PAC Learnability of a Concept Class under Non-atomic Measures: A Problem by Vidyasagar

Vladimir Pestov

Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Ontario, Canada K1N 6N5

Abstract. In response to a 1997 problem of M. Vidyasagar, we state a necessary and sufficient condition for distribution-free PAC learnability of a concept class $\mathscr{C}$ under the family of all non-atomic (diffuse) measures on the domain Ω. Clearly, finiteness of the classical Vapnik–Chervonenkis dimension of $\mathscr{C}$ is a sufficient, but no longer necessary, condition. Besides, learnability of $\mathscr{C}$ under non-atomic measures does not imply the uniform Glivenko–Cantelli property with regard to non-atomic measures. Our learnability criterion is stated in terms of a combinatorial parameter $\mathrm{VC}(\mathscr{C} \bmod \omega_1)$ which we call the VC dimension of $\mathscr{C}$ modulo countable sets. The new parameter is obtained by "thickening up" single points in the definition of VC dimension to uncountable "clusters". Equivalently, $\mathrm{VC}(\mathscr{C} \bmod \omega_1) \leq d$ if and only if every countable subclass of $\mathscr{C}$ has VC dimension $\leq d$ outside a countable subset of Ω. The new parameter can be also expressed as the classical VC dimension of $\mathscr{C}$ calculated on a suitable subset of a compactification of Ω. We do not make any measurability assumptions on $\mathscr{C}$, assuming instead the validity of Martin's Axiom (MA).

1 Introduction

A fundamental result of statistical learning theory says that for a concept class $\mathscr{C}$ the three conditions are equivalent: (1) $\mathscr{C}$ is distribution-free PAC learnable over the family $P(\Omega)$ of all probability measures on the domain Ω, (2) $\mathscr{C}$ is a uniform Glivenko–Cantelli class with regard to $P(\Omega)$, and (3) the Vapnik–Chervonenkis dimension of $\mathscr{C}$ is finite [VC, BEHW]. In this paper we are interested in the problem, discussed by Vidyasagar in both editions of his book [V1, V2] as problem 12.8, of giving a similar combinatorial description of concept classes $\mathscr{C}$ which are PAC learnable under the family $P_{na}(\Omega)$ of all non-atomic probability measures on Ω. (A measure μ is *non-atomic*, or *diffuse,* if every set A of strictly positive measure contains a subset B with $0 < \mu(B) < \mu(A)$.)

The condition $\mathrm{VC}(\mathscr{C}) < \infty$, while of course sufficient for $\mathscr{C}$ to be learnable under $P_{na}(\Omega)$, is not necessary. Let a concept class $\mathscr{C}$ consist of all finite and all cofinite subsets of a standard Borel space Ω. Then $\mathrm{VC}(\mathscr{C}) = \infty$, and moreover $\mathscr{C}$ is clearly not a uniform Glivenko-Cantelli class *with regard to non-atomic measures.* At the same time, $\mathscr{C}$ is PAC learnable under non-atomic measures: any learning rule $\mathcal{L}$ consistent with the subclass $\{\emptyset, \Omega\}$ will learn $\mathscr{C}$. Notice that $\mathscr{C}$ is not *consistently* learnable under non-atomic measures: there are consistent learning rules mapping every training sample to a finite set, and they will not learn any cofinite subset of Ω.

The point of this example is that PAC learnability of a concept class $\mathscr{C}$ under non-atomic measures is not affected by adding to $\mathscr{C}$ symmetric differences $C \bigtriangleup N$ for each $C \in \mathscr{C}$ and every countable set N.

A version of VC dimension oblivious to this kind of set-theoretic "noise" is obtained from the classical definition by "thickening up" individual points and replacing them with uncountable clusters (Figure 1).
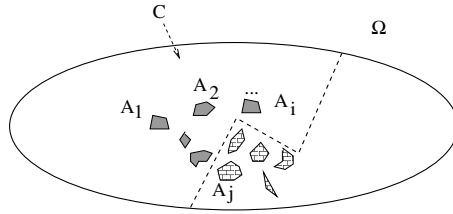

Fig. 1. A family $A_1, A_2, \ldots, A_n$ of uncountable sets shattered by $\mathscr{C}$

Define the *VC dimension of a concept class $\mathscr{C}$ modulo countable sets* as the supremum of natural n for which there exists a family of n uncountable sets, $A_1, A_2, \ldots, A_n \subseteq \Omega$, shattered by $\mathscr{C}$ in the sense that for each $J \subseteq \{1, 2, \ldots, n\}$, there is $C \in \mathscr{C}$ which contains all sets A_i, $i \in J$, and is disjoint from all sets A_j, $j \notin J$. Denote this parameter by $\text{VC}(\mathscr{C} \bmod \omega_1)$. Clearly, for every concept class $\mathscr{C}$

$$\text{VC}(\mathscr{C} \bmod \omega_1) \leq \text{VC}(\mathscr{C}).$$

In our example above, one has $\text{VC}(\mathscr{C} \bmod \omega_1) = 1$, even as $\text{VC}(\mathscr{C}) = \infty$.

Here is our main result.

Theorem 1. *Let $(\Omega, \mathscr{A})$ be a standard Borel space, and let $\mathscr{C} \subseteq \mathscr{A}$ be a concept class. Under the Martin's Axiom (MA), the following are equivalent.*

1. *$\mathscr{C}$ is PAC learnable under the family of all non-atomic measures.*
2. *$VC(\mathscr{C} \bmod \omega_1) = d < \infty$.*
3. *Every countable subclass $\mathscr{C}' \subseteq \mathscr{C}$ has finite VC dimension on the complement to some countable subset of Ω (which depends on $\mathscr{C}'$).*
4. *There is d such that for every countable $\mathscr{C}' \subseteq \mathscr{C}$ one has $VC(\mathscr{C}') \leq d$ on the complement to some countable subset of Ω (depending on $\mathscr{C}'$).*
5. *Every countable subclass $\mathscr{C}' \subseteq \mathscr{C}$ is a uniform Glivenko–Cantelli class with regard to the family of non-atomic measures.*
6. *Same, with sample complexity $s(\epsilon, \delta)$ which only depends on $\mathscr{C}$ and not on $\mathscr{C}'$.*

If $\mathscr{C}$ is universally separable [P], the above are also equivalent to:

7. *VC dimension of $\mathscr{C}$ is finite outside of a countable subset of Ω.*
8. *$\mathscr{C}$ is a uniform Glivenko-Cantelli class with respect to the family of non-atomic probability measures.*

Martin's Axiom (MA) [F] is one of the most often used and best studied additional set-theoretic assumptions beyond the standard Zermelo-Frenkel 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+¬CH that is really interesting.

The concept class in our initial simple example (which is even image admissible Souslin [D]) shows that in general (7) and (8) are not equivalent to the remaining conditions. Notice that for universally separable classes, (1), (7) and (8) are equivalent without additional set-theoretic assumptions.

The core of the theorem — and the main technical novelty of our paper — is the proof of the implication (3)$\Rightarrow$(1). It is based on a special choice of a consistent learning rule $\mathcal{L}$ having the property that for every concept $C \in \mathscr{C}$, 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.

Most of the remaining implications are relavely straightforward adaptations of the standard techniques of statistical learning. Nevertheless, (2)$\Rightarrow$(3) requires a certain technical dexterity, and we study this implication in the setting of Boolean algebras.

We begin the paper by reviewing a general formal setting, followed by a dicussion of Boolean algebras which seem like a natural framework for the problem at hand, especially in view of possible generalizations to learning under other intermediate families of measures.

In particular, we will show that our version of the VC dimension modulo countable sets, VC($\mathscr{C}$ mod ω_1), is just the usual VC dimension of the class $\mathscr{C}$ of concepts extended over a suitable compactification of Ω and restricted to a certain subdomain of the compactification.

Now the part of Theorem 1 for universally separable concept classes follows easily. Afterwards, we discuss Martin's Axiom, prove the existence of a learning rule with the above special property, and deduce Theorem 1 for arbitrary concept classes.

2   The Setting

We need to fix a precise setting, which is mostly standard. The *domain* (*instance space*) $\Omega = (\Omega, \mathscr{A})$ is a *measurable space,* that is, a set Ω equipped with a sigma-algebra of subsets $\mathscr{A}$. Typically, Ω is assumed to be a *standard Borel space,* that is, a complete separable metric space equipped with the sigma-algebra of Borel subsets. We will clarify the assumption whenever necessary.

A *concept class* is a family, $\mathscr{C}$, of measurable subsets of Ω. (Equivalently, $\mathscr{C}$ can be viewed as a family of measurable $\{0, 1\}$-valued functions on Ω.).

In the learning model, a set $\mathcal{P}$ of probability measures on Ω 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 family $\mathcal{P} = P_{na}(\Omega)$ of all non-atomic measures.

Every probability measure μ on Ω defines a distance d_μ on $\mathscr{A}$ as follows:

$$d_\mu(A, B) = \mu\,(A \bigtriangleup B)\,.$$

We will not distinguish between a measure μ and its Lebesgue completion, that is, an extension of μ over the larger sigma-algebra of Lebesgue measurable subsets of Ω. Consequently, we will sometimes use the term *measurability* meaning *Lebesgue measurability*. No confusion can arise here.

Often it is convenient to approximate the concepts from $\mathscr{C}$ with elements of the *hypothesis space, $\mathscr{H}$*, which is, technically, a subfamily of $\mathscr{A}$ whose closure with regard to each (pseudo)metric d_μ, $\mu \in \mathcal{P}$, contains $\mathscr{C}$. However, in our article we make no distinction between $\mathscr{H}$ and $\mathscr{C}$.

A *learning sample* is a pair $s = (\sigma, \tau)$ of finite subsets of Ω, where $\tau \subseteq \sigma$. It is convenient to assume that elements $x_1, x_2, \ldots, x_n \in \sigma$ are ordered, and thus the set of all samples (σ, τ) with $|\sigma| = n$ can be identified with $(\Omega \times \{0, 1\})^n$. A *learning rule* (for $\mathscr{C}$) is a mapping

$$\mathcal{L} \colon \bigcup_{n=1}^{\infty} \Omega^n \times \{0, 1\}^n \to \mathscr{C}$$

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

$$\Omega \ni \sigma \mapsto \mu\left(\mathcal{L}(\sigma, C \cap \sigma) \triangle C\right) \in \mathbb{R} \tag{1}$$

is measurable.

A learning rule $\mathcal{L}$ is *consistent* (with $\mathscr{C}$) if for every $C \in \mathscr{C}$ and each $\sigma \in \Omega^n$ 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 $\epsilon > 0$

$$\sup_{\mu \in \mathcal{P}} \sup_{C \in \mathscr{C}} \mu^{\otimes n} \{\sigma \in \Omega^n : \mu\left(\mathcal{L}(\sigma, C \cap \sigma) \triangle C\right) > \epsilon\} \to 0 \text{ as } n \to \infty. \tag{2}$$

Here $\mu^{\otimes n}$ denotes the (Lebesgue extension of the) product measure on Ω^n. Now the origin of the measurability condition (1) on the mapping $\mathcal{L}$ is clear: it is implicit in (2).

Equivalently, there is a function $s(\epsilon, \delta)$ (*sample complexity of $\mathcal{L}$*) such that for each $C \in \mathscr{C}$ and every $\mu \in \mathcal{P}$ an i.i.d. sample σ with $\geq s(\epsilon, \delta)$ points has the property $\mu(C \triangle \mathcal{L}(\sigma, C \cap \sigma)) < \epsilon$ with confidence $\geq 1 - \delta$.

A concept class $\mathscr{C}$ consisting of measurable sets is *PAC learnable under* $\mathcal{P}$, if there exists a PAC learning rule for $\mathscr{C}$ under $\mathcal{P}$. A class $\mathscr{C}$ is *consistently learnable* (under $\mathcal{P}$) if every learning rule consistent with $\mathscr{C}$ is PAC under $\mathcal{P}$. If $\mathcal{P} = P(\Omega)$ is the set of all probability measures, then $\mathscr{C}$ is said to be (distribution-free) *PAC learnable*. At the same time, learnability under intermediate families of measures on Ω has received considerable attention, cf. Chapter 7 in [V2].

Notice that in this paper, we only talk of *potential* PAC learnability, adopting a purely information-theoretic viewpoint.

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

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

(Cf. [D], Ch. 3; [M].) Here μ_n stands for the empirical (uniform) measure on n points, sampled in an i.i.d. fashion from Ω according to the distribution μ. One also says that $\mathscr{C}$ has the property of *uniform convergence of empirical measures* (*UCEM* property) *with regard to* $\mathcal{P}$ [V2].

Every uniform Glivenko–Cantelli class (with regard to $\mathcal{P}$) is PAC learnable (under $\mathcal{P}$), and in the distribution-free situation, the converse is true as well. Already in the case of learning under a single measure, it is not so: a PAC learnable class under a single distribution μ need not be uniform Glivenko-Cantelli with regard to μ (cf. Chapter 6 in [V2]). Not every PAC learnable class under non-atomic measures is uniform Glivenko–Cantelli with regard to non-atomic measures either: the class consisting of all finite and all cofinite subsets of Ω is a counter-example.

We say, following Pollard [P], that a concept class $\mathscr{C}$ consisting of measurable sets is *universally separable* if it contains a countable subfamily $\mathscr{C}'$ with the property that every $C \in \mathscr{C}$ is a pointwise limit of a suitable sequence $(C_n)_{n=1}^{\infty}$ of sets from $\mathscr{C}'$: for every $x \in \Omega$ there is N with the property that, for all $n \geq N$, $x \in C_n$ if $x \in C$, and $x \notin C_n$ if $x \in C$. Such a family $\mathscr{C}'$ is said to be *universally dense* in $\mathscr{C}$.

Probably the main source of uniform Glivenko–Cantelli classes is the finiteness of VC dimension. Assume that $\mathscr{C}$ satisfies a suitable measurability condition, for instance, $\mathscr{C}$ is image admissible Souslin, or else universally separable. (In particular, a countable $\mathscr{C}$ satisfies either condition.) If $\mathrm{VC}(\mathscr{C}) = d < \infty$, then $\mathscr{C}$ is uniform Glivenko–Cantelli, with a sample complexity bound that does not depend on $\mathscr{C}$, but only on ϵ, δ, and d. The following is a typical (and far from being optimal) such estimate, which can be deduced, for instance, along the lines of [M]:

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

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

A subset $N \subseteq \Omega$ is *universal null* if for every non-atomic probability measure μ on $(\Omega, \mathscr{A})$ one has $\mu(N') = 0$ for some Borel set N' containing N. Universal null Borel sets are just countable sets.

3   VC Dimension and Boolean Algebras

Recall that a *Boolean algebra*, $B = \langle B, \wedge, \vee, \neg, 0, 1 \rangle$, consists of a set, B, equipped with two associative and commutative binary operations, $\wedge$ ("meet") and $\vee$ ("join"), which are distributive over each other and satisfy the absorption principles $a \vee (a \wedge b) = a$, $a \wedge (a \vee b) = a$, as well as a unary operation $\neg$ (complement), and two elements 0 and 1, satisfying $a \vee \neg a = 1$, $a \wedge \neg a = 0$.

For instance, the family 2^{Ω} of all subsets of a set Ω, with the union as join, intersection as meet, the empty set as 0 and Ω as 1, as well as the set-theoretic complement $\neg A = A^c$, forms a Boolean algebra. In fact, every Boolean algebra can be realized as an algebra of subsets of a suitable Ω. Even better, according to the Stone representation theorem, a Boolean algebra B is isomorphic to the Boolean algebra formed by all

open-and-closed subsets of a suitable compact space, $S(B)$, called the *Stone space* of B, where the Boolean algebra operations are interpreted set-theoretically as above.

The space $S(B)$ can be obtained in different ways. For instance, one can think of elements of $S(B)$ as Boolean algebra homomorphisms from B to the two-element Boolean algebra $\{0, 1\}$ (the algebra of subsets of a singleton). In this way, $S(B)$ is a closed topological subspace of the compact zero-dimensional space $\{0, 1\}^B$ with the usual Tychonoff product topology.

The Stone space of the Boolean algebra $B = 2^{\Omega}$ is known as the *Stone-Čech compactification of* Ω, and is denoted $\beta\Omega$. The elements of $\beta\Omega$ are *ultrafilters* on Ω. A collection ξ of non-empty subsets of Ω is an ultrafilter if it is closed under finite intersections and if for every subset $A \subseteq \Omega$ either $A \in \xi$ or $A^c \in \xi$. To every point $x \in \Omega$ there corresponds a *trivial* (*principal*) *ultrafilter,* $\bar{x}$, consisting of all sets A containing x. However, if Ω is infinite, the Axiom of Choice assures that there exist non-principal ultrafilters on Ω. Basic open sets in the space $\beta\Omega$ are of the form $\bar{A} = \{\zeta \in \beta\Omega \colon A \in \zeta\}$, where $A \subseteq \Omega$. It is interesting to note that each $\bar{A}$ is at the same time closed, and in fact $\bar{A}$ is the closure of A in $\beta\Omega$. Moreover, every open and closed subset of $\beta\Omega$ is of the form $\bar{A}$.

A one-to-one correspondence between ultrafilters on Ω and Boolean algebra homomorphisms $2^{\Omega} \to \{0, 1\}$ is this: think of an ultrafilter ξ on Ω as its own indicator function χ_ξ on 2^{Ω}, sending $A \subseteq \Omega$ to 1 if and only if $A \in \xi$. It is not difficult to verify that χ_ξ is a Boolean algebra homomorphism, and that every homomorphism arises in this way.

The book [Jo] is a standard reference to the above topics.

Given a subset $\mathscr{C}$ of a Boolean algebra B, and a subset X of the Stone space $S(B)$, one can regard $\mathscr{C}$ as a set of binary functions restricted to X, and compute the VC dimension of $\mathscr{C}$ over X. We will denote this parameter $\mathrm{VC}(\mathscr{C} \restriction X)$.

A subset I of a Boolean algebra B is an *ideal* if, whenever $x, y \in I$ and $a \in B$, one has $x \vee y \in I$ and $a \wedge x \in I$. Define a *symmetric difference* on B by the formula $x \bigtriangleup y = (x \vee y) \vee \neg(x \wedge y)$. The *quotient Boolean algebra* B/I consists of all equivalence classes modulo the equivalence relation $x \sim y \iff x \bigtriangleup y \in I$. It can be easily verified to be a Boolean algebra on its own, with operations induced from B in a unique way.

The Stone space of B/I can be identified with a compact topological subspace of $S(B)$, consisting of all homomorphisms $B \to \{0, 1\}$ whose kernel contains I. For instance, if $B = 2^{\Omega}$ and I is an ideal of subsets of Ω, then the Stone space of $2^{\Omega}/I$ is easily seen to consist of all ultrafilters on Ω which do not contain sets from I.

Theorem 2. *Let $\mathscr{C}$ be a concept class on a domain Ω, and let I be an ideal of sets on Ω. The following conditions are equivalent.*

1. *The VC dimension of the (family of closures of the) concept class $\mathscr{C}$ restricted to the Stone space of the quotient algebra $2^{\Omega}/I$ is at least n: $\mathrm{VC}(\mathscr{C} \restriction S(2^{\Omega}/I)) \geq n$.*
2. *There exists a family $A_1, A_2, \ldots, A_n$ of measurable subsets of Ω not belonging to I, which is shattered by $\mathscr{C}$ in the sense that if $J \subseteq \{1, 2, \ldots, n\}$, then there is $C \in \mathscr{C}$ which contains all sets A_i, $i \in J$, and is disjoint from all sets A_i, $i \notin J$.*

Proof. $(1) \Rightarrow (2)$. Choose ultrafilters $\xi_1, \ldots, \xi_n$ in the Stone space of the Boolean algebra $2^{\Omega}/I$, whose collection is shattered by $\mathscr{C}$. For every $J \subseteq \{1, 2, \ldots, n\}$, select $C_J \in \mathscr{C}$

which carves the subset $\{\xi_i \colon i \in J\}$ out of $\{\xi_1, \ldots, \xi_n\}$. This means $C_J \in \xi_i$ if and only if $i \in J$. For all $i = 1, 2, \ldots, n$, set

$$A_i = \bigcap_{J \ni i} C_J \bigcap \bigcap_{J \not\ni i} C_J^c. \tag{5}$$

Then $A_i \in \xi_i$ and hence $A_i \notin I$. Furthermore, if $i \in J$, then clearly $A_i \subseteq C_J$, and if $i \notin J$, then $A_i \cap C_J = \emptyset$. The sets A_i are measurable by their definition.

$(2) \Rightarrow (1)$. Let $A_1, A_2, \ldots, A_n$ be a family of subsets of Ω not belonging to the set ideal I and shattered by $\mathscr{C}$ in sense of the lemma. For every i, the family of sets of the form $A_i \cap B^c$, $B \in I$ is a filter and so is contained in some free ultrafilter ξ_i, which is clearly disjoint from I and contains A_i. If $J \subseteq \{1, 2, \ldots, n\}$ and $C_J \in \mathscr{C}$ contains all sets A_i, $i \in J$ and is disjoint from all sets A_i, $i \notin J$, then the closure $\bar{C}_J$ of C_J in the Stone space contains ξ_i if and only if $i \in J$. We conclude: the collection of ultrafilters ξ_i, $i = 1, 2, \ldots, n$, which are all contained in the Stone space of $2^\Omega / I$, is shattered by the closed sets $\bar{C}_J$.

It follows in particular that the VC dimension of a concept class does not change if the domain Ω is compactified.

Corollary 1. $VC(\mathscr{C} \restriction \Omega) = VC(\mathscr{C} \restriction \beta\Omega)$.

Proof. The inequality $VC(\mathscr{C} \restriction \Omega) \le VC(\mathscr{C} \restriction \beta\Omega)$ is trivial. To establish the converse, assume there is a subset of $\beta\Omega$ of cardinality n shattered by $\mathscr{C}$. Choose sets A_i as in Theorem 2,(2). Clearly, any subset of Ω meeting each A_i at exactly one point is shattered by $\mathscr{C}$.

Definition 1. *Given a concept class $\mathscr{C}$ on a domain Ω and an ideal I of subsets of Ω, we define the VC dimension of $\mathscr{C}$ modulo I,*

$$VC(\mathscr{C} \bmod I) = VC(\mathscr{C} \restriction S(2^\Omega / I)).$$

That is, $VC(\mathscr{C} \bmod I) \ge n$ if and only if any of the equivalent conditions of Theorem 2 are met.

Definition 2. *Let $\mathscr{C}$ be a concept class on a domain Ω. If I is the ideal of all countable subsets of Ω, we denote the $VC(\mathscr{C} \bmod I)$ by $VC(\mathscr{C} \bmod \omega_1)$ and call it the* VC dimension modulo countable sets.

4 Finiteness of VC Dimension Modulo Countable Sets Is Necessary for Learnability

Lemma 1. *Every uncountable Borel subset of a standard Borel space supports a non-atomic Borel probability measure.*

Proof. Let A be an uncountable Borel subset of a standard Borel space Ω, that is, Ω is a Polish space equipped with its Borel structure. According to Souslin's theorem (see e.g. Theorem 3.2.1 in [A]), there exists a Polish (complete separable metric) space X

and a continuous one-to-one mapping $f : X \to A$. The Polish space X must be therefore uncountable, and so supports a diffuse probability measure, ν. The direct image measure $f_*\nu = \nu(f^{-1}(B))$ on Ω is a Borel probability measure supported on A, and it is diffuse because the inverse image of every singleton is a singleton in X and thus has measure zero.

The following result makes no measurability assumptions on the concept class.

Theorem 3. *Let $\mathscr{C}$ be a concept class on a domain $(\Omega, \mathscr{B})$ which is a standard Borel space. If $\mathscr{C}$ is PAC learnable under non-atomic measures, then the VC dimension of $\mathscr{C}$ modulo countable sets is finite.*

Proof. This is just a minor variation of a classical result for distribution-free PAC learnability (Theorem 2.1(i) in [BEHW]; we will follow the proof as presented in [V2], Lemma 7.2 on p. 279).

Suppose $\mathrm{VC}(\mathscr{C} \bmod \omega_1) \geq d$. According to Theorem 2, there is a family of uncountable Borel sets A_i, $i = 1, 2, \ldots, d$, shattered by $\mathscr{C}$ in our sense. Using Lemma 1, select for every $i = 1, 2, \ldots, d$ a non-atomic probability measure μ_i supported on A_i, and let $\mu = \frac{1}{d} \sum_{i=1}^{d} \mu_i$. This μ is a non-atomic Borel probability measure, giving each A_i equal weight $1/d$.

For every d-bit string σ there is a concept $C_\sigma \in \mathscr{C}$ which contains all A_i with $\sigma_i = 1$ and is disjoint from A_i with $\sigma_i = 0$. If A and B take constant values on all the sets A_i, $i = 1, 2, \ldots, d$, then $d_\mu(A, B)$ is just the normalized Hamming distance between the corresponding d-bit strings. Now, given $A \in \mathscr{C}$ and $0 \leq k \leq d$, there are

$$\sum_{k \leq 2\epsilon d} \binom{d}{k}$$

concepts B with $d_\mu(A, B) \leq 2\epsilon$. This allows to get the following lower bound on the number of pairwise 2ϵ-separated concepts:

$$\frac{2^d}{\sum_{k \leq 2\epsilon d} \binom{d}{k}}.$$

The Chernoff–Okamoto bound allows to estimate the above expression from below by $\exp[2(0.5 - 2\epsilon)^2 d]$. We conclude: the metric entropy of $\mathscr{C}$ with regard to μ is bounded below as:

$$M(2\epsilon, \mathscr{C}, \mu) \geq \exp[2(0.5 - 2\epsilon)^2 d].$$

The assumption $\mathrm{VC}(\mathscr{C} \bmod \omega_1) = \infty$ now implies that for every $0 < \epsilon < 0.25$,

$$\sup_{P \in \mathcal{P}} M(2\epsilon, \mathscr{C}, \mu) = \infty,$$

where $\mathcal{P}$ denotes the family of all non-atomic measures on Ω. By Lemma 7.1 in [V2], p. 278, the class $\mathscr{C}$ is not PAC learnable under $\mathcal{P}$.

5 The Universally Separable Case

Lemma 2. *Let $\mathscr{C}$ be a universally separable concept class, and let $\mathscr{C}'$ be a universally dense countable subset of $\mathscr{C}$. Then*

$$VC(\mathscr{C}) = VC(\mathscr{C}').$$

Proof. For every $C \in \mathscr{C}$ there is a sequence (C_n) of elements of $\mathscr{C}'$ with the property that for each $x \in \Omega$ there is N such that if $n \geq N$ and $x \in C$, then $x \in C_n$, and if $x \notin C$, then $x \notin C_n$. Equivalently, for every finite $A \subseteq \Omega$, there is an N so that whenever $n \geq N$, one has $C_n \cap A = C \cap A$. This means that if A is shattered by $\mathscr{C}$, it is equally well shattered by $\mathscr{C}'$. This established the inequaity $VC(\mathscr{C}) \leq VC(\mathscr{C}')$, while the converse inequality is obviously true.

Theorem 4. *For a universally separable concept class $\mathscr{C}$, the following conditions are equivalent.*

1. *$VC(\mathscr{C} \bmod \omega_1) \leq d$.*
2. *There exists a countable subset $A \subseteq \Omega$ such that $VC(\mathscr{C} \upharpoonright (\Omega \setminus A)) \leq d$.*

Proof. (1)$\Rightarrow$(2): Choose a countable universally dense subfamily $\mathscr{C}'$ of $\mathscr{C}$. Let $\mathscr{B}$ be the smallest Boolean algebra of subsets of Ω containing $\mathscr{C}'$. Denote by A the union of all elements of $\mathscr{B}$ that are countable sets. Clearly, $\mathscr{B}$ is countable, and so A is a countable set.

Let a finite set $B \subseteq \Omega \setminus A$ be shattered by $\mathscr{C}$. Then, by Lemma 2, it is shattered by $\mathscr{C}'$. Select a family $\mathscr{S}$ of $2^{|B|}$ sets in $\mathscr{C}'$ shattering B. For every $b \in B$ the set

$$[b] = \bigcap_{b \in C \in \mathscr{S}} C \bigcap \bigcap_{b \notin C \in \mathscr{S}} C^c$$

is uncountable (for it belongs to $\mathscr{B}$ yet is not contained in A), and the collection of sets $[b]$, $b \in B$ is shattered by $\mathscr{C}'$. This establishes the inequality $VC(\mathscr{C} \upharpoonright (\Omega \setminus A)) \leq VC(\mathscr{C} \bmod \omega_1)$.

(2)$\Rightarrow$(1): Fix an $A \subseteq \Omega$ so that $VC(\mathscr{C} \bmod A^c) \leq d$. Suppose a collection of n uncountable sets A_i, $i = 1, 2, \ldots, n$ is shattered by $\mathscr{C}$ in our sense. The sets $A_i \setminus A$ are non-empty; pick a representative $a_i \in A_i \setminus A$, $i = 1, 2, \ldots, n$. The resulting set $\{a_i\}_{i=1}^n$ is shattered by $\mathscr{C}$, meaning $n \leq d$.

Corollary 2. *Let $\mathscr{C}$ be a universally separable concept class on a Borel domain Ω. If $d = VC(\mathscr{C} \bmod \omega_1) < \infty$, then $\mathscr{C}$ is a universal Glivenko-Cantelli class with regard to non-atomic measures and consistently PAC learnable under non-atomic measures.*

Proof. The class $\mathscr{C}$ has finite VC dimension in the complement to a suitable countable subset A of Ω, hence $\mathscr{C}$ is a universal Glivenko-Cantelli class (in the classical sense) in the standard Borel space $\Omega \setminus A$. But A is a universal null set in Ω, hence clearly $\mathscr{C}$ is universal Glivenko-Cantelli with regard to non-atomic measures.

The class $\mathscr{C}$ is distribution-free consistently PAC learnable in the domain $\Omega \setminus A$, with the standard sample complexity $s(\epsilon, \delta, d)$. Let $\mathcal{L}$ be any consistent learning rule for $\mathscr{C}$ in Ω. The restriction of $\mathcal{L}$ to $\Omega \setminus A$ (more exactly, to $\cup_{n=1}^{\infty} ((\Omega \setminus A)^n \times \{0, 1\}^n)$) is a consistent learning rule for $\mathscr{C}$ restricted to the standard Borel space $\Omega \setminus A$, and together with the fact that A has measure zero with regard to any non-atomic measure, it implies that $\mathcal{L}$ is a PAC learning rule for $\mathscr{C}$ under non-atomic measures, with the same sample complexity function $s(\epsilon, \delta, d)$.

6 Martin's Axiom and Learnability

Martin's Axiom (MA) in one of its equivalent forms 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 can be seen as a strengthening of the statement of the Baire Category Theorem. In particular, the Continuum Hypothesis (CH) 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 will be using just one particular consequence of MA.

Theorem 5 (Martin-Solovay). *Let* (Ω, μ) *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_α, $\alpha < \kappa$ *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 less than continuum null subsets of Ω *is a null subset.* □

For the proof and more on MA, see [K], Theorem 2.21, or [F], or [Je], pp. 563–565.

Lemma 3. *Let* $\mathscr{C}$ *be an infinite concept class on a measurable space* Ω. *Denote* $\kappa = |\mathscr{C}|$ *the cardinality of* $\mathscr{C}$. *There exists a consistent learning rule* $\mathcal{L}$ *for* $\mathscr{C}$ *with the property that for every* $C \in \mathscr{C}$ *and each* n, *the set*

$$\{\mathcal{L}(\sigma, C \cap \sigma) : \sigma \in \Omega^n\} \subseteq \mathscr{C} \tag{6}$$

has cardinality $< \kappa$. *Under MA the rule* $\mathcal{L}$ *satisfies the measurability condition (1).*

Proof. Choose a minimal well-ordering of elements of $\mathscr{C}$:

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

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

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

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

$$\mathcal{L}(\sigma, C_\alpha \cap \sigma) \subseteq \{C_\beta : \beta \le \alpha\},$$

which assures (6). Besides, the learning rule $\mathcal{L}$ is consistent.

Fix $C = C_\alpha \in \mathscr{C}$, $\alpha < \kappa$. For every $\beta \leq \alpha$ define $D_\beta = \{\sigma \in \Omega^n : C \cap \sigma = C_\beta \cap \sigma\}$. The sets D_β 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 $\mu(C_\beta \triangle C_\alpha)$ on each set $D_\beta \setminus \cup_{\gamma < \beta} D_\gamma$, $\beta \leq \alpha$. Such sets, as well as all their possible unions, are measurable under MA by force of Martin–Solovay's Theorem 5, and their union is Ω^n. This implies the condition (1) for $\mathcal{L}$.

We again recall that a set $A \subseteq \Omega$ is *universal null* if it is Lebesgue measurable with regard to every non-atomic Borel probability measure μ on Ω and $\mu(A) = 0$.

Lemma 4 (Assuming MA). *Let $\mathscr{C}$ be a class of Borel subsets on a standard Borel space Ω. Suppose there is a natural d such that every countable subclass $\mathscr{C}' \subseteq \mathscr{C}$ has VC dimension $\leq d$ outside of an universal null set (which depends on $\mathscr{C}$). Then every subclass of $\mathscr{C}$ of cardinality $< 2^{\aleph_0}$ has the same property.*

Proof. By induction on the cardinality of $\mathscr{C}$, which we denote α (notice that it never exceeds $2^{\aleph_0}$, and so the proof only makes sense under the negation of the Continuum Hypothesis). Suppose the result is true for all β, $\aleph_0 \leq \beta < \alpha$. Choose a minimally well-ordered chain $\mathscr{C}_\gamma$, $\gamma < \alpha$ of subclasses of $\mathscr{C}$ whose union is $\mathscr{C}$. For every γ, let $\mathcal{N}_\gamma$ be a universal null subset of Ω with the property that $\mathscr{C}_\gamma$ has VC dimension $\leq d$ outside of $\mathcal{N}_\gamma$. Martin–Sollovay's Theorem implies that $\mathcal{N} = \cup_{\gamma < \alpha} \mathcal{N}_\gamma$ is universal null. Consequently, each $\mathscr{C}_\gamma$ has VC dimension $\leq d$ outside of $\mathcal{N}$, and the same applies to the union of the chain.

Lemma 5 (Assuming MA). *Let $\mathscr{C}$ be a concept class of cardinality $\kappa = |\mathscr{C}| < 2^{\aleph_0}$ on a standard Borel space Ω. If $d = VC(\mathscr{C})$ is finite, then $\mathscr{C}$ is a uniform Glivenko–Cantelli class, with a standard sample complexity estimate $s(\epsilon, \delta, d)$.*

Proof. A transfinite induction on κ. For $\kappa = \aleph_0$ the result is classical. Else, represent $\mathscr{C}$ as a union of an increasing transfinite chain of concept classes $\mathscr{C}_\alpha$, $\alpha < \kappa$, for each of which the statement of Lemma holds. For every $\epsilon > 0$ and $n \in \mathbb{N}$, the set

$$\left\{\sigma \in \Omega^n : \sup_{C \in \mathscr{C}} |\mu_n(\sigma) - \mu(C)| < \epsilon\right\} = \bigcap_{\alpha < \kappa} \left\{\sigma \in \Omega^n : \sup_{C \in \mathscr{C}_\alpha} |\mu_n(\sigma) - \mu(C)| < \epsilon\right\}$$

is measurable by Martin-Solovay's Theorem 5. Given $\delta > 0$ and $n \geq s(\epsilon, \delta, d)$, another application of the same result leads to conclude that for every $\mu \in P(\Omega)$:

$$\mu^{\otimes n}\left\{\sigma \in \Omega^n : \sup_{C \in \mathscr{C}} |\mu_n(\sigma) - \mu(C)| < \epsilon\right\} = \mu^{\otimes n}\left(\bigcap_{\alpha < \kappa} \left\{\sigma \in \Omega^n : \sup_{C \in \mathscr{C}_\alpha} |\mu_n(\sigma) - \mu(C)| < \epsilon\right\}\right)$$

$$= \inf_{\alpha < \kappa} \mu^{\otimes n}\left\{\sigma \in \Omega^n : \sup_{C \in \mathscr{C}_\alpha} |\mu_n(\sigma) - \mu(C)| < \epsilon\right\}$$

$$\geq 1 - \delta,$$

as required.

The following is an immediate consequence of two previous lemmas.

Lemma 6 (Assuming MA). *Under the assumptions of Lemma 4, every subclass of $\mathscr{C}$ of cardinality $< 2^{\aleph_0}$ is uniform Glivenko-Cantelli with regard to the family of non-atomic measures on Ω. The sample complexity of this class is the usual sample complexity $s(\delta, \epsilon, d)$ of concept classes of VC dimension $\leq d$.*

Lemma 7 (Assuming MA). *Let $\mathscr{C}$ be a concept class consisting of Borel subsets of a standard Borel space Ω. Assume that for some natural d, every countable subclass of $\mathscr{C}$ has VC dimension $\leq d$ outside of some universal null subset of Ω. Then the class $\mathscr{C}$ is PAC learnable under the family of all non-atomic measures on Ω, with the usual sample complexity $s(\delta, \epsilon)$ of distribution-free PAC learning concept classes of VC dimension $\leq d$.*

Proof. Using Lemma 3, choose a learning rule $\mathcal{L}$ for $\mathscr{C}$ with the property in Eq. (6). Since the family of all Borel subsets of Ω is well-known to have cardinality continuum, for every concept C and each n the cardinality of the image $\mathscr{L}_C = \mathcal{L}\{C \cap \sigma \colon \sigma \in \Omega^n\} \subseteq \mathscr{C}$ is strictly less than $2^{\aleph_0}$. By Lemma 6, $\mathscr{L}_C$ is a uniform Glivenko-Cantelli class with regard to non-atomic measures on Ω, satisfying the standard sample complexity bound. The proof is now concluded in a standard way.

7   The Proof of the Main Theorem

(1)$\Rightarrow$(2): this is Theorem 3.

(2)$\Rightarrow$(3): follows from Theorem 4.

(3)$\Rightarrow$(4): assume that for every d there is a countable subclass $\mathscr{C}_d$ of $\mathscr{C}$ with the property that the VC dimension of $\mathscr{C}_d$ is $\geq d$ after removing any countable subset of Ω. Clearly, the countable class $\cup_{d=1}^{\infty} \mathscr{C}_d$ will have infinite VC dimension outside of every countable subset of Ω, a contradiction.

(4)$\Rightarrow$(6): as a consequence of a classical result of Vapnik and Chervonenkis, every countable subclass $\mathscr{C}'$ is universal Glivenko-Cantelli with regard to all probability measures supported outside of some countable subset of Ω, and a standard bound for the sample complexity $s(\delta, \epsilon)$ only depends on d, from which the statement follows.

(6)$\Rightarrow$(5): trivial.

(5)$\Rightarrow$(3): modelling the classical argument that the uniform Glivenko-Cantelli property implies finite VC dimension, in exactly the same spirit as in the proof of our Theorem 3, one shows that the uniform Glivenko-Cantelli property of a concept class with regard to non-atomic measures implies a finite VC dimension modulo countable sets. But for a countable (more generally, universally separable) class $\mathscr{C}'$ this means finite VC dimension after a removal of a countable set, cf. Theorem 4.

(3)$\Rightarrow$(1): this is Lemma 7, and the only implication requiring Martin's Axiom.

The equivalence of (1), (7) and (8) in the universally separable case follows from Theorem 4 and Corollary 2.   □

8   Conclusion

We have characterized concept classes $\mathscr{C}$ that are distribution-free PAC learnable under the family of all non-atomic probability measures on the domain. The criterion is

obtained without any measurability conditions on the concept class, but at the expense of making a set-theoretic assumption in the form of Martin's Axiom. In fact, assuming MA makes things easier, and as this axiom is very natural, perhaps it deserves its small corner within the foundations of statistical learning.

It seems that generalizing the result from concept to function classes, using a version of the fat shattering dimension modulo countable sets, will not pose particular technical difficulties, and we plan to perform this extension in a full journal version of the paper, in order to keep the conference submission short. The Boolean algebras will however have to give way to commutative C^*-algebras [A].

It would be still interesting to know if the present results hold without Martin's Axiom, under the assumption that the concept class $\mathscr{C}$ is image admissible Souslin ([D], pages 186–187). The difficulty here is selecting a measurable learning rule $\mathcal{L}$ with the property that the images of all learning samples $(\sigma, C \cap \sigma)$, $\sigma \in \Omega^n$, are uniform Glivenko-Cantelli. An obvious route to pursue is the recursion on the Borel rank of $\mathscr{C}$, but we were unable to follow it through.

Now, a concept class $\mathscr{C}$ will be learnable under diffuse measures provided there is a hypothesis class $\mathscr{H}$ which has finite VC dimension and such that every $C \in \mathscr{C}$ differs from a suitable $H \in \mathscr{H}$ by a null set. If $\mathscr{C}$ consists of all finite and all cofinite subsets of Ω, this $\mathscr{H}$ is given by $\{\emptyset, \Omega\}$. One may conjecture that $\mathscr{C}$ is learnable under diffuse measures if and only if it admits such a "core" $\mathscr{H}$ having finite VC dimension. Is this true?

Another natural question is: can one characterize concept classes that are uniformly Glivenko–Cantelli with regard to all non-atomic measures? Apparently, this task requires yet another version of shattering dimension, which is strictly intermediate between Talagrand's "witness of irregularity" [T] and our VC dimension modulo countable sets. We do not have a viable candidate.

Finally, our investigation open up a possibility of linking learnability and VC dimension to Boolean algebras and their Stone spaces. This could be a glib exercise in generalization for its own sake, or maybe something deeper if one manages to invoke model theory and forcing.

References

[A]   Arveson, F.: An Invitation to C^*-Algebras. Graduate Texts in Mathematics, vol. 39. Springer, Heidelberg (1976)

[BEHW] Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM 36(4), 865–929 (1989)

[D]   Dudley, R.M.: Uniform Central Limit Theorems. In: Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge (1999)

[F]   Fremlin, D.H.: Consequences of Martin's Axiom. Cambridge Tracts in Mathematics, vol. 84. Cambridge University Press, Cambridge (1984)

[Je]   Jech, T.: Set theory. Academic Press, New York (1978)

[Jo]   Johnstone, P.T.: Stone Spaces. Cambridge Studies in Advanced Mathematics, vol. 3. Cambridge University Press, Cambridge (1986), Reprint of the 1982 edition

[K]   Kunen, K.: Set Theory. North-Holland, Amsterdam (1980)

[M]   Mendelson, S.: A few notes on statistical learning theory. In: Mendelson, S., Smola, A.J. (eds.) Advanced Lectures on Machine Learning. LNCS (LNAI), vol. 2600, pp. 1–40. Springer, Heidelberg (2003)

[P]   Pollard, D.: Convergence of Stochastic Processes. Springer, New York (1984)

[T]   Talagrand, M.: The Glivenko-Cantelli problem, ten years later. J. Theoret. Probab. 9(1996), 371–384 (1996)

[VC]   Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. 16(2), 264–280 (1971)

[V1]   Vidyasagar, M.: A theory of learning and generalization. With applications to neural networks and control systems. Communications and Control Engineering Series. Springer, London (1997)

[V2]   Vidyasagar, M.: Learning and Generalization, with Applications to Neural Networks, 2nd edn. Springer, Heidelberg (2003)