algorithm when the target concept is $f$ and the multisample is $\mathbf{x}$. We say that the hypothesis $h_m$ *agrees with* $f$ *on* $\mathbf{x}$ if

$$h_m(x_i) = f(x_i), \qquad i = 1, \cdots, m.$$

The algorithm is said to be *consistent* if $h_m(f; \mathbf{x})$ agrees with $f$ on $\mathbf{x}$ for every function $f \in \mathcal{F}$ and every multisample $\mathbf{x} \in X^m$, for every $m \geq 1$. To put it into words, an algorithm is consistent if the hypothesis produced by the algorithm always matches the data points. (Note that this usage of the word "consistent" follows that in the learning theory literature and does *not* refer to asymptotic properties of the estimator as used in statistics.)

An alternative definition of a consistent algorithm is that it *minimizes empirical risk*, in the following sense. Given two functions $f, g \in \mathcal{F}$, we can define their "true" distance as $d_P(f, g)$ and their "empirical" distance based on the multisample $x$ as

$$\hat{d}(f, g; \mathbf{x}) := \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - g(x_i)|. \tag{6}$$

With this definition, we can see that an algorithm is consistent if and only if the hypotheses generated by the algorithm satisfy

$$\hat{d}[f, h_m(f; \mathbf{x}); \mathbf{x}] = 0, \qquad \forall m, \ \forall \mathbf{x} \in X^m, \ \forall f \in \mathcal{F}.$$

Because zero is the minimum possible value for the empirical distance, a consistent algorithm is one that minimizes the empirical estimate of the "risk" involved in assuming that $h_m(f; \mathbf{x})$ is the same as $f$.

*Definition 5:* A concept class $\mathcal{C}$ is said to be MER learnable if every consistent algorithm is PUAC.

In the computational learning theory literature, we encounter the notions of "solid learnability" [26] or "potential learnability" [1]. These two notions are equivalent to each other, and the definition of these notions is that every consistent algorithm must be PAC. The present definition of "MER" learnability is more stringent in that every consistent algorithm is required to be PUAC, and not merely PAC. As shown subsequently, PUAC learnability is a stronger requirement than PAC learnability. Hence, MER learnability is a stronger requirement than solid or potential learnability.

*Example 2 [4]:* The objective of this example is to demonstrate a concept class that is PAC learnable, but not MER learnable. Let $X = [0, 1]$, and let $\mathcal{C}$ consist of all finite subsets of $X$ together with $X$ itself. Given any multisample $\mathbf{x} \in X^m$, let the corresponding hypothesis $H_m$ equal the finite set consisting of all positive examples, i.e., all $x_j$ for which the oracle returns the value $I_T(x_j) = 1$. This algorithm is certainly consistent. If the target concept $T$ equals $X$, however, then for *every* multisample $\mathbf{x}$, we get $H_m(T; \mathbf{x}) = \{x_1, \cdots, x_m\}$. As a result, $d_P[T, H_m(T; \mathbf{x})] = 1$ for every $\mathbf{x} \in X^m$; consequently, the quantity $r(m, \epsilon)$ defined in (1) equals one for every $m$, whenever $\epsilon < 1$. Thus, the present algorithm is not even PAC, let alone PUAC. On the other hand, the pair $\{\emptyset, X\}$ is an $\epsilon$-cover of $\mathcal{C}$ with respect to the pseudometric $d_P$ for *every* $\epsilon > 0$. Hence, $\mathcal{C}$ is learnable using the minimum empirical risk algorithm of [4] applied to this pair, to zero accuracy and zero confidence using *just one sample*. To see this, suppose the target concept $T$ is a finite set. Pick an $x \in X$ at random. Then, $x \notin T$ with probability one. Hence, the minimum risk algorithm returns the hypothesis $H = \emptyset$ with probability one, which is at a distance of zero from the target concept. If, on the other hand, the target concept $T = X$, then any $x$ belongs to $T$, and as a result, the minimum risk algorithm returns the hypothesis $H = X$, which happens to be correct.

## IV. MER Learnability and the Shrinking Width Property

In this section, it is shown that both MER learnability and PUAC learnability are in fact equivalent properties and, in turn, equivalent to another property referred to here as the "shrinking width" property. The equivalence of MER and PUAC learnabilities is somewhat surprising, because the situation is in sharp contrast to the fact that if we drop the "uniformity" requirement, then PAC learnability is a strictly weaker requirement than solid learnability (as brought out in Example 2).

*Definition 6:* Given a family of functions $\mathcal{F}$, define

$$w(m, \epsilon) := P^m\{\mathbf{x} \in X^m: \exists f, g \in \mathcal{F} \text{ s.t. } \hat{d}(f, g; \mathbf{x}) = 0 \text{ and } d_P(f, g) > \epsilon\}.$$

The family $\mathcal{F}$ is said to have the shrinking width property if $w(m, \epsilon) \to 0$ as $m \to \infty$.

In the case of *concept* classes, the shrinking width property is equivalent to the "empirical coverability" property defined in [6, Definition 4.1].

The shrinking width property can also be interpreted by the convergence of a stochastic process. Given an $\mathbf{x}^* \in X^\infty$, define

$$\phi_m(\mathbf{x}^*) := \sup\Big\{d_P(f, g): f, g \in \mathcal{F} \text{ and } \hat{d}_m(f, g; \mathbf{x}^*) = 0\Big\} \tag{7}$$

[2]This particular definition of the sets $I_i$ plays no role in the argument below and is intended only for illustrative purposes.

where $\hat{d}_m(f, g; \mathbf{x}^*)$ denotes the empirical distance between the functions $f$ and $g$ based on the first $m$ components of $\mathbf{x}^*$, i.e.,

$$\hat{d}_m(f, g; \mathbf{x}^*) := \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - g(x_i)|.$$

Because $\phi_m(\mathbf{x}^*)$ depends only on the first $m$ components of $\mathbf{x}^*$, we can also write $\phi_m(\mathbf{x}_m)$ instead of $\phi_m(\mathbf{x}^*)$, where $\mathbf{x}_m \in X^m$ consists of the first $m$ components of $\mathbf{x}^*$. It is easy to see that the shrinking width property is equivalent to the requirement that the stochastic process $\phi_m(\cdot)$ converges *in probability* to the zero function. The ultimate behavior of this stochastic process is the topic of the next lemma.

*Lemma 1:* Given any family of functions $\mathcal{F} \subseteq [0, 1]^X$, a constant $c = c(\mathcal{F})$ exists such that the stochastic process $\{\phi_m(\cdot)\}$ converges almost surely to $c$ as $m \to \infty$.

*Proof:* For each fixed $\mathbf{x}^* \in X^\infty$, the sequence $\{\phi_m(\mathbf{x}^*)\}$ is a nonincreasing sequence of real numbers and bounded below by zero. Hence, it converges to a limit, call it $c(\mathbf{x}^*)$. It only remains to show that $c(\mathbf{x}^*)$ is a constant almost everywhere. This result is a consequence of the Kolmogorov 0-1 law and the fact that an i.i.d. sequence is ergodic. ■

In view of Lemma 1, it is clear that the shrinking width property is equivalent to the requirement that the stochastic process $\{\phi_m(\cdot)\}$ converges *almost surely* to zero as $m \to \infty$, that is, to the requirement:

$$P^\infty\{\mathbf{x}^* \in X^\infty\colon \sup\{d_P(f, g)\colon f, g \in \mathcal{F} \text{ and } \hat{d}_m(f, g; \mathbf{x}^*) = 0\} \to 0 \text{ as } m \to \infty\} = 1. \quad (8)$$

Thus, we might wonder why the shrinking width property is not at once defined by (8). The reason is that the shrinking width property as defined here makes sense even when the probability measure $P$ is not known, but belongs to a *family* $\overline{P}$ of probability measures, merely by taking a supremum with respect to $P \in \mathcal{P}$; in contrast, the condition in (8) cannot be readily extended when $\mathcal{P}$ is not a singleton set. For a discussion of the shrinking width property in the case in which $P$ varies over a family $\overline{P}$, see [36, Theorem 8.2].

Now, we come to the main result of this section.

*Theorem 4:* Given a family of functions $\mathcal{F}$, the following statements are equivalent:

- family $\mathcal{F}$ has the shrinking width property;
- family $\mathcal{F}$ is MER learnable;
- family $\mathcal{F}$ is PUAC learnable.

*Proof:* i) $\Rightarrow$ ii) Suppose $\mathcal{F}$ has the shrinking width property, and define the stochastic processes $\{\phi_m(\cdot)\}$ and $\{b_m(\cdot)\}$ in accordance with (7) and (5), respectively. Suppose an algorithm is consistent. Then, by definition

$$\hat{d}_m[f, h_m(f; \mathbf{x}^*); \mathbf{x}^*] = 0, \qquad \forall m, \forall \mathbf{x}^* \in X^\infty, \forall f \in \mathcal{F}.$$

Hence, it follows from the definition of $\phi_m(\cdot)$ that

$$d_P[f, h_m(f; \mathbf{x}^*)] \le \phi_m(\mathbf{x}^*), \qquad \forall f \in \mathcal{F}.$$

Taking the supremum with respect to $f \in \mathcal{F}$ and comparing with (5) shows that

$$b_m(\mathbf{x}^*) \le \phi_m(\mathbf{x}^*), \qquad \forall \mathbf{x}^* \in X^\infty.$$

Because by assumption the right side converges to zero in probability, so does the left side. Hence, the algorithm is PUAC. Because this argument can be applied to *any* consistent algorithm, it follows that $\mathcal{F}$ is MER learnable.

ii) $\Rightarrow$ iii) In order to prove this implication, it is shown that, if we ignore issues of effective computability, computational complexity, etc., then a consistent algorithm exists in every learning problem, if we assume the axiom of choice. With this assumption, it follows that it is possible to order the function class $\mathcal{F}$ in every situation. Thus, given a labeled multisample $(x_1, f(x_1)), \cdots, (x_m, f(x_m))$, we can simply scan through all functions in $\mathcal{F}$ until we find a function consistent with the labeled multisample. Such a function surely exists, because the data is assumed to be generated by a target function belonging to the function class. This algorithm is well defined and consistent. Of course, this "algorithm" is also purely conceptual and not claimed to be implementable in any way.[3] Now, suppose $\mathcal{F}$ is MER learnable. By assumption, this algorithm is PUAC; hence, $\mathcal{F}$ is PUAC learnable.

iii) $\Rightarrow$ i)[4] Suppose $\mathcal{F}$ fails to have the shrinking width property. Then, numbers $\epsilon$, $\delta$ and a sequence $\{m_i\}$ exist approaching infinity such that

$$P^{m_i}\{\mathbf{x} \in X^{m_i}\colon \phi_{m_i}(\mathbf{x}) > \epsilon\} \ge \delta, \qquad \forall i.$$

Temporarily drop the subscript "$i$" on $m_i$, and examine the above inequality. From the definition of $\phi_m(\cdot)$, this inequality is equivalent to

$$P^m\{\mathbf{x} \in X^m\colon \exists f, g \in \mathcal{F} \text{ s.t. } \hat{d}(f, g; \mathbf{x}) = 0 \text{ and } d_P(f, g) > \epsilon\} \ge \delta.$$

For convenience, define the set $S \subseteq X^m$ by

$$S = \{\mathbf{x} \in X^m\colon \exists f, g \in \mathcal{F} \text{ s.t. } \hat{d}(f, g; \mathbf{x}) = 0 \text{ and } d_P(f, g) > \epsilon\}.$$

Suppose $\mathbf{x} \in S$, and choose $f, g \in \mathcal{F}$ such that

$$\hat{d}(f, g; \mathbf{x}) = 0 \quad \text{and} \quad d_P(f, g) > \epsilon.$$

Because $\hat{d}(f, g; \mathbf{x}) = 0$, the labeled samples $(x_i, f(x_i))$, $i = 1, \cdots, m$ and $(x_i, g(x_i))$, $i = 1, \cdots, m$ are identical. Thus, *every* algorithm returns the same hypothesis on the multisample $\mathbf{x}$ irrespective of whether the target function is $f$ or $g$. Let $h_m$ denote the hypothesis returned by an algorithm. Because, $d_P(f, g) > \epsilon$, it follows from the triangle inequality that

$$\text{either } d_P(f, h_m) > \epsilon/2 \quad \text{or} \quad d_P(g, h_m) > \epsilon/2.$$

In any case, it follows that

$$\sup_{f \in \mathcal{F}} d_P[f, h_m(f; \mathbf{x})] > \epsilon/2.$$

[3]If the function class $\mathcal{F}$, however, is *recursively enumerable*, then the above procedure would indeed satisfy most persons as being a true algorithm.

[4]This part of the proof is taken from [16] and was independently suggested by one of the reviewers of the original version of the paper.

This result is true for *every* algorithm. Because the argument can be repeated for *every* $\mathbf{x} \in S$, it follows that for *every* algorithm we have

$$\sup_{f \in \mathcal{F}} d_P[f, h_m(f; \mathbf{x})] > \epsilon/2, \qquad \forall \mathbf{x} \in S.$$

Now, restore the subscript "$i$" on $m_i$, and label the set $S$ as $S_i \subseteq X^{m_i}$. Because $P^{m_i}(S_i) \geq \delta$ for all $i$, it follows that *no algorithm* can be PUAC. Hence, $\mathcal{F}$ is not PUAC learnable. ■

In the process of proving the implication i) $\Rightarrow$ ii) above, we have actually shown that if $\mathcal{F}$ satisfies the shrinking width property and if an algorithm is consistent, then the "worst-case error" stochastic process $\{b_m(\cdot)\}$ actually converges to zero *almost surely*, and not merely in probability.

## V. Sufficient Conditions for PUAC Learnability

In this section, we give a simple sufficient condition for a function class to be PUAC learnable. Specifically, it is shown that if a family of functions $\mathcal{F}$ has the property that empirical means converge uniformly to their true values, then $\mathcal{F}$ is PUAC learnable. Though this result is implicit in earlier literature, it is worthwhile to state it explicitly. As a means of enhancing the applicability of this result, it is shown that if a collection of sets $\mathcal{A}$ has the UCEP property, then so does another collection obtained by performing a finite number of Boolean operations on $\mathcal{A}$. This result is apparently new and of independent interest. In the course of establishing this result, we derive a bound on the VC-dimension of a collection of sets obtained by performing Boolean operations on another collection of sets. This is a significant extension of a previous result by Dudley [14].

### A. UCEM Property Implies PUAC Learnability

Suppose $f, g$: $X \rightarrow [0, 1]$ and are measurable; then, the function $x \mapsto |f(x) - g(x)|$ is measurable and maps $X$ into $[0, 1]$. Moreover, if $A, B \in \mathcal{S}$ are measurable sets and $a, b$ are their corresponding indicator functions, then the function $x \mapsto |a(x) - b(x)|$ is the indicator function of the symmetric difference $A \Delta B$. In view of this, it is reasonable to denote the function $x \mapsto |f(x) - g(x)|$ by $f \Delta g$. With this background, given a function class $\mathcal{F}$, define the set $\mathcal{F}\Delta\mathcal{F}$ as follows:

$$\mathcal{F}\Delta\mathcal{F} := \{f \Delta g\colon f, g \in \mathcal{F}\}.$$

Note that $\mathcal{F}\Delta\mathcal{F}$ also consists of measurable functions mapping $X$ into $[0, 1]$. Now, an important result from [34] states that if the family $\mathcal{F}$ has the UCEM property, then so does the family $\mathcal{F}\Delta\mathcal{F}$. This result is considerably generalized in [36, Theorem 5.10 and Corollary 5.10]. The fact that $\mathcal{F}\Delta\mathcal{F}$ has the UCEM property if $\mathcal{F}$ does imply that it is possible to empirically estimate *distances* between functions in $\mathcal{F}$. Given $f, g \in \mathcal{F}$, let $d_P(f, g)$ denote the pseudometric distance between them defined previously. Similarly, given i.i.d. samples $x_1, \cdots, x_m \in X$ drawn in accordance with $P$, define $\hat{d}(f, g; \mathbf{x})$ as in (6) as the "empirical distance" between $f$ and $g$. Note that $\hat{d}(f, g; \mathbf{x})$ is just the empirical mean of the function $x \mapsto |f(x) - g(x)| \in [0, 1]^X$. Hence, as $m \rightarrow \infty$, the empirical estimates $\hat{d}(f, g; \mathbf{x})$ converge uniformly to their true values $d_P(f, g)$. Precisely, define

$$q_d(m, \epsilon) = P^m\Big\{\mathbf{x} \in X^m\colon \exists f, g \in \mathcal{F} \text{ s.t. } |\hat{d}(f, g; \mathbf{x}) - d_P(f, g)| > \epsilon\Big\}. \tag{9}$$

Because $\mathcal{F}\Delta\mathcal{F}$ has the UCEM property, it follows that $q_d(m, \epsilon) \rightarrow 0$ as $m \rightarrow \infty$ for each fixed $\epsilon > 0$.

In Section III, the notion of a consistent algorithm was introduced. In some applications, the requirement that an algorithm be consistent is strict. In the next few paragraphs, we introduce several less-restrictive versions of "consistency" that are in some sense "good enough" to ensure learnability of various types.

We begin by defining an "asymptotically" consistent algorithm. Let $h_m(f; \mathbf{x})$ denote the output of the algorithm when the target function is $f$ and the multisample is $\mathbf{x}$. Then, the algorithm is said to be asymptotically consistent if

$$\sup_{f \in \mathcal{F}} P^m\{\mathbf{x} \in X^m\colon \hat{d}[f, h_m(f; \mathbf{x}); \mathbf{x}] > \epsilon\} \rightarrow 0 \quad \text{as } m \rightarrow \infty, \ \forall \epsilon > 0.$$

Thus, we can think of an asymptotically consistent algorithm as one that produces hypotheses "nearly" consistent with the data "with high probability" as more and more samples are drawn. Similarly, we can define an algorithm to be asymptotically uniformly consistent if

$$P^m\{\mathbf{x} \in X^m\colon \sup_{f \in \mathcal{F}} \hat{d}[f, h_m(f; \mathbf{x}); \mathbf{x}] > \epsilon\} \rightarrow 0 \quad \text{as } m \rightarrow \infty, \ \forall \epsilon > 0.$$

In other words, an algorithm is asymptotically uniformly consistent if the maximal empirical distance between a target function and the hypothesis approaches zero *in probability*.

Now, we come to the main result of this section.

*Theorem 5:* Suppose a family $\mathcal{F} \subseteq [0, 1]^X$ of measurable functions has the property that empirical means converge uniformly. Then, the family is PUAC learnable, as follows:

- every asymptotically consistent algorithm is PAC;
- every asymptotically uniformly consistent algorithm is PUAC.

*Proof of Theorem:* Suppose the family $\mathcal{F}$ has the UCEM property, and let $\{A_m\}$ be any asymptotically consistent algorithm. Define $r(m, \epsilon)$ as in (1); it is desired to show that $r(m, \epsilon) \rightarrow 0$ as $m \rightarrow \infty$. For this purpose, let $\epsilon, \delta > 0$ be specified, and define

$$\tilde{q}(m, \epsilon) := \sup_{f \in \mathcal{F}} P^m\{\mathbf{x} \in X^m\colon \exists g \in \mathcal{F} \text{ s.t. } |\hat{d}(f, g; \mathbf{x}) - d_P(f, g)| > \epsilon\},$$

where $\hat{d}(f, g; \mathbf{x})$ is defined in (6). Note that $\tilde{q}(m, \epsilon) \leq q_d(m, \epsilon)$, where $q_d(m, \epsilon)$ is defined in (9). Also, from [34], the fact that the family $\mathcal{F}$ has the UCEM property implies that $q_d(m, \epsilon) \rightarrow 0$ as $m \rightarrow \infty$, which, in turn, implies that $\tilde{q}(m, \epsilon) \rightarrow 0$ as $m \rightarrow \infty$. Finally, the algorithm $\{A_m\}$ is assumed to be asymptotically consistent. Hence, it is possible to choose $m_0$ large enough that

$$\tilde{q}(m, \epsilon/2) \leq \delta/2 \ \forall m \geq m_0,$$

and

$$\sup_{f\in\mathcal{F}} P^m\{\mathbf{x}\in X^m\colon \hat{d}[f, h_m(f;\mathbf{x});\mathbf{x}] > \epsilon\} \le \delta/2\ \forall m \ge m_0.$$

It is now shown that

$$r(m,\epsilon)\le\delta\ \forall m\ge m_0.$$

To establish this inequality, fix $f\in\mathcal{F}$ and draw a multisample $\mathbf{x}=[x_1\cdots x_m]^t\in X^m$. Then, with probability at least $1-\delta/2$ with respect to $\mathbf{x}$, it is true that

$$\hat{d}(f, h_m;\mathbf{x})\le\epsilon/2.$$

where $h_m$ is a shorthand for $h_m(f;\mathbf{x})$. Also, with probability $1-\tilde{q}(m,\epsilon/2)\ge 1-\delta/2$ with respect to $\mathbf{x}$, it is true that

$$\left|\hat{d}(f, h_m;\mathbf{x}) - d_P(f, h_m)\right|\le\epsilon/2.$$

Hence, with probability at least $1-\delta$ with respect to $\mathbf{x}$, it is true that

$$d_P(f, h_m)\le\epsilon.$$

This is the same as saying that $r(m,\epsilon)\le\delta$, which is precisely the PAC inequality. This process shows that every asymptotically consistent algorithm is PAC.

The proof that every asymptotically uniformly consistent algorithm is PUAC is entirely similar and is left to the reader. ■

Theorem 5 has an interesting intuitive appeal. Suppose a family of functions has the property that, by repeatedly drawing i.i.d. samples, we can estimate the *mean value* of each function with high accuracy and high confidence; then, in fact, it is possible, not merely to make an accurate assessment of the mean value of the function, but of *the function itself*. In the case of estimating probabilities empirically, this result can be interpreted as follows. If a family of measurable sets has the property that the *size* of each set can be estimated with accuracy and confidence by drawing i.i.d. samples, then it is possible to estimate the *set itself*.

We can ask whether the sufficient conditions given in Theorem 5 are also necessary. In other words, is the UCEM property *necessary* for a function class to be either PAC or PUAC (MER) learnable? The next two examples show that this is not so—the UCEM property is *not necessary* for either PAC or PUAC learnability. This result is achieved by showing that the shrinking width property is strictly weaker than the UCEM property, so that a function class can be PUAC (MER) learnable without satisfying the UCEM property. Because PUAC learnability implies PAC learnability, UCEM is not necessary for PAC learnability either.

*Example 3:* Let $X=[0, 1]$, $\mathcal{S}=$ the Borel $\sigma$-algebra on $X$, and $P$ = the uniform probability measure on $X$. Let $\mathcal{C}_1$ = the collection of all finite subsets of $X$. Then, it is easy to see that $\mathcal{C}_1$ does *not* have the UCEP property. In fact, given any multisample $\mathbf{x}\in X^m$, the set $S(\mathbf{x}):=\{x_1,\cdots,x_m\}$ has empirical probability one, but true probability zero. On the other hand, since $d_P(A, B)=0$ for every pair $A, B\in\mathcal{S}$, it follows that $w(m,\epsilon)=0$ for every integer $m$ and every $\epsilon>0$. Hence, $\mathcal{C}_1$ *does* have the shrinking width property. As a result, every consistent algorithm is PUAC. This example shows that the shrinking width property is strictly weaker than the UCEP (or UCEM) property. See Example 4 below for a less trivial example of a collection of sets that does not have the UCEP property, but does have the shrinking width property.

Now, define $\mathcal{C}_2=\mathcal{C}_1\cup\{X\}$. Thus, $\mathcal{C}_2$ consists of all finite subsets of $X$ together with $X$ itself. It is claimed that $w(m,\epsilon)=1$ for every $\epsilon<1$ and every integer $m$. To see this, let $m\ge 1$, and let $\mathbf{x}\in X^m$ be arbitrary. Define $S=\{x_1,\cdots,x_m\}$ after deleting repeated elements if necessary. Then, $\hat{d}(S, X;\mathbf{x})=0$, because each $x_i$ belongs to both $S$ and $X$. On the other hand, $d_P(S, X)=1>\epsilon$ if $\epsilon<1$. This establishes the claim. So clearly $\mathcal{C}_2$ does *not* have the shrinking width property, and from Theorem 4, it follows that not every consistent algorithm is PUAC.

In this simple example, it is easy to construct a consistent algorithm that fails to be PUAC. Given a labeled sample $[((x_1, I_T(x_1)),\cdots,((x_m, I_T(x_m))]$, define

$$H_m(T;\mathbf{x}):=\bigcup_{I_T(x_i)=1}\{x_i\}.$$

In other words, $H_m(T;\mathbf{x})$ consists of those $x_i$ classified as belonging to $T$ by the oracle or, equivalently, all "positive" examples of the unknown target concept. The algorithm is clearly consistent. Suppose now the target concept $T$ equals $X$. Then

$$H_m(T;\mathbf{x})=\{x_1,\cdots,x_m\}$$

and $d_P[T, H_m(T;\mathbf{x})]=1$. Because this is true for *every* $\mathbf{x}\in X^m$, it follows that the quantity $r(m,\epsilon)$ defined in (1) satisfies

$$r(m,\epsilon)=1\ \forall m,\qquad \text{if } \epsilon<1.$$

Thus, the algorithm is not even PAC, let alone PUAC.

The preceding example is adapted from [4].

*Example 4:* Let $X=[0, 1]$, $\mathcal{S}=$ the Borel $\sigma$-algebra on $X$, and let $P$ = the uniform probability measure on $X$. Let $\mathcal{C}$ consist of all unions of the form $[0, a]\cup F$, where $a\le 0.5$ and $F$ is a *finite* subset of $(0.5, 1]$. It is claimed that $\mathcal{C}$ fails to have the UCEP property, but does have the shrinking width property.

It is shown first that $\mathcal{C}$ *does not* have the UCEP property. The proof of this claim is based on Theorem 1. It is clear that every subset of $(0.5, 1]$ is shattered by $\mathcal{C}$. Hence, given a multisample $\mathbf{x}\in X^m$, the restricted VC-dimension $d(x)$ is at least equal to the number of components of $x$ that lie in $(0.5, 1]$, which equals $m/2$ on average. Hence

$$\lim_{m\to\infty}\frac{E_{P^m}[d(\mathbf{x})]}{m}\ge\frac{1}{2},\ \forall m\ge 2.$$

Because the condition (3) of Theorem 1 is violated, the collection of sets $\mathcal{C}$ *fails* to have the UCEP property.

Now, it is claimed that $\mathcal{C}$ has the shrinking width property. Suppose $A=[0, a]\cup F$, $B=[0, b]\cup G$ belong to $\mathcal{C}$, where $0\le a, b\le 0.5$, and $F, G$ are finite subsets of $(0.5, 1]$. Suppose $\hat{d}(A, B;\mathbf{x})=0$ for some $\mathbf{x}\in X^m$; i.e., suppose $A, B$ agree on a multisample $\mathbf{x}$. Then, in particular, $A$ and $B$ also agree on all components of $\mathbf{x}$ lying in $[0, 0.5]$. Given $\mathbf{x}\in X^m$, let $\phi(\mathbf{x})$ denote the number of components of $\mathbf{x}$ lying in $[0, 0.5]$. Because these are also uniformly distributed in $[0, 0.5]$, the probability that $\hat{d}(A, B;\mathbf{x})=0$ is no larger than $(0.5-|a-b|)^{\phi(\mathbf{x})}=[0.5-d_P(A, B)]^{\phi(\mathbf{x})}$, because $0.5-d_P(A, B)$ is the probability that a randomly selected $x\in[0, 1]$ belongs to $[0, 0.5]$, but not

to $A \Delta B$. Therefore, the probability that $\hat{d}(A, B; \mathbf{x}) = 0$ for a random $\mathbf{x} \in X^m$ is at most

$$\sum_{l=0}^{m} [0.5 - d_P(A, B)]^l \cdot \Pr\{\phi(\mathbf{x}) = l\}.$$

Because the map $\lambda \mapsto (0.5 - \lambda)^l$ is a decreasing function of $\lambda \in [0, 0.5]$ for each $l$, it follows that the probability that $\hat{d}(A, B; \mathbf{x}) = 0$ given $d_P(A, B) > \epsilon$ is at most

$$\sum_{l=0}^{m} (0.5 - \epsilon)^l \cdot \Pr\{\phi(\mathbf{x}) = l\}.$$

This quantity is an upper bound for $w(m, \epsilon)$. Now, note that $(0.5 - \epsilon)^l \leq 0.5^l$ for each $l$, and that $0.5^l$ is a decreasing function of $l$. Hence, for each $m$, we have

$$w(m, \epsilon) \leq \Pr\{\phi(\mathrm{x}) < m/3\} + 0.5^{m/3} \Pr\{\phi(\mathbf{x}) \geq m/3\}.$$

Because $P$ is the uniform measure, $\phi(x)$ has the binomial distribution. Hence, $\Pr\{\phi(x) < m/3\} \to 0$ as $m \to \infty$. Also, $0.5^{m/3} \to 0$ as $m \to \infty$. This process leads to the conclusion that $w(m, \epsilon) \to 0$ as $m \to \infty$, for each $\epsilon > 0$, i.e., that $\mathcal{C}$ has the shrinking width property.

### *B. VC-Dimension of Collections of Sets Obtained by Boolean Operations*

In this subsection, we prove a new bound on the VC-dimension of a collection of sets $\mathcal{U}$ obtained by performing Boolean operations on given collection of sets $\mathcal{A}_1, \cdots, \mathcal{A}_k$. In [14], it is shown only that $\mathcal{U}$ has finite VC-dimension if each $\mathcal{A}_i$ has finite VC-dimension. The present result gives a considerable extension of Dudley's result. This result is used in the next subsection to show that, if a collection of sets $\mathcal{A}$ has the UCEP property, then so does another collection $\mathcal{U}(\mathcal{A})$ obtained from $\mathcal{A}$ by performing Boolean operations. This latter result (Theorem 7 proved in Section V-C) considerably enhances the applicability of Theorem 5.

Observe that a natural identification exists between collections of sets and families of binary-valued functions. Specifically, if $A \in \mathcal{S}$, then its indicator function $I_A(\cdot)$ is a binary-valued function. Conversely, if $f: X \to \{0, 1\}$ is measurable, then its support

$$\mathrm{supp}(f) := \{x \in X: f(x) = 1\}$$

is a measurable set. For this paper, it is more convenient to work with families of binary-valued functions.

Suppose $k \geq 2$ is a given integer, and that $u: \{0, 1\}^k \to \{0, 1\}$ is a given function.[5] Suppose $f_1, \cdots, f_k: X \to \{0, 1\}$ are binary-valued functions. Then, we define $u(f_1, \cdots, f_k): X \to \{0, 1\}$ to be the binary-valued function

$$x \mapsto u[f_1(x), \cdots, f_k(x)].$$

Finally, if $\mathcal{A}_1, \cdots, \mathcal{A}_k$ are families of binary-valued functions, we define $\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)$ to be the family of binary-valued functions

$$\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k) := \{u(f_1, \cdots, f_k): f_i \in \mathcal{A}_i \ \forall i\}.$$

Now, we can ask: what if anything can be said about the VC-dimension of the family $\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)$ in terms of the VC-dimensions of the individual families $\mathcal{A}_1, \cdots, \mathcal{A}_k$? Dudley [14, Theorem 9.2.3, p. 85] shows that if each of the $\mathcal{A}_i$ has finite VC-dimension, then so does $\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)$. For the present purposes, however, this is not enough.

*Theorem 6:* Suppose $\mathcal{A}_1, \cdots, \mathcal{A}_k$ are families of binary-valued functions, and that $u: \{0, 1\}^k \to \{0, 1\}$ is arbitrary. Finally, suppose VC-dim$(\mathcal{A}_i)$ is finite for each $i$. Then, $\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)$ also has finite VC-dimension, and in fact

$$\text{VC-dim}[\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)] < \alpha(k) \max_{1 \leq i \leq k} \text{VC-dim}(\mathcal{A}_i)$$

where $\alpha(k) =: \alpha$ is the smallest integer that satisfies

$$k < \frac{\alpha}{\log(e\alpha)}. \tag{10}$$

In particular,[6]

$$\alpha(k) \leq \lceil 2k \log(ek) \rceil$$

so that

$$\text{VC-dim}\,[\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)] \leq \lceil 2k \log(ek) \rceil \max_{1 \leq i \leq k} \text{VC-dim}(\mathcal{A}_i).$$

*Proof:* The proof is based on the following simple observation. Let $\mathcal{A}$ be a family of binary-valued functions, and define the integer $\pi(n; \mathcal{A})$ as follows. For a fixed finite set $S \subseteq X$, define $\pi(S; \mathcal{A})$ to be the number of distinct subsets of $S$ of the form $S \cap A$ for some $A \in \mathcal{A}$. Now, define, for each integer $n \geq 1$

$$\pi(n; \mathcal{A}) := \max_{|S| = n} \pi(S; \mathcal{A}).$$

If VC-dim$(\mathcal{A}) = n$, then $\pi(n; \mathcal{A}) = 2^n$ for every $n \leq d$. Therefore, if $\pi(n; \mathcal{A}) < 2^n$, then VC-dim$(\mathcal{A}) < n$.

To apply this observation to the problem at hand, we derive an upper bound on the integer $\pi(n; \mathcal{U})$, where $\mathcal{U} := \mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)$. Recall the following well-known result [5], [31], [33]. Given integers $d, n \geq 1$, define

$$\phi(n, d) := \sum_{i=0}^{d} \binom{n}{i} \qquad \text{if } n > d, \ 2^n \text{ if } n \leq d,$$

where

$$\binom{n}{i} = \frac{n!}{(n-i)!i!}$$

is the binomial coefficient. If VC-dim$(\mathcal{A}) = d$, then for $n > d$

$$\pi(n; \mathcal{A}) \leq \phi(n, d) \leq 2\frac{n^d}{d!} \leq \left(\frac{en}{d}\right)^d, \qquad \forall\, n \geq d \geq 1.$$

Now, let $S = \{x_1, \cdots, x_n\}$ be an arbitrary set of cardinality $n$, fix an index $i \in \{1, \cdots, k\}$, and examine the set of binary vectors of the form

$$[f(x_1) \cdots f(x_n)]^t \in \{0, 1\}^n$$

generated by varying $f$ over $\mathcal{A}_i$. By definition, it follows that the total number of such vectors is no larger than $\pi(S; \mathcal{A}_i)$. Next, given the function $u: \{0, 1\}^k \to \{0, 1\}$, we can define a corresponding function $\mathrm{u}_n: [\{0, 1\}^n]^k \to \{0, 1\}^n$ as follows. Given $\mathbf{v}^1, \cdots, \mathbf{v}^k \in \{0, 1\}^n$, let

$$\mathrm{u}_n(\mathbf{v}^1, \cdots, \mathbf{v}^k) := \left[u(v_1^1, \cdots, v_1^k) \cdots u(v_n^1, \cdots, v_n^k)\right]^t \in \{0, 1\}^n.$$

[5]It is common to refer to such functions as *Boolean functions.*

[6]Note that $\lceil a \rceil$ denotes the smallest integer greater than or equal to $a$.

Now, we come to the key point. Suppose it is desired to estimate the integer $\pi(n; \mathcal{U})$. By definition, $\pi(n; \mathcal{U})$ is the number of distinct vectors of the form

$$[u(f_1(x_1), \cdots, f_k(x_1)) \cdots u(f_1(x_n), \cdots, f_k(x_n))]^t \in \{0, 1\}^n$$

obtained by varying each function $f_i$ over the corresponding family $\mathcal{A}_i$. This is the same as the number of distinct vectors $\mathrm{u}_n(\mathbf{v}^1, \cdots, \mathbf{v}^k)$ obtained by varying each vector $\mathbf{v}^i$ over a set of cardinality $\pi(S; \mathcal{A}_i)$. Hence

$$\pi(S; \mathcal{U}) \leq \prod_{i=1}^{k} \pi(S; \mathcal{A}_i) \leq \prod_{i=1}^{k} \pi(n; \mathcal{A}_i).$$

Now

$$\pi(n; \mathcal{A}_i) \leq \phi(n, d_i) \leq \phi(n, d), \ \forall i$$

where $d_i :=$ VC-dim$(\mathcal{A}_i)$, and as before

$$d := \max_i d_i.$$

Therefore

$$\pi(n; \mathcal{A}_i) \leq \left(\frac{en}{d}\right)^d, \qquad \forall n \geq d, \ \forall i.$$

As a result

$$\pi(S; \mathcal{U}) \leq \left(\frac{en}{d}\right)^{kd}, \qquad \forall n \geq d.$$

Finally, because $S$ is arbitrary, it follows that

$$\pi(n; \mathcal{U}) \leq \left(\frac{en}{d}\right)^{kd}, \qquad \forall n \geq d.$$

Hence, if we can find an integer $n$ such that $\pi(n; \mathcal{U}) < 2^n$, then VC-dim$(\mathcal{U}) < n$. Thus, we look for a solution for $n$ to the inequality

$$\left(\frac{en}{d}\right)^{kd} < 2^n.$$

Let us simplify the problem by looking for solutions of the form $n = \alpha\, d$, where $\alpha = \alpha(k)$. Then, the above inequality becomes

$$(e\alpha)^{kd} < 2^{\alpha\, d},$$

or, after taking the log of both sides and dividing by $d$,

$$k \log(e\alpha) < \alpha.$$

Hence, if $\alpha$ satisfies (10), then $\pi(\alpha\, d; \mathcal{U}) < 2^{\alpha\, d}$, and VC-dim$(\mathcal{U}) < \alpha\, d$.

A quick estimate for $\alpha(k)$ can be obtained as follows. Rewrite the inequality (10) as

$$k \log e \ln(e\alpha) < \alpha.$$

Then, this inequality is satisfied provided

$$\alpha = \lceil 2k \log(ek) \rceil.$$

The proof is easy and left to the reader. Hence, one can rewrite the conclusion of Theorem 6 as

$$\text{VC-dim}[\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)] \leq \lceil 2k \log(ek) \rceil \max_{1 \leq i \leq k} \text{VC-dim}(\mathcal{A}_i).$$

This is the desired conclusion. ■

The table below shows, for $k$ between 2 and 10, the *smallest* integer $\alpha(k)$ such that $k \log(e\alpha) < \alpha$, as well as the number $\lceil 2k \log(ek) \rceil$, i.e., the number $2k \log(ek)$ rounded upward. From this table, it can be seen that the estimate $\alpha(k) = \lceil 2k \log(ek) \rceil$ is not too conservative, and has the advantage of being "in closed form."

| $k$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha(k)$ | 10 | 17 | 25 | 33 | 41 | 50 | 59 | 68 | 78 |
| $\lceil 2k \log(ek) \rceil$ | 10 | 19 | 28 | 38 | 49 | 60 | 72 | 84 | 96. |

Theorem 6 is a considerable extension of previously known results. As stated above, it is shown in [14] that if VC-dim$(\mathcal{A}_i)$ is finite for every $i$, then so is VC-dim$[\mathcal{U}(\mathcal{A}_1, \cdots, \mathcal{A}_k)]$; but no explicit estimate is given for the latter VC-dimension. The only other related result we are aware of is from [6, Lemma A.2], which states that

$$\text{VC-dim}(\mathcal{A}\Delta\mathcal{A}) \leq 10\text{VC-dim}(\mathcal{A})$$

where

$$\mathcal{A}\Delta\mathcal{A} := \{A\Delta B\colon A, B \in \mathcal{A}\}.$$

This bound is now obtained as a special case of Theorem 6.

### C. Uniform Convergence Properties of Iterated Families

In the preceding subsection, we studied the VC-dimension of Boolean functions of sets. These results are used in the present subsection to show that if a collection of sets $\mathcal{A}$ has the property that empirical probabilities converge uniformly, then (roughly speaking) every Boolean function of $\mathcal{A}$ also has the UCEP property. Similarly, it can be shown that if a family of functions $\mathcal{F}$ has the property that empirical means converge uniformly, then every uniformly continuous function of $\mathcal{F}$ also has the UCEM property; however, the case of function families is not studied here, and refered instead to [36, Section 5.8.2].

Given a measurable space $(X, \mathcal{S})$, suppose $\mathcal{A} \subseteq \mathcal{S}$ is a given collection of sets. By a slight abuse of notation, we can also think of $\mathcal{A}$ as a family of functions mapping $X$ into $\{0, 1\}$. Suppose $k$ is an integer and that $u\colon \{0, 1\}^k \to \{0, 1\}$ is a given function. We can define a corresponding collection of sets $\mathcal{U}(\mathcal{A})$ as follows. Suppose $f_1, \cdots, f_k\colon X \to \{0, 1\}$ are binary-valued functions. Then, we define $u(f_1, \cdots, f_k)\colon X \to \{0, 1\}$ to be the binary-valued function

$$x \mapsto u[f_1(x), \cdots, f_k(x)].$$

Finally, $\mathcal{U}(\mathcal{A})$ is defined as

$$\mathcal{U}(\mathcal{A}) := \{u(f_1, \cdots, f_k)\colon f_i \in \mathcal{A}\ \forall i\}.$$

This equation defines $\mathcal{U}(\mathcal{A})$ as a family of binary-valued functions, but there is an obvious interpretation of $\mathcal{U}(\mathcal{A})$ as a collection of subsets of $X$. A few examples serve to illustrate the definition.

Given $\mathcal{A} \subseteq \mathcal{S}$, define

$$\mathcal{A} \oplus \mathcal{A} := \{A \cup B\colon A, B \in \mathcal{A}\}$$
$$\mathcal{A} \odot \mathcal{A} := \{A \cap B\colon A, B \in \mathcal{A}\}$$
$$\mathcal{A}\Delta\mathcal{A} := \{A\Delta B\colon A, B \in \mathcal{A}\}.$$

These collections of sets can be formed from $\mathcal{A}$ by defining

$$u(a, b) = \max\{a, b\},\ a \cdot b,\ |a - b|,$$

respectively.

*Theorem 7:* Suppose $\mathcal{A} \subseteq \mathcal{S}$ has the property of uniform convergence of empirical probabilities, and that $u$: $\{0, 1\}^k \to \{0, 1\}$ is a given function. Then, $\mathcal{U}(\mathcal{A})$ also has the UCEP property.

*Proof:* The proof consists of showing that the collection $\mathcal{U}(\mathcal{A})$ satisfies the condition (3) with $\mathcal{A}$ replaced by $\mathcal{U}(\mathcal{A})$, and then appealing to Theorem 1. By assumption, $\mathcal{A}$ has the UCEP property. Hence, by Theorem 1

$$\frac{E_{P^m}[d(\mathbf{x};\ \mathcal{A})]}{m} = 0,$$

where we use $d(\mathbf{x};\ \mathcal{A})$ instead of $d(\mathbf{x})$ to make clear which collection of sets we are talking about. Now, by Theorem 6, a constant $\alpha(k)$ exists that depends only on $k$ (and not on $\mathcal{A}$ or $\mathbf{x}$ or $m$) such that

$$d(\mathbf{x};\ \mathcal{U}(\mathcal{A})) \le \alpha(k)\, d(\mathbf{x};\ \mathcal{A}),\ \forall \mathbf{x} \in X^m,\ \forall m \ge 2.$$

Hence

$$\lim_{m\to\infty} \frac{E_{P^m}\left[d(\mathbf{x};\ \mathcal{U}(\mathcal{A}))\right]}{m} \le \alpha(k) \lim_{m\to\infty} \frac{E_{P^m}[d(\mathbf{x};\ \mathcal{A})]}{m} = 0.$$

Hence, by Theorem 1, it follows that $\mathcal{U}(\mathcal{A})$ also has the UCEP property. ∎

*Example 5:* Consider the problem of learning the family of convex polygons inside the unit square $[0, 1]^2$. It is known [29, pp. 22–24] that this family has the UCEP (and, hence, the ASCEP) property. Hence, by Theorem 5, it follows that this family is also learnable, and that *every* consistent algorithm is PUAC. For instance, we could simply choose $H_m$ to be the *smallest* convex polygon that correctly classifies all of the sample points, i.e., the convex hull of all the positive examples [all $x_i$ such that $I_T(x_i) = 1$]. This algorithm is PUAC. An examination of the proof in [29] reveals that the claim holds with $[0, 1]^2$ replaced by $[0, 1]^k$ for every integer $k$.

More generally, let $k, l$ be fixed positive integers. Suppose $X = [0, 1]^k$, $\mathcal{S}$ = the Borel $\sigma$-algebra on $X$, and let $P$ equal the uniform probability measure on $X$. Let $\mathcal{C}$ consist of all unions of $l$ *or fewer* convex sets in $X$. Then, we can write

$$\mathcal{C} = \bigcup_{s=1}^{l} \mathcal{C}_s$$

where the collection $\mathcal{C}_s$ consists of all unions of *exactly* $s$ convex sets in $X$. By Theorem 7, it follows that each $\mathcal{C}_s$ has the UCEP property. Hence, their finite union $\mathcal{C}$ also has the UCEP property. Now, it follows from Theorem 5 that the collection $\mathcal{C}$ is PUAC learnable, and that in fact every consistent algorithm is PUAC. In contrast to the case, however, in which $\mathcal{C}$ consists of all convex sets in $X$, finding a consistent hypothesis is no longer as straightforward as taking the convex hull of all positive examples. ∎

## VI. Model-Free Learning

In this section, we study the model-free learning problem described in Section II-C. Some previously known results for the problem of learning an unknown target function are extended to this case. Throughout the notation is as in Section II.

### A. A Sufficient Condition for Learnability

In the case in which the data develops from a target function, it is known (see [4] and its extension to function learning discussed in Section II-C) that $\mathcal{F}$ is PAC learnable if it has finite metric entropy. In this subsection, an analogous result is proved in the case of model-free learning, with the assumption that the loss function satisfies a uniform Lipschitz condition.

Specifically, it is assumed that the "decision space" $U$ is a subset of $\Re$, and that a finite constant $\mu$ exists such that

$$|\ell(y, u_1) - \ell(y, u_2)| \le \mu |u_1 - u_2|,\quad \forall u_1, u_2 \in \Re,\ \forall y \in Y. \tag{11}$$

The minimum empirical risk algorithm in the case of model-free learning is a natural extension of that introduced previously in [4]. Let $\{g_1, \cdots, g_k\}$ be a finite subset of $\mathcal{H}$. Once samples $(x_1, y_1), \cdots, (x_m, y_m)$ are drawn, define

$$\hat{J}_i := \frac{1}{m} \sum_{j=1}^{m} \ell\left[y_j, g_i(x_j)\right], \qquad 1 \le i \le k.$$

Then, the hypothesis $h_m$ is chosen as a $g_{i_0}$ such that

$$\hat{J}_{i_0} = \min_{1\le i\le k} \hat{J}_i.$$

Now, we can state a result analogous to Theorem 3.

*Theorem 8:* Suppose

- family of probabilities $\overline{\mathcal{P}}$ has the property that every $\overline{P} \in \overline{\mathcal{P}}$ has the same marginal measure on $X$, call it $P$;
- hypothesis class $\mathcal{H}$ has the property that

$$N(\epsilon, \mathcal{H}, d_P) < \infty,\ \forall \epsilon > 0;$$

- loss function $\ell$ satisfies the uniform Lipschitz condition (11) above.

Then, the triple $(\mathcal{H}, \overline{\mathcal{P}}, \ell)$ is PAC learnable. In particular, given any $\epsilon > 0$, choose $\{g_1, \cdots, g_k\}$ to be an $\epsilon_0/2\mu$-cover of $\mathcal{H}$ with respect to $d_P$ for some $\epsilon_0 < \epsilon$. Then, the minimum empirical risk algorithm applied to $\{g_1, \cdots, g_k\}$ is PAC to accuracy $\epsilon$, and

$$r_{\mathrm{mf}}(m, \epsilon) \le k \exp(-m\epsilon^2/8).$$

Hence, the algorithm is PAC to accuracy $\epsilon$ and confidence $\delta$ provided at least

$$m \ge \frac{8}{\epsilon^2} \ln \frac{k}{\delta}$$

samples are drawn.

*Proof:* The first step is to show that, for every $\overline{P} \in \overline{\mathcal{P}}$ and every $f, g \in \mathcal{H}$, we have

$$|J(f, \overline{P}) - J(g, \overline{P})| \leq \mu\, d_P(f, g). \tag{12}$$

This is a ready consequence of (11), because

$$\begin{aligned}
&|J(f, \overline{P}) - J(g, \overline{P})| \\
&\quad = \left| \int_{X\times Y} [\ell(y, f(x)) - \ell(y, g(x))]\, \overline{P}(dx, dy) \right| \\
&\quad \leq \int_{X\times Y} |\ell(y, f(x)) - \ell(y, g(x))|\, \overline{P}(dx, dy) \\
&\quad \leq \mu \int_{X\times Y} |f(x) - g(x)|\, \overline{P}(dx, dy) \\
&\quad = \mu \int_{X} |f(x) - g(x)|\, P(dx) \\
&\quad = \mu d_P(f, g).
\end{aligned}$$

Another useful way of expressing the above inequality is the following: given a function $f\colon X \to U$, define the corresponding function $\ell_f\colon X \times Y \to [0, 1]$ by

$$\ell_f(x, y) := \ell(y, f(x)).$$

Then, the above argument in fact shows that

$$d_{\overline{P}}(\ell_f, \ell_g) \leq \mu\, d_P(f, g)$$

and of course

$$|J(f, \overline{P}) - J(g, \overline{P})| \leq d_{\overline{P}}(\ell_f, \ell_g).$$

To prove that the minimum empirical risk algorithm applied to an $\epsilon_0/2\mu$-cover of $\mathcal{H}$ is PAC to accuracy $\epsilon$, let $\overline{P} \in \overline{\mathcal{P}}$ be arbitrary, and select an $h = h(\epsilon, \overline{P})$ such that

$$J(h, \overline{P}) \leq J^*(\overline{P}) + \frac{\epsilon - \epsilon_0}{2}.$$

Such an $h$ exists, by the definition of $J^*(\overline{P})$. Now, it is known that $h$ is within a distance $\epsilon_0/2\mu$ (with respect to $d_P$) of one of the $g_i$'s, though it is not known which one. Assume without loss of generality that the $g_i$'s are renumbered such that $d_P(h, g_k) \leq \epsilon_0/2\mu$, which in turn implies that

$$J(g_k, \overline{P}) \leq J(h, \overline{P}) + \epsilon_0/2 \leq J^*(\overline{P}) + \epsilon/2.$$

Assume that the renumbering is such that

$$J(g_i, \overline{P}) > J^*(\overline{P}) + \epsilon, \qquad \text{for } 1 \leq i \leq l$$

and

$$J(g_i, \overline{P}) \leq J^*(\overline{P}) + \epsilon, \qquad \text{for } l+1 \leq i \leq k.$$

Note that $l \leq k - 1$. Suppose i.i.d. samples $(x_1, y_1), \cdots, (x_m, y_m)$ are drawn in accordance with $\overline{P}$, and as before, let

$$\mathrm{z} := [(x_1, y_1) \cdots (x_m, y_m)]^t.$$

Note that the inequality $J(h_m, \overline{P}) \leq J^*(\overline{P}) + \epsilon$ is satisfied if $h_m$ is one of $g_{l+1}, \cdots, g_k$. This will be the case if

$$\hat{J}(g_k; \mathrm{z}) \leq J^*(\overline{P}) + 3\epsilon/4$$

and

$$\hat{J}(g_i; \mathrm{z}) > J^*(\overline{P}) + 3\epsilon/4, \qquad \text{for } 1 \leq i \leq l.$$

Hence, in order for the inequality $J(h_m, \overline{P}) \leq J^*(\overline{P}) + \epsilon$ to be *violated*, it is necessary that

$$\hat{J}(g_k; \mathrm{z}) > J^*(\overline{P}) + 3\epsilon/4 \geq J(g_k, \overline{P}) + \epsilon/4$$

or

$$\hat{J}(g_i; \mathrm{z}) \leq J^*(\overline{P}) + 3\epsilon/4 < J(g_i, \overline{P}) - \epsilon/4, \quad \text{for some } i \leq l.$$

Note that $J(g_i, \overline{P})$ is just the expected value of the function $\ell_{g_i}$, although $\hat{J}(g_i; \mathrm{z})$ is its empirical mean based on the multisample z. Hence, by Hoeffding's inequality, each of the above events has a probability no larger than $\exp(-m\epsilon^2/8)$. Hence

$$\begin{aligned}
&r_{\mathrm{mf}}(m, \epsilon) \\
&\quad = \Pr\left\{ J(h_m, \overline{P}) > J^*(\overline{P}) + \epsilon \right\} \leq k \exp(-m\epsilon^2/8).
\end{aligned}$$

Setting

$$k \exp(-m\epsilon^2/8) \leq \delta$$

and solving for $m$ leads to the sample complexity estimate. ■

As a specific application of the above approach, consider the problem of learning a binary concept class with a noisy oracle. Thus, a probability space $(X, \mathcal{S}, P)$ and a concept class $\mathcal{C} \subseteq \mathcal{S}$ exist. Given a target concept $T \in \mathcal{C}$ and a random sample $x \in X$, a noisy oracle outputs $I_T(x)$ with a probability $1 - \alpha$ and $1 - I_T(x)$ with a probability of $\alpha$, where the error probability $\alpha \in [0, 0.5)$ is known. The hypothesis class $\mathcal{H}$ is taken as $\mathcal{C}$ itself, and the collection of probability measures $\overline{\mathcal{P}}$ is taken as $\{\overline{P}_T, T \in \mathcal{C}\}$, where each $\overline{P}_T(\cdot)$ has the marginal $P$ on $X$, and

$$\overline{P}_T(1|x) = (1 - \alpha) I_T(x) + \alpha I_T(x),$$

and

$$\overline{P}_T(0|x) = \alpha I_T(x) + (1 - \alpha)(1 - I_T(x)).$$

The above definition of $\overline{P}_T$ ensures that if $I_T(x) = 1$, then $y = 1$ with probability $1 - \alpha$, and $y = 0$ with probability $\alpha$; the situation is reversed if $I_T(x) = 0$. Finally, the loss function $\ell(y, u)$ is taken as $|y - u|$.

An easy calculation shows that, for each $A \subseteq \mathcal{S}$, we have

$$\overline{P}_T(A \times \{0\}) := (1 - \alpha) P(A) - (1 - 2\alpha) P(A \cap T),$$

and

$$\overline{P}_T(A \times \{1\}) := \alpha P(A) + (1 - 2\alpha) P(A \cap T).$$

Also, for each $H \in \mathcal{H}$ and each $P_T \in \overline{\mathcal{P}}$, it is easy to show that

$$\begin{aligned}
J(H, \overline{P}_T) &= \int_{X\times\{0,1\}} |y - H(x)|\, P(dx) \\
&= \alpha + (1 - 2\alpha) P(H \Delta T) \\
&= \alpha + (1 - 2\alpha)\, d_P(H, T).
\end{aligned}$$

Finally

$$J^*(\overline{P}_T) = \inf_{H \in \mathcal{C}} J(H, \overline{P}_T) = \alpha$$

which is achieved by the choice $H = T$.

To apply Theorem 8, we begin by estimating the Lipschitz constant of the function $J$. Clearly, the loss function $\ell(y, u) = |y-u|$ satisfies a Lipschitz condition with the Lipschitz constant of one, because

$$\big|\, |y-u_1| - |y-u_2| \,\big| = |u_1 - u_2|, \ \forall y, u_1, u_2 \in \{0, 1\}.$$

By taking advantage of the special nature of the function $J$, however, it is possible to obtain a lower Lipschitz constant. Recall that

$$J(H, \overline{P}_T) = \alpha + (1-2\alpha)\, d_P(H, T).$$

Now, the claim is that

$$|J(H_1, \overline{P}_T) - J(H_2, \overline{P}_T)| \le (1-2\alpha)\, d_P(H_1, H_2).$$

In other words, (12) is satisfied with $\mu = (1-2\alpha)$. Note that

$$\begin{aligned} &|J(H_1, \overline{P}_T) - J(H_2, \overline{P}_T)| \\ &\quad = (1-2\alpha)\, |d_P(H_1, T) - d_P(H_2, T)|. \end{aligned}$$

So, the claim is established if it can be shown that

$$|d_P(H_1, T) - d_P(H_2, T)| \le d_P(H_1, H_2).$$

This last inequality, however, is immediate, because from the triangle inequality, we have

$$d_P(H_1, T) - d_P(H_2, T) \le d_P(H_1, H_2),$$

and

$$d_P(H_2, T) - d_P(H_1, T) \le d_P(H_1, H_2).$$

Thus, in order for the minimal empirical risk algorithm to be PAC to accuracy $\epsilon$, it is enough to apply the algorithm to an $\epsilon/2(1-2\alpha)$-cover of $\mathcal{C}$. Note that, because the infimum $J^*(\overline{P}_T)$ is actually attained for each $T \in \mathcal{C}$ (by choosing $H = T$), it is not necessary to choose an $\epsilon_0 < \epsilon$ as in the proof of Theorem 8. Now, by the theorem, it follows that

$$r_{\mathrm{mf}}(m, \epsilon) \le k(\epsilon/2(1-2\alpha)) \exp(-m\epsilon^2/8)$$

where the notation $k(\epsilon/2(1-2\alpha))$ serves to remind us that $k$ is the cardinality of an $\epsilon/2(1-2\alpha)$-cover of $\mathcal{C}$.

Because

$$J(H, \overline{P}_T) = \alpha + (1-2\alpha)\, d_P(H, T)$$

and

$$J^*(\overline{P}_T) = \alpha$$

it follows that

$$d_P(H, T) > \epsilon \Leftrightarrow J(H, \overline{P}_T) > J^*(\overline{P}_T) + (1-2\alpha)\epsilon.$$

Recall the definition of the quantity $r(m, \epsilon)$ from (1). The previous relationship now means that

$$r_{\mathrm{mf}}(m, (1-2\alpha)\epsilon) = r(m, \epsilon).$$

From the discussion in Section II-D, we have that, in case the minimum empirical risk algorithm is applied to *noise-free* oracle outputs

$$r(m, \epsilon) \le k(\epsilon/2) \exp(-m\epsilon^2/8)$$

whereas from the above inequality

$$r_{\mathrm{mf}}(m, (1-2\alpha)\epsilon) \le k(\epsilon/2) \exp[-m(1-2\alpha)^2\epsilon^2/8].$$

The effect of the oracle noise can be gauged from the above two bounds. In the noise-free case, in order to ensure that the hypothesis $H_m$ produced by the algorithm satisfies $d_P(T, H_m) \le \epsilon$ with probability at least $1-\delta$, it is enough to take

$$m_{\text{noise-free}} = \frac{8}{\epsilon^2} \ln \frac{k(\epsilon/2)}{\delta}$$

samples. In contrast, in the case of noisy oracles, it is enough to take

$$m_{\text{noisy}} = \frac{8}{(1-2\alpha)^2\epsilon^2} \ln \frac{k(\epsilon/2)}{\delta}$$

samples. If $\alpha = 0$ so that the oracle is noise-free, the above bound reduces to its predecessor, whereas if $\alpha \to 0.5^-$, the above bound approaches infinity.

### B. A Necessary Condition

In this subsection, we state and prove a necessary condition for model-free learnability that generalizes an earlier result from [4]. The previously known result states the following.

*Theorem 9 [4]:* Suppose $\mathcal{C}$ is a given concept class, that $P$ is a given probability measure, and that a $2\epsilon$-separated subset $\{B_1, \cdots, B_M\}$ in $\mathcal{C}$ exists with respect to $d_P$; that is

$$d_P(B_i, B_j) > 2\epsilon, \qquad \forall\, i \ne j.$$

Then, for each $\delta > 0$, any algorithm that PAC learns the concept class $\mathcal{C}$ to accuracy $\epsilon$ and confidence $\delta$ requires at least $\lg M(1-\delta)$ samples.

This result is now generalized to the following:

*Theorem 10:* Suppose $Y = \{0, 1\}$, and consider the model-free learning problem. Given the family $\overline{\mathcal{P}}$, suppose a single fixed probability $P$ on $X$ exists such that the marginal probability of every $\overline{P} \in \overline{\mathcal{P}}$ equals $P$. Suppose probabilities $\overline{P}_1, \cdots, \overline{P}_M \in \overline{\mathcal{P}}$ exist such that no $H \in \mathcal{H}$ can satisfy the inequality

$$J(H, \overline{P}_j) \le J^*(\overline{P}_j) + \epsilon$$

for more than one value of the index $j$. Then, any algorithm that is PAC to accuracy $\epsilon$ and confidence $\delta$ requires at least $\lg [M(1-\delta)]$ samples.

*Remarks:* To see that the above theorem is a true generalization of Theorem 9, formulate the standard PAC learning

problem as a model-free learning problem. This can be achieved by taking $Y = U = \{0, 1\}$, $\ell(y, u) = |y - u|$, and defining the family of probability measures $\overline{\mathcal{P}}$ as $\{\overline{P}_T: T \in \mathcal{H}\}$, where

$$\overline{P}_T(1|x) = I_T(x), \quad \overline{P}_T(0|x) = 1 - I_T(x).$$

This formula is analogous to those in the preceding subsection with the error rate $\alpha$ set equal to zero. In this case

$$J(H, \overline{P}_T) = d_P(H, T).$$

Now, if $\{B_1, \cdots, B_M\}$ is a $2\epsilon$-separated subset of $\mathcal{H}$, then clearly no $H \in \mathcal{H}$ can satisfy $d_P(H, B_j) \leq \epsilon$ for more than one index $j$. Hence, the corresponding set of probabilities $\{\overline{P}_{B_1}, \cdots, \overline{P}_{B_M}\}$ satisfies the hypothesis of the theorem, and the theorem itself reduces to Theorem 9. Thus, the requirement that no hypothesis $H$ can be $\epsilon$-optimal for more than one probability measure is a natural generalization of the $2\epsilon$-separation requirement.

*Proof:* Let $A_m$: $X^m \times \{0, 1\}^m \to \mathcal{H}$ denote the algorithm, and let $H_m(\mathbf{x}, \mathbf{L}) \in \mathcal{H}$ denote the hypothesis produced by the algorithm with the input $(\mathbf{x}, \mathbf{L}) \in X^m \times \{0, 1\}^m$. Then, the "PAC" assumption implies that

$$\begin{aligned}\overline{P}_j^m\{(\mathbf{x}, \mathbf{L}) \in X^m \times \{0, 1\}^m: J[H_m(\mathbf{x}, \mathbf{L}), \overline{P}_j] \\ \leq J^*(\overline{P}_j) + \epsilon\} \\ \geq 1 - \delta, \qquad \text{for } j = 1, \cdots, M. \end{aligned} \tag{13}$$

Define a function

$$g: \{1, \cdots, M\} \times X^m \times \{0, 1\}^m \to \{0, 1\}$$

as follows. Suppose $1 \leq j \leq M$, $\mathbf{x} \in X^m$, and $\mathbf{L} \in \{0, 1\}^m$. Then

$$g(j, \mathbf{x}, \mathbf{L}) = \begin{cases} 1, & \text{if } J[H_m(\mathbf{x}, \mathbf{L}), \overline{P}_j] \leq J^*(\overline{P}_j) + \epsilon, \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

Then, the hypothesis of the theorem implies that, for each $\mathbf{x}$, $\mathbf{L}$, the function $g(j, \mathbf{x}, \mathbf{L})$ can equal one for at most one value of $j$. Hence,

$$\sum_{i=1}^{M} g(j, \mathbf{x}, \mathbf{L}) \leq 1, \qquad \forall \mathbf{x}, \mathbf{L}. \tag{14}$$

With each probability $\overline{P}_j$, we can associate the conditional probabilities $q_j(1|x)$ and $q_j(0|x)$. Given $\mathbf{L} \in \{0, 1\}^m$ and $\mathbf{x} \in X^m$, define

$$Q_j(\mathbf{x}, \mathbf{L}) := \prod_{i=1}^{m} q_j(b_i|x_i), \qquad \text{where } \mathbf{L} = b_1 \cdots b_m.$$

Thus, given any function $f$: $X^m \times \{0, 1\}^m \to \Re$, we have

$$\begin{aligned}\int_{X^m \times \{0,1\}^m} f(\mathbf{x}, \mathbf{L}) \overline{P}_j^m (d\mathbf{x}, d\mathbf{L}) \\ = \int_{X^m} \sum_{\mathbf{L} \in \{0,1\}^m} f(\mathbf{x}, \mathbf{L})\, Q_j(\mathbf{x}, \mathbf{L})\, P^m(dx).\end{aligned}$$

In particular

$$\begin{aligned}\int_{X^m \times \{0,1\}^m} \sum_{i=1}^{M} g(j, \mathbf{x}, \mathbf{L}) \overline{P}_j^m (d\mathbf{x}, d\mathbf{L}) \\ = \int_{X^m} \sum_{\mathbf{L} \in \{0,1\}^m} \sum_{i=1}^{M} g(j, \mathbf{x}, \mathbf{L})\, Q_j(\mathbf{x}, \mathbf{L})\, P^m(d\mathbf{x}) \\ \leq \int_{X^m} \sum_{\mathbf{L} \in \{0,1\}^m} P^m(d\mathbf{x}) = 2^m\end{aligned}$$

from (14) and the fact that $Q_j(\mathbf{x}, \mathbf{L}) \leq 1$ for all $j$, $\mathbf{x}$, $\mathbf{L}$.

On the other hand, the PAC assumption (13) implies that

$$E_{\overline{P}_j^m}[g(j, \mathbf{x}, \mathbf{L})] \geq 1 - \delta, \qquad \text{for } j = 1, \cdots, M$$

or, equivalently

$$\begin{aligned}\int_{X^m \times \{0,1\}^m} g(j, \mathbf{x}, \mathbf{L}) \overline{P}_j^m (d\mathbf{x}, d\mathbf{L}) \\ \geq 1 - \delta, \qquad \text{for } j = 1, \cdots, M.\end{aligned}$$

Consequently

$$\sum_{i=1}^{M} \int_{X^m \times \{0,1\}^m} g(j, \mathbf{x}, \mathbf{L}) \overline{P}_j^m (d\mathbf{x}, d\mathbf{L}) \geq M(1 - \delta).$$

Combining the preceding two inequalities shows that

$$2^m \geq M(1 - \delta)$$

which is the desired conclusion. ■

## VII. The Issue of Representation

In this section, we study the importance of the issue of *representation* in determining whether a given collection of sets has the UCEP property. This is done through example.

*Example 6:* Let $X = [0, 1]$, $\mathcal{S}$ = the Borel $\sigma$-algebra on $X$, and let $P$ be the uniform distribution on $X$. Define $\mathcal{A}$ to be the collection of all unions of the form $[0, a] \cup G$, where $a \in [0, 0.5]$ and $G$ is a finite set. Then, it can be shown that $\mathcal{A}$ *does not* have the UCEP property. More generally, *any* collection $\mathcal{A}$ fails to have the UCEP property, provided two conditions are satisfied: i) A number $\alpha < 1$ exists such that $P(A) \leq \alpha$ for all $A \in \mathcal{A}$ and ii) for *every* finite set $G$, a corresponding set $A \in \mathcal{A}$ exists such that $G \subseteq A$. The proof of this statement is easy and left to the reader. The collection of sets in the present example satisfies these conditions with $\alpha = 0.5$. Hence, $\mathcal{A}$ does not have the UCEP property.

Let us now define the equivalence relation $\sim$ on $\mathcal{A}$ by defining $A \sim B$ if $d_P(A, B) = 0$. Let $\mathcal{A}/\sim$ denote the collection of equivalence classes under this relation. The objective of the example is to show that, if the representative elements of the family of equivalence classes $\mathcal{A}/\sim$ are selected on one way, then the resulting reduced collection has the UCEP property, whereas if they are selected in another way, then the resulting reduced collection *fails* to have the UCEP property. Therefore, whether a collection has the UCEP property is very much dependent on the

choice of the representative elements of the equivalence classes. In this sense, the UCEP property is rather "fragile."

Suppose we define the collection $\mathcal{A}_1$ as $\{[0, a], a \in [0, 0.5]\}$. Then, each set in $\mathcal{A}_1$ belongs to a different equivalence class in $\mathcal{A}$; moreover, every set in $\mathcal{A}$ is equivalent to exactly one set in $\mathcal{A}_1$. Thus, $\mathcal{A}_1$ is a "reduced" version of $\mathcal{A}$ consisting of one representative element from each equivalence class in $\mathcal{A}/\sim$. This choice might be considered as the natural choice. By the Glivenko–Cantelli lemma, it follows that $\mathcal{A}_1$ does indeed have the UCEP property.

Now, consider another collection $\mathcal{A}_2$, defined next. Let $\mathcal{G}$ denote the collection of all finite subsets of $X$, and suppose $\tau: \mathcal{G} \to [0, 0.5)$ is a *one-to-one* (but not necessarily onto) map. (An example of such a map $\tau$ is given in Example 1.) For each number $a \in [0, 0.5)$, if $a$ belongs to the range of the map $\tau$, so that $a = \tau(F)$ for a *unique* finite set $G$, choose $[0, a] \cup F$ to be the representative of (the equivalence class of) all unions of the form $[0, a] \cup G$, $G \in \mathcal{G}$. If $a$ does not belong to the range of $\tau$, then choose $[0, a]$ to be the representative of all unions of the form $[0, a] \cup G$, $G \in \mathcal{G}$. As $a$ varies over $[0, 0.5)$, this defines a collection $\mathcal{A}_2$ that also represents $\mathcal{A}/\sim$. Now, the previous argument can be applied to show that $\mathcal{A}_2$ *does not* have the UCEP property. Clearly, $P(A) \leq 0.5$ for all $A \in \mathcal{A}_2$, and every finite set $F$ is contained in the set $[0, \tau(F)] \cup F$, which belongs to $\mathcal{A}_2$. ■

The significance of the preceding example lies in the fact that, whereas the choice of representative elements from each equivalence class affects whether the class has the UCEP property, it *does not* affect whether the class is PAC learnable. This latter observation follows from the results of [4], wherein it is shown that a concept class is learnable if and only if it has a finite $\epsilon$-covering number with respect to the pseudometric $d_P$ for each $\epsilon > 0$ (cf. Theorem 2). It is obvious that $\mathcal{C}$ and $\mathcal{C}/\sim$ have the same $\epsilon$-covering numbers for each $\epsilon$. Hence, irrespective of how we choose the elements of $\mathcal{C}/\sim$, the resulting collection of sets is PAC learnable if and only if the original class $\mathcal{C}$ is PAC learnable. Thus, in contrast to the UCEP property, PAC learnability is "robust."

## VIII. Conclusions

In this paper, two new notions of learnability have been introduced, namely, PUAC learnability and MER learnability. PUAC learnability is a stronger requirement than PAC learnability, and MER learnability is a stronger requirement than "solid" learnability. It has been shown through example that an algorithm can be PAC without being PUAC. In spite of the different motivations for introducing these two notions of learnability, it has been shown that both notions are in fact equivalent. A new property called the shrinking width property has been defined, and it has been shown that the shrinking width property is equivalent to PUAC (MER) learnability. It has been shown that, if a function class has the property has the property that empirical means converge uniformly to their true values, then such a class is PUAC (MER) learnable. A new bound on the VC-dimension of the collection of sets obtained by performing Boolean operations on a given collection of sets has been proven, which might be of independent interest. Using this newly derived bound in conjunction with standard results in empirical process theory, it has been shown that if a collection of sets $\mathcal{A}$ has the property and empirical probabilities converge uniformly to their true values, then any other collection $\mathcal{U}(\mathcal{A})$ obtained by performing a number of Boolean operations on elements of $\mathcal{A}$ also has this property. Various known necessary and sufficient conditions for PAC learnability have been extended to model-free learning problems. Finally, it has been shown by example that the choice of representative elements for each equivalence class is important when we study the uniform convergence property, but not so when we study PAC learnability.

An interesting question left open by the results in this paper is the following. How can the notion of PUAC learnability be extended to model-free learning, and what are necessary and sufficient conditions for this type of learnability?

The notion of a PUAC algorithm makes sense even when the underlying probability measure is itself unknown. It turns out that, if the underlying concept class has finite VC-dimension, then not only is every consistent algorithm PAC (as shown in [5]), but in fact every consistent algorithm is PUAC. Similarly, the shrinking width property is also meaningful when $P$ is unknown, and it can be shown that the shrinking width property is equivalent to the property that every consistent algorithm is PUAC. These and other allied results are proved in [36].

## References


[1] M. Anthony and N. Biggs, *Computational Learning Theory*. Cambridge, U.K.: Cambridge University Press, 1992.

[2] P. L. Bartlett and S. R. Kulkarni, "The complexity of model classes, and smoothing of noisy data," in *Proc. 35th Conf. Decision Contr.*, Dec. 1996, pp. 2312–2317.

[3] "Special Issue on Learning Theory," *Syst. Contr. Lett.*, vol. 34, no. 3, June 1998.

[4] G. M. Benedek and A. Itai, "Learnability by fixed distributions," in *Proc. 1st Workshop Computat. Learning Theory*. San Mateo, CA, 1988, pp. 80–90.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Learnability and the Vapnik–Chervonenkis dimension," *J. ACM*, vol. 36, no. 4, pp. 929–965, 1989.

[6] K. L. Buescher and P. R. Kumar, "Simultaneous learning of concepts and simultaneous estimation of probabilities," in *Proc. 4th Workshop Computational Learning Theory*. San Mateo, CA, 1991, pp. 33–42.

[7] ——, "Learning by canonical smooth estimation—Part I: Simultaneous estimation," *IEEE Trans. Automat. Control.*, vol. 42, pp. 545–556, Apr. 1996.

[8] ——, "Learning by canonical smooth estimation—Part II: Learning and choice of model complexity," *IEEE Trans. Automat. Control.*, vol. 42, pp. 557–569, Apr. 1996.

[9] M. Campi and P. R. Kumar, "Learning dynamical systems in a stationary environment," in *Proc. 35th Conf. Decision Contr.*, Dec. 1996, pp. 2308–2311.

[10] ——, "Learning dynamical systems in a stationary environment," *Syst. Contr. Lett.*, vol. 34, no. 3, pp. 125–132, June 1998.

[11] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, vol. 23, pp. 493–507, 1952.

[12] M. A. Dahleh, E. D. Sontag, D. N. C. Tse, and J. N. Tsitsiklis, "Worst-case identification of nonlinear fading memory systems," *Automatica*, vol. 31, pp. 303–308, 1995.

[13] B. DasGupta and E. D. Sontag, "Sample complexity for learning recurrent perceptron mappings," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1479–1487, Sept. 1996.

[14] R. M. Dudley, *A Course on Empirical Processes*, 1984, Lecture Notes in Mathematics, pp. 1–142.

[15] W. Greblicki and M. Pawlak, "Dynamic system identification of nonlinear fading memory systems," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1474–1489, 1994.

[16] B. Hammer, "On the learnability of recursive data," Math. Contr., Signals Syst., to be published.
[17] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Computat.*, vol. 100, pp. 78–150, 1992.
[18] D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth, "Equivalence of models for polynomial learnability," *Inform. Computat.*, vol. 95, pp. 129–161, 1991.
[19] M. Kearns and U. Vazirani, *Introduction to Computational Learning Theory*. Cambridge, MA: MIT Press, 1994.
[20] A. N. Kolmogorov and V. M. Tikhomirov, "$\epsilon$-Entropy and $\epsilon$-capacity of sets in functional spaces," *Am. Math. Soc. Transl.*, vol. 17, pp. 277–364, 1961.
[21] A. Krzyżak, "On estimation of a class of nonlinear systems by the kernel regression estimate," *IEEE Trans. Inform. Theory*, vol. 36, pp. 141–152, Jan. 1990.
[22] S. R. Kulkarni and S. E. Posner, "Universal prediction of nonlinear systems," in *Proc. 34th Conf. Decision Contr.*, Dec. 1995, pp. 4024–4029.
[23] ——, "Nonparametric output prediction for nonlinear fading memory systems," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 29–37, Jan. 1999.
[24] M. Loève, *Probability Theory*. Princeton, NJ: Van Nostrand, 1963.
[25] B. K. Natarajan, "Learning over families of distributions," in *Proc. 1st Workshop Computat. Learning Theory*. San Mateo, CA, 1988, pp. 408–409.
[26] ——, *Machine Learning: A Theoretical Approach*. San Mateo, CA: Morgan-Kaufmann, 1991.
[27] ——, "Probably approximate learning over classes of distributions," *SIAM J. Comput.*, vol. 21, no. 3, pp. 438–449, 1992.
[28] K. Poolla and A. Tikku, "On the time complexity of worst-case identification," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 944–950, May 1994.
[29] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.
[30] ——, Empirical Processes: Theory and Applications, Inst. Math. Stat., vol. 2, ser. NSF-CBMS Regional Conference Series in Probability and Statistics, 1990.
[31] N. Sauer, "On the densities of families of sets," *J. Combin. Theory*, ser. A , vol. 13, pp. 145–147, 1972.
[32] J. M. Steele, "Empirical discrepancies and subadditive processes," *Ann. Prob.*, vol. 6, pp. 118–127, 1978.
[33] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies to their probabilities," *Theory Prob. Applicat.*, vol. 16, no. 2, pp. 264–280, 1971.
[34] ——, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory Prob. Applicat.*, vol. 26, no. 3, pp. 532–553, 1981.
[35] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag, 1982.
[36] M. Vidyasagar, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. London: Springer-Verlag, 1997.
[37] G. Zames, "On the metric complexity of causal linear systems: $\epsilon$-entropy and $\epsilon$-dimension for continuous time," *IEEE Trans. Automat. Contr.*, vol. 24, no. 4, pp. 222–230, Apr. 1979.
[38] G. Zames and J. G. Owen, "A note on metric dimension and feedback in discrete time," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 664–667, 1993.

**M. Vidyasagar** (S'69–M'69–SM'78–F'83) was born in Guntur, Andhra Pradesh, India, on September 29, 1947. He received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the University of Wisconsin, Madison, in 1965, 1967, and 1969, respectively.

He has taught at Marquette University, Milwaukee, WI, from 1969 to 1970, Concordia University, Montreal, Canada, from 1970 to 1980, and the University of Waterloo, Ontario, Canada, from 1980 to 1989. Since June 1989, he has been the Director of the Centre for Artificial Intelligence and Robotics (under the Defence Research and Development Organization), Bangalore, India. In addition to the above, the has held visiting positions at several universities including the Massachusetts Institute of Technology, the University of California (Berkeley, Los Angeles), C.N.R.S., Toulouse, France, the Indian Institute of Science, the University of Minnesota, Minneapolis, and Tokyo Institute of Technology, Japan. He is the author or coauthor of seven books and more than 120 papers in archival journals.

Dr. Vidyasagar has received several honors in recognition of his research activities, including the Distinguished Service Citation from his Alma Mater, The University of Wisconsin. In addition, he is a Fellow of the Indian Academy of Sciences, the Indian National Science Academy, the Indian National Academy of Engineering, and the Third World Academy of Sciences.

**Sanjeev R. Kulkarni** received the B.S. degree in mathematics, the B.S. degree in electrical engineering, and the M.S. degree in mathematics from Clarkson University in 1983, 1984, and 1985, respectively. He also received the M.S. degree in electrical engineering from Stanford University in 1985, and the Ph.D. degree in electrical engineering from M.I.T. in 1991.

From 1985 to 1991, he was a member of the technical staff at M.I.T. Lincoln Laboratory working on the modeling and processing of laser radar measurements. In the spring of 1986, he was a part-time faculty member at the University of Massachusetts, Boston. Since 1991, he has been with Princeton University where he is currently an Associate Professor of electrical engineering. In January 1996, he was a Research Fellow at the Australian National University, and in 1998 was with Susquehanna Investment Group.

Prof. Kulkarni received an ARO Young Investigator Award in 1992, an NSF Young Investigator Award in 1994, and several teaching awards at Princeton University. He served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. His research interests include statistical pattern recognition, nonparametric estimation, learning and adaptive systems, information theory, and image/video processing.