Scale-sensitive Dimensions, Uniform Convergence, and Learnability

Noga Alon*
Department of Mathematics
R. and B. Sackler Faculty of Exact Sciences
Tel Aviv University, Israel

Shai Ben-David†
Dept. of Computer Science
Technion
Haifa 32000, Israel

Nicolò Cesa-Bianchi‡
DSI, Università di Milano
Via Comelico 39
20135 Milano, Italy

David Haussler§
Dept. of Computer Science
UC Santa Cruz
Santa Cruz, CA 95064, USA

Abstract

Learnability in Valiant's PAC learning model has been shown to be strongly related to the existence of uniform laws of large numbers. These laws define a distribution-free convergence property of means to expectations uniformly over classes of random variables. Classes of real-valued functions enjoying such a property are also known as uniform Glivenko-Cantelli classes. In this paper we prove, through a generalization of Sauer's lemma that may be interesting in its own right, a new characterization of uniform Glivenko-Cantelli classes. Our characterization yields Dudley, Giné, and Zinn's previous characterization as a corollary. Furthermore, it is the first based on a simple combinatorial quantity generalizing the Vapnik-Chervonenkis dimension. We apply this result to obtain the weakest combinatorial condition known to imply PAC learnability in the statistical regression (or "agnostic") framework. Furthermore, we show a characterization of learnability in the probabilistic concept model, solving an open problem posed by Kearns and Schapire. These results show that the accuracy parameter plays a crucial role in determining the effective complexity of the learner's hypothesis class.

Keywords: Uniform laws of large numbers, Glivenko-Cantelli classes, Vapnik-Chervonenkis dimension, PAC learning.

An extended abstract of this paper appeared in the Proceedings of the 34th Annual Symposium on the Foundations of Computer Science, IEEE Press, 1993.

*Research supported in part by a USA-Israeli BSF grant. Email: noga@math.tau.ac.il.

†Email: shai@cs.technion.ac.il.

‡This is the author to whom all correspondence should be sent. Part of this research was done while this author was visiting UC Santa Cruz partially supported by the "Progetto finalizzato sistemi informatici e calcolo parallelo" of CNR under grant 91.00884.69.115.09672. The author also acknowledges support of ESPRIT Working Group in Neural and Computational Learning, NeuroCOLT 8556. Email: cesabian@dsi.unimi.it

§Email: haussler@cse.ucsc.edu

1 Introduction

In typical learning problems, the learner is presented with a finite sample of data generated by an unknown source and has to find, within a given class, the model yielding best predictions on future data generated by the same source. In a realistic scenario, the information provided by the sample is incomplete, and therefore the learner might settle for approximating the actual best model in the class within some given accuracy. If the data source is probabilistic and the hypothesis class consists of functions, a sample size sufficient for a given accuracy has been shown to be dependent on different combinatorial notions of "dimension", each measuring, in a certain sense, the complexity of the learner's hypothesis class.

Whenever the learner is allowed a low degree of accuracy, the complexity of the hypothesis class might be measured on a coarse "scale" since, in this case, we do not need the full power of the entire set of models. This position can be related to Rissanen's MDL principle [17], Vapnik's structural minimization method [22], and Guyon et al.'s notion of effective dimension [11]. Intuitively, the "dimension" of a class of functions decreases as the coarseness of the scale at which it is measured increases. Thus, by measuring the complexity at the right "scale" (i.e., proportional to the accuracy) the sample size sufficient for finding the best model within the given accuracy might dramatically shrink.

As an example of this philosophy, consider the following scenario.[1] Suppose a meteorologist is requested to compute a daily prediction of the next day's temperature. His forecast is based on a set of presumably relevant data, such as the temperature, barometric pressure, and relative humidity over the past few days. On some special events, such as the day before launching a Space Shuttle, his prediction should have a high degree of accuracy, and therefore he analyzes a larger amount of data to finely tune the parameters of his favorite mathematical meteorological model. On regular days, a smaller precision is tolerated, and thus he can afford to tune the parameters of the model on a coarser scale, saving data and computational resources.

In this paper we demonstrate quantitatively how the accuracy parameter plays a crucial role in determining the effective complexity of the learner's hypothesis class.[2]

We work within the decision-theoretic extension of the PAC framework, introduced in [12] and also known as *agnostic learning.* In this model, a finite sample of pairs (x, y) is obtained through independent draws from a fixed distribution $\boldsymbol{P}$ over $X \times [0,1]$. The goal of the learner is to be able to estimate the conditional expectation of y given x. This quantity is defined by a function $f : X \to [0,1]$, called the *regression function* in statistics. The learner is given a class $\mathcal{H}$ of candidate regression functions, which may or may not include the true regression function f. This class $\mathcal{H}$ is called *ϵ-learnable* if there is a learner with the property that for any distribution $\boldsymbol{P}$ and corresponding regression function f, given a large enough random sample from $\boldsymbol{P}$, this learner can find an ϵ-close approximation[3] to f within the class $\mathcal{H}$, or if f is not in $\mathcal{H}$, an ϵ-close approximation to a function in $\mathcal{H}$ that best approximates f. (This analysis of learnability is purely information-theoretic, and does not take into account computational complexity.) Throughout the paper, we assume that $\mathcal{H}$ (and later $\mathcal{F}$) satisfies some mild measurability conditions. A suitable such condition is the "image admissible Suslin" property (see [8, Section 10.3.1, page 101].)

[1] Adapted from [14].

[2] Our philosophy can be compared to the approach studied in [13], where the range of the functions in the hypothesis class is discretized in a number of elements proportional to the accuracy. In this case, one is interested in bounding the complexity of the discretized class through the dimension of the original class. Part of our results builds on this discretization technique.

[3] All notions of approximation are with respect to mean square error.

The special case where the distribution $\boldsymbol{P}$ is taken over $X \times \{0,1\}$ was studied in [14] by Kearns and Schapire, who called this setting *probabilistic concept learning.* If we further demand that the functions in $\mathcal{H}$ take only values in $\{0,1\}$, it turns out that this reduces to one of the standard PAC learning frameworks for learning deterministic concepts. In this case it is well known that the learnability of $\mathcal{H}$ is completely characterized by the finiteness of a simple combinatorial quantity known as the Vapnik-Chervonenkis (VC) dimension of $\mathcal{H}$ [24, 6]. An analogous combinatorial quantity for the probabilistic concept case was introduced by Kearns and Schapire. We call this quantity the *P_γ-dimension* of $\mathcal{H}$, where $\gamma > 0$ is a parameter that measures the "scale" to which the dimension of the class $\mathcal{H}$ is measured. They were only able to show that finiteness of this parameter was necessary for probabilistic concept learning, leaving the converse open. We solve this problem showing that this condition is also sufficient for learning in the harder agnostic model.

This last result has been recently complemented by Bartlett, Long, and Williamson [4], who have shown that the P_γ-dimension characterizes agnostic learnability with respect to the mean absolute error. In [20], Simon has independently proven a partial characterization of (nonagnostic) learnability using a slightly different notion of dimension.

As in the pioneering work of Vapnik and Chervonenkis [24], our analysis of learnability begins by establishing appropriate uniform laws of large numbers. In our main theorem, we establish the first combinatorial characterization of those classes of random variables whose means uniformly converge to their expectations for all distributions. Such classes of random variables have been called *Glivenko-Cantelli classes* in the empirical processes literature [9]. Given the usefulness of related uniform convergence results in combinatorics and randomized algorithms, we feel that this result may have many applications beyond those we give here. In addition, our results rely on a combinatorial result that generalizes Sauer's Lemma [18, 19]. This new lemma considerably extends some previously known results concerning $\{0,1,*\}$ tournament codes [21, 7]. As other related variants of Sauer's Lemma were proven useful in different areas, such as geometry and Banach space theory (see, e.g., [15, 1]), we also have hope to apply this result further.

2 Uniform Glivenko-Cantelli classes

The uniform, distribution-free convergence of empirical means to true expectations for classes of real-valued functions has been studied by Dudley, Giné, Pollard, Talagrand, Vapnik, Zinn, and others in the area of empirical processes. These results go under the general name of *uniform laws of large numbers.* We give a new combinatorial characterization of this phenomenon using methods related to those pioneered by Vapnik and Chervonenkis.

Let $\mathcal{F}$ be a class of functions from a set X into $[0,1]$. (All the results presented in this section can be generalized to classes of functions taking values in any bounded real range.) Let $\boldsymbol{P}$ denote a probability distribution over X such that f is $\boldsymbol{P}$-measurable for all $f \in \mathcal{F}$. By $\boldsymbol{P}(f)$ we denote the $\boldsymbol{P}$-mean of f, i.e., its integral w.r.t. $\boldsymbol{P}$. By $\boldsymbol{P}_n(f)$ we denote the random variable $\frac{1}{n}\Sigma_{i=1}^n f(x_i)$, where $x_1, x_2, \ldots, x_n$ are drawn independently at random according to $\boldsymbol{P}$.

Following Dudley, Giné and Zinn [9], we say that $\mathcal{F}$ is an *ϵ-uniform Glivenko-Cantelli class* if

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}} \Pr\left\{ \sup_{m\geq n} \sup_{f\in\mathcal{F}} |\boldsymbol{P}_m(f) - \boldsymbol{P}(f)| > \epsilon \right\} = 0. \tag{1}$$

Here Pr denotes the probability with respect to the points $x_1, x_2, \ldots,$ drawn independently at random according to $\boldsymbol{P}$.[4] The supremum is understood with respect to all distributions $\boldsymbol{P}$ over X

[4] Actually Dudley *et al.* use outer measure here, to avoid some measurability problems in certain cases.

(with respect to some suitable σ-algebra of subsets of X; see [9]).

We say that $\mathcal{F}$ satisfies a *distribution-free uniform strong law of large numbers*, or more briefly, that $\mathcal{F}$ is a *uniform Glivenko-Cantelli class*, if $\mathcal{F}$ is an ϵ-uniform Glivenko-Cantelli class for all $\epsilon > 0$.

We now recall the notion of VC-dimension, which characterizes uniform Glivenko-Cantelli classes of $\{0,1\}$-valued functions.

Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions on some domain set, X. We say $\mathcal{F}$ *VC-shatters* a set $A \subseteq X$ if, for every $E \subseteq A$, there exists some $f_E \in \mathcal{F}$ satisfying: For every $x \in A \setminus E$, $f_E(x) = 0$, and, for every $x \in E$, $f_E(x) = 1$. Let the *VC-dimension* of $\mathcal{F}$, denoted VC-dim$(\mathcal{F})$, be the maximal cardinality of a set $A \subseteq X$ that is VC-shattered by $\mathcal{F}$. (If $\mathcal{F}$ VC-shatters sets of unbounded finite sizes, then let VC-dim$(\mathcal{F}) = \infty$).

The following was established by Vapnik and Chervonenkis [24] for the "if" part and (in a stronger version) by Assouad and Dudley [2] (see [9, proposition 11, page 504].)

Theorem 2.1 *Let $\mathcal{F}$ be a class of functions from X into $\{0,1\}$. Then $\mathcal{F}$ is a uniform Glivenko-Cantelli class if and only if VC-dim$(\mathcal{F})$ is finite.*

Several generalizations of the VC-dimension to classes of real-valued functions have been previously proposed: Let $\mathcal{F}$ be a class of $[0,1]$-valued functions on some domain set X.

- (Pollard [16], see also [12]): We say $\mathcal{F}$ *P-shatters* a set $A \subseteq X$ if there exists a function $s : A \to \mathbb{R}$ such that, for every $E \subseteq A$, there exists some $f_E \in \mathcal{F}$ satisfying: For every $x \in A \setminus E$, $f_E(x) < s(x)$ and, for every $x \in E$, $f_E(x) \geq s(x)$.

  Let the *P-dimension* (denoted by P-dim) be the maximal cardinality of a set $A \subseteq X$ that is P-shattered by $\mathcal{F}$. (If $\mathcal{F}$ P-shatters sets of unbounded finite sizes, then let P-dim$(\mathcal{F}) = \infty$.)

- (Vapnik [23]): We say $\mathcal{F}$ *V-shatters* a set $A \subseteq X$ if there exists a constant $\alpha \in \mathbb{R}$ such that, for every $E \subseteq A$, there exists some $f_E \in \mathcal{F}$ satisfying: For every $x \in A \setminus E$, $f_E(x) < \alpha$ and, for every $x \in E$, $f_E(x) \geq \alpha$.

  Let the *V-dimension* (denoted by V-dim) be the maximal cardinality of a set $A \subseteq X$ that is V-shattered by $\mathcal{F}$. (If $\mathcal{F}$ V-shatters sets of unbounded finite sizes, then let V-dim$(\mathcal{F}) = \infty$.)

It is easily verified (see below) that the finiteness of neither of these combinatorial quantities provides a characterization of uniform Glivenko-Cantelli classes (more precisely, they both provide only a sufficient condition.)

Kearns and Schapire [14] introduced the following parametrized variant of the P-dimension. Let $\mathcal{F}$ be a class of $[0,1]$-valued functions on some domain set X and let γ be a positive real number. We say $\mathcal{F}$ *P_γ-shatters* a set $A \subseteq X$ if there exists a function $s : A \to [0,1]$ such that for every $E \subseteq A$ there exists some $f_E \in \mathcal{F}$ satisfying: For every $x \in A \setminus E$, $f_E(x) \leq s(x) - \gamma$ and, for every $x \in E$, $f_E(x) \geq s(x) + \gamma$.

Let the *P_γ-dimension* of $\mathcal{F}$, denoted P_γ-dim$(\mathcal{F})$, be the maximal cardinality of a set $A \subseteq X$ that is P_γ-shattered by $\mathcal{F}$. (If $\mathcal{F}$ P_γ-shatters sets of unbounded finite sizes, then let P_γ-dim$(\mathcal{F}) = \infty$).

A parametrized version of the V-dimension, which we'll call V_γ-dimension, can be defined in the same way we defined the P_γ-dimension from the P-dimension. The first lemma below follows directly from the definitions. The second lemma is proven through the pigeonhole principle.

Lemma 2.1 *For any $\mathcal{F}$ and any $\gamma > 0$, P_γ-dim$(\mathcal{F}) \leq P$-dim$(\mathcal{F})$ and V_γ-dim$(\mathcal{F}) \leq V$-dim$(\mathcal{F})$.*

Lemma 2.2 *For any class $\mathcal{F}$ of* $[0,1]$*-valued functions and for all* $\gamma > 0$,

$$V_\gamma\text{-}dim(\mathcal{F}) \le P_\gamma\text{-}dim(\mathcal{F}) \le \left(2\left\lceil\frac{1}{2\gamma}\right\rceil - 1\right) V_{\frac{\gamma}{2}}\text{-}dim(\mathcal{F}).$$

The P_γ and the V_γ dimensions have the advantage of being sensitive to the scale at which differences in function values are considered significant.

Our main result of this section is the following new characterization of uniform Glivenko-Cantelli classes, which exploits the scale-sensitive quality of the P_γ and the V_γ dimensions.

Theorem 2.2 *Let $\mathcal{F}$ be a class of functions from X into* $[0,1]$.

1. *There exist constants $a, b > 0$ (independent of $\mathcal{F}$) such that for any $\gamma > 0$*
   (a) *If P_γ-dim($\mathcal{F}$) is finite, then $\mathcal{F}$ is an ($a\gamma$)-uniform Glivenko-Cantelli class.*
   (b) *If V_γ-dim($\mathcal{F}$) is finite, then $\mathcal{F}$ is a ($b\gamma$)-uniform Glivenko-Cantelli class.*
   (c) *If P_γ-dim($\mathcal{F}$) is infinite, then $\mathcal{F}$ is not a ($\gamma - \tau$)-uniform Glivenko-Cantelli class for any $\tau > 0$.*
   (d) *If V_γ-dim($\mathcal{F}$) is infinite, then $\mathcal{F}$ is not a ($2\gamma - \tau$)-uniform Glivenko-Cantelli class for any $\tau > 0$.*
2. *The following are equivalent:*
   (a) *$\mathcal{F}$ is a uniform Glivenko-Cantelli class.*
   (b) *P_γ-dim($\mathcal{F}$) is finite for all $\gamma > 0$.*
   (c) *V_γ-dim($\mathcal{F}$) is finite for all $\gamma > 0$.*

(In the proof we actually show that $a \le 24$ and $b \le 48$, however these values are likely to be improved through a more careful analysis.)

The proof of this theorem is deferred to the next section. Note however that part 1 trivially implies part 2.

The following simple example (a special case of [9, Example 4, page 508], adapted to our purposes) shows that the finiteness of neither P-dim nor V-dim yields a characterization of Glivenko-Cantelli classes. (Throughout the paper we use ln to denote the natural logarithm and log to denote the logarithm in base 2.)

Example 2.1 *Let $\mathcal{F}$ be the class of all* $[0,1]$*-valued functions f defined on the positive integers and such that $f(x) \le e^{-x}$ for all $x \in \mathbb{N}$ and all $f \in \mathcal{F}$. Observe that, for all $\gamma > 0$, P_γ-dim($\mathcal{F}$) = V_γ-dim($\mathcal{F}$) = $\left\lfloor \ln \frac{1}{2\gamma} \right\rfloor$. Therefore, $\mathcal{F}$ is a uniform Glivenko-Cantelli class by Theorem 2.2. On the other hand, it is not hard to show that the P-dimension and the V-dimension of $\mathcal{F}$ are both infinite.*

Theorem 2.2 provides the first characterization of Glivenko-Cantelli classes in terms of a simple combinatorial quantity generalizing the Vapnik-Chervonenkis dimension to real-valued functions. Our results extend previous work by Dudley, Giné, and Zinn, where an equivalent characterization is shown to depend on the asymptotic properties of the metric entropy. Before stating the metric-entropy characterization of Glivenko-Cantelli classes we recall some basic notions from the theory of metric spaces.

Let (X, d) be a (pseudo) metric space, let A be a subset of X and $\epsilon > 0$.

- A set $B \subseteq A$ is an *ϵ-cover* for A if, for every $a \in A$, there exists some $b \in B$ such that $d(a, b) < \epsilon$. The *ϵ-covering number* of A, $\mathcal{N}_d(\epsilon, A)$, is the minimal cardinality of an ϵ-cover for A (if there is no such finite cover then it is defined to be ∞).
- A set $A \subseteq X$ is *ϵ-separated* if, for any distinct $a, b \in A$, $d(a, b) \geq \epsilon$. The *ϵ-packing number* of A, $\mathcal{M}_d(\epsilon, A)$, is the maximal size of an ϵ-separated subset of A.

The following is a simple, well-known fact.

Lemma 2.3 *For every (pseudo) metric space (X, d), every $A \subseteq X$, and $\epsilon > 0$*

$$\mathcal{M}_d(2\epsilon, A) \leq \mathcal{N}_d(\epsilon, A) \leq \mathcal{M}_d(\epsilon, A).$$

For a sequence of n points $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$ and a class $\mathcal{F}$ of real-valued functions defined on X, let $l^\infty_{\boldsymbol{x}_n}(f, g)$ denote the l^∞ distance between $f, g \in \mathcal{F}$ on the points $\boldsymbol{x}_n$, that is

$$l^\infty_{\boldsymbol{x}_n}(f, g) \overset{\mathrm{def}}{=} \max_{1 \leq i \leq n} |f(x_i) - g(x_i)|.$$

As we will often use the $l^\infty_{\boldsymbol{x}_n}$ distance, let us introduce the notation $\mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x}_n)$ and $\mathcal{M}(\epsilon, \mathcal{F}, \boldsymbol{x}_n)$ to stand for, respectively, the ϵ-covering and the ϵ-packing number of $\mathcal{F}$ with respect to $l^\infty_{\boldsymbol{x}_n}$.

A notion of metric entropy H_n, defined by

$$H_n(\epsilon, \mathcal{F}) \overset{\mathrm{def}}{=} \sup_{\boldsymbol{x}_n \in X^n} \log \mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x}_n),$$

has been used by Dudley, Giné and Zinn to prove the following.

Theorem 2.3 ([9, Theorem 6, page 500]) *Let $\mathcal{F}$ be a class of functions from X into* $[0, 1]$. *Then*

1. *$\mathcal{F}$ is a uniform Glivenko-Cantelli class if and only if* $\lim_{n \to \infty} H_n(\epsilon, \mathcal{F})/n = 0$ *for all $\epsilon > 0$.*
2. *For all $\epsilon > 0$, if* $\lim_{n \to \infty} H_n(\epsilon, \mathcal{F})/n = 0$ *then $\mathcal{F}$ is an (8ϵ)-uniform Glivenko-Cantelli class.*

The results by Dudley *et al.* also give similar characterizations using l^p norms in place of the l^∞ norm.

Related results were proved earlier by Vapnik and Chervonenkis [24, 25]. In particular, they proved an analogue of Theorem 2.3, where the convergence of means to expectations is characterized for a single distribution $\boldsymbol{P}$. Their characterization is based on $H_n(\epsilon, \mathcal{F})$ averaged with respect to samples drawn from $\boldsymbol{P}$.

3 Proof of the main theorem

We wish to obtain a characterization of uniform Glivenko-Cantelli classes in terms of their P_γ-dimension. By using standard techniques, we just need to bound the γ-packing numbers of sets of real-valued functions by an appropriate function of their $P_{c\gamma}$-dimension, for some positive constant c. Our line of attack is to reduce the problem to an analogous problem in the realm of finite-valued functions. Classes of functions into a discrete and finite range can then be analyzed using combinatorial tools.

We shall first introduce the discrete counterparts of the definitions above. Our next step will be to show how the real-valued problem can be reduced to a combinatorial problem. The final, and most technical part of our proof, will be the analysis of the combinatorial problem through a new generalization of Sauer's Lemma.

Let X be any set and let $B = \{1, \dots, b\}$. We consider classes $\mathcal{F}$ of functions f from X to B. Two such functions f and g are *separated* if they are 2-separated in the l^∞ metric, *i.e.*, if there exists some $x \in X$ such that $|f(x) - g(x)| \geq 2$. The class $\mathcal{F}$ is *pairwise separated* if f and g are separated for all $f \neq g$ in $\mathcal{F}$.

$\mathcal{F}$ *strongly shatters* a set $A \subseteq X$ if A is nonempty and there exists a function $s : A \to B$ such that, for every $E \subseteq A$, there exists some $f_E \in \mathcal{F}$ satisfying: For every $x \in A \setminus E$, $f_E(x) \leq s(x) - 1$ and, for every $x \in E$, $f_E(x) \geq s(x) + 1$. If s is any function witnessing the shattering of A by $\mathcal{F}$, we shall also say that $\mathcal{F}$ strongly shatters A according to s. Let the *strong dimension* of $\mathcal{F}$, S-dim$(\mathcal{F})$, be the maximal cardinality of a set $A \subseteq X$ that is strongly shattered by $\mathcal{F}$. (If $\mathcal{F}$ strongly shatters sets of unbounded finite size, then let S-dim$(\mathcal{F}) = \infty$).

For a function $f : X \to \mathbb{R}$, $f \geq 0$, and a real number $\rho > 0$, the *ρ-discretization* of f, denoted by f^ρ, is the function $f^\rho(x) \stackrel{\text{def}}{=} \lfloor \frac{f(x)}{\rho} \rfloor$, i.e. $f^\rho(x) = \max\{i \in \mathbb{N} : i\rho \leq f(x)\}$. For a class $\mathcal{F}$ of nonnegative real-valued functions let $\mathcal{F}^\rho \stackrel{\text{def}}{=} \{f^\rho : f \in \mathcal{F}\}$.

We need the following lemma.

Lemma 3.1 *For any class $\mathcal{F}$ of $[0,1]$-valued functions on a set X and for any $\rho > 0$,*

1. *for every $\gamma \leq \rho/2$, S-dim$(\mathcal{F}^\rho) \leq P_\gamma$-dim$(\mathcal{F})$;*

2. *for every $\epsilon \geq 2\rho$ and every $\boldsymbol{x}_n \in X^n$, $\mathcal{M}(\epsilon, \mathcal{F}, \boldsymbol{x}_n) \leq \mathcal{M}(2, \mathcal{F}^\rho, \boldsymbol{x}_n)$.*

Proof. To prove part 1 we show that any set strongly shattered by $\mathcal{F}^\rho$ is also $P_{\rho/2}$-shattered by $\mathcal{F}$. If $A \subseteq X$ is strongly shattered by $\mathcal{F}^\rho$, then there exists a function s such that for every $E \subseteq A$ there exists some $f_{(E)} \in \mathcal{F}$ satisfying: for every $x \in A \setminus E$, $f^\rho_{(E)}(x) + 1 \leq s(x)$ and for every $x \in E$, $f^\rho_{(E)}(x) \geq s(x) + 1$.

Assume first $f^\rho_{(E)}(x) + 1 \leq s(x)$. Then $\rho \cdot f^\rho_{(E)}(x) + \rho \leq \rho \cdot s(x)$ holds and, by definition of $f^\rho_{(E)}$, we have $f_{(E)}(x) < \rho \cdot f^\rho_{(E)}(x) + \rho$, which implies $f_{(E)}(x) < \rho \cdot s(x)$. Now assume $f^\rho_{(E)}(x) \geq s(x) + 1$. Then $\rho \cdot f^\rho_{(E)}(x) \geq \rho \cdot s(x) + \rho$ and, by definition of $f^\rho_{(E)}$, we have $f_{(E)}(x) \geq \rho f^\rho_{(E)}(x)$, which implies $f_{(E)}(x) \geq \rho \cdot s(x) + \rho$. Thus A is $P_{\rho/2}$-shattered by $\mathcal{F}$, as can be seen using the function $s' : A \to [0,1]$ defined by $s'(x) \stackrel{\text{def}}{=} \rho \cdot s(x) + \rho/2$ for all $x \in X$.

To prove part 2 of the lemma it is enough to observe that, by the definition of $\mathcal{F}^\rho$, for all $f, g \in \mathcal{F}$ and all $x \in X$, $|f(x) - g(x)| \geq 2\rho$ implies $|f^\rho(x) - g^\rho(x)| \geq 2$. □

We now prove our main combinatorial result which gives a new generalization of Sauer's Lemma. Our result extends some previous work concerning $\{0, 1, *\}$ tournament codes, proven in a completely different way (see [21, 7]).

The lemma concerns the l^∞ packing numbers of classes of functions into a finite range. It shows that, if such a class has a finite strong dimension, then its 2-packing number is bounded by a subexponential function of the cardinality of its domain. For simplicity, we arbitrarily fix a sequence $\boldsymbol{x}_n$ of n points in X and consider only the restriction of $\mathcal{F}$ to this domain, dropping the subscript $\boldsymbol{x}_n$ from our notation.

Lemma 3.2 *If $\mathcal{F}$ is a class of functions from a finite domain X of cardinality n to a finite range, $B = \{1, 2, \dots, b\}$, and S-dim$(\mathcal{F}) = d$, then $\mathcal{M}_{l^\infty}(2, \mathcal{F}) < 2(nb^2)^{\lceil \log y \rceil}$, where $y = \sum_{i=1}^{d} \binom{n}{i} b^i$.*

Note that for fixed d the bound in Lemma 3.2 is $n^{O(\log n)}$ even if b is not a constant but a polynomial in n.

Proof of Lemma 3.2. Fix $b \geq 3$ (the case $b < 3$ is trivial.) Let us say that a class $\mathcal{F}$ as above strongly shatters a pair (A, s) (for a nonempty subset A of X and a function $s : A \rightarrow B$) if $\mathcal{F}$ strongly shatters A according to s. For all integers $h \geq 2$ and $n \geq 1$, let $t(h, n)$ denote the maximum number t such that for every set $\boldsymbol{F}$ of h pairwise separated functions f from X to B, $\boldsymbol{F}$ strongly shatters at least t pairs (A, s) where $A \subseteq X$, $A \neq \emptyset$, and $s : A \rightarrow B$. If no such $\boldsymbol{F}$ exists, then $t(h, n)$ is infinite.

Note that the number of possible pairs (A, s) for which the cardinality of A does not exceed $d \geq 1$ is less than $y = \sum_{i=1}^{d} \binom{n}{i} b^i$ (as for A of size $i > 0$ there are strictly less than b^i possibilities to choose s.) It follows that, if $t(h, n) \geq y$ for some h, then $\mathcal{M}_{l^\infty}(2, \mathcal{F}) < h$ for all sets $\mathcal{F}$ of functions from X to B and such that $S\text{-dim}(\mathcal{F}) \leq d$. Therefore, to finish the proof, it suffices to show that $t(2(nb^2)^{\lceil \log y \rceil}, n) \geq y$ for all $d \geq 1$ and $n \geq 1$.

We claim that $t(2, n) = 1$ for all $n \geq 1$, and $t(2mnb^2, n) \geq 2t(2m, n-1)$ for all $m \geq 1$ and $n \geq 2$. The first part of the claim is readily verified. For the second part, first note that if no set of $2mnb^2$ pairwise separated functions from X to B exists, then $t(2mnb^2, n) = \infty$ and hence the claim holds. Assume then that there is a set $\boldsymbol{F}$ of $2mnb^2$ pairwise separated functions from X to B. Split it arbitrarily into mnb^2 pairs. For each pair (f, g) find a coordinate $x \in X$ where $|f(x) - g(x)| > 1$. By the pigeonhole principle, the same coordinate x is picked for at least mb^2 pairs. Again by the pigeonhole principle, there are at least $mb^2/\binom{b}{2} > 2m$ of these pairs (f, g) for which the (unordered) set $\{f(x), g(x)\}$ is the same. This means that there are two sub-classes of $\boldsymbol{F}$, call them $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, and there are $x \in X$ and $i, j \in B$, with $j > i + 1$, so that for each $f \in \boldsymbol{F}_1$, $f(x) = i$ and for each $g \in \boldsymbol{F}_2$ $g(x) = j$, and $|\boldsymbol{F}_1| = |\boldsymbol{F}_2| = 2m$. Obviously, the members of $\boldsymbol{F}_1$ are pairwise separated on $X \setminus \{x\}$ and the same holds for the members of $\boldsymbol{F}_2$. Hence, by the definition of the function t, $\boldsymbol{F}_1$ strongly shatters at least $t(2m, n-1)$ pairs (A, s) with $A \subseteq X \setminus \{x\}$, and the same holds for $\boldsymbol{F}_2$. Clearly $\boldsymbol{F}$ strongly shatters all pairs strongly shattered by $\boldsymbol{F}_1$ or $\boldsymbol{F}_2$. Moreover, if the same pair (A, s) is strongly shattered both by $\boldsymbol{F}_1$ and by $\boldsymbol{F}_2$, then $\boldsymbol{F}$ also strongly shatters the pair $(A \cup \{x\}, s')$, where $s'(y) = s(y)$ for $y \in A$ and $s'(x) = \lfloor \frac{i+j}{2} \rfloor$. It follows that $t(2mnb^2, n) \geq 2t(2m, n-1)$, establishing the claim.

Now suppose $n > r \geq 1$. Let $h = 2(nb^2)((n-1)b^2) \cdots ((n-r+1)b^2)$. By repeated application of the above claim, it follows that $t(h, n) \geq 2^r$. Since t is clearly monotone in its first argument, and $2(nb^2)^r \geq h$, this implies $t(2(nb^2)^r, n) \geq 2^r$ for all $n > r \geq 1$. Now set $r = \lceil \log_2 y \rceil$, where $y = \sum_{i=1}^{d} \binom{n}{i} b^i$. If $n \leq r$, then $2(nb^2)^r > b^n$. However, since the total number of functions from X to B is b^n, there are no sets of pairwise separated functions of size larger than this, and hence $t(2(nb^2)^r, n) = t(2(nb^2)^{\lceil \log_2 y \rceil}, n) = \infty > y$ in this case. On the other hand, when $n > r$, the result above yields $t(2(nb^2)^{\lceil \log_2 y \rceil}, n) \geq 2^{\lceil \log_2 y \rceil} \geq y$. Thus in either case $t(2(nb^2)^{\lceil \log_2 y \rceil}, n) \geq y$, completing the proof. □

Before proving Theorem 2.2, we need two more lemmas. The first one is a straightforward adaptation of [22, Section A.6, p. 223].

Lemma 3.3 *Let $\mathcal{F}$ be a class of functions from X into $[0, 1]$ and let $\boldsymbol{P}$ be a distribution over X. Then, for all $\epsilon > 0$ and all $n \geq 2/\epsilon^2$,*

$$\Pr\left\{\sup_{f \in \mathcal{F}} |\boldsymbol{P}_n(f) - \boldsymbol{P}(f)| > \epsilon\right\} \leq 12n \cdot \boldsymbol{E}\left[\mathcal{N}(\epsilon/6, \mathcal{F}, \boldsymbol{x}'_{2n})\right] e^{-\epsilon^2 n/36} \qquad (2)$$

where $\Pr$ *denotes the probability w.r.t. the sample $x_1, \ldots, x_n$ drawn independently at random according to $\boldsymbol{P}$, and $\boldsymbol{E}$ the expectation w.r.t. a second sample $\boldsymbol{x}'_{2n} = x'_1, \ldots, x'_{2n}$ also drawn independently at random according to $\boldsymbol{P}$.*

Proof. A well-known result (see e.g. [8, Lemma 11.1.5] or [10, Lemma 2.5]) shows that, for all $n \geq 2/\epsilon^2$,

$$\Pr\left\{\sup_{f\in\mathcal{F}} |\boldsymbol{P}_n(f) - \boldsymbol{P}(f)| > \epsilon\right\} \leq 2\Pr\left\{\sup_{f\in\mathcal{F}} |\boldsymbol{P}_{n'}(f) - \boldsymbol{P}_{n''}(f)| > \frac{\epsilon}{2}\right\},$$

where $\boldsymbol{P}_{n'}(f) = \frac{1}{n}\Sigma_{i=1}^{n} f(x_i')$, $\boldsymbol{P}_{n''}(f) = \frac{1}{n}\Sigma_{i=n+1}^{2n} f(x_i')$.

We combine this with a result by Vapnik [22, pp. 225-228] showing that for all $\epsilon > 0$

$$\Pr\left\{\sup_{f\in\mathcal{F}} |\boldsymbol{P}_{n'}(f) - \boldsymbol{P}_{n''}(f)| > \epsilon\right\} \leq 6n \cdot \boldsymbol{E}\left[\mathcal{N}(\epsilon/3, \mathcal{F}, \boldsymbol{x}_{2n}')\right] e^{-\epsilon^2 n/9}.$$

This concludes the proof. □

The next result applies Lemma 3.2 to bound the expected covering number of a class $\mathcal{F}$ in terms of P_γ-dim$(\mathcal{F})$.

Lemma 3.4 *Let $\mathcal{F}$ be a class of functions from X into $[0,1]$ and $\boldsymbol{P}$ a distribution over X. Choose $0 < \epsilon < 1$ and let $d = P_{\epsilon/4}$-dim$(\mathcal{F})$. Then*

$$\boldsymbol{E}\left[\mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x}_n)\right] \leq 2\left(\frac{4n}{\epsilon^2}\right)^{d\log(2en/(d\epsilon))}$$

where the expectation $\boldsymbol{E}$ is taken w.r.t. a sample $x_1, \ldots, x_n$ drawn independently at random according to $\boldsymbol{P}$.

Proof. By Lemma 2.3, Lemmas 3.1 and 3.2, and Stirling's approximation,

$$\begin{aligned} \boldsymbol{E}\left[\mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x}_n)\right] &\leq \sup_{\boldsymbol{x}_n} \mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x}_n) \\ &\leq \sup_{\boldsymbol{x}_n} \mathcal{M}(\epsilon, \mathcal{F}, \boldsymbol{x}_n) \\ &\leq \sup_{\boldsymbol{x}_n} \mathcal{M}(2, \mathcal{F}^{\epsilon/2}, \boldsymbol{x}_n) \\ &\leq 2\left(\frac{4n}{\epsilon^2}\right)^{d\log(2en/(d\epsilon))} \end{aligned} \tag{3}$$

□

We are now ready to prove our characterization of uniform Glivenko-Cantelli classes.

Proof of Theorem 2.2. We begin with part 1.d: If V_γ-dim$(\mathcal{F}) = \infty$ for some $\gamma > 0$, then we will show that $\mathcal{F}$ is not a $(2\gamma - \tau)$-uniform Glivenko-Cantelli class for any $\tau > 0$. To see this, assume V_γ-dim$(\mathcal{F}) = \infty$. For any sample size n and any $d > n$, find in X a set S of d points that are V_γ-shattered by $\mathcal{F}$. Then there exists $\alpha > 0$ such that for every $E \subseteq S$ there exists some $f_E \in \mathcal{F}$ satisfying: For every $x \in A \setminus E$, $f_E(x) \leq \alpha - \gamma$ and, for every $x \in E$, $f_E(x) \geq \alpha + \gamma$. Let $\boldsymbol{P}$ be the uniform distribution on S. For any sample $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ from S there is a function $f \in \mathcal{F}$ such that $f(x_i) \leq \alpha - \gamma$, $1 \leq i \leq n$, and $f(x) \geq \alpha + \gamma$ for all $x \in S \setminus \{x_1, \ldots, x_n\}$. Thus, for any $\tau > 0$, if $d = |S|$ is large enough we can find some $f \in \mathcal{F}$ such that $|\boldsymbol{P}(f) - \boldsymbol{P}_n(f)| \geq 2\gamma - \tau$. This proves part 1.d. Part 1.c follows from Lemma 2.2.

To prove part 1.a we use inequality (2) from Lemma 3.3. Then, to bound the expected covering number we apply Lemma 3.4. This shows that

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}} \Pr\left\{\sup_{f\in\mathcal{F}} |\boldsymbol{P}_n(f) - \boldsymbol{P}(f)| > a\gamma\right\} = 0 \tag{4}$$

for some $a > 0$ whenever P_γ-dim$(\mathcal{F})$ is finite.

Equation (4) shows that $\boldsymbol{P}_n(f) \to \boldsymbol{P}(f)$ in probability for all $f \in \mathcal{F}$ and all distributions $\boldsymbol{P}$. Furthermore, as Lemma 3.3 and Lemma 3.4 imply that $\sum_{n=1}^{\infty} \Pr\{\sup_{f\in\mathcal{F}} |\boldsymbol{P}_n(f) - \boldsymbol{P}(f)| > a\gamma\} < \infty$, one may apply the Borel-Cantelli lemma and strengthen (4) to almost sure convergence, i.e.

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}} \Pr\left\{\sup_{m\geq n}\sup_{f\in\mathcal{F}} |\boldsymbol{P}_m(f) - \boldsymbol{P}(f)| > a\gamma\right\} = 0.$$

This completes the proof of part 1.a. The proof of part 1.b follows immediately from Lemma 2.2. □

The proof of Theorem 2.2, in addition to being simpler than the proof in [9] (see Theorem 2.3 in this paper), also provides new insights into the behaviour of the metric entropy used in that characterization. It shows that there is a large gap in the growth rate of the metric entropy $H_n(\epsilon, \mathcal{F})$: either $\mathcal{F}$ is a uniform Glivenko-Cantelli class, and hence, by (3) and by definition of H_n, for all $\epsilon > 0$, $H_n(\epsilon, \mathcal{F}) = O(\log^2 n)$; or $\mathcal{F}$ is not a uniform Glivenko-Cantelli class, and hence there exists $\epsilon > 0$ such that P_ϵ-dim$(\mathcal{F}) = \infty$, which is easily seen to imply that $H_n(\epsilon, \mathcal{F}) = \Omega(n)$. It is unknown if $\log^2 n$ can be replaced by $\log^\alpha n$ where $1 \leq \alpha < 2$.

From the proof of Theorem 2.2 we can obtain bounds on the sample size sufficient to guarantee that, with high probability, in a class of $[0,1]$-valued random variables each mean is close to its expectation.

Theorem 3.1 *Let $\mathcal{F}$ be a class of functions from X into $[0,1]$. Then for all distributions $\boldsymbol{P}$ over X and all $\epsilon, \delta > 0$*

$$\Pr\left\{\sup_{f\in\mathcal{F}} |\boldsymbol{P}_n(f) - \boldsymbol{P}(f)| > \epsilon\right\} \leq \delta \tag{5}$$

for

$$n = O\left(\frac{1}{\epsilon^2}\left(d \ln^2 \frac{d}{\epsilon} + \ln \frac{1}{\delta}\right)\right)$$

where d is the $P_{\epsilon/24}$-dimension of $\mathcal{F}$.

Theorem 3.1 is proven by applying Lemma 3.3 and Lemma 3.4 along with standard approximations. We omit the proof of this theorem and mention instead that an improved sample size bound has been shown by Bartlett and Long [3, Equation (5), Theorem 9]. In particular, they show that if the $P_{(1/4-\tau)\epsilon}$-dimension d' of $\mathcal{F}$ is finite for some $\tau > 0$, then a sample size of order

$$O\left(\frac{1}{\epsilon^2}\left(d' \ln^2 \frac{1}{\epsilon} + \ln \frac{1}{\delta}\right)\right) \tag{6}$$

is sufficient for (5) to hold.

4 Applications to Learning

In this section we define the notion of *learnability up to accuracy* ϵ, or ϵ*-learnability*, of statistical regression functions. In this model, originally introduced in [12] and also known as "agnostic learning", the learning task is to approximate the regression function of an unknown distribution. The probabilistic concept learning of Kearns and Schapire [14] and the real-valued function learning with noise investigated by Bartlett, Long, and Williamson [4] are special cases of this framework.

We show that a class of functions is ϵ-learnable whenever its $P_{a\epsilon}$-dimension is finite for some constant $a > 0$. Moreover, combining this result with those of Kearns and Schapire, who show that a similar condition is necessary for the weaker probabilistic concept learning, we can conclude that the finiteness of the P_γ-dimension for all $\gamma > 0$ characterizes learnability in the probabilistic concept framework. This solves an open problem from [14].

Let us begin by briefly introducing our learning model. The model examines learning problems involving statistical regression on $[0,1]$-valued data. Assume X is an arbitrary set (as above), and $Y = [0,1]$. Let $Z = X \times Y$, and let $\boldsymbol{P}$ be an unknown distribution on Z. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be random variables respectively distributed according to the marginal of $\boldsymbol{P}$ on X and Y. The *regression function* f for distribution $\boldsymbol{P}$ is defined, for all $x \in X$, by

$$f(x) = \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X} = x).$$

The general goal of regression is to approximate f in the mean square sense (i.e. in L_2-norm) when the distribution $\boldsymbol{P}$ is unknown, but we are given $\boldsymbol{z}_n = (z_1, z_2, \ldots, z_n)$, where each $z_i = (x_i, y_i)$ is independently generated from the distribution $\boldsymbol{P}$.

In general we cannot hope to approximate the regression function f for an arbitrary distribution $\boldsymbol{P}$. Therefore we choose a *hypothesis space* $\mathcal{H}$, which is a family of mappings $h : X \to [0,1]$, and settle for a function in $\mathcal{H}$ that is close to the best approximation to f in the hypothesis space $\mathcal{H}$. To this end, for each hypothesis $h \in \mathcal{H}$, let the function $\ell_h : Z \to [0,1]$ be defined by: $\ell_h(x,y) = (h(x) - y)^2$, for all $x \in X$ and $y \in [0,1]$. Thus $\boldsymbol{P}(\ell_h)$ is the mean square loss of h. The goal of learning in the present context is to find a function $\widehat{h} \in \mathcal{H}$ such that

$$\boldsymbol{P}(\ell_{\widehat{h}}) \leq \inf_{h \in \mathcal{H}} \boldsymbol{P}(\ell_h) + \epsilon$$

for some given accuracy $\epsilon > 0$. It is easily verified that if $\inf_{h\in\mathcal{H}} \boldsymbol{P}(\ell_h)$ is achieved by some $h \in \mathcal{H}$, then h is the function in $\mathcal{H}$ closest to the true regression function f in the L_2 norm.

A *learning procedure* is a mapping A from finite sequences in Z to $\mathcal{H}$. A learning procedure produces a hypothesis $\widehat{h} = A(\boldsymbol{z}_n)$ for any *training sample* $\boldsymbol{z}_n$. For given accuracy parameter ϵ, we say that $\mathcal{H}$ is *ϵ-learnable* if there exists a learning procedure A such that

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}} \Pr\left\{\boldsymbol{P}(\ell_{A(\boldsymbol{z}_n)}) > \inf_{\mathcal{H}} \boldsymbol{P}(\ell_h) + \epsilon\right\} = 0. \tag{7}$$

Here Pr denotes the probability with respect to the random sample $\boldsymbol{z}_n \in Z^n$, each z_i drawn independently according to $\boldsymbol{P}$, and the supremum is over all distributions $\boldsymbol{P}$ defined on a suitable σ-algebra of subsets of Z. Thus $\mathcal{H}$ is ϵ-learnable if, given a large enough training sample, we can reliably find a hypothesis $\widehat{h} \in \mathcal{H}$ with mean square error close to that of the best hypothesis in $\mathcal{H}$. Finally, we say $\mathcal{H}$ is *learnable* if and only if it is ϵ-learnable for all $\epsilon > 0$.

If $Z = X \times \{0,1\}$ the above definitions of learnability yield the probabilistic concept learning model. In this case, if (7) holds for some $\epsilon > 0$ and some class $\mathcal{H}$, we say that $\mathcal{H}$ is *ϵ-learnable in the p-concept model.*

We now state and prove the main results of this section. We start by establishing sufficient conditions for ϵ-learnability and learnability in terms of the P_γ-dimension.

Theorem 4.1 *There exist constants $a, b > 0$ such that for any $\gamma > 0$:*

1. *If P_γ-dim($\mathcal{H}$) is finite, then $\mathcal{H}$ is $(a\gamma)$-learnable.*
2. *If V_γ-dim($\mathcal{H}$) is finite, then $\mathcal{H}$ is $(b\gamma)$-learnable.*

3. *If P_γ-dim$(\mathcal{H})$ is finite for all $\gamma > 0$ or V_γ-dim$(\mathcal{H})$ is finite for all $\gamma > 0$, then $\mathcal{H}$ is learnable.*

We then prove the following, which characterizes p-concept learnability.

Theorem 4.2

1. *If P_γ-dim$(\mathcal{H})$ is infinite, then $\mathcal{H}$ is not $(\gamma^2/8 - \tau)$-learnable in the p-concept model for any $\tau > 0$.*
2. *If V_γ-dim$(\mathcal{H})$ is infinite, then $\mathcal{H}$ is not $(\gamma^2/2 - \tau)$-learnable in the p-concept model for any $\tau > 0$.*
3. *The following are equivalent:*
   (a) *$\mathcal{H}$ is learnable in the p-concept model.*
   (b) *P_γ-dim$(\mathcal{H})$ is finite for all $\gamma > 0$.*
   (c) *V_γ-dim$(\mathcal{H})$ is finite for all $\gamma > 0$.*
   (d) *$\mathcal{H}$ is a uniform Glivenko-Cantelli class.*

Proof of Theorem 4.1. It is clear that part 3 follows from part 1 using Theorem 2.2. Also, by Lemma 2.2, part 1 is equivalent to part 2. Thus, to prove Theorem 4.1 it suffices to establish part 1. We do so via the next two lemmas.

Let $\ell_{\mathcal{H}} = \{\ell_h : h \in \mathcal{H}\}$.

Lemma 4.1 *If $\ell_{\mathcal{H}}$ is an ϵ-uniform Glivenko-Cantelli class, then $\mathcal{H}$ is (3ϵ)-learnable.*

Proof. The proof uses the method of *empirical risk minimization*, analyzed by Vapnik [22]. As above, let $\boldsymbol{P}_n(\ell_h)$ denote the empirical loss on the given sample $\boldsymbol{z}_n = (z_1, z_2, \ldots, z_n)$, that is

$$\boldsymbol{P}_n(\ell_h) = \frac{1}{n}\sum_{i=1}^{n} \ell_h(z_i) = \frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2.$$

A learning procedure, A^*_ϵ, ϵ-minimizes the empirical risk if $A^*_\epsilon(\boldsymbol{z}_n)$ is any $\widehat{h} \in \mathcal{H}$ such that $\boldsymbol{P}_n(\ell_{\widehat{h}}) \leq \inf_{h\in\mathcal{H}} \boldsymbol{P}_n(\ell_h) + \epsilon$. Let us show that any such procedure is guaranteed to 3ϵ-learn $\mathcal{H}$.

Fix any $n \in \mathbb{N}$. If

$$|\boldsymbol{P}_n(\ell_h) - \boldsymbol{P}(\ell_h)| \leq \epsilon$$

for all $h \in \mathcal{H}$, then

$$\begin{aligned} \boldsymbol{P}(\ell_{A^*_\epsilon(\boldsymbol{z}_n)}) &\leq \boldsymbol{P}_n(\ell_{A^*_\epsilon(\boldsymbol{z}_n)}) + \epsilon \\ &\leq \boldsymbol{P}_n(\ell_h) + 2\epsilon \qquad \forall h \in \mathcal{H} \\ &\leq \boldsymbol{P}(\ell_h) + 3\epsilon \qquad \forall h \in \mathcal{H}, \end{aligned}$$

and thus $\boldsymbol{P}(\ell_{A^*_\epsilon(\boldsymbol{z}_n)}) \leq \inf_{h\in\mathcal{H}} \boldsymbol{P}(\ell_h) + 3\epsilon$. Hence, since we chose n and ϵ arbitrarily,

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}} \Pr\left\{\sup_{m\geq n}\sup_{h\in\mathcal{H}} |\boldsymbol{P}_m(\ell_h) - \boldsymbol{P}(\ell_h)| > \epsilon\right\} = 0$$

implies

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}} \Pr\left\{\boldsymbol{P}(\ell_{A^*_\epsilon(\boldsymbol{z}_n)}) > \inf_{h\in\mathcal{H}} \boldsymbol{P}(\ell_h) + 3\epsilon\right\} = 0.$$

□

The following lemma shows that bounds on the covering numbers of a family of functions $\mathcal{H}$ can be applied to the induced family of loss functions $\ell_{\mathcal{H}}$. We formulate the lemma in terms of the square loss but it may be readily generalized to other loss functions. A similar result was independently proven by Bartlett, Long, and Williamson in [4] for the absolute loss $L(x, y) = |x - y|$ (and with respect to the l^1 metric rather than the l^∞ metric used here).

Lemma 4.2 *For all $\epsilon > 0$, all $\mathcal{H}$, and any $\boldsymbol{z}_n = (z_1, \ldots, z_n)$, where $z_i = (x_i, y_i)$, $i = 1, \ldots, n$,*

$$\mathcal{N}(\epsilon, \ell_{\mathcal{H}}, \boldsymbol{z}_n) \leq \mathcal{N}(\epsilon/2, \mathcal{H}, \boldsymbol{x}_n)$$

where $\boldsymbol{x}_n = (x_1, \ldots, x_n)$.

Proof. It suffices to show that, for any $f, g \in \mathcal{H}$ and any $1 \leq i \leq n$, if $|f(x_i) - g(x_i)| \leq \epsilon/2$ then $|(f(x_i) - y_i)^2 - (g(x_i) - y_i)^2| \leq \epsilon$. This follows by noting that, for every $s, t, w \in [0, 1]$, $|(s - w)^2 - (t - w)^2| \leq 2|s - t|$. □

We end the proof of Theorem 4.1 by proving part 1. By Lemma 4.1, it suffices to show that $\ell_{\mathcal{H}}$ is $(a\gamma)$-uniform Glivenko-Cantelli for some $a > 0$. To do so we use (2) from Lemma 3.3. Then, to bound the expected covering number, we apply first Lemma 4.2 and then Lemma 3.4. This establishes

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}} \Pr\left\{\sup_{h\in\mathcal{H}} |\boldsymbol{P}_n(\ell_h) - \boldsymbol{P}(\ell_h)| > a\gamma\right\} = 0$$

for some $a > 0$ whenever P_γ-dim$(\mathcal{H})$ is finite. An application of the Borel-Cantelli lemma to get almost sure convergence yields the proof. □

We conclude this section by proving our characterization of p-concept learnability.

Proof of Theorem 4.2. As ϵ-learnability implies ϵ-learnability in the p-concept model, we have that part 3 follows from part 1, part 2, and from Theorem 4.1 using Theorem 2.2.

The proof of part 2 uses arguments similar to those used to prove part 1.d of Theorem 2.2. Finally note that part 1 follows from part 2 by Lemma 2.2 (we remark that a more restricted version of part 1 was proven in Theorem 11 of [14].) □

5 Conclusions and open problems

In this work we have shown a characterization of uniform Glivenko-Cantelli classes based on a combinatorial notion generalizing the Vapnik-Chervonenkis dimension. This result has been applied to show that the same notion of dimension provides the weakest combinatorial condition known to imply agnostic learnability and, furthermore, characterizes learnability in the model of probabilistic concepts under the square loss. Our analysis demonstrates how the accuracy parameter in learning plays a central role in determining the effective dimension of the learner's hypothesis class.

An open problem is what other notions of dimension may characterize uniform Glivenko-Cantelli classes. In fact, for classes of functions with finite range, the same characterization is achieved by each member of a family of several notions of dimension (see [5]).

A second open problem is the asymptotic behaviour of the metric entropy: we have already shown that for all $\epsilon > 0$, $H_n(\epsilon, \mathcal{F}) = O(\log^2 n)$ if $\mathcal{F}$ is a uniform Glivenko-Cantelli class and $H_n(\epsilon, \mathcal{F}) = \Omega(n)$ otherwise. We conjecture that for all $\epsilon > 0$, $H_n(\epsilon, \mathcal{F}) = O(\log n)$ whenever $\mathcal{F}$ is a uniform Glivenko-Cantelli class. A positive solution of this conjecture would also affect the sample complexity bound (6) of Bartlett and Long. In fact, suppose that Lemma 3.4 is improved by showing that $\sup_{\boldsymbol{x}_n} \mathcal{M}(\epsilon, \mathcal{F}, \boldsymbol{x}_n) \leq (n/\epsilon^2)^{cd}$ for some positive constant c and for $d = P_{\epsilon/4}\text{-dim}(\mathcal{F})$ (note that this implies our conjecture.) Then, combining this with [3, Lemma 10–11], we can easily show a sample complexity bound of

$$O\left(\frac{1}{\epsilon^2}\left(d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}\right)\right),$$

for any $0 < \tau < 1/8$ for which $d = P_{(1/8-\tau)\epsilon}\text{-dim}(\mathcal{F})$ is finite. It is not clear how to bring the constant 1/8 down to 1/4 as in (6), which was proven using l^1 packing numbers.

Acknowledgments

We would like to thank Michael Kearns, Yoav Freund, Ron Aharoni and Ron Holzman for fruitful discussions, and Alon Itai for useful comments concerning the presentation of the results.

Thanks also to an anonymous referee for the many valuable comments, suggestions, and references.

References

[1] N. Alon and V.D. Milman. Embedding of ℓ_∞^k in finite dimensional Banach spaces. *Israel Journal of Mathematics*, 45:265–280, 1983.

[2] P. Assouad and R.M. Dudley. Minimax nonparametric estimation over classes of sets. Preprint, 1989.

[3] P.L. Bartlett and P.M. Long. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, pages 392–401. ACM Press, 1995.

[4] P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the 7th Annual ACM Workshop on Computational Learning Theory*, pages 299–310. ACM Press, 1994. To appear in *Machine Learning.*

[5] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P.M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and Systems Sciences*, 50(1):74–86, 1995.

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

[7] K.L. Collins, P.W. Shor, and J.R. Stembridge. A lower bound for $\{0, 1, *\}$ tournament codes. *Discrete Mathematics*, 63:15–19, 1987.

[8] R.M. Dudley. A course on empirical processes. In *Lecture Notes in Mathematics*, volume 1097, pages 2–142. Springer, 1984.

[9] R.M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4:485–510, 1991.

[10] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12:929–989, 1984.

[11] I. Guyon, V. Vapnik, B. Boser, L. Bottou, and S.A. Solla. Structural risk minimization for character recognition. In *Proceedings of the 1991 Conference on Advances in Neural Information Processing Systems*, pages 471–479, 1991.

[12] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[13] D. Haussler and P.M. Long. A generalization of Sauer's lemma. *J. Combinatorial Theory (A)*, 71:219–240, 1995.

[14] M. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and Systems Sciences*, 48(3):464–497, 1994. An extended abstract appeared in the Proceedings of the 30th Annual Symposium on the Foundations of Computer Science.

[15] V.D. Milman. Some remarks about embedding of ℓ_1^k in finite dimensional spaces. *Israel Journal of Mathematics*, 43:129–138, 1982.

[16] D. Pollard. *Empirical Processes : Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[17] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[18] N. Sauer. On the density of families of sets. *J. Combinatorial Theory (A)*, 13:145–147, 1972.

[19] S. Shelah. A combinatorial problem: Stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.

[20] H.U. Simon. Bounds on the number of examples needed for learning functions. In *Proceedings of the First Euro-COLT Workshop*, pages 83–94. The Institute of Mathematics and its Applications, 1994.

[21] J.H. van Lint. $\{0,1,*\}$ distance problems in combinatorics. In *Lecture Notes of the London Mathematical Society*, volume 103, pages 113–135. Cambridge University Press, 1985.

[22] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, 1982.

[23] V.N. Vapnik. Inductive principles of the search for empirical dependencies. In *Proceedings of the 2nd Annual Workshop on Computational Learning Theory*, pages 3–21, 1989.

[24] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[25] V.N. Vapnik and A.Y. Chervonenkis. Necessary and sufficient conditions for uniform convergence of means to mathematical expectations. *Theory of Probability and its Applications*, 26(3):532–553, 1981.