Learning Decision Rules for Pattern Classification Under a Family of Probability Measures

Sanjeev R. Kulkarni, *Senior Member, IEEE*, and Mathukumalli Vidyasagar, *Fellow, IEEE*

***Abstract*—In this paper, uniformly consistent estimation (learnability) of decision rules for pattern classification under a family of probability measures is investigated. In particular, it is shown that uniform boundedness of the metric entropy of the class of decision rules is both necessary and sufficient for learnability under each of two conditions: i) the family of probability measures is *totally bounded*, with respect to the total variation metric, and ii) the family of probability measures *contains an interior point*, when equipped with the same metric. In particular, this shows that insofar as uniform consistency is concerned, when the family of distributions contains a total variation neighborhood, nothing is gained by this knowledge about the distribution. Then two sufficient conditions for learnability are presented. Specifically, it is shown that learnability with respect to each of a finite collection of families of probability measures implies learnability with respect to their union; also, learnability with respect to each of a finite number of measures implies learnability with respect to the convex hull of the corresponding families of uniformly absolutely continuous probability measures.**

***Index Terms*—Pattern classification, estimation, uniform consistency, PAC learning, decision rules, class of distributions, metric entropy, VC dimension.**

I. Introduction

THE standard pattern classification problem is generally formulated as follows. (In this paper, for simplicity, we consider only the two-class case, although all our results can be easily extended to any finite number of classes.) There are two classes ω_0 and ω_1 with prior probabilities $P(\omega_0)$ and $P(\omega_1)$, respectively. We wish to decide whether an unknown object belongs to class ω_0 or ω_1 based on some measured features of the object. The observed feature vector x belongs to a set X (which is typically taken to be $\boldsymbol{R}^d$). Given that the unknown object belongs to class ω_i, the observed feature vector is distributed according to a conditional distribution $P(x|\omega_i)$ on X. If the prior probabilities and conditional distributions are known, then it is well known that the optimal decision rule (in the sense of minimum probability of error, or more generally minimum risk) is the Bayes decision rule. Of course, in many applications these distributions may be unknown or only partially known. In this case, it is often assumed that in addition to the observed feature vector x, one has previous labeled observations $(x_1, y_1), \cdots, (x_m, y_m)$. In the two-class case, $y_k \in \{0, 1\}$ corresponds to an object in class ω_0 or ω_1, respectively, and x_k is a feature vector drawn from $P(x|\omega_0)$ or $P(x|\omega_1)$, respectively. Thus the (x_k, y_k) pairs are drawn independently from the (unknown) distributions $P(x|\omega_i), P(\omega_i)$ characterizing the problem. Using this data, there are various parametric and nonparametric techniques that can be used to provide classification rules (e.g., see [4] for a somewhat dated but excellent and still relevant treatment).

One common approach in pattern classification is to restrict the form of the classifier to belong to some class $\mathcal{D}$ of decision rules. For example, in neural networks, fixing the architecture and size of the network imposes a restriction on the class of decision rules. In the two-class case, each decision rule $D \in \mathcal{D}$ is a map $D : X \to \{0, 1\}$, indicating the classification based on the observed feature x. Of course, with a restriction on the decision rule, in general we cannot hope to perform as well as the optimal Bayes decision rule, even with a very large number of samples. Hence, we should attempt only to try to find the best rule from within the class $\mathcal{D}$. Moreover, with a finite amount of random data, we cannot hope to always learn the optimal rule exactly. Therefore, for finite sample sizes it is natural to require only that with high probability we find a near-optimal decision rule. However, we will require that for a given accuracy and confidence there is a sample size that works uniformly over a class of underlying data distributions. Thus we are really seeking estimators for the optimal decision rule that are uniformly consistent in probability. Much of the early and fundamental work related to this approach to pattern classification was done in the probability and statistics literature—e.g., by Dudley [6], Pollard [14], and Vapnik and Chervonenkis [16]–[18]. The paper of Valiant [15] spurred recent work in the computer science community in this area as well using the terminology PAC learning (for probably-approximately-correct). Haussler [9] has recently refined and consolidated some of the work in the statistics and computer science communities.

In this setting, it is useful to view the data as having x drawn from some (unconditional) distribution P on X, and the label for the class being drawn according to the posterior probabilities $P(0|x)$ and $P(1|x)$, corresponding to class ω_0 and ω_1, respectively. This is completely equivalent to the formulation mentioned earlier. In dealing with prior information on the distributions, most authors have studied one of two extreme cases, namely: i) P is a *fixed and known* probability measure, or ii) P is *arbitrary and unknown* (distribution-free learning). Generally, the posterior probabilities $P(0|x)$ and $P(1|x)$ are allowed to be arbitrary (although some work has also been done for arbitrary P and $P(1|x)$ in some fixed class). In the former case ([1, pp. 149–151] and [18]), $\mathcal{D}$ is learnable (i.e.,

Manuscript received November 1, 1993; revised June 10, 1996. The work of S. R. Kulkarni was supported in part by the National Science Foundation under Grant IRI-9209577 and NYI Award IRI-9457645 and by the U.S. Army Research Office under Grant DAAL03-92-G-0320.

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

M. Vidyasagar is with the Centre for Artificial Intelligence and Robotics, Bangalore 560 001, India.

Publisher Item Identifier S 0018-9448(97)00162-4.

0018–9448/97$10.00 © 1997 IEEE

there is a uniformly consistent estimator) if and only if $\mathcal{D}$ has a finite ϵ-cover for each positive ϵ with respect to the metric $d_P(A, B) = P(A \Delta B)$ as defined following (5) below. In the latter case ([2, pp. 152–161] and [18]), $\mathcal{D}$ is learnable if and only if it has finite Vapnik–Chervonenkis (VC) dimension. Thus the two extreme situations are well-understood, but very little is known about learnability with respect to other families of probability measures. A *necessary* condition for learnability with respect to an arbitrary family of probability measures is that the class of decision rules should have uniformly bounded covering numbers (or, equivalently, uniformly bounded metric entropy)—see Lemma 1 below. However, it has been recently shown that this condition is *not* sufficient [8].

Unless some structure is imposed on the family of distributions under which learning is to take place, it appears that the problem of characterizing when a uniformly consistent estimator exists is intractable. The present paper contributes to this theory by characterizing the existence of uniformly consistent estimators under certain structural assumptions on the family of distributions. Our approach is based on finite covering ideas, which are at the heart of statistical consistency work going back to Wald (e.g., see [19]). First, it is shown that uniform boundedness of the metric entropy of $\mathcal{D}$ is both necessary and sufficient for learnability under each of two conditions: i) the family of probability measures is *totally bounded*, or else ii) the family of probability measures contains an *interior point*, where the topology on the set of all probability measures is that induced by the total variation metric (i.e., the distance between two probability measures is defined as the maximum difference between the probabilities assigned to a measurable set). In the latter case, uniform boundedness of the metric entropy over a family containing an interior point is shown to be equivalent to uniform boundedness of the metric entropy over the class of all distributions, which is equivalent to finite VC dimension. Thus an arbitrarily small total variation neighborhood is still a sufficiently rich family of probability measures that makes uniform learning with respect to this family a stringent requirement. In fact, this shows that insofar as uniform consistency is concerned, nothing is gained from restricting P to a total variation neighborhood, and so one may as well allow arbitrary P. Second, two sufficient conditions for learnability are presented. i) It is shown that learnability with respect to each of a finite collection of families of probability measures implies learnability with respect to the *union* of these families. ii) It is shown that learnability with respect to each of a finite number of probability measures implies learnability with respect to the convex hull of the families of all "commensurate" probability measures—i.e., all measures that are absolutely continuous with respect to a nominal distribution and have a uniform bound on the Radon–Nikodym derivative.

II. Formulation and Previous Results

Recall the basic ingredients of the two-class pattern classification problem introduced in Section I, in which we restrict our decision rules to belong to a certain class of decision rules.

- A set X which is the feature space together with a σ-algebra $\mathcal{S}$ of subsets of X (usually taken to be $\boldsymbol{R}^n$ with the Borel σ-algebra).
- A class $\mathcal{D}$ of decision rules, such that for each $D \in \mathcal{D}$, $D : X \to \{0, 1\}$ is measurable with respect to $\mathcal{S}$. Each $D \in \mathcal{D}$ corresponds in a natural way to a subset of X, namely those $x \in X$ for which $D(x) = 1$. For simplicity, we often let D refer to both the binary-valued function as well as a subset, and the interpretation will be clear from the context.
- An unknown (and arbitrary) conditional distribution $P(y|x)$ with $y \in \{0, 1\}$, so that $P(0|x)$ and $P(1|x)$ are the posterior probabilities of the two classes conditioned on x. (One could restrict $f(x) = P(1|x)$ to belong to a family of functions $\mathcal{F}$ as is often done in regression and pattern recognition. Here we make no such restriction other than measurability.)
- A distribution P on X with $P \in \mathcal{P}$ where $\mathcal{P}$ is a known family of probability measures on $(X, \mathcal{S})$.

Of course, a generalization to the last two items is to let $\mathcal{P}$ be a family of joint distributions on $X \times \{0, 1\}$ and require uniform consistency with respect to $\mathcal{P}$. However, the tools used and results obtained here on uniform consistency are in terms of the marginal distributions on X, and so the formulation given is quite natural and simpler. It allows arbitrary conditional distributions, while allowing restrictions/prior knowledge about the marginal distribution, which might arise, for instance, from observing a number of unlabeled samples.

For the present purposes, an *estimator* is a family of maps $A_m : [X \times \{0, 1\}]^m \to \mathcal{D}$, for each integer $m \geq 1$. "Learning" takes place as follows: A sequence of independent labeled samples $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ is observed with the x_i drawn independent and identically distributed (i.i.d.) according to $P \in \mathcal{P}$ and the y_i drawn according to $P(y|x_i)$. After m observations, one generates a decision rule

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

The hope is that with high probability H_m is close to the optimal decision rule in $\mathcal{D}$, as made precise next.

Conditioned on $x \in X$, the probability of error of a decision rule $D \in \mathcal{D}$ is given by

$$R(D|x) = 1 - P(D(x)|x). \tag{1}$$

The overall average performance of the decision rule D is then given by

$$R(D) = ER(D|x) = 1 - EP(D(x)|x) \tag{2}$$

where the expectation is with respect to the distribution P on X. The optimal performance over all decision rules in $\mathcal{D}$ is given by

$$R^* = \inf_{D \in \mathcal{D}} R(D). \tag{3}$$

Note that, in general, there may be no particular $D \in \mathcal{D}$ that actually achieves the infimum error rate R^*, or there may be several distinct $D \in \mathcal{D}$ all of which achieve error rate R^*. The risk of a decision rule D and the optimal risk, $R(D)$ and R^*, depend on the (possibly unknown) input distribution P and the unknown conditional distribution $P(y|x)$, $y \in \{0, 1\}$.

Definition 1: Given $\epsilon \in (0, 1)$, an estimator $\{A_m\}$ is said to be *uniformly consistent to accuracy* ϵ if, for each $\delta \in (0, 1)$ there exists an integer $m = m(\epsilon, \delta)$ such that

$$\Pr\{(x_1, y_1), \cdots, (x_m, y_m) \in (X \times \{0, 1\})^m : R(H_m) \leq R^* + \epsilon\} \geq 1 - \delta \quad (4)$$

for all probability measures $P \in \mathcal{P}$ and all conditional distributions $P(y|x)$ where $y \in \{0, 1\}$, and the probability in (4) is with respect to the m-fold product distribution on $(X \times \{0, 1\})^m$. An estimator is said to *uniformly consistent* if it is uniformly ϵ-consistent for each $\epsilon > 0$. The class $\mathcal{D}$ is said to be *learnable to accuracy* ϵ with respect to the family $\mathcal{P}$ if there exists an estimator for $\mathcal{D}$ that is uniformly consistent to accuracy ϵ. Finally, the class $\mathcal{D}$ is said to be *learnable* with respect to the family $\mathcal{P}$, if there exists an estimator for $\mathcal{D}$ that is uniformly consistent for each $\epsilon \in (0, 1)$ (i.e., if there is an estimator that is uniformly consistent in probability).

This definition simply requires convergence in probability of $R_P(\hat{\mathcal{D}}_m)$ to R_P^* uniformly for $P \in \mathcal{P}$, where $\hat{\mathcal{D}}_m$ denotes the estimate after m observations. This is also the definition of probably approximately correct (PAC) learnability used in [9] and elsewhere. Note that if a class $\mathcal{D}$ is learnable to accuracy ϵ with respect to a family $\mathcal{P}$, then $\mathcal{D}$ is also learnable to accuracy ϵ with respect to any subset of $\mathcal{P}$.

We now introduce some standard definitions and summarize the known conditions for learnability in the cases of fixed distribution and distribution-free learnability.

Definition 2: A set $\Gamma = \{x_1, \cdots, x_n\} \subset X$ is said to be *shattered* by the class $\mathcal{D}$ if, for every subset A of Γ, there exists a decision rule $D \in \mathcal{D}$ such that $\Gamma \cap D = A$. The *Vapnik–Chervonenkis (VC) dimension* of the class $\mathcal{D}$ is the largest integer n such that there exists a set of cardinality n that is shattered by $\mathcal{D}$. If there exist sets of arbitrarily large (integer) cardinality that are shattered by $\mathcal{D}$, then $\mathcal{D}$ is said to have infinite VC dimension.

The following result characterizes distribution-free learnability. Certain additional, but mild, measurability conditions are actually required (e.g., see [2] or [18, p. 168]), but we ignore these conditions throughout and assume that the decision rules satisfy the required measurability conditions.

Theorem 1 ([2] and [18, pp. 152–161]): Let $\mathcal{P}^*$ denote the set of all probability measures on $\mathcal{S}$. Then a class $\mathcal{D}$ is learnable with respect to $\mathcal{P}^*$ if and only if the VC dimension of $\mathcal{D}$ is finite.

The main idea in learning with respect to a fixed distribution is that since we only need to approximate the optimal decision rule, if we can replace $\mathcal{D}$ with a finite number of decision rules (one of which is close to optimal), we can simplify the problem to one of deciding between a finite number of alternatives. This motivates the following definition.

Definition 3: Suppose d is a pseudometric on $\mathcal{D}$, and let $\epsilon > 0$. A finite set $\{B_1, \cdots, B_n\}$ where each $B_i \in \mathcal{D}$ is said to be an *ϵ-cover* of $\mathcal{D}$ with respect to d if, for each $D \in \mathcal{D}$, there exists an index i such that $d(B_i, D) \leq \epsilon$. The smallest integer $N = N(\epsilon, \mathcal{D}, d)$ such that $\mathcal{D}$ has an ϵ-cover of cardinality N is called the *ϵ-covering number* of $\mathcal{D}$ with respect to d. If $\mathcal{D}$ does not have a finite ϵ-cover, then N is taken to be infinite.

For the problem of learning with respect to a fixed distribution, a natural candidate for the pseudometric $d(\cdot, \cdot)$ is to take

$$d(D_1, D_2) = |R(D_1) - R(D_2)| = |E[P(D_2(x)|x)] - E[P(D_1(x)|x)]|.$$

Unfortunately, since $P(y|x)$ is unknown, with this choice for $d(\cdot, \cdot)$, we have no way to evaluate the distance between two decision rules. Therefore, a more appropriate choice for the pseudometric is

$$d_P(D_1, D_2) = E|D_1(x) - D_2(x)|. \quad (5)$$

With D_1, D_2 regarded as subsets of X, an equivalent definition of d_P is $d_P(D_1, D_2) = P(D_1 \Delta D_2)$, where $D_1 \Delta D_2$ is the symmetric difference of the sets D_1, D_2, i.e., $D_1 \Delta D_2 = (D_1 \cap D_2^c) \cup (D_1^c \cap D_2)$. It is easy to show that $d_P(\cdot, \cdot)$ is in fact a pseudometric on $\mathcal{D}$ (and actually on the set of all measurable sets $\mathcal{S}$). It is also straightforward to show that

$$|R(D_1) - R(D_2)| \leq d_P(D_1, D_2)$$

so that if a decision rule is close to an optimal rule in $d_P(\cdot, \cdot)$ then it is approximately optimal in the sense required by the definition of learnability.

Theorem 2 ([1] and [18, pp. 149–151]): Suppose P_0 is a probability measure on $\mathcal{S}$, and $\mathcal{D}$ is a class of decision rules. Define the associated pseudometric d_{P_0} on $\mathcal{D}$ as in (5). Let $\mathcal{P}$ equal the singleton set $\{P_0\}$. Then $\mathcal{D}$ is learnable to accuracy ϵ with respect to $\{P_0\}$ if $\mathcal{D}$ has a finite $\epsilon/2$-cover with respect to the pseudometric d_{P_0}. $\mathcal{D}$ is learnable to accuracy ϵ with respect to $\{P_0\}$ only if $\mathcal{D}$ has a finite 2ϵ-cover with respect to the pseudometric d_{P_0}. Finally, $\mathcal{D}$ is learnable with respect to $\{P_0\}$ if and only if $N(\epsilon, \mathcal{D}, d_{P_0})$ is finite for each $\epsilon \in (0, 1)$.

Sufficiency for a more general formulation is shown in [18, pp. 149–151] and for a more restricted formulation in [1]. Necessity for the more restricted formulation is shown in [1] which implies necessity for the present formulation. The basic estimator given in [1] and [18] that satisfies the conditions of the theorem is as follows. Suppose $\mathcal{D}$ has a finite $\epsilon/2$-cover with respect to d_{P_0}. Let $N := N(\epsilon/2, \mathcal{D}, d_{P_0})$, and choose an $\epsilon/2$-cover $B_1, \cdots, B_N$ for $\mathcal{D}$. Then the following "minimum empirical risk" estimator is uniformly consistent to accuracy ϵ: Given δ, define $m(\epsilon, \delta)$ as the least integer such that

$$m(\epsilon, \delta) \geq \frac{32}{\epsilon^2} \ln \frac{N}{\delta}.$$

Select i.i.d. samples $x_1, \cdots, x_m \in X$, and choose the hypothesis H_m as the decision rule B_k that misclassifies the smallest number of samples.

The above results naturally suggest that there may be connections between the covering numbers of a class $\mathcal{D}$ with respect to various distributions and the VC dimension of $\mathcal{D}$. The known results typically provide upper and lower bounds on $\sup_{P \in \mathcal{P}^*} N(\epsilon, \mathcal{D}, d_P)$ in terms of the VC dimension of $\mathcal{D}$ (e.g., some known results are summarized in [13]). The upper bounds are the deeper results and the fundamental result along this line was obtained by Dudley [6], which was subsequently refined by Pollard [14], and more recently by Haussler [9]. These bounds imply the following result which is sufficient

for the present paper. Again, the result requires some weak measurability conditions which we ignore here.

Theorem 3: For any class of decision rules $\mathcal{D}$,

$$\sup_{P \in \mathcal{P}^*} N(\epsilon, \mathcal{D}, d_P) < \infty$$

for all $\epsilon > 0$ if and only if the VC dimension of $\mathcal{D}$ is finite.

For future convenience, one more concept is defined.

Definition 4: Suppose $\mathcal{D}$ is a class of decision rules, and $\mathcal{P}$ is a family of probability measures. Then the class $\mathcal{D}$ is said to have *uniformly bounded metric entropy* with respect to the family $\mathcal{P}$, or to satisfy the *UBME condition* with respect to $\mathcal{P}$, if

$$\overline{N}(\epsilon, \mathcal{D}, \mathcal{P}) := \sup_{P \in \mathcal{P}} N(\epsilon, \mathcal{D}, d_P) < \infty \text{ for each } \epsilon > 0. \quad (6)$$

Now suppose $\mathcal{P}$ is an *arbitrary* family of probability measures. For each $P \in \mathcal{P}$, define d_P as in (5). Then, by a slight extension of the results in [1] and [18, pp. 149–151], one can prove the following:

Lemma 1: Suppose $\mathcal{D}$ is a class of decision rules and $\mathcal{P}$ is a family of probability measures. Suppose $\mathcal{D}$ is learnable with respect to $\mathcal{P}$. Then $\mathcal{D}$ satisfies the *UBME* condition with respect to $\mathcal{P}$.

Lemma 1 shows that a *necessary* condition for $\mathcal{D}$ to be learnable with respect to $\mathcal{P}$ is that $\mathcal{D}$ satisfies the UBME condition with respect to $\mathcal{P}$. In other words, $\mathcal{D}$ has a finite ϵ-cover for each $\epsilon > 0$, and each distribution $P \in \mathcal{P}$; moreover, the ϵ-covering number is uniformly bounded with respect to P for each ϵ. It is clear from the results presented so far that the UBME condition is also *sufficient* for learnability in the two extreme cases, namely: i) $\mathcal{P}$ is a singleton set, and ii) $\mathcal{P} = \mathcal{P}^*$. At this point, it is natural to ask whether the condition (6) is sufficient for learnability when $\mathcal{P}$ is an *arbitrary* family of probability measures. The answer is "No" as shown in [8]. Thus the problem of deriving necessary and sufficient conditions for learnability under a general family of probability measures is still open. The present paper derives a few results in this direction.

While Theorem 2 is for learnability with respect to a *single* measure, a careful examination of the proof of this result shows that it is actually possible to prove a rather stringent *sufficient* condition for learnability.

Corollary 1: Suppose $\mathcal{D} \subseteq \mathcal{S}$, $\mathcal{P} \subseteq \mathcal{P}^*$, and define the pseudometric $\overline{d}_{\mathcal{P}}$ as follows: For each $A, B \in \mathcal{S}$, let

$$\overline{d}_{\mathcal{P}} = \sup_{P \in \mathcal{P}} d_P(A, B). \quad (7)$$

Then $\mathcal{D}$ is learnable to an accuracy ϵ with respect to $\mathcal{P}$ if $\mathcal{D}$ has an $\epsilon/2$-cover with respect to the pseudometric $\overline{d}_{\mathcal{P}}$. In particular, the minimum empirical risk estimator described above is uniformly consistent to accuracy ϵ.

The reason why (6) is *not* sufficient in general for learnability, whereas the condition of Corollary 1 *is* sufficient, lies in the order of the quantifiers. In Corollary 1, the *same* sets $B_1, \cdots, B_N$ can serve as the "centers" of the N "balls" that provide an $\epsilon/2$-cover of $\mathcal{D}$, for *every* distribution $P \in \mathcal{P}$. In contrast, if (6) holds, then the *number* of balls in an ϵ-cover of $\mathcal{D}$ is independent of $P \in \mathcal{P}$, but the "centers" of these balls may depend on P (as it does in the counterexample of [8]).

Let us now examine the question: How close is the condition of Corollary 1 to being necessary for learnability? To shed some light on this question, let us study the case where $\mathcal{P} = \mathcal{P}^*$, the set of *all* probability measures on $\mathcal{S}$, and compare the sufficient condition of Corollary 1 with the known necessary condition provided by Theorem 1.

Suppose $\mathcal{P}$ is the set of all probability measures on $\mathcal{S}$, and define $\overline{d} := \overline{d}_{\mathcal{P}^*}$ as in (7). Of course, if $A = B$ then $\overline{d}(A, B) = 0$. However, if $A \neq B$ then $A \Delta B$ is nonempty. In this case, if P is taken to be the purely atomic measure concentrated at a point $x \in A \Delta B$ then $d_P(A, B) = 1$. It follows that $\overline{d}(A, B) = 1$ whenever $A \neq B$. Hence, if $\mathcal{P}$ is the set of all distributions on $\mathcal{S}$, then any two distinct decision rules in $\mathcal{D}$ are a distance 1 apart with respect to $\overline{d}$. Therefore, $\mathcal{D}$ has a finite ϵ-cover with respect to $\overline{d}$ for each $\epsilon > 0$ if and only if $\mathcal{D}$ is finite.

Thus in the extreme case of distribution-free learning, the sufficient condition of Corollary 1 reduces to the requirement that the class $\mathcal{D}$ be finite, which is much more stringent than the requirement that the VC dimension of $\mathcal{D}$ be finite.

III. Families of Measures with a Nonempty Interior

The objective of this section is to show that the UBME condition is sufficient as well as necessary in the case where the family of probability measures $\mathcal{P}$ has a nonempty interior. Actually, we will see that as long as there is some positive amount of nonparametric uncertainty (measured in various ways) around some nominal P_0, then a uniformly consistent estimator exists iff there is an estimator that is uniformly consistent for all distributions.

Let ρ denote the *total variation* metric on $\mathcal{P}^*$. That is, given $P, Q \in \mathcal{P}^*$, define

$$\rho(P, Q) = \sup_{A \in \mathcal{S}} |P(A) - Q(A)|. \quad (8)$$

Let $\mathcal{B}(P, \lambda)$ denote the open sphere of radius λ centered at P, and let $\overline{\mathcal{B}}(P, \lambda)$ denote the closed sphere of radius λ centered at P, with respect to the total variation metric ρ. Then P_0 is an interior point of $\mathcal{P}$ if (and only if) there exists a $\lambda > 0$ such that $\mathcal{B}(P_0, \lambda) \subseteq \mathcal{P}$.

Suppose P_1, P_2 are probability measures on $\mathcal{S}$; then their *convex combination* $\lambda P_1 + (1 - \lambda) P_2$ is also a probability measure on $\mathcal{S}$ for each $\lambda \in [0, 1]$. Now, for $P_0 \in \mathcal{P}^*$ and $\lambda \in [0, 1]$, define

$$\mathcal{P}_c(P_0, \lambda) = \{(1 - \eta) P_0 + \eta P : \eta \in [0, \lambda], P \in \mathcal{P}^*\}. \quad (9)$$

The collection $\mathcal{P}_c(P_0, \lambda)$ can be thought of as those distributions nominally equal to P_0, but with some nonparametric uncertainty with respect to mixtures up to λ in extent. Note that

$$\mathcal{P}_c(P_0, \lambda) \subseteq \overline{\mathcal{B}}(P_0, \lambda).$$

To see this, suppose $Q = (1 - \eta) P_0 + \eta P$ for some $P \in \mathcal{P}^*$, $\eta \in [0, \lambda]$. Then, for each $A \in \mathcal{S}$, we have

$$\begin{aligned} |Q(A) - P(A)| &= |(1 - \eta) P_0(A) + \eta P(A) - P_0(A)| \\ &= \eta \, |P(A) - P_0(A)| \leq \eta \leq \lambda. \end{aligned}$$

Also, observe that

$$\mathcal{P}_c(P_0, 0) = \overline{\mathcal{B}}(P_0, 0) = \{P_0\}$$
$$\mathcal{P}_c(P_0, 1) = \overline{\mathcal{B}}(P_0, 1) = \mathcal{P}^*.$$

Theorem 4: Let $\mathcal{D}$ be a class of decision rules, P_0 a fixed probability measure, and $0 < \lambda \leq 1$. Then the following are equivalent.

1) $\mathcal{D}$ is learnable with respect to $\mathcal{P}_c(P_0, \lambda)$.
2) $\mathcal{D}$ is learnable with respect to $\overline{\mathcal{B}}(P_0, \lambda)$.
3) $\mathcal{D}$ is learnable with respect to $\mathcal{P}^*$.
4) $\mathcal{D}$ has finite VC dimension.

Proof:

"2) $\Rightarrow$ 1)" Obvious, because

$$P_c(P_0, \lambda) \subseteq \overline{\mathcal{B}}(P_0, \lambda).$$

"3) $\Rightarrow$ 2)" Obvious, because $\overline{\mathcal{B}}(P_0, \lambda) \subseteq \mathcal{P}^*$.

"4) $\Rightarrow$ 3)" This is a consequence of Theorem 1.

"1) $\Rightarrow$ 4)" If $\mathcal{D}$ is learnable with respect to $\mathcal{P}_c(P_0, \lambda)$, then $\mathcal{D}$ satisfies the UBME condition with respect to $P_c(P_0, \lambda)$. Now choose $P \in \mathcal{P}^*$ arbitrarily, and define

$$Q = (1 - \lambda)P_0 + \lambda P \in \mathcal{P}_c(P_0, \lambda).$$

If $A, B \in \mathcal{S}$ are any measurable sets, then

$$d_Q(A, B) = Q(A \Delta B) = (1 - \lambda)P_0(A \Delta B) + \lambda P(A \Delta B)$$
$$\geq \lambda P(A \Delta B) = \lambda d_P(A \Delta B).$$

Therefore, $N(\lambda\epsilon, \mathcal{D}, Q) \geq N(\epsilon, \mathcal{D}, P)$ and so

$$\sup_{P \in \mathcal{P}^*} N(\epsilon, \mathcal{D}, d_P) \leq \sup_{P \in \mathcal{P}^*} N(\lambda\epsilon, \mathcal{D}, d_{(1-\lambda)P_0 + \lambda P})$$
$$= \sup_{Q \in \mathcal{P}_c(P_0, \lambda)} N(\lambda\epsilon, \mathcal{D}, d_Q) < \infty.$$

Hence, from Theorem 3, C has finite VC dimension. □

Corollary 2: Let $\mathcal{D}$ be a class of decision rules, and suppose the family $\mathcal{P}$ has a nonempty interior. Then $\mathcal{D}$ is learnable with respect to $\mathcal{P}$ if and only if $\mathcal{D}$ has finite VC dimension.

Proof:

"*If*" This is a consequence of Theorem 1 and the observation after Definition 1.

"*Only if*" Since $\mathcal{P}$ has a nonempty interior, it is possible to select $P_0 \in \mathcal{P}$, $\lambda > 0$ such that $\overline{\mathcal{B}}(P_0, \lambda) \subseteq \mathcal{P}$. By hypothesis, $\mathcal{D}$ is learnable with respect to $\overline{\mathcal{B}}(P_0, \lambda)$ as well. The desired conclusion now follows from Theorem 4. □

Note that Theorem 4 can actually be improved by considering the class $\mathcal{P}_a^*$ of all purely atomic probability measures. We can define $\mathcal{P}_{ac}(P_0, \lambda)$ in analogy with (9) where convex combinations are taken only with atomic measures. It turns out that Theorem 3 can be improved by taking the sup only over all $P \in \mathcal{P}_a^*$. Then using this result, it is easy to see that learnability with respect to $\mathcal{P}_{ac}(P_0, \lambda)$ is equivalent to the other four conditions in Theorem 4. Another interesting equivalent condition involves all measures close to some measure P_0 in relative entropy. Namely, let

$$\overline{\mathcal{B}}_I(P_0, \lambda) = \{Q : I(P_0 \| Q) \leq \lambda\} \tag{10}$$

where

$$I(P\|Q) = \begin{cases} \int f \log f \, dQ(x), & \text{if } f = dP/dQ \text{ exists} \\ \infty, & \text{otherwise} \end{cases}$$

where dP/dQ denotes the Radon–Nykodym derivative of P with respect to Q when it exists. Then, if $Q = (1-\eta)P_0 + \eta P$ we have $I(P_0\|Q) \leq -\log(1 - \eta)$. Thus for any $\lambda > 0$, we have $\overline{\mathcal{B}}_I(P_0, \lambda) \supset \mathcal{P}_c(P_0, 1 - e^{-\lambda})$, so that learnability with respect to $\overline{\mathcal{B}}_I(P_0, \lambda)$ for any $\lambda > 0$ is also equivalent to all four conditions of Theorem 4. This result may be useful in analyzing the robustness of statistical procedures that assume a certain P_0, and may also be useful in analyzing the tradeoff between labeled and unlabeled examples (e.g., see [3]).

IV. Totally Bounded Families of Measures

We begin with a preliminary result that serves as a counterpoint to Theorem 4.

Lemma 2: Suppose $\mathcal{D}$ is a class of decision rules, P_0 is a given probability measure, and that $N(\epsilon/4, \mathcal{D}, d_{P_0})$ is finite. Then $\mathcal{D}$ is learnable with respect to $\overline{\mathcal{B}}(P_0, \epsilon/4)$ to accuracy ϵ.

Proof: Let $N := N(\epsilon/4, \mathcal{D}, d_{P_0})$, and let $B_1, \cdots, B_N$ be an $\epsilon/4$-cover for $\mathcal{D}$ with respect to the pseudometric d_{P_0}. Let $P \in \overline{\mathcal{B}}(P_0, \epsilon/4)$ be arbitrary. Then, for each $A \in \mathcal{D}$, we have

$$d_P(A, B_i) = P(A \Delta B_i) \leq P_0(A \Delta B_i) + \epsilon/4 \leq \epsilon/2 \text{ for some } i.$$

Hence the collection $\{B_1, \cdots, B_N\}$ is also an $\epsilon/2$-cover for $\mathcal{D}$ with respect to the pseudometric $d_{\overline{\mathcal{B}}(P_0, \epsilon/4)}$. The result now follows from Corollary 1. □

In fact, an estimator for learning $\mathcal{D}$ to accuracy ϵ with respect to $\overline{\mathcal{B}}(P_0, \epsilon)$ can be readily constructed by adapting Corollary 1. Given $\delta > 0$, choose

$$m = \frac{32}{\epsilon^2} \ln \frac{N}{\delta}$$

i.i.d. samples, and choose H_m to be the "minimum empirical risk" hypothesis from $B_1, \cdots, B_N$.

The seemingly contradictory results of Theorem 4 and Lemma 2 can be reconciled as follows: In both cases, the family $\mathcal{P}$ is a sphere of probability measures, representing some nonparametric uncertainty about a nominal distribution P_0. Theorem 4 states that, if it is desired to learn a class of decision rules to *arbitrarily small* accuracy in spite of the uncertainty regarding the probability measure, then the problem reduces to one of distribution-free learning. In contrast, Lemma 2 states that, if we are content to learn a class of decision rules to a *fixed finite* accuracy, then a correspondingly small amount of nonparametric uncertainty in the distribution can be tolerated, so long as the class is learnable with respect to the nominal distribution. Together, Theorem 4 and Lemma 2 suggest that there is a tradeoff between the amount of nonparametric uncertainty under which learning is to take place, and the specified accuracy of learning.

The focus of attention in this section is the case where the set $\mathcal{P}$ is totally bounded. The study is commenced with a few relevant definitions and properties. Recall that the metric on $\mathcal{P}$ is the total variation metric defined by (8).

Definition 5: A family $\mathcal{P} \subseteq \mathcal{P}^*$ is said to be *totally bounded* if for each $\epsilon > 0$ the ϵ-covering number of $\mathcal{P}$ (with respect to ρ) is finite—that is, if for every $\epsilon > 0$ there exists a finite set $\{P_1, \cdots, P_M\}$ with $P_i \in \mathcal{P}$ for all i, such that for all $P \in \mathcal{P}$ there exists an index i such that $\rho(P, P_i) \leq \epsilon$.

The next result shows that the UBME condition is sufficient as well as necessary for learnability, when the family of probability measures is a totally bounded set.

Theorem 3: Suppose $\mathcal{D}$ is a class of decision rules and $\mathcal{P}$ is a totally bounded family of probability measures. Then $\mathcal{D}$ is learnable with respect to $\mathcal{P}$ if and only if $\mathcal{D}$ satisfies the UBME condition with respect to $\mathcal{P}$.

Proof:

"*Only if*" This is a consequence of Lemma 1.

"*If*" Given ϵ, δ, choose $\{P_1, \cdots, P_M\}$ such that each $P \in \mathcal{P}$ is within $\epsilon/8$ of some P_i. For each i, choose an $\epsilon/8$-cover $\{B_1^i, \cdots, B_{N_i}^i\}$ for $\mathcal{D}$ with respect to the pseudometric d_{P_i}. Draw

$$m_0 = \max_{1 \leq i \leq M} \frac{32}{\epsilon^2/4} \ln \frac{N_i}{\delta/2}$$

i.i.d. samples according to the unknown probability measure P, and for each $i = 1, \cdots, M$ generate a "minimum empirical risk" hypothesis $H_i \in \{B_1^i, \cdots, B_{N_i}^i\}$ such that the corresponding indicator function misclassifies the fewest of the m_0 samples. Now run off another

$$m_{\text{extra}} = \frac{32}{\epsilon^2} \ln \frac{M}{\delta}$$

samples, and let H be a hypothesis among $H_1, \cdots, H_M$ that misclassifies the fewest of these last samples. It is now shown that H satisfies (4), establishing that the estimator is uniformly consistent.

By Lemma 2, it follows that if $P \in \overline{\mathcal{B}}(P_i, \epsilon/8)$ for some i, then the corresponding hypothesis H_i satisfies $R(H_i) \leq R^* + \epsilon/2$ with probability at least $1 - \delta/2$, where R^* is as defined in (3). Thus among $H_1, \cdots, H_M$, at least one H_i satisfies $R(H_i) \leq R^* + \epsilon/2$ with probability at least $1 - \delta/2$. Now, one can repeat the line of reasoning in [1] and [18] used to establish Theorem 2 to show that the probability that the minimum empirical risk hypothesis H satisfies $R(H) \leq R^* + \epsilon$ is at least $1 - \delta$. □

Taken together, Theorems 4 and 5 show that the UBME condition is both sufficient as well as necessary under each of two conditions: i) the family $\mathcal{P}$ has an interior point, and ii) the family $\mathcal{P}$ is totally bounded. To put it another way, in any counterexample along the lines of [8] showing that the UBME condition is not sufficient for learnability, the family of probability measures cannot be totally bounded, and yet, at the same time, must have an empty interior.

The proof of Theorem 5 also gives an upper bound on the sample complexity of a uniformly consistent estimator. In analogy with Definition 3, let $M(\epsilon)$ denote the ϵ-covering number of the set $\mathcal{P}$ with respect to the metric ρ, and observe that each of the integers N_i in the proof above is at most equal to $\overline{N}(\epsilon/8) := \overline{N}(\epsilon/8, \mathcal{D}, \mathcal{P})$ as defined in Definition 4. Therefore

$$m(\epsilon, \delta) \leq \frac{128}{\epsilon^2} \ln \frac{2\overline{N}(\epsilon/8)}{\delta} + \frac{32}{\epsilon^2} \ln \frac{M(\epsilon/8)}{\delta}.$$

The next result shows that, in some sense, Theorem 5 is just Corollary 1 in disguise.

Theorem 6: Suppose $\mathcal{D}$ is a class of decision rules, $\mathcal{P}$ is a totally bounded family of probability measures, and define the pseudometric $\overline{d}_{\mathcal{P}}$ as in (7). Then $\mathcal{D}$ satisfies the UBME condition with respect to $\mathcal{P}$ if and only if $\mathcal{D}$ has a finite ϵ-cover with respect to the pseudometric $\overline{d}_{\mathcal{P}}$ for each $\epsilon > 0$.

Proof:

"*If*" Obvious.

"*Only If*" Given $\epsilon > 0$, select an $\epsilon/2$-cover $\{P_1, \cdots, P_M\}$ for $\mathcal{P}$ with respect to the metric ρ. For each index i, select an $\epsilon/4$-cover $\{B_1, \cdots, B_{\overline{N}}\}$ for $\mathcal{D}$ with respect to the pseudometric d_{P_i}, where $\overline{N} = \overline{N}(\epsilon/4)$. It is shown next that $\mathcal{D}$ has an $\epsilon/2$-cover of cardinality at most $N_0 = \overline{N}^M$ with respect to the pseudometric $\overline{d}_\epsilon$ defined by

$$\overline{d}_\epsilon(A, B) = \max_{1 \leq i \leq M} d_{P_i}(A, B).$$

Assume for a moment that the above claim has been established, and let $\{A_1, \cdots, A_{N_0}\}$ denote an $\epsilon/2$-cover for $\mathcal{D}$ with respect to $\overline{d}_\epsilon$. Then $\{A_1, \cdots, A_{N_0}\}$ is also an ϵ-cover for $\mathcal{D}$ with respect to $\overline{d}_{\mathcal{P}}$. To see this, let $A \in \mathcal{D}$, $P \in \mathcal{P}$ be arbitrary, and select indices i, j such that $\rho(P, P_i) \leq \epsilon/2$, and $\overline{d}_\epsilon(A, A_j) \leq \epsilon/2$. Then certainly $d_{P_i}(A, A_j) \leq \epsilon/2$. Therefore

$$d_P(A, A_j) = P(A \Delta A_j) \leq P_i(A \Delta A_j) + \rho(P, P_i) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus in order to complete the proof, it remains only to establish the existence of an $\epsilon/2$-cover for $\mathcal{D}$ with respect to the pseudometric $\overline{d}_\epsilon$. Such a cover can be constructed as follows: For each $i \in \{1, \cdots, M\}$, $j \in \{1, \cdots, \overline{N}\}$, define

$$\mathcal{D}_{ij} = \{D \in \mathcal{D} : d_{P_i}(D, B_j^i) \leq \epsilon/2\}.$$

Next, for each vector $\boldsymbol{k} = [k_1 \cdots k_M] \in \{1, \cdots, \overline{N}\}^M$, define

$$\mathcal{A}_{\boldsymbol{k}} = \bigcap_{i=1}^{M} \mathcal{D}_{i,k_i}.$$

Then each (nonempty) set $\mathcal{A}_{\boldsymbol{k}}$ has "diameter" at most $\epsilon/2$ in each of the pseudometrics d_{P_i}, and hence in the pseudometric $\overline{d}_\epsilon$. Also, it is easy to see that the collection $\{\mathcal{A}_{\boldsymbol{k}}\}$ covers $\mathcal{D}$. Now choose some $A_{\boldsymbol{k}} \in \mathcal{A}_{\boldsymbol{k}}$ for each $\boldsymbol{k}$, provided of course that $\mathcal{A}_{\boldsymbol{k}}$ is nonempty. This is the desired $\epsilon/2$-cover for $\mathcal{D}$. □

The proof of Theorem 6 enables us to obtain a different upper bound for the sample complexity $m(\epsilon, \delta)$ than the proof of Theorem 5. The proof of Theorem 6 shows that, for each $\epsilon > 0$, the class $\mathcal{D}$ has an ϵ-cover of cardinality at most $[N(\epsilon/4)]^{M(\epsilon/2)}$ with respect to the pseudometric $\overline{d}_{\mathcal{P}}$. Hence $\mathcal{D}$ has an $\epsilon/2$-cover of cardinality at most $N = [\overline{N}(\epsilon/8)]^{M(\epsilon/4)}$. Therefore, the "minimum empirical risk" estimator described in Corollary 1 is uniformly consistent, provided

$$m(\epsilon, \delta) = \frac{32}{\epsilon^2} \ln \frac{N}{\delta} = \frac{32}{\epsilon^2} [M(\epsilon/4) \ln \overline{N}(\epsilon/8) + \ln(1/\delta)].$$

The next set of results is aimed at showing that, if a class of decision rules is learnable with respect to a totally bounded family of probability measures, then the class is also learnable with respect to the *closed convex hull* of the family.

Definition 6: Suppose $\mathcal{P} \subseteq \mathcal{P}^*$. Then the *convex hull* of $\mathcal{P}$ is defined as the set of all convex combinations

$$P = \sum_{i=1}^{n} \lambda_i P_i$$

where

$$n \geq 1, P_i \in \mathcal{P} \, \forall i, \lambda_i \geq 0 \, \forall i, \sum_{i=1}^{n} \lambda_i = 1$$

and is denoted by $\boldsymbol{C}(\mathcal{P})$. The *closed convex hull* of $\mathcal{P}$ is defined as the closure of $\boldsymbol{C}(\mathcal{P})$ and is denoted by $\overline{\boldsymbol{C}}(\mathcal{P})$.

Observe that $C(\mathcal{P})$ and $\overline{C}(\mathcal{P})$ are also families of probability measures and are thus subsets of $\mathcal{P}^*$. The following lemma is absolutely standard, and holds in any topological linear vector space (e.g., see [10, Theorem 3, p. 70 (or p. 644 of the 1964 edition)]). This lemma will be used in the subsequent theorem which is the result of interest for the present purposes.

Lemma 3: Suppose $\mathcal{P} \subseteq \mathcal{P}^*$ is totally bounded. Then $\overline{\boldsymbol{C}}(\mathcal{P})$ is totally bounded.

Theorem 7: Suppose $\mathcal{D}$ is a class of decision rules, $\mathcal{P}$ is a totally bounded family of probability measures, and that $\mathcal{D}$ satisfies the UBME condition with respect to $\mathcal{P}$. Then $\mathcal{D}$ satisfies the UBME condition with respect to $\overline{\boldsymbol{C}}(\mathcal{P})$.

Proof: By Lemma 3, it is enough to establish that $\mathcal{D}$ satisfies the UBME condition with respect to $\boldsymbol{C}(\mathcal{P})$.

Accordingly, suppose $\mathcal{D}$ satisfies the UBME condition (6). From Theorem 6, it follows that $\mathcal{D}$ also satisfies the finite metric entropy condition with respect to the pseudometric $\overline{d}_{\mathcal{P}}$. Specifically, given any $\epsilon > 0$, there exists a set $\{A_1, \cdots, A_N\}$ where $N \leq [\overline{N}(\epsilon/4)]^{M(\epsilon/2)}$, that forms an ϵ-cover for $\mathcal{D}$ with respect to $\overline{d}_{\mathcal{P}}$. It is shown now that the *same* set is also an ϵ-cover for $\mathcal{D}$ with respect to the pseudometric $\overline{d}_{\boldsymbol{C}(\mathcal{P})}$. To see this, suppose $P \in \boldsymbol{C}(\mathcal{P})$, and suppose to be specific that

$$P = \sum_{i=1}^{n} \lambda_i P_i, \; P_i \in \mathcal{P}, \; \lambda_i \geq 0, \; \sum_{i=1}^{n} \lambda_i = 1.$$

Suppose $B \in \mathcal{D}$. By assumption, there exists an index j such that

$$\overline{d}_{\mathcal{P}}(A_j, B) \leq \epsilon$$

or, in other words,

$$d_Q(A_j, B) \leq \epsilon, \quad \forall \boldsymbol{Q} \in \mathcal{P}.$$

Therefore

$$d_P(A_j, B) = P(A_j \Delta B) = \sum_{i=1}^{n} \lambda_i d_{P_i}(A_j, B) \leq \epsilon \sum_{i=1}^{n} \lambda_i = \epsilon.$$

As this inequality holds for *each* $P \in \boldsymbol{C}(\mathcal{P})$, it follows that $\overline{d}_{\boldsymbol{C}(\mathcal{P})}(A_j, B) \leq \epsilon$. Therefore, $\{A_1, \cdots, A_n\}$ is an ϵ-cover for $\mathcal{D}$ with respect to the pseudometric $\overline{d}_{\boldsymbol{C}(\mathcal{P})}$. □

Corollary 3: Suppose $\mathcal{D}$ is a class of decision rules, $\mathcal{P}$ is a totally bounded family of probability measures, and that $\mathcal{D}$ satisfies the UBME condition with respect to $\mathcal{P}$. Then, for each $\epsilon > 0$

$$N(\epsilon, \mathcal{D}, \overline{d}_{\mathcal{P}}) = N(\epsilon, \mathcal{D}, \overline{d}_{\boldsymbol{C}(\mathcal{P})}).$$

Note that Theorem 7 is *false*, in general, if $\mathcal{P}$ is not totally bounded. In fact, the counterexample of [8] is also a counterexample to Theorem 7 in case $\mathcal{P}$ is not totally bounded. To avoid disrupting the flow of the paper, the counterexample is given in the Appendix.

V. Two Sufficient Conditions

Thus far, the emphasis has been on finding families of probability measures for which the UBME condition is both sufficient and necessary for learnability. In this section, two *sufficient* conditions for learnability are presented.

The first result shows that learnability of a class of decision rules is retained under finite unions of families of distributions.

Theorem 8: Let $\mathcal{D}$ be a class of decision rules, and let $\mathcal{P}_1, \cdots, \mathcal{P}_n$ be n families of probability measures. If $\mathcal{D}$ is learnable with respect to $\mathcal{P}_i$ for $i = 1, \cdots, n$ then $\mathcal{D}$ is learnable with respect $\cup_{i=1}^{n} \mathcal{P}_i$.

Proof: Let f_i be an estimator which learns $\mathcal{D}$ with respect to $\mathcal{P}_i$, and let $m_i(\epsilon, \delta)$ be the number of samples required by f_i to learn with accuracy ϵ and confidence δ. Define an estimator f as follows. Ask for

$$m(\epsilon, \delta) = \max_{1 \leq i \leq n} m_i\left(\frac{\epsilon}{2}, \frac{\delta}{2}\right) + \frac{32}{\epsilon^2} \ln \frac{n}{\delta/2}$$

samples. Using the first $\max_i m_i(\epsilon/2, \delta/2)$ samples, form hypotheses $H_1, \cdots, H_n$ using estimators $f_1, \cdots, f_n$, respectively. Then, using the last $(32/\epsilon^2) \ln \frac{n}{\delta/2}$ samples, let f output the hypothesis H_i which is inconsistent with the fewest number of this second group of samples, and call it H. It is claimed that f is a learning algorithm for $\mathcal{D}$ with respect to $\cup_{i=1}^{n} \mathcal{P}_i$.

Let $P \in \cup_{i=1}^{n} \mathcal{P}_i$. Then $P \in \mathcal{P}_k$ for some k. Since the f_i are learning algorithms with respect to the $\mathcal{P}_i$, at least one H_i satisfies $R(H_i) \leq R^* + \epsilon/2$ with probability (with respect to product measures of P) greater than $1 - \delta/2$. Then as long as $R(H_i) \leq R^* + \epsilon/2$ for some i, the arguments from [1] and [18] used to establish Theorem 2 show that the "minimum empirical risk" estimator above will return a hypothesis H that satisfies $R(H) \leq R^* + \epsilon$ with a probability of at least $1 - \delta/2$. Thus the probability that $R(H) \leq R^* + \epsilon$ is at least $1 - \delta$. □

Note that the above result is not true in general for an infinite number of families of distributions since the sample complexity of the corresponding estimators may be unbounded (i.e., we may have $\sup_i N(\epsilon, \mathcal{D}, \mathcal{P}_i) = \infty$). However, even if $N(\epsilon, \mathcal{D}, \mathcal{P}_i)$ is uniformly bounded the result is not true. In fact, this is the case for the counterexample provided in [8].

The second set of results pertains to learnability with respect to a family consisting of probability measures that are "commensurate" with respect to a given nominal distribution. In Section III, the uncertainty regarding the probability

measure P under which learning is to take place was modeled by representing $\mathcal{P}$ as a sphere centered at a nominal measure P_0. However, there was no restriction about the nature of the "perturbation" about P_0. In particular, even if P_0 is a nonatomic measure, a sphere centered about P_0 may contain measures with an atomic part. As shown in Theorem 4, learning to arbitrarily small accuracy under such a family of distributions is equivalent (in terms of feasibility) to distribution-free learning. The notion of "commensurate" measures is intended to permit the family $\mathcal{P}$ to contain a large variety of distributions, while at the same time excluding "malicious" choices of probability measures. To motivate the definition, recall the proof of the fact that if a class $\mathcal{D}$ does not have finite VC dimension, then it is not learnable if $\mathcal{P}$ is the class of *all* probability measures on $\mathcal{S}$ (e.g., see [2]). Given an arbitrarily large integer m, one selects a set of cardinality $2m$ that is shattered by $\mathcal{D}$ (call it Γ), and chooses P to be a purely atomic measure concentrated on Γ. With this choice of P and a suitable choice of the conditional distribution $P(y|x)$, the inequality (4) is shown to be violated for suitable ϵ, δ. Choosing P to be purely atomic, and choosing the support set of P to be a set that will cause difficulties, may be thought of as "malicious." The notion of a commensurate family is intended specifically to exclude such extreme choices of P.

Let P_0 be a given fixed probability measure, referred to as the nominal distribution, and suppose $b \geq 1$. We define $\mathcal{M}(b, P_0)$ to be the set of all probability measures P on $\mathcal{S}$ such that

$$P(S) \leq b\, P_0(S), \quad \forall\, S \in \mathcal{S}.$$

Thus $\mathcal{M}(b, P_0)$ consists of all distributions that are absolutely continuous with respect to the nominal distribution P_0, whose Radon–Nikodym derivatives are (essentially) bounded by b. The family $\mathcal{M}(b, P_0)$ is said to be *commensurate* with P_0. A distribution P in $\mathcal{M}(b, P_0)$ is "nonmalicious" in the sense that it will never assign a positive measure to a set that has zero measure with respect to P_0. For example, if X is a subset of $\Re^l$ for some integer l and P_0 is the Lebesgue measure on X (suitably normalized such that $P_0(X) = 1$), then every purely atomic measure is "malicious" and does not belong to $\mathcal{M}(b, P_0)$ for any b; we may say therefore that such a measure is "noncommensurate" with P_0. Similarly, the family of distributions studied in [8] to show that (6) is not sufficient for learnability is also *not* of the form $\mathcal{M}(b, P_0)$ for any b and P_0. The reason is that, given any number $\epsilon > 0$, one can find measures P_a and P_b in this family and sets A and B such that

$$\frac{P_a(A)}{P_b(A)} > \epsilon \qquad \frac{P_b(B)}{P_a(B)} > \epsilon.$$

Theorem 9: Suppose $\mathcal{D}$ is a class of decision rules, P_0 is a probability measure, and $b \geq 1$. Then the following statements are equivalent.

1) The class $\mathcal{D}$ is learnable with respect to $\mathcal{M}(b, P_0)$.
2) The class $\mathcal{D}$ is learnable with respect to P_0.
3) The class $\mathcal{D}$ has a finite ϵ-cover with respect to the pseudometric d_{P_0} on $\mathcal{S}$ induced by P_0, for all $\epsilon > 0$.

Proof:

"1) $\Rightarrow$ 2)" Obvious, because $P_0 \in \mathcal{M}(b, P_0)$.

"2) $\Rightarrow$ 3)" This follows from Theorem 2.

"3) $\Rightarrow$ 1)" Define $\overline{d}_{\mathcal{M}} := \overline{d}_{\mathcal{M}(b, P_0)}$ as in (7). Then it is easy to see that

$$\overline{d}_{\mathcal{M}}(A, B) \leq b\, d_{P_0}(A, B), \ \forall\, A, B \in \mathcal{S}.$$

Hence an ϵ/b-cover of $\mathcal{D}$ with respect to d_{P_0} is also an ϵ-cover of $\mathcal{D}$ with respect to $\overline{d}_{\mathcal{M}}$. Now Statement 1 follows from Corollary 1. □

Note that, in general, the family $\mathcal{M}(b, P_0)$ is not totally bounded (though it can be, for specific choices of P_0). Also, $\mathcal{M}(b, P_0)$ has an empty interior unless the measurable space $(X, \mathcal{S})$ is rather trivial. Thus in spite of its simplicity, Theorem 9 cannot be derived as a consequence of earlier results.

By combining Theorem 9 and the method of proof used in Section IV, it is possible to prove the following result; the proof is omitted as it is simple.

Theorem 10: Suppose $P_1, \cdots, P_l$ are probability measures on $\mathcal{S}$, and $b_1, \cdots, b_l$ are constants such that $b_i \geq 1$ for all i. Suppose $\mathcal{D} \subseteq \mathcal{S}$ is a class of decision rules. Then $\mathcal{D}$ is learnable with respect to the family

$$\mathcal{P} = \overline{\boldsymbol{C}}(\cup_{i=1}^{l} \mathcal{M}(b_i, P_i)) \tag{11}$$

if and only if $\mathcal{D}$ is learnable with respect to each P_i, for $i = 1, \cdots, l$.

VI. Examples

In this section, we give several examples that illustrate the application of the results presented thus far to specific families of decision rules.

Example 1: In this example, we examine decision rules obtained by "thresholding" Lipschitz-continuous functions. Specifically, let $X = [0, 1]$, let P_0 be the uniform distribution on X, and let $\mathcal{F}$ denote the family of Lipschitz-continuous functions f on X such that $f(0) = 0$, and the Lipschitz constant of f is at most one. Then, it is clear that each such f maps X into $[-1, 1]$. One can associate a decision rule d with each such f by defining

$$d(x) = \text{sat}\,[f(x)], \quad \forall\, x \in X$$

where the "sat" (for saturation) function is defined by

$$\text{sat}\,(\sigma) = \begin{cases} 1, & \text{if } \sigma > 0 \\ 0, & \text{if } \sigma \leq 0. \end{cases}$$

Now let $\mathcal{D} = \{d[f(\cdot)] : f \in \mathcal{F}\}$. Thus $\mathcal{D}$ consists of all decision rules obtained by thresholding the family of Lipschitz-continuous functions $\mathcal{F}$.

It is easy to see that the following set of decision rules, known as the "dyadics," are contained in $\mathcal{D}$: For each $x \in [0, 1)$, define the nth dyadic function $d_n(x)$ to equal the nth bit in the binary representation of x in the form

$$x = \sum_{i=1}^{\infty} d_i(x)\, 2^{-i}.$$

The definition of $d_n(1)$ is unimportant. These rules $d_n(x)$ can be obtained by thresholding corresponding triangular Lipschitz functions $f_n(\cdot)$. Now it is routine to verify that

$$\int_X |d_n(x) - d_m(x)|\, dx = 0.5, \quad \forall n \neq m.$$

Hence the set of rules $\{d_n(\cdot)\}$ is pairwise separated by a distance of 0.5. As this set is infinite and is a subset of $\mathcal{D}$, it follows that

$$N(\epsilon, \mathcal{D}, P_0) = \infty, \quad \forall \epsilon < 0.5.$$

Hence, by Theorem 2,it follows that this set of decision rules is *not* learnable.

In contrast, if we restrict $P(1|x) - P(0|x)$ to also be in the Lipschitz class $\mathcal{F}$, then using the same class $\mathcal{D}$ of decision rules we do have uniform learnability, i.e., $R_{P_{X,Y}}(H_m) \to R^*_{P_{X,Y}}$ uniformly in probability over all $P_X \in \mathcal{P}^*$ and all $P(1|x) - P(0|x) \in \mathcal{F}$. This holds for various estimators H_m, for instance, thresholding a suitably defined regression histogram. The crucial point here that allows learnability in this case is the smoothness assumption on the conditional distributions, which guarantees that whenever two conditional distributions are close then the corresponding risks will also be close. On the other hand, the corresponding optimal decision rules may be far in terms of d_{P_0}.

Note that the above example can be readily modified to the case where P_0 is *any* nonatomic probability measure with a continuous density. In order for the above reasoning to work, it is only necessary to find points $x_0 = 0, x_1, \cdots, x_{2^n} = 1$ for each integer $n \geq 1$ such that

$$P_0([x_i, x_{i+1}]) = 2^{-n}, \quad i = 0, \cdots, 2^n - 1.$$

The above example can also be modified by allowing the thresholding to be performed with a "dead zone," or multilevel thresholding in general. Let $\theta \in (0, 1)$ (the "tolerance") be a fixed number, and now define the function $s : [-1, 1] \to \{-1, 0, 1\}$ by

$$s(\sigma) = \begin{cases} 1, & \text{if } \sigma > \theta \\ 0, & \text{if } -\theta \leq \sigma \leq \theta \\ -1, & \text{if } \sigma < -\theta. \end{cases}$$

One can think of $s(\cdot)$ as the "sign" function with a "dead zone" of width 2θ. Let $\mathcal{D} = \{s[f(\cdot)] : f \in \mathcal{F}\}$.

Note that the present set of decision rules can assume one of *three* values, in contrast to the binary-valued decision rules discussed earlier. Nevertheless, since the range is a finite set, the finite metric entropy condition

$$N(\epsilon, \mathcal{D}, P_0) < \infty, \quad \forall \epsilon > 0$$

continues to be a necessary and sufficient condition for $\mathcal{D}$ to be learnable.

By adapting the reasoning in the preceding example, it can be shown that the present set of decision rules is also *not* learnable with respect to the uniform distribution. The idea is to construct a family of functions $\{f_n(\cdot)\}$ as follows: For each n, let $f_n(x) = x$ for $0 \leq x \leq \theta$. Thus $f_n(\theta) = \theta$, for each n. Over the interval $[\theta, 1]$, the function $f_n(x)$ crosses the value θ exactly 2^n times, just like the dyadics but over the interval $[\theta, 1]$. By passing these functions through $s(\cdot)$, one obtains a set of *binary-valued* decision rules $\{d_n(\cdot)\}$ such that

$$\int_X |d_n(x) - d_m(x)|\, dx = \frac{1-\theta}{2}, \quad \forall n \neq m.$$

Hence the finite metric entropy condition is violated whenever $\epsilon < (1 - \theta)/2$, and as a result, $\mathcal{D}$ is once again *not* learnable.

Example 2: In this example, we will define a class of decision rules on the unit square $[0, 1] \times [0, 1]$ by taking subgraphs of a class of functions. Specifically, consider the set F of all Lipschitz functions on $[0, 1]$ with Lipschitz constant 1, and taking values between 0 and 1. Then consider the set of decision rules $\mathcal{D}$ (subsets) of $[0, 1] \times [0, 1]$ which are the subgraphs of such function. That is, given such a Lipschitz function $f \in F$, let the corresponding decision rule be

$$D_f = \{(x, y): y \leq f(x)\}$$

and let $\mathcal{D} = \{D_f : f \in F\}$.

This class clearly has infinite VC dimension, and so is not distribution-free learnable. However, it is well known that the class of such Lipschitz functions has finite metric entropy with respect to the L^1 norm

$$||f(x) - g(x)|| = \int_0^1 |f(x) - g(x)|\, dx$$

(e.g., [12, p. 286], gives a result with respect to the sup norm, which implies a result with respect to L^1).

Now, if P is the uniform distribution on $[0, 1] \times [0, 1]$, we see that $d_P(D_f, D_g) = ||f - g||$. Therefore, the class $\mathcal{D}$ has finite covering number with respect to d_P, and hence is learnable with respect to all measures commensurate with P.

This argument actually goes through for any class of decision rules obtained as the subgraph of a function class with finite covering number with respect to L^1 (e.g., functions of bounded variation, restrictions of bandlimited functions, etc.).

Example 3: Let n be any fixed integer, and let $\mathcal{D}$ consist of all closed convex polygons in $[0, 1]^n$, where n is some fixed integer denoting the dimension of the feature vector. Then $\mathcal{D}$ has infinite VC dimension because, for example, any finite set of points on the surface of a sphere can be shattered. Thus $\mathcal{D}$ is *not* distribution-free learnable.

Now let P_0 denote the uniform distribution on $[0, 1]^n$, let $\lambda > 0$, and let $\overline{\mathcal{B}}(P_0, \lambda)$ denote the "ball" of all probabilities P such that $\rho(P, P_0) \leq \lambda$, where ρ denotes the total variation metric on the space of probability measures on $[0, 1]^n$. Then, by Theorem 4, it follows that $\mathcal{D}$ is *not* learnable with respect to $\overline{\mathcal{B}}(P_0, \lambda)$, no matter how small λ is.

On the other hand, if $\lambda = 0$ so that $\overline{\mathcal{B}}(P_0, \lambda) = \{P_0\}$, then Dudley ([5] and [7, Sec. 7.3]) shows that $N(\epsilon, \mathcal{D}, P_0)$ is finite for each ϵ, and in fact provides explicit upper bounds for $N(\epsilon, \mathcal{D}, P_0)$. Hence $\mathcal{D}$ *is* learnable with respect to the uniform distribution. The fact that $\mathcal{D}$ is *not* learnable with respect to $\overline{\mathcal{B}}(P_0, \lambda)$, no matter how small λ is, means that while $\mathcal{D}$ is learnable with respect to P_0, the slightest amount of *nonparametric* uncertainty around P_0 is enough to destroy this learnability.

Now let $b < \infty$ be a given constant, and let $\mathcal{M}(b, P_0)$ consist of all probabilities on $[0, 1]^n$ that have a density bounded by b. Then it follows from Theorem 9 that $\mathcal{D}$ *is* learnable with respect to $\mathcal{M}(b, P_0)$.

The class of decision rules realizable by closed convex polygons in $[0, 1]^n$ can be given a ready interpretation in terms of neural networks, by exploiting the fact that every closed convex polygon in $[0, 1]^n$ can be expressed as an intersection of a finite number of closed half-planes; conversely, every such intersection is a closed convex polygon. Therefore, the present set of decision rules can be realized by taking the outputs of a finite (but not *a priori* bounded) number of perceptrons, and then passing these outputs through an "and" operation (which can also be realized by another perceptron). By adjusting both the number of "hidden layer" perceptrons as well as the corresponding weights of the neural network, it is possible to realize every decision rule in $\mathcal{D}$; conversely, every such neural network realizes a decision rule in $\mathcal{D}$.

Example 4: As a generalization of Example 3, let $\mathcal{D}$ consist of all sets that can be expressed as a union of at most k closed convex polygons in $[0, 1]^n$, where k is a specified integer. For technical reasons, let us treat the empty set as a closed convex set, so that the phrase "at most" in the preceding sentence can be replaced by "exactly." Note that the empty set can indeed be realized using a neural network of the form shown discussed in Example 3. Thus the present class of decision rules consists of those realizable by a three-layer neural network, where each neuron in the second hidden layer performs an "and" operation, and the final output neuron performs an "or" operation.

Since the present class $\mathcal{D}$ contains the class of Example 3, it follows that the present class is also *not* distribution-free learnable, and thus not learnable in the presence of nonparametric uncertainty in the underlying probability. On the other hand, if P_0 is the uniform distribution on $[0, 1]^n$, then $\mathcal{D}$ *is* learnable. The proof of this fact is only sketched here in the interests of brevity. First, the argument of Pollard [14, pp. 22–24], is extended from $[0, 1]^2$ to $[0, 1]^n$ by replacing the factor $8/9$ used by Pollard with $1 - 3^{-n}$; this shows that the class of all closed convex polygons in $[0, 1]^n$ has the property known as the *uniform convergence of empirical probabilities (UCEP)*. Next, the proof of Dudley [7, p. 84] is sharpened to show that, if a class of decision regions has the UCEP property, then so does the class obtained by performing any set of k Boolean operations (e.g., unions) on these regions. Finally, it is shown that a class of decision regions having the UCEP property is learnable. Each of these steps is fairly straightforward, but giving all the details would take us too far afield.

Example 5: Let l, k, n be fixed integers, and let $\mathcal{D}$ consist of all unions (or intersections, or indeed any Boolean operations) of up to k sets in $\Re^n$, each of which is a level set of a polynomial in $x_1, \cdots, x_n$ of degree l or less. In this case, it can be shown that the $\mathcal{D}$ has finite VC dimension. The proof follows in two steps. First, one can write a polynomial in $x_1, \cdots, x_n$ as *a linear functional* in the various powers and cross-products of $x_1, \cdots, x_n$; then one can appeal to the result (see, e.g., [2]) which states that the set of closed half-planes (over a space $\Re^\nu$ where ν is fixed) has VC dimension $\nu + 1$. Then one can invoke the theorem of Dudley [7, p. 84], to the effect that the class formed by performing k Boolean operations on collections of finite VC dimension is itself of finite VC dimension. Thus such a class of decision rules is distribution-free learnable.

In particular, decision regions consisting of finite unions of ellipsoids, hyperbolas, and rectangles, are distribution-free learnable, provided the number of such unions is bounded *a priori*.

VII. Conclusions

In the present paper, the problem of learning a decision rule for two-class pattern classification with respect to an *arbitrary* family of probability measures has been studied. A metric is defined on the set of all probability measures on a given measurable space, whereby the distance between two probability measures is defined as the maximum variation between the two. The uniform boundedness of the metric entropy of the class of decision rules, which is known to be a *necessary* condition for learnability always, has been shown to be *sufficient* as well, under each of two conditions: i) the family of probability measures is *totally bounded*, or else ii) the family of probability measures contains an *interior point*. These two results establish that, in any counterexample along the lines of [8] (showing that the uniform boundedness of the metric entropy is *not* sufficient for learnability), the family of probability measures cannot be totally bounded, and yet at the same time must have an empty interior. Second, two sufficient conditions for learnability are presented. i) It is shown that learnability with respect to each of a finite collection of families of probability measures implies learnability with respect to the *union* of these families. ii) It is shown that learnability with respect to each of a finite number of probability measures implies learnability with respect to the convex hull of the families of all "commensurate" probability measures.

There are a number of interesting questions that are left open by the present paper. First, one can conjecture that the uniform boundedness of the metric entropy of the class of decision rules is sufficient for learnability whenever the family of distributions under which learning is to take place is *closed and convex*. The counterexample given in [8] to the Benedek–Itai conjecture (namely, that uniform boundedness of the metric entropy is sufficient for learnability for an *arbitrary* family of probability measures) is not a counterexample to this more restricted conjecture, as shown in the Appendix. Second, one can ask whether the result of Lemma 2 holds for an *arbitrary* family of distributions. That is, is it true that, whenever a class $\mathcal{D}$ can be learned to an accuracy ϵ with respect to a given family of probability measures $\mathcal{P}$, it can also be learned to some degraded accuracy $\nu_1(\epsilon)$ for all P that lie within some distance $\nu_2(\epsilon)$ of $\mathcal{P}$? In case $\mathcal{P}$ is a singleton set, Lemma 2 states (roughly) that the above statement is indeed true with $\nu_1(\epsilon) = 2\epsilon$ and $\nu_2(\epsilon) = \epsilon/2$; but it would be interesting to determine whether an analogous statement is true for more general families $\mathcal{P}$. If such a statement were to be true, then it would provide a tradeoff between the accuracy

of learning, and the accuracy to which the family $\mathcal{P}$ (under which learning is to take place) needs to be known.

APPENDIX
A COUNTEREXAMPLE

In this appendix, it is shown by example that Theorem 7 is *not* true in general if the family of distributions $\mathcal{P}$ is not totally bounded.

Our example is the same as in [8]. The set X equals $\{0, 1\}^{\mathcal{N}}$, the set of all binary sequences indexed over the set of natural numbers $\mathcal{N}$ (beginning with 1). $\mathcal{S}$ equals the Borel σ-field over X. Define the sequence

$$p_i = \frac{1}{\lg (i+1)}$$

where lg denotes the logarithm to the base 2. A product measure P_I can be induced on X by identifying $p_i = P(x_i = 1)$. Let $\sigma : \mathcal{N} \to \mathcal{N}$ denote a permutation (possibly infinite) of the integers; thus σ is a one-to-one and onto map on $\mathcal{N}$. Let Σ denote the set of all such permutations. Let P_σ denote the probability measure on X defined by $P_\sigma(x_{\sigma(i)} = 1) = p_i$. Now let $\mathcal{P} = \{P_\sigma, \sigma \in \Sigma\}$. This specifies the family of probability measures. Next, let $C_i = \{x \in X : x_i = 1\}$, and define $\mathcal{D} = \{C_i, i \in N\}$. Since any C_i with $p_{\sigma^{-1}(i)} < \epsilon$ satisfies $d_{P_\sigma}(C_i, \emptyset) < \epsilon$, it is easy to see that the sets $\{C_{\sigma(1)}, \cdots, C_{\sigma(n)}, \emptyset\}$ form an ϵ-cover for $\mathcal{D}$ with respect to the pseudometric d_{P_σ} provided $n \geq 2^{1/\epsilon}$. It follows, therefore, that the class $\mathcal{D}$ satisfies the UBME condition with respect to the family $\mathcal{P}$, and that

$$N(\epsilon, \mathcal{D}, \mathcal{P}) \leq 2^{1/\epsilon}.$$

It is shown in [8] that the class $\mathcal{D}$ is *not* learnable with respect to $\mathcal{P}$. Our objective here is somewhat different: It is shown here that $\mathcal{D}$ *does not* satisfy the UBME condition with respect to the convex hull of $\mathcal{P}$. In order to do this, a bit of terminology is introduced. Suppose S is a set and ρ is a pseudometric on S. A subset $M \subseteq S$ is said to be *ϵ-separated* with respect to the pseudometric ρ if the distance between every pair of nonidentical points in M equals or exceeds ϵ. It is clear that the cardinality of such a set M is a lower bound on the $\epsilon/2$-covering number of the set S.

Lemma 4: Define

$$\alpha = 1 - \frac{1}{\lg 3} \approx 0.36907, \quad d = 2^\alpha \approx 1.2915.$$

For each sufficiently small $\epsilon < \alpha$ and each integer n, there exists a probability measure $P \in \boldsymbol{C}(\mathcal{P})$ such that $\mathcal{D}$ contains a set of cardinality $nd^{1/\epsilon}$ that is ϵ-separated with respect to d_P. Therefore, for each sufficiently small $\epsilon < \alpha$

$$\sup_{P \in \boldsymbol{C}(\mathcal{P})} N(\epsilon, \mathcal{D}, \boldsymbol{C}(\mathcal{P})) = \infty.$$

The proof of the lemma makes use of the following preliminary result.

Lemma 5: For each sufficiently small $\delta > 0$ and each sufficiently large integer n, there exists another integer $M = 2^{c(\delta, n)/\delta}$, where $c(\delta, n) \to 1$ as $n \to \infty$, $\delta \to 0$, such that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lg iM} \geq \delta.$$

Proof: Let $x = \lg M$. Then the above summation can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lg iM} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x + \lg i} = \frac{N(x)}{D(x)}$$

where $N(x)$ and $D(x)$ are polynomials in x. Specifically

$$\begin{aligned} D(x) &= n \prod_{i=1}^{n} (x + \lg i) \\ &= nx^n + n\left(\sum_{i=1}^{n} \lg i\right) x^{n-1} + \cdots + \left(\prod_{i=2}^{n} \lg i\right) x \end{aligned}$$

after observing that $\lg 1 = 0$. Note that there is no constant term (x^0) in $D(x)$. Similarly

$$\begin{aligned} N(x) &= \sum_{i=1}^{n} \prod_{j \neq i} (x + \lg j) \\ &= nx^{n-1} + \left(\sum_{i=1}^{n} \sum_{j \neq i} \lg j\right) x^{n-2} + \cdots + \prod_{i=2}^{n} \lg i. \end{aligned}$$

Now note that

$$\sum_{i=1}^{n} \lg i = \lg n!.$$

If we define

$$\beta_n = \sum_{i=1}^{n} \sum_{j \neq i} \lg j$$

then $\beta_n < n \lg n!$, because

$$\sum_{j \neq i} \lg j < \lg n! \text{ for all } i > 1.$$

Now rewrite the desired inequality as

$$N(x) \geq \delta D(x).$$

Observe that $D(0) = 0$, while $N(0) > 0$. Hence the polynomial

$$\phi(x) := \delta D(x) - N(x)$$

satisfies $\phi(0) < 0$, and $\phi(x) \to \infty$ as $x \to \infty$ (because the degree of $D(x)$ is higher than that of $N(x)$). Let $r(\delta, n)$ denote the smallest positive root of the equation $\phi(x) = 0$. It is claimed that

$$r(\delta, n) \approx 1/\delta$$

for sufficiently large n and sufficiently small δ. To show this, we proceed by establishing that i) there is a root of the form

$$x_0 = \frac{c(\delta, n)}{\delta}$$

where $c(\delta, n) \to 1$ as $\delta \to 0, n \to \infty$, and ii) $\phi(x) < 0 \; \forall x < x_0$. To prove i), substitute $x = c/\delta$ into $\phi(x)$. This gives

$$\phi(c/\delta) = n\frac{c^n}{\delta^{n-1}} + n \lg n! \frac{c^{n-1}}{\delta^{n-2}} + \cdots - n\frac{c^{n-1}}{\delta^{n-1}} - \beta_n \frac{c^{n-2}}{\delta^{n-2}} \cdots.$$

For $\delta \to 0$, these are the dominant terms. Observe first that

$$\phi(1/\delta) = \frac{n \lg n! - \beta_n}{\delta^{n-2}} + \cdots > 0$$

because $\beta_n < n \lg n!$. So the root x_0 is less than $1/\delta$ as $\delta \to 0$. However, as $\delta \to 0$

$$\phi(c/\delta) = \frac{1}{\delta^{n-1}} n(c^n - c^{n-1}) + \cdots$$

equals zero when $c \approx 1$. The same argument shows that if $c < 1$, then $c^n < c^{n-1}$, so that $\phi(c/\delta) < 0$ when $\delta \to 0$, $n \to \infty$. So the *smallest* positive root of the equation $\phi(x) = 0$ roughly equals $1/\delta$. □

Proof: Now to get back to the proof of Lemma 4: Observe that if $C_i, C_j \in \mathcal{D}$, then

$$C_i \Delta C_j = \{x \in X : x_i = 1 \text{ and } x_j = 0, \text{ or } x_i = 0 \text{ and } x_j = 1\}.$$

Therefore

$$P(C_i \Delta C_j) = \phi_i(1 - \phi_j) + (1 - \phi_i)\phi_j$$

where

$$\phi_i = P(x_i = 1).$$

Let us use d_σ as an abbreviation for d_{P_σ}. If $P = P_I$, then

$$d_I(C_i, C_j) = p_i(1 - p_j) + (1 - p_i)p_j.$$

If $P = P_\sigma$, then

$$d_\sigma(C_i, C_j) = p_{\sigma(i)}(1 - p_{\sigma(j)}) + (1 - p_{\sigma(i)})p_{\sigma(j)}.$$

Given ϵ, n, choose $M \approx 2^{\alpha/\epsilon} = d^{1/\epsilon}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lg i(M+1)} \geq \frac{\epsilon}{\alpha}.$$

This is possible by Lemma 5. Now define a permutation σ on the natural numbers as follows:

$$\begin{aligned} \sigma(i) &= i + M && \text{for } 1 \leq i \leq (n-1)M \\ \sigma(i) &= i - (n-1)M && \text{for } (n-1)M < i \leq nM \\ \sigma(i) &= i && \text{for } i > nM. \end{aligned}$$

In other words, σ is a "block"-cyclic permutation, and $\sigma^n = I$. Now define

$$P = \frac{1}{n}\sum_{l=0}^{n-1} P_{\sigma^l} \in \boldsymbol{C}(\mathcal{P})$$

where σ^0 is taken as the identity permutation. It is now shown that the set $\{C_1, \cdots, C_{nM}\}$ is ϵ-separated with respect to the metric d_P. For this purpose, let us compute $d_P(C_i, C_j)$. There are two cases to consider, namely: i) i and j belong to the same "block" of length M, that is, $(k-1)M + 1 \leq i, j \leq kM$ for some integer k, and ii) i and j belong to different "blocks."

Case i): Suppose i, j belong to the same block. In this case, it is easy to see that $\sigma^l(i)$, $\sigma^l(j)$ also belong to the same block for each l. Also, as l varies from 0 to $n-1$, $\sigma^l(i)$ and $\sigma^l(j)$ will visit each of the n blocks

$$\{1, \cdots, M\}, \cdots, \{(n-1)M + 1, \cdots, nM\}$$

exactly once. So it may be assumed, without loss of generality, that $i, j \in \{1, \cdots, M\}$. In this case, we have

$$d_I(C_i, C_j) = p_i(1 - p_j) + (1 - p_i)p_j.$$

If $i = 1$, $j > 1$, then

$$\begin{aligned} d_I(C_i, C_j) &= 1 - p_j = 1 - \frac{1}{\lg(j+1)} \\ &\geq 1 - \frac{1}{\lg(M+1)} \geq \frac{2\alpha}{\lg(M+1)} \end{aligned}$$

whenever $M \geq 2^{2\alpha+1} = 2d^2 \approx 3.3360$. If $i, j > 1$, then

$$\begin{aligned} 1 - p_i, 1 - p_j &\geq 1 - p_2 = \alpha \\ p_i, p_j &\geq \frac{1}{\lg(M+1)} \\ d_I(C_i, C_j) &\geq \frac{2\alpha}{\lg(M+1)}. \end{aligned}$$

Next, $\sigma(i), \sigma(j) \in \{M+1, \cdots, 2M\}$. So

$$d_\sigma(C_i, C_j) = p_{\sigma(i)}(1 - p_{\sigma(j)}) + (1 - p_{\sigma(i)})p_{\sigma(j)} \geq \frac{2\alpha}{\lg(2M+1)}.$$

Similarly

$$d_{\sigma^l}(C_i, C_j) \geq \frac{2\alpha}{\lg[(l+1)M+1]}, \quad l = 0, 1, \cdots, n-1.$$

Hence

$$d_P(C_i, C_j) = \frac{1}{n}\sum_{l=0}^{n-1} d_{\sigma^l}(C_i, C_j) \geq \frac{1}{n}\sum_{i=1}^{n} \frac{2\alpha}{\lg(iM+1)} \geq 2\epsilon.$$

Case ii): Suppose i, j belong to different blocks. By the same logic as in Case i), it can be assumed without loss of generality that $i \in \{1, \cdots, M\}$ and $j \in \{M+1, \cdots, nM\}$. In this case

$$d_I(C_i, C_j) \geq p_i(1 - p_j) \geq p_M(1 - p_2) = \frac{\alpha}{\lg(M+1)}$$

because $p_i \geq p_M$ and $1 - p_j \geq 1 - p_2$. Similarly

$$d_{\sigma^l}(C_i, C_j) \geq \frac{\alpha}{\lg[(l+1)M+1]}$$

and as a consequence

$$d_P(C_i, C_j) \geq \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha}{\lg(iM+1)} \geq \epsilon.$$

This shows that the set $\{C_1, \cdots, C_{nM}\}$ is ϵ-separated with respect to the pseudometric d_P. □

ACKNOWLEDGMENT

The authors wish to thank A. Barron for many helpful suggestions.

REFERENCES

[1] G. M. Benedek and A. Itai, "Learnability with respect to fixed distributions," *Theor. Comput. Sci.*, vol. 86, no. 2, pp. 377–390, 1991.

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

[3] V. Castelli and T. M. Cover, "Classification rules in the unknown mixture parameter case: Relative value of labeled and unlabeled samples," in *Proc. 1994 IEEE Int. Symp. on Information Theory*, 1994, p. 111.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[5] R. M. Dudley, "Metric entropy of some classes of sets with differentiable boundaries," *J. Approx. Theory*, vol. 10, no. 3, pp. 227–236, 1974.

[6] ——, "Central limit theorems for empirical measures," *Ann. Probab.*, vol. 6, no. 6, pp. 899–929, 1978.

[7] ——, "A course on empirical processes," *Lecture Notes in Mathematics*, no. 1097, 1984, pp. 1–142.

[8] R. M. Dudley, S. R. Kulkarni, T. J. Richardson, and O. Zeitouni, "A metric entropy bound is not sufficient for learnability," *IEEE Trans. Inform. Theory*, vol. 40, pp. 883–885, 1994.

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

[10] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, 2nd ed. New York: Pergamon, 1982.

[11] M. Kearns and U. Vazirani, *Introduction Computational Learning Theory.* Cambridge, MA: MIT Press, 1994.

[12] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in functional spaces," *Amer. Math. Soc. Transl.*, vol. 17, pp. 277–364, 1961.

[13] S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis, "Active learning using arbitrary binary valued queries," *Mach. Learn.*, vol. 11, pp. 23–35, 1993.

[14] D. Pollard, *Convergence of Stochastic Processes.* New York: Springer-Verlag, 1984.

[15] L. G. Valiant, "A theory of the learnable," *Commun. Assoc. Comput. Mach.*, vol. 27, no. 11, pp. 1134–1142, 1984.

[16] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies to their probabilities," *Theory of Prob. and its Appl.*, vol. 16, no. 2, pp. 264–280, 1971.

[17] ——, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory of Prob. and its Appl.*, vol. 26, no. 3, pp. 532–553, 1981.

[18] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data.* New York: Springer-Verlag, 1982.

[19] A. Wald, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, vol. 20, no. 4, pp. 595–601, 1949.