

# PAC learnability versus VC dimension: a footnote to a basic result of statistical learning

Vladimir Pestov

**Abstract**—A fundamental result of statistical learning theory states that a concept class is PAC learnable if and only if it is a uniform Glivenko–Cantelli class if and only if the VC dimension of the class is finite. However, the theorem is only valid under special assumptions of measurability of the class, in which case the PAC learnability even becomes consistent. Otherwise, there is a classical example, constructed under the Continuum Hypothesis by Dudley and Durst and further adapted by Blumer, Ehrenfeucht, Haussler, and Warmuth, of a concept class of VC dimension one which is neither uniform Glivenko–Cantelli nor consistently PAC learnable. We show that, rather surprisingly, under an additional set-theoretic hypothesis which is much milder than the Continuum Hypothesis (Martin’s Axiom), PAC learnability is equivalent to finite VC dimension for every concept class.

## I. INTRODUCTION

The following is a fundamental result of statistical learning theory.

*Theorem 1:* For a concept class  $\mathcal{C}$  the following three conditions are equivalent:

- 1)  $\mathcal{C}$  is distribution-free PAC learnable,
- 2)  $\mathcal{C}$  is a uniform Glivenko–Cantelli class, and
- 3) the Vapnik–Chervonenkis dimension of  $\mathcal{C}$  is finite.

It is in this form that the theorem is usually stated in textbooks on the subject, see [1], [2]. The condition 1) means the existence of a learning rule for  $\mathcal{C}$  which is probably approximately correct.

However, strictly speaking, the result is only true under a suitable measurability assumption on the concept class  $\mathcal{C}$ . One such assumption is that of  $\mathcal{C}$  being *image admissible Souslin*: the class  $\mathcal{C}$  can be parametrized with elements of the unit interval so that pairs  $(x, t)$ ,  $x \in C_t$ ,  $t \in [0, 1]$  form an analytic subset of  $\Omega \times [0, 1]$  [3]. Another measurability assumption, more difficult to state, is that of a *well-behaved class*  $\mathcal{C}$  [2]. Under either of those conditions, the statement (1) in Theorem 1 can be replaced with

- 1')  $\mathcal{C}$  is distribution-free consistently PAC learnable,

meaning that every consistent learning rule  $\mathcal{L}$  for  $\mathcal{C}$  is distribution-free probably approximately correct. In the proof, a measurability hypothesis on  $\mathcal{C}$  has to be invoked twice, in order to deduce implications (3)  $\Rightarrow$  (2) and (1'  $\Rightarrow$  (1).

Vladimir Pestov is with Departamento de Matemática, Universidade Federal de Santa Catarina, Campus Universitário Trindade, CEP 88.040-900 Florianópolis-SC, Brasil (CNPq Visiting Researcher) and the Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Ontario, K1N 6N5 Canada (permanent address, phone: 613-562-5800 ext. 3523, fax: 613-562-5776, email: vpest283@uottawa.ca).

In particular, Theorem 1 holds for every countable class  $\mathcal{C}$  or, more generally, for every *universally separable* class [4]. It is arguable that every concept class emerging in either theory or applications of statistical learning will be measurable in a sufficiently strong sense. For this reason, a measurability condition on  $\mathcal{C}$  is typically not even mentioned.

The fact remains that Theorem 1 cannot be derived in full generality. An example of a concept class  $\mathcal{C}$  of Vapnik–Chervonenkis (VC) dimension one which is not uniform Glivenko–Cantelli was constructed by Durst and Dudley [5], and a further modification of this example, also of VC dimension one, fails consistent PAC learnability [2].

This example has been constructed under Continuum Hypothesis (CH), which is arguably not a natural assumption in a probabilistic context [6]. However, the example remains valid under much more relaxed and natural set-theoretic hypothesis: Martin’s Axiom (MA). It is one of the most often used and best studied additional set-theoretic assumptions beyond the standard Zermelo–Fraenkel set theory with the Axiom of Choice (ZFC). In particular, Martin’s Axiom follows from the Continuum Hypothesis (CH), but it is also compatible with the negation of CH, and in fact it is namely the combination  $MA + \neg CH$  that is really interesting [7], [8], [9].

In this note we make the following, somewhat astonishing, observation: under the same assumption (Martin’s Axiom), the conditions (1) and (3) in Theorem 1 are equivalent. Here is our main result.

*Theorem 2:* Assume the validity of Martin’s Axiom (MA). Then the following are equivalent for every concept class  $\mathcal{C}$  consisting of universally measurable subsets of a Borel domain  $\Omega$ :

- 1)  $\mathcal{C}$  is distribution-free PAC learnable, and
- 2) the Vapnik–Chervonenkis dimension of  $\mathcal{C}$  is finite.

Of course it is only the implication (2) $\Rightarrow$ (1) that needs proving, because (1) $\Rightarrow$ (2) is a well-known classical result from [2] which does not require any assumptions on  $\mathcal{C}$ .

We review a precise formal setting for learnability, after which we proceed to analysis of a counter-example from [5], [2]. We observe that the concept class  $\mathcal{C}$  in the example is in fact PAC learnable, and this observation provides a clue to a general result.

The construction of the learning rule  $\mathcal{L}$  can be described as a “first in, first served” approach. The concept class  $\mathcal{C}$  is given a minimal well-ordering,  $\prec$ , and  $\mathcal{L}$  is constructed recursively, by assigning to a learning sample the  $\prec$ -smallest consistent concept  $C$  with regard to the the ordering. As a

consequence, for every concept  $C \in \mathcal{C}$ , the image of all learning samples of the form  $(\sigma, C \cap \sigma)$  under  $\mathcal{L}$  forms a uniform Glivenko–Cantelli class. It is for establishing this property of  $\mathcal{L}$  that we need Martin’s Axiom. Now the probable approximate correctness of  $\mathcal{L}$  is straightforward.

The present approach goes back to present author’s earlier work [10], but the results are new and have never been stated explicitly before.

## II. THE SETTING

For obvious reasons, we need to be quite precise when fixing a general setting for learnability. The *domain (instance space)*  $\Omega = (\Omega, \mathcal{A})$  is a *standard Borel space*, that is, a complete separable metric space equipped with the sigma-algebra of Borel subsets (the smallest family of sets containing all open balls and closed under complements and countable intersections).

*Measures* on  $\Omega$  mean *Borel probability measures*, that is, countably additive functions on  $\mathcal{A}$  with values in the unit interval  $[0, 1]$ , having the property  $\mu(\Omega) = 1$ . We will not distinguish between a measure  $\mu$  and its Lebesgue completion, that is, an extension of  $\mu$  over a larger sigma-algebra of Lebesgue  $\mu$ -measurable subsets of  $\Omega$ . Furthermore, recall that a subset  $A \subseteq \Omega$  is *universally measurable* if it is Lebesgue  $\mu$ -measurable for every probability measure  $\mu$  on  $\Omega$ .

With this caveat, a *concept class*,  $\mathcal{C}$ , is a family of universally measurable subsets of  $\Omega$ .

In the learning model, a set  $\mathcal{P}$  of probability measures on  $\Omega$  is fixed. Usually either  $\mathcal{P} = P(\Omega)$  is the set of all probability measures (distribution-free learning), or  $\mathcal{P} = \{\mu\}$  is a single measure (learning under fixed distribution). In our article, the case of interest is the former, although some of our results are valid in the case of a general family  $\mathcal{P} \subseteq P(\Omega)$ .

A *learning sample* is a pair  $(\sigma, \tau)$  of finite subsets of  $\Omega$ , where  $\tau \subseteq \sigma$  is thought of as the set of points belonging to an unknown concept,  $C$ . The set of all samples of size  $n$  is usually identified with  $(\Omega \times \{0, 1\})^n$ .

A *learning rule* (for  $\mathcal{C}$ ) is a mapping

$$\mathcal{L}: \bigcup_{n=1}^{\infty} \Omega^n \times \{0, 1\}^n \rightarrow \mathcal{C}$$

which satisfies the following measurability condition: for every  $C \in \mathcal{C}$ ,  $n \in \mathbb{N}$  and  $\mu \in \mathcal{P}$ , the function

$$\Omega^n \ni \sigma \mapsto \mu(\mathcal{L}(\sigma, C \cap \sigma) \triangle C) \in \mathbb{R} \quad (1)$$

is measurable.

A learning rule  $\mathcal{L}$  is *consistent* (with a concept class  $\mathcal{C}$ ) if for all  $C \in \mathcal{C}$ ,  $n \in \mathbb{N}$  and  $\sigma \in \Omega^n$  one has

$$\mathcal{L}(\sigma, C \cap \sigma) \cap \sigma = C \cap \sigma.$$

A learning rule  $\mathcal{L}$  is *probably approximately correct (PAC) under  $\mathcal{P}$*  if for every  $\epsilon > 0$

$$\mu^{\otimes n} \{ \sigma \in \Omega^n : \mu(\mathcal{L}(\sigma, C \cap \sigma) \triangle C) > \epsilon \} \rightarrow 0 \quad (2)$$

as  $n \rightarrow \infty$ , uniformly over all  $C \in \mathcal{C}$  and  $\mu \in \mathcal{P}$ . Here  $\mu^{\otimes n}$  denotes the product measure on  $\Omega^n$ .

In terms of sample complexity function  $s(\epsilon, \delta)$ , a learning rule  $\mathcal{L}$  is PAC if for each  $C \in \mathcal{C}$  and every  $\mu \in \mathcal{P}$  an independent identically distributed (i.i.d.) sample  $\sigma = (x_1, x_2, \dots, x_n)$  with  $n \geq s(\epsilon, \delta)$  points has the property  $\mu(C \triangle \mathcal{L}(\sigma, C \cap \sigma)) < \epsilon$  with confidence  $\geq 1 - \delta$ .

A concept class  $\mathcal{C}$  is *PAC learnable under  $\mathcal{P}$* , if there exists a PAC learning rule for  $\mathcal{C}$  under  $\mathcal{P}$ . A class  $\mathcal{C}$  is *consistently learnable* (under  $\mathcal{P}$ ) if every learning rule consistent with  $\mathcal{C}$  is PAC under  $\mathcal{P}$ . If  $\mathcal{P} = P(\Omega)$  is the set of all probability measures, then  $\mathcal{C}$  is said to be *distribution-free PAC learnable*. If  $\mathcal{P} = \{\mu\}$  is a single probability measure, one is talking of *learning under a single distribution*. Learnability under intermediate families  $\mathcal{P}$  is also receiving considerable attention, cf. Chapter 7 in [11].

Notice that in this paper, we only talk of *potential* PAC learnability, adopting a purely information-theoretic viewpoint. As a consequence, our statements about learning rules are existential rather than constructive, and building learning rules by transfinite recursion is perfectly acceptable.

A concept class  $\mathcal{C}$  is *uniform Glivenko–Cantelli with regard to a family of measures  $\mathcal{P}$* , if for each  $\epsilon > 0$

$$\sup_{\mu \in \mathcal{P}} \mu^{\otimes n} \left\{ \sup_{C \in \mathcal{C}} |\mu(C) - \mu_n(C)| \geq \epsilon \right\} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (3)$$

Here  $\mu_n$  stands for the empirical (uniform) measure on  $n$  points, sampled in an i.i.d. fashion from  $\Omega$  according to the distribution  $\mu$ . In this case, one also says that  $\mathcal{C}$  has the property of *uniform convergence of empirical measures (UCEM property)* (with regard to  $\mathcal{P}$ ) [11].

Every uniform Glivenko–Cantelli concept class (with regard to  $\mathcal{P}$ ) is consistently PAC learnable (under  $\mathcal{P}$ ), as is easy to verify. In the distribution-free situation ( $\mathcal{P} = P(\Omega)$ ) the converse holds under additional measurability conditions on the class mentioned in the Introduction, but, as we will see, not always.

More precisely, every distribution-free PAC learnable class has finite VC dimension (it was proved in [2], Theorem 2.1(i); see also e.g. [11], Lemma 7.2 on p. 279). Now the measurability conditions on  $\mathcal{C}$  assure that a class  $\mathcal{C}$  of finite VC dimension  $d$  is uniform Glivenko–Cantelli, with a sample complexity bound that does not depend on  $\mathcal{C}$ , but only on  $\epsilon$ ,  $\delta$ , and  $d$ . The following is a typical (and far from being optimal) such estimate, which can be deduced, for instance, along the lines of [12]:

$$s(\epsilon, \delta, d) \leq \frac{128}{\epsilon^2} \left( d \log \left( \frac{2e^2}{\epsilon} \log \frac{2e}{\epsilon} \right) + \log \frac{8}{\delta} \right). \quad (4)$$

For our purposes, we will fix any such bound and refer to it as a “*standard*” sample complexity estimate for  $s(\epsilon, \delta, d)$ .

Now the consistent learnability for  $\mathcal{C}$ , with the same sample complexity, follows. Of course in order to conclude that  $\mathcal{C}$  is PAC learnable, it is necessary to prove the existence of a consistent learning rule satisfying Eq. (1). This is usually being done using subtle measurable selection theorems using the same measurability assumptions on  $\mathcal{C}$  yet again.

Finally, recall that a subset  $N \subseteq \Omega$  is *universal null* if for every non-atomic probability measure  $\mu$  on  $(\Omega, \mathcal{A})$  one has

$\mu(N') = 0$  for some Borel set  $N'$  containing  $N$ . Universal null Borel sets are just countable sets.

### III. REVISITING AN EXAMPLE OF DURST AND DUDLEY

The proof of the implication (3) $\Rightarrow$ (2) in Theorem 1 depends in an essential way on the Fubini theorem, which is why some measurability restrictions on the class  $\mathcal{C}$  are unavoidable. Without them, the conclusion is not true in general. Here is a classical example of a concept class having finite VC dimension which is not uniform Glivenko–Cantelli.

*Example 3 (Durst and Dudley [5], Proposition 2.2):*

Assume the validity of the Continuum Hypothesis (CH). Let  $\Omega$  be an uncountable standard Borel space, that is, up to an isomorphism, a Borel space associated to the unit interval  $[0, 1]$ . The statement of CH is equivalent to the existence of a total order  $\prec$  on  $\Omega$  with the property that every half-open initial segment  $I_y = \{x \in \Omega: x \prec y\}$ ,  $y \in \Omega$  is countable, and  $\prec$  is a well-ordering: every non-empty subset of  $\Omega$  has the smallest element. Fix such an order.

Let  $\mathcal{C}$  consist of all half-open initial segments  $I_y$ ,  $y \in \Omega$  as above. Clearly, the VC dimension of the class  $\mathcal{C}$  is one.

Now let  $\mu$  be a non-atomic Borel probability measure on  $\Omega$  (e.g., the Lebesgue measure on  $[0, 1]$ ). Under CH, every element of  $\mathcal{C}$  is a countable set, therefore Borel measurable of measure zero. At the same time, for every  $n$  and each i.i.d. random  $n$ -sample  $\sigma$ , there is a countable initial segment  $C = I_y \in \mathcal{C}$  containing all elements of  $\sigma$ . The empirical measure of  $C$  with regard to  $\sigma$  is one. Thus, no finite sample guesses the measure of all elements of  $\mathcal{C}$  to within an accuracy  $\epsilon < 1$  with a non-vanishing confidence.

See also [13], p. 314; [3], pp. 170–171.

A further modification of this construction gives an example of a concept class of finite VC dimension which is not consistently PAC learnable.

*Example 4 (Blumer et al. [2], p. 953):* Again, assume CH. Add to the concept class  $\mathcal{C}$  from Example 3 the set  $\Omega$  as an element, forming a new concept class  $\mathcal{C}' = \mathcal{C} \cup \{\Omega\}$ . One still has  $\text{VC}(\mathcal{C}') = 1$ . For a finite labelled sample  $(\sigma, \tau)$  define

$$\mathcal{L}(\sigma, \tau) = I_z, \quad z = \min\{y \in (\Omega, \prec): \tau \subseteq I_y\}. \quad (5)$$

The learning rule  $\mathcal{L}$  is consistent with the class  $\mathcal{C}'$ . At the same time,  $\mathcal{L}$  is not probably approximately correct. Indeed, for the concept  $C = \Omega$  the value of the learning rule  $\mathcal{L}(\sigma, \Omega \cap \sigma) = \mathcal{L}(\sigma, \sigma)$  will always return a countable concept  $I_y$  for some  $y \in \Omega$ , and if  $\mu$  is a non-atomic Borel probability measure on  $\Omega$ , then  $\mu(C \Delta I_y) = 1$ . The concept  $C = \Omega$  cannot be learned to accuracy  $\epsilon < 1$  with a non-zero confidence.

*Remark 5:* It is important to note that — again, under CH — the class  $\mathcal{C}'$  is distribution-free PAC learnable.

Indeed, redefine a well-ordering on  $\mathcal{C} = \{I_x: x \in \Omega\} \cup \{\Omega\}$  by making  $\Omega$  the smallest element (instead of the largest one) and keeping the order relation between other elements the same. Denote the new order relation by  $\prec_1$ , and define a learning rule  $\mathcal{L}_1$  similarly to Eq. (5), but this time

understanding the minimum with regard to the well-ordering  $\prec_1$ :

$$\mathcal{L}_1(\sigma, \tau) = \min_{(\prec_1)} \left\{ C \in \mathcal{C}: C \cap \sigma = \bigcap_{\tau \subseteq D} D \right\}. \quad (6)$$

In essence,  $\mathcal{L}_1$  examines all the concepts following a transfinite order on them, and returns the first encountered concept consistent with the sample, provided it exists.

To see what difference it makes with Example 4, let  $\mu$  be again a non-atomic probability measure on  $\Omega$ . If  $C = \Omega$ , then for every sample  $\sigma$  consistently labelled with  $C$  the rule  $\mathcal{L}_1$  will return  $C$ , because this is the smallest consistent concept encountered by the algorithm. If  $C \neq \Omega$ , then for  $\mu$ -almost all samples  $\sigma$  the labelling on  $\sigma$  produced by  $C$  will be empty, and the concept  $\mathcal{L}_1(\sigma, \emptyset)$  returned by  $\mathcal{L}_1$ , while possibly different from  $C$ , will be again a countable concept, meaning that  $\mu(C \Delta \mathcal{L}(\sigma, \emptyset)) = 0$ .

To give a formal proof that  $\mathcal{L}_1$  is PAC, notice that for every  $C \in \mathcal{C}'$  and each  $n \in \mathbb{N}$  the collection of pairwise distinct concepts  $\mathcal{L}_1(\sigma \cap C)$ ,  $\sigma \in \Omega^n$  is only countable (under CH), because they are all contained in the  $\prec_1$ -initial segment of a minimally ordered set  $\mathcal{C}$  of cardinality continuum, bounded by  $C$  itself. As a consequence, the concept class

$$\mathcal{L}_1^C = \{\mathcal{L}_1(\sigma \cap C): \sigma \in \Omega^n, n \in \mathbb{N}\} \subseteq \mathcal{C}' \quad (7)$$

is also countable (assuming CH). The VC dimension of the family  $\mathcal{L}_1^C \cup \{C\}$  is  $\leq 1$ , and being countable, it is a uniform Glivenko–Cantelli class with a standard sample complexity as in Eq. (4). Consequently, given  $\epsilon, \delta > 0$ , and assuming that  $n$  is sufficiently large, one has for each probability measure  $\mu$  on  $\Omega$  and every  $\sigma \in \Omega^n$

$$\mu(C \Delta \mathcal{L}(\sigma, C \cap \sigma)) < \epsilon$$

provided  $n \geq s(\epsilon, \delta, 1)$ , as required.

*Remark 6:* Notice that the role of the Continuum Hypothesis in the above examples was merely to assure that every initial segment  $I_y$ ,  $y \in \Omega$  is a universally measurable set. As we will see, it can be achieved under a much milder assumption of Martin’s Axiom.

*Remark 7:* Thus, under the Continuum Hypothesis, the example of Dudley and Durst as modified by Blumer, Ehrenfeucht, Haussler, and Warmuth gives an example of a PAC learnable concept class which is not uniform Glivenko–Cantelli (even if having finite VC dimension). As it will become clear in the next Section, the assumption of CH can be weakened to Martin’s Axiom. Still, it would be interesting to know whether an example with the same combination of properties can be constructed without additional set-theoretic assumptions.

A basic observation of this Section is that in order for a learning rule  $\mathcal{L}$  to be PAC, the assumption on  $\mathcal{C}$  being uniform Glivenko–Cantelli can be weakened as follows.

*Lemma 8:* Let  $\mathcal{C}$  be a concept class and  $\mathcal{P}$  a family of probability measures on the domain  $\Omega$ . Suppose there exists a function  $s(\epsilon, \delta)$  and a learning rule  $\mathcal{L}$  for  $\mathcal{C}$  with the property

that for every  $C \in \mathcal{C}$ , the set  $\mathcal{L}^C \cup \{C\}$  is Glivenko–Cantelli with regard to  $\mathcal{P}$  with the sample complexity  $s(\epsilon, \delta)$ , where

$$\mathcal{L}^C = \{\mathcal{L}(C \cap \sigma) : \sigma \in \Omega^n, n \in \mathbb{N}\}.$$

Then  $\mathcal{L}$  is probably approximately correct under  $\mathcal{P}$  with sample complexity  $s(\epsilon, \delta)$ . ■

This simple fact becomes useful in combination with the technique of well-orderings. Of course the Continuum Hypothesis is a particularly unnatural assumption in a probabilistic context (cf. [6]). But it is unnecessary. Martin’s Axiom (MA) is a much weaker and natural additional set-theoretic axiom, which works just as well.

#### IV. LEARNABILITY UNDER MARTIN’S AXIOM

Martin’s Axiom (MA) says that no compact Hausdorff topological space with the countable chain condition is a union of strictly less than continuum nowhere dense subsets. Thus, it is a stronger statement than the Baire Category Theorem. In particular, the Continuum Hypothesis implies MA. However, MA is compatible with the negation of CH, and this is where the most interesting applications of MA are to be found. We need the following consequence of MA.

*Theorem 9 (Martin-Solovay):* Let  $(\Omega, \mu)$  be a standard Lebesgue non-atomic probability space. Under MA, the Lebesgue measure is  $2^{\aleph_0}$ -additive, that is, if  $\kappa < 2^{\aleph_0}$  and  $A_\alpha$ ,  $\alpha < \kappa$  is family of pairwise disjoint measurable sets, then  $\cup_{\alpha < \kappa} A_\alpha$  is Lebesgue measurable and

$$\mu \left( \bigcup_{\alpha < \kappa} A_\alpha \right) = \sum_{\alpha < \kappa} \mu(A_\alpha).$$

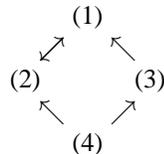
In particular, the union of strictly less than continuum null subsets of  $\Omega$  is a null subset. ■

For the proof and more on MA, see [9], Theorem 2.21, or [7], or [8], pp. 563–565.

*Lemma 10:* Let  $\mathcal{C}$  be a concept class and  $\mathcal{P}$  a family of probability measures on a standard Borel domain  $\Omega$ . Consider the following properties.

- 1) Every countable subclass of  $\mathcal{C}$  is uniform Glivenko–Cantelli with regard to  $\mathcal{P}$ .
- 2) There is a function  $s(\epsilon, \delta)$  so that every countable subclass of  $\mathcal{C}$  is uniform Glivenko–Cantelli with regard to  $\mathcal{P}$  with sample complexity  $s(\epsilon, \delta)$ .
- 3) Every subclass  $\mathcal{C}'$  of  $\mathcal{C}$  having cardinality  $< 2^{\aleph_0}$  is uniform Glivenko–Cantelli with regard to  $\mathcal{P}$ .
- 4) There is a function  $s(\epsilon, \delta)$  so that every subclass  $\mathcal{C}'$  of  $\mathcal{C}$  having cardinality  $< 2^{\aleph_0}$  is uniform Glivenko–Cantelli with regard to  $\mathcal{P}$  with sample complexity  $s(\epsilon, \delta)$ .

Then



Under Martin’s Axiom, all four conditions are equivalent.

*Proof:* The implications (2)  $\Rightarrow$  (1), (3)  $\Rightarrow$  (1), (4)  $\Rightarrow$  (2) and (4)  $\Rightarrow$  (3) are trivially true. To show (1)  $\Rightarrow$  (2), let  $\delta, \epsilon > 0$  be arbitrary but fixed. For each countable subclass  $\mathcal{C}'$ , choose the smallest value of sample complexity  $s = s(\mathcal{C}', \epsilon, \delta)$ . The function  $\mathcal{C}' \mapsto s(\mathcal{C}', \epsilon, \delta)$  is monotone under inclusions: if  $\mathcal{C}' \subseteq \mathcal{C}''$ , then  $s(\mathcal{C}', \epsilon, \delta) \leq s(\mathcal{C}'', \epsilon, \delta)$ . If  $\mathcal{C}'_n$  is a sequence of countable classes, then the union  $\cup_{n=1}^{\infty} \mathcal{C}'_n$  is a countable class, whose sample complexity value bounds from above  $s(\mathcal{C}'_n, \epsilon, \delta)$ ,  $n = 1, 2, \dots$ . Thus, the function  $\mathcal{C}' \mapsto s(\mathcal{C}', \epsilon, \delta)$  for  $\delta, \epsilon > 0$  fixed is bounded on countable sets of inputs, and therefore bounded.

Now assume (MA). It is enough to prove (2)  $\Rightarrow$  (4). This is done by a transfinite induction on the cardinality  $\kappa = |\mathcal{C}'|$ , which never exceeds  $2^{\aleph_0}$  because  $\mathcal{C}'$  consists of Borel subsets of a standard Borel domain. For  $\kappa = \aleph_0$  there is nothing to prove. Else, represent  $\mathcal{C}$  as a union of an increasing transfinite chain of concept classes  $\mathcal{C}_\alpha$ ,  $\alpha < \kappa$ , for each of which the statement of (4) holds. For every  $\epsilon > 0$  and  $n \in \mathbb{N}$ , the set

$$\begin{aligned} & \{\sigma \in \Omega^n : \sup_{C \in \mathcal{C}} |\mu_n(\sigma)(C) - \mu(C)| < \epsilon\} \\ &= \bigcap_{\alpha < \kappa} \{\sigma \in \Omega^n : \sup_{C \in \mathcal{C}_\alpha} |\mu_n(\sigma)(C) - \mu(C)| < \epsilon\} \end{aligned}$$

is measurable by Martin-Solovay’s Theorem 9. Given  $\delta > 0$  and  $n \geq s(\epsilon, \delta, d)$ , another application of the same result leads to conclude that for every  $\mu \in P(\Omega)$ :

$$\begin{aligned} & \mu^{\otimes n} \left\{ \sigma \in \Omega^n : \sup_{C \in \mathcal{C}} |\mu_n(\sigma)(C) - \mu(C)| < \epsilon \right\} \\ &= \mu^{\otimes n} \left( \bigcap_{\alpha < \kappa} \left\{ \sigma \in \Omega^n : \sup_{C \in \mathcal{C}_\alpha} |\mu_n(\sigma)(C) - \mu(C)| < \epsilon \right\} \right) \\ &= \inf_{\alpha < \kappa} \mu^{\otimes n} \left\{ \sigma \in \Omega^n : \sup_{C \in \mathcal{C}_\alpha} |\mu_n(\sigma)(C) - \mu(C)| < \epsilon \right\} \\ &\geq 1 - \delta, \end{aligned}$$

as required. ■

*Lemma 11:* Let  $\mathcal{C}$  be a concept class whose countable subclasses are uniform Glivenko–Cantelli with regard to a family of probability measures  $\mathcal{P}$ . Let  $\mathcal{L}$  be a learning rule for  $\mathcal{C}$  with the property that for every  $C \in \mathcal{C}$ , the set

$$\mathcal{L}^{C,n} = \{\mathcal{L}(C \cap \sigma) : \sigma \in \Omega^n\} \quad (8)$$

has cardinality strictly less than continuum. Under Martin’s Axiom, the rule  $\mathcal{L}$  is probably approximately correct under  $\mathcal{P}$ . The common sample complexity bound of countable subclasses of  $\mathcal{C}$  becomes the sample complexity bound for the learning rule  $\mathcal{L}$ .

*Proof:* Recall that  $2^{\aleph_0}$  is a regular cardinal, and thus admits no countable cofinal subset. Therefore, under the assumptions of Lemma, the cardinality of  $\mathcal{L}^C = \cup_{n=1}^{\infty} \mathcal{L}^{C,n}$  is still strictly less than continuum. Applying now Lemma 10 and then Lemma 8, we conclude. ■

The following result establishes existence of learning rules with the required property.

*Lemma 12:* Let  $\mathcal{C}$  be an infinite concept class on a measurable space  $\Omega$ . Denote  $\kappa = |\mathcal{C}|$  the cardinality of  $\mathcal{C}$ . There exists a consistent learning rule  $\mathcal{L}$  for  $\mathcal{C}$  with the

property that for every  $C \in \mathcal{C}$  and each  $n$ , the set  $\mathcal{L}^{C,n}$  (cf. Eq. (8)) has cardinality  $< \kappa$ . Under MA the rule  $\mathcal{L}$  satisfies the condition in Eq. (1).

*Proof:* Choose a minimal well-ordering of elements of  $\mathcal{C}$ :

$$\mathcal{C} = \{C_\alpha : \alpha < \kappa\},$$

and set for every  $\sigma \in \Omega^n$  and  $\tau \in \{0, 1\}^n$  the value  $\mathcal{L}(\sigma, \tau)$  equal to  $C_\beta$ , where

$$\beta = \min\{\alpha < \kappa : C_\alpha \cap \sigma = \tau\},$$

provided such a  $\beta$  exists. Clearly, for each  $\alpha < \kappa$  one has

$$\mathcal{L}(\sigma, C_\alpha \cap \sigma) \in \{C_\beta : \beta \leq \alpha\},$$

which assures (8). Besides, the learning rule  $\mathcal{L}$  is consistent.

Fix  $C = C_\alpha \in \mathcal{C}$ ,  $\alpha < \kappa$ . For every  $\beta \leq \alpha$  define  $D_\beta = \{\sigma \in \Omega^n : C \cap \sigma = C_\beta \cap \sigma\}$ . The sets  $D_\beta$  are measurable, and the function

$$\Omega^n \ni \sigma \mapsto \mu(\mathcal{L}(C \cap \sigma) \triangle C) \in \mathbb{R}$$

takes a constant value  $\mu(C \triangle C_\alpha)$  on each set  $D_\beta \setminus \bigcup_{\gamma < \beta} D_\gamma$ ,  $\beta \leq \alpha$ . Such sets, as well as all their possible unions, are measurable under MA by force of Martin–Solovay’s Theorem 9, and their union is  $\Omega^n$ . This implies the validity of Eq. (1) for  $\mathcal{L}$ . ■

Lemma 11 and Lemma 12 lead to the following result.

*Theorem 13 (Assuming MA):* Let  $\mathcal{C}$  be a concept class consisting of Borel measurable subsets of a standard Borel domain  $\Omega$ , and let  $\mathcal{P}$  be a family of probability measures on  $\Omega$ . Suppose that every countable subclass of  $\mathcal{C}$  is uniform Glivenko–Cantelli with regard to  $\mathcal{P}$ . Then the concept class  $\mathcal{C}$  is PAC learnable under  $\mathcal{P}$ . In addition, there exists a common sample complexity bound for countable subclasses of  $\mathcal{C}$ , and any such bound gives a sample complexity bound for PAC learnability of  $\mathcal{C}$ . ■

Finally, we can deduce our main result.

*Proof of (2)⇒(1) in Theorem 2*

The implication follows from Theorem 13 with  $\mathcal{P} = P(\Omega)$  and the common complexity bound (4). ■

## V. CONCLUSION

As a footnote to the fundamental theorem of statistical learning, we have proved that in the presence of a mild set-theoretic axiom (Martin’s Axiom), PAC learnability of a concept class  $\mathcal{C}$  is equivalent to finiteness of VC dimension of  $\mathcal{C}$ , without any extra assumptions on the measurability of the class  $\mathcal{C}$ . The price to pay is giving up consistent PAC learnability, as well as constructive choice of a learning rule.

It would be interesting to know to what extent the results remain true in the usual ZFC model of set theory. In particular, can an example of a concept class  $\mathcal{C}$  on a standard Borel domain which has finite VC dimension and still is not consistently PAC learnable, be constructed without any additional set-theoretic axioms?

## REFERENCES

- [1] V.N. Vapnik and A.Ya. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory Probab. Appl. **16**, issue 2 (1971), 264–280.
- [2] A. Blumer, A. Ehrenfeucht, D. Haussler and M.K. Warmuth, *Learnability and the Vapnik-Chervonenkis dimension*, Journal of the ACM, **36**(4) (1989), 929–865.
- [3] R.M. Dudley, *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, **63**, Cambridge University Press, Cambridge, 1999.
- [4] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.
- [5] M. Durst and R.M. Dudley, *Empirical processes, Vapnik–Chervonenkis classes, and Poisson processes*, Prob. and Math. Statistics **1** (1980), 109–115.
- [6] C. Freiling, *Axioms of symmetry: throwing darts at the real number line*, J. Symbolic Logic **51** (1986), 190–200.
- [7] D.H. Fremlin, *Consequences of Martin’s Axiom*, Cambridge Tracts in Mathematics, **84**, Cambridge University Press, Cambridge, 1984.
- [8] T. Jech, *Set Theory*, Academic Press, New York–London, 1978.
- [9] K. Kunen, *Set Theory*, North-Holland, Amsterdam, 1980.
- [10] V. Pestov, *PAC learnability of a concept class under non-atomic measures: a problem by Vidyasagar*, in: Proc. 21st Intern. Conference on Algorithmic Learning Theory (ALT’2010), Lect. Notes in Artificial Intelligence **6331**, Springer, 2010, pp. 134–147.
- [11] M. Vidyasagar, *Learning and Generalization, with Applications to Neural Networks*, 2nd Ed., Springer-Verlag, 2003.
- [12] S. Mendelson, *A few notes on statistical learning theory*, in: S. Mendelson, A.J. Smola, Eds., *Advanced Lectures in Machine Learning*, LNCS 2600, Springer, 2003, pp. 1–40.
- [13] R.S. Wenokur and P.M. Dudley, *Some special Vapnik–Chervonenkis classes*, Discrete Math. **33** (1981), 313–318.