ARTICLES
https://doi.org/10.1038/s42256-018-0002-3
nature machine intelligence

Learnability can be undecidable

Shai Ben-David[1], Pavel Hrubeš[2], Shay Moran[3], Amir Shpilka[4] and Amir Yehudayoff[5]*

The mathematical foundations of machine learning play a key role in the development of the field. They improve our understanding and provide tools for designing new learning paradigms. The advantages of mathematics, however, sometimes come with a cost. Gödel and Cohen showed, in a nutshell, that not everything is provable. Here we show that machine learning shares this fate. We describe simple scenarios where learnability cannot be proved nor refuted using the standard axioms of mathematics. Our proof is based on the fact the continuum hypothesis cannot be proved nor refuted. We show that, in some cases, a solution to the 'estimating the maximum' problem is equivalent to the continuum hypothesis. The main idea is to prove an equivalence between learnability and compression.

Identifying the learnable is a fundamental goal of machine learning. To achieve this goal, one should first choose a mathematical framework that allows a formal treatment of learnability. This framework should be rich enough to capture a wide variety of learning problems. Then, one should find concrete ways to characterize learnability within this framework.

This paradigm has been successfully applied in many contexts of machine learning. In this work, however, we show that this paradigm fails in a well studied learning model. We exhibit a simple problem where learnability cannot be decided using the standard axioms of mathematics (that is, of Zermelo–Fraenkel set theory with the axiom of choice, or ZFC set theory). We deduce that there is no dimension-like quantity that characterizes learnability in full generality.

Standard learning models

Machine learning deals with various kinds of statistical problems, such as pattern recognition, regression and clustering. These problems share important properties in common. Perhaps the most basic similarity is in their goal, which can be roughly stated as:

> Approximate a target concept given a bounded amount of data about it.

Another similarity lies in the ideas and techniques that are used to study them. One example is the notion of generalization, which quantifies the quality of the approximation of the target concept. Other notable examples include algorithmic principles such as ensemble methods, which combine multiple algorithms in ways that improve on their individual performances, and optimization techniques such as gradient descent.

The affinity between these learning contexts leads to a pursuit of a unified theory. Such a theory would expose common structures, and enable a fruitful flow of ideas and techniques between the different contexts as well as into new learning problems that may arise in the future.

A unified theory exists in the context of classification problems, which includes problems like speech and spam recognition. This theory is called probably approximately correct (PAC) learning[1] or Vapnik–Chervonenkis (VC) theory[2,3]. A profound discovery within this theory, which is known as the 'fundamental theorem of PAC learning', is the characterization of PAC learnability in terms of VC dimension[2,4]. This result provides tight upper and lower bounds on the statistical complexity—the number of examples needed for learning—of arbitrary binary classification problems.

This characterization is remarkable in that it reduces the notion of PAC learnability to a simple combinatorial parameter. In some cases, it can even provide insights for designing learning algorithms. For example, it is useful in quantifying tradeoffs between expressivity and generalization capabilities.

Wide extensions of PAC learnability include Vapnik's statistical learning setting[5,6] and the equivalent general learning setting by Shalev-Shwartz and colleagues[7]. These rich frameworks capture many well studied settings, such as binary classification, multiclass classification, regression as well as some clustering problems. The existence of a VC dimension-like parameter that characterizes learnability in these frameworks has attracted considerable attention (see, for example, refs. [8–11]).

A corollary of our results is that there is no VC dimension-like parameter that generally characterizes learnability. We offer a formal definition of the term 'dimension'. All notions of dimension that have been proposed in statistical learning comply with this definition. We show that there can be no such notion of dimension whose finiteness characterizes learnability in general models of learning. This is discussed in more detail in the section 'Dimensions for learning'.

Our focus is on a specific learning problem we call 'estimating the maximum' (EMX). The EMX problem belongs to both models discussed above. Here is a motivating example. Imagine a website that is being visited by a variety of users. Denote by X the set of all potential visitors to the website. The owner of the website wishes to post ads on it. The posted ads are to be chosen from a given pool of ads. Each ad A in the pool targets a certain population of users $F_A \subseteq X$. For example, if A is a sports ad then F_A is the collection of sports fans. The goal is to place an ad whose target population visits the site most frequently. The challenge is that it is not known in advance which visitors are to visit the site.

More formally, we assume access to a training sample of visitors drawn from an (unknown) distribution P. The collection of ads corresponds to the family of sets $\mathcal{F} = \{F_A : A \text{ is an ad in the pool}\}$. The ad problem above becomes an instance of the following EMX problem:

[1]School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada. [2]Institute of Mathematics of the Academy of Sciences of the Czech Republic, Prague, Czech Republic. [3]Department of Computer Science, Princeton University, Princeton, NJ, USA. [4]Department of Computer Science, Tel Aviv University, Tel Aviv, Israel. [5]Department of Mathematics, Technion-IIT, Haifa, Israel. *e-mail: amir.yehudayoff@gmail.com

Given a family $\mathcal{F}$ of subsets of some domain X, find a set in $\mathcal{F}$ whose measure with respect to an unknown probability distribution P is close to maximal. This should be done based on a finite sample generated i.i.d. from P.

For more details and definitions, see section 'Estimating the maximum'.

Independence of learnability

Our main conclusion is rather surprising. We describe a simple family of sets so that its EMX learnability cannot be proved or disproved. Namely, deciding whether or not the EMX problem is solvable over this family is independent of the standard axioms of mathematics.

Our proof utilizes one of the most revolutionary mathematical discoveries in the past century: Gödel's incompleteness theorems. Roughly speaking, they state that there are mathematical questions that cannot be resolved. Theorems stating the impossibility of resolving mathematical questions are called independence results.

The family of sets $\mathcal{F}^*$ we consider is the family of all finite subsets of the interval [0, 1]. The class of probability distributions $\mathcal{P}^*$ we consider is the class of all distributions over [0, 1] with finite support.

Theorem. *The EMX learnability of $\mathcal{F}^*$ with respect to $\mathcal{P}^*$ is independent of the ZFC axioms.*

In other words, we are faced with the following scenario. There is an unknown probability distribution P over some finite subset of the interval [0, 1]. We get to see m i.i.d. (independent and identically distributed) samples from P for m of our choice. We then need to find a finite subset of [0, 1] whose P-measure is at least 2/3. The theorem says that the standard axioms of mathematics cannot be used to prove that we can solve this problem, nor can they be used to prove that we cannot solve this problem.

How can one prove an independence result? We briefly describe the main idea behind forcing, which is the tool Cohen[12,13] developed to prove independence (see also refs. [14,15]). To prove that a statement T is independent of given axioms $\mathcal{A}$, one constructs two 'worlds' (that is, two models of set theory). In both worlds the axioms $\mathcal{A}$ hold, but in one world T is true and in the second T is false. These two worlds manifest that both T and its negation $\neg T$ are consistent with axioms $\mathcal{A}$.

Gödel[16] and Cohen[12,13] proved the independence of the continuum hypothesis. The continuum hypothesis states that there are no sets whose cardinality lies strictly between the cardinalities of the integers and the continuum. By now, we know of quite a few independence results, mostly for set theoretic questions like the continuum hypothesis, but also for results in algebra, analysis, infinite combinatorics and more. Machine learning, so far, has escaped this fate.

Coming back to learnability, our arguments yield that there are two worlds that are consistent with the standard axioms of mathematics, but in one world $\mathcal{F}^*$ is EMX-learnable and in the other world it is not. Our approach is to show that the EMX learnability of $\mathcal{F}^*$ is captured by the cardinality of the continuum. In a nutshell, the class $\mathcal{F}^*$ is EMX-learnable if and only if there are only finitely many distinct cardinalities in the gap between the integers and the continuum. The latter is a variant of the continuum hypothesis that is known to be independent of ZFC. For more details, see section 'Monotone compressions and cardinalities'.

Learning and compression

Learning and compression are known to be deeply related. The learning–compression relationship is central and fruitful in machine learning (see refs. [17–20] and references within).

A central concept in our analysis is a notion of compression that we term a 'monotone compression scheme'. We show that, for classes satisfying certain closure properties, the existence of monotone compression is equivalent to EMX learnability (Lemma 1). This equivalence allows to reduce EMX learnability to the following collaborative two-player game.

The finite superset reconstruction game. *There are two players: Alice ('the compressor') and Bob ('the reconstructor'). Alice gets as input a finite set $S \subseteq X$. She sends Bob a subset $S' \subseteq S$ according to a pre-agreed strategy. Bob then outputs a finite set $\eta(S') \subseteq X$. Their goal is to find a strategy for which $S \subseteq \eta(S')$ for every S.*

It may be helpful to think of the players as abstractions of two different functionalities of a learning algorithm. Alice is the part of the algorithm that gets a sample of points as input, and needs to carefully choose a 'meaningful' subset of the points. Bob is the part of the learning algorithm that takes the compressed data and translates it to a decision.

Alice can, of course, always send $S' = S$ to Bob, which he reconstructs to $\eta(S') = S$. We focus on the following question:

Can Alice send Bob strict subsets?

It turns out that this depends on the cardinality of X. For example, if X is finite then Alice can send the empty set $\varnothing$ to Bob, which Bob reconstructs to the finite set $\eta(\varnothing) = X$. A more interesting example is when $X = \mathbb{N}$. In this case Alice can send the maximal element in her input $x_{\max} = \max S$ to Bob, which Bob successfully reconstructs to the interval $\{0, 1, \ldots, x_{\max}\}$. Alice can send a single point to Bob, and they still achieve their goal.

What about the case when X is uncountable? In the section 'Monotone compressions and cardinalities' we show that the parameters in an optimal strategy for the game over X are captured by the cardinality of X.

We are now able to see the high level structure of the proof. In the learning framework we consider there is an equivalence between the three notions: learnability, compression and cardinalities. Many statements concerning cardinalities cannot be proved nor refuted. Learnability, therefore, sometimes shares this fate.

Dimensions for learning

We now present a more concrete application. As discussed above, a fundamental result of statistical learning theory is the characterization of PAC learnability in terms of VC dimension[2,4]. Variants of the VC dimension similarly characterize other natural learning set-ups. The Natarajan and Graph dimensions characterize multi-class classificiation when the number of classes is small. For learning real valued functions (in noisy or agnostic settings), the fat-shattering dimension provides a similar characterization[21–23]. The aforementioned dimensions actually provide useful and often sharp bounds on the sample complexity needed for learning[4,8,24].

We show that there does not exist a VC dimension-like parameter that characterizes EMX learnability. In the following we provide a high-level description of the main ideas (for more details see section 'No general dimension for learning'). First-order logic allows us to formally define a notion of 'dimension' that we can investigate. Our definition says that a dimension $D(\mathcal{F})$ of a family of sets $\mathcal{F}$ is a quantity that can be defined by asking questions concerning any finite number of sets in $\mathcal{F}$ and points in X. All notions of dimension mentioned above satisfy this definition. Using this definition, we show that there is no dimension such that $D(\mathcal{F})$ is finite if and only if $\mathcal{F}$ is learnable in the EMX setting (unless the standard axioms of mathematics are inconsistent).

Learnability and compression

In this section we describe the equivalence between learning and compression that is central in our work. We start by providing the relevant background and definitions, and then state the equivalence (Lemma 1).

Estimating the maximum. The EMX problem was (implicitly) introduced in ref. [25] in the context of proper learning when the labelling rule is known to the learner. The definition of the EMX problem is similar to that of PAC learning. Let X be some domain set, and let $\mathcal{F}$ be a family of functions from X to $\{0, 1\}$ (we often think of each function $f \in \mathcal{F}$ as a subset of X and vice versa). Given a sample S of elements drawn i.i.d. from some unknown distribution P over X, the EMX problem is about finding a function $f \in \mathcal{F}$ that approximately maximizes the expectation $\mathbb{E}_P(f)$ with respect to P.

Remark. *To make sense of $\mathbb{E}_P(f)$, we need f to be measurable with respect to P. To solve this measurability issue, we make the following assumption: All distributions in this text are finitely supported over the σ-algebra of all subsets of X.*

A learner for the family $\mathcal{F}$ in our setting is a function $G: \bigcup_{k\in\mathbb{N}} X^k \to \mathcal{F}$ that takes a finite sequence of points as input, and outputs an element of $\mathcal{F}$. It is important to restrict learners to being proper (that is, to output functions in $\mathcal{F}$), since otherwise the algorithm may simply output the all-ones function, which trivially maximizes this expectation. The goal of an EMX learner is to find a function in $\mathcal{F}$ whose expectation is approximately $\mathrm{Opt}_P(\mathcal{F}) = \sup_{h\in\mathcal{F}} \mathbb{E}_P(h)$. This is captured by the following definition.

Definition 1 (EMX learner). *A learner G is an (ϵ, δ)-EMX learner for $\mathcal{F}$ if for some integer $d = d(\epsilon, \delta)$,*

$$\Pr_{S\sim P^d}[\mathbb{E}_P(G(S)) \le \mathrm{Opt}_P(\mathcal{F}) - \epsilon] \le \delta$$

for every (finitely supported) probability distribution P over X.

Monotone compression schemes. A standard notion of compression in machine learning is 'sample compression schemes'[18]. Several natural learning algorithms, such as support vector machines, can be viewed as implementing sample compression schemes. The existence of compression schemes implies learnability[18]. The reverse direction is also true. Learnable classes have compression-based learners[17,19].

Here we define a monotone version of compression schemes. Before reading the definition below it is worth recalling the 'finite superset reconstruction game' introduced in the section 'Learning and compression', and the interpretation of the two players as two components of a learning algorithm.

For integers $d \le m$, an $m \to d$ monotone compression scheme corresponds to a strategy that allows playing the game where Alice gets a set $x_1, \ldots, x_m$ as input and sends to Bob a subset $x_{i_1}, \ldots, x_{i_d}$ of size d. The intuition is that after observing m points that belong to some unknown set $h \in \mathcal{F}$, there is a way to choose d of the points so that the reconstruction η of the d points contains all the m observed examples.

Definition 2 (monotone compression schemes). *An $m \to d$ monotone compression scheme for $\mathcal{F}$ is a function $\eta: X^d \to \mathcal{F}$ such that for every $h \in \mathcal{F}$ and $x_1, \ldots, x_m \in h$, there exist $i_1, \ldots i_d$ so that*

$$\{x_1, \ldots, x_m\} \subseteq \eta[(x_{i_1}, \ldots x_{i_d})]$$

The function η is called the reconstruction function.

EMX learnability and compression. Here we state and prove the equivalence between learning and compression. Our focus is on families satisfying the following closure property.

Definition 3 (union bounded). *A family $\mathcal{F}$ of sets is union bounded if for every $h_1, h_2 \in \mathcal{F}$ there exists $h_3 \in \mathcal{F}$ such that $h_1 \cup h_2 \subseteq h_3$.*

Every class that is closed under finite unions is also union bounded. However, many natural classes that are not closed under unions are union bounded, like the class of all convex polygons.

The learnability–compression connection is summarized as follows.

Lemma 1. *The following are equivalent for a union bounded family $\mathcal{F}$ of finite sets:*

- **Weak learnability** *The family $\mathcal{F}$ is (1/3, 1/3)-EMX learnable.*
- **Weak compressibility** *There exists an $(m+1) \to m$ monotone compression scheme for $\mathcal{F}$ for some $m \in \mathbb{N}$.*

The lemma shows that, for some classes, learnability is equivalent to the weakest type of compression: removing just a single point from the input set.

Proof idea. We first explain why weak compressibility implies learnability. Assume we have an $(m+1) \to m$ monotone compression scheme for $\mathcal{F}$. The argument consists of two parts: 'boosting' and 'compression ⇒ generalization'. In the boosting part, we show how to build an $M \to m$ monotone compression scheme for all $M > m$. The second part is standard. An $M \to m$ compression for large enough M implies learnability[18].

We now explain how to achieve boosting. Start by building an $(m+2) \to m$ monotone compression scheme. Let $S = (x_1, \ldots, x_{m+2})$ be a collection of $m+2$ points that belong to some set in $\mathcal{F}$. To compress S to a collection of m points, apply the given $(m+1) \to m$ compression 'twice' as follows. First, compress the first $m+1$ points $x_1, \ldots, x_{m+1}$ to a collection $x_{i_1}, \ldots, x_{i_m}$ of m points. Then, compress the $m+1$ points $x_{i_1}, \ldots, x_{i_m}, x_{m+2}$ to a set of m points $x_{j_1}, \ldots, x_{j_m}$. This collection of m points is the compression of S. The reconstruction function is obtained by applying the given reconstruction function 'twice' as follows. Given a collection S' of m points, apply the given reconstruction once to get $\eta(S')$. Apply η a second time; let R be a set in $\mathcal{F}$ that contains S' and all sets of the form $\eta(T)$ for a collection T of m points that belong to $\eta(S')$. The set R exists since $\mathcal{F}$ is union bounded and since $\eta(S')$ is finite. The reconstruction of S' is defined to be R.

It is easy to verify that the above construction yields a $(m+2) \to m$ monotone compression scheme. By repeating this process we get an $M \to m$ compression for all $M > m$.

It remains to explain why learnability implies compressibility. We explain how to transform a learner G with sample size $d = d(1/3, 1/3)$ into an $(m+1) \to m$ monotone compression scheme for $m = \lceil 3d/2 \rceil$.

We first define the reconstruction function η and then explain how to compress a given sample. Given S' of size m, let $\eta(S')$ be a set in $\mathcal{F}$ that contains S' and also all sets of the form $G(T)$ for a collection T of d points that belong to S' (we allow repetitions in T).

We now explain how to compress a collection S of $m+1$ points that belong to some set in $\mathcal{F}$. It suffices to prove that there is a subset S' of S of size m so that all points in S belong to $\eta(S')$. Assume towards a contradiction that there is no such S'. This means that for each x in S, for every collection T of d points in S that does not contain x we have $x \notin G(T)$. Now, let P be the uniform distribution on S. On the one hand, $\mathrm{Opt}_P(\mathcal{F}) = 1$. On the other hand, for every collection T of d points from S we have $\mathbb{E}_P(G(T)) \le \frac{d}{m+1} < \frac{2}{3}$. This contradicts the fact that G is a learner.

Monotone compressions and cardinalities

We have shown that EMX learnability is equivalent to monotone compression. To complete the argument, it remains to explain the connection between monotone compressions and cardinalities.

We start with a brief overview of cardinal numbers. Cardinals are used to measure the size of sets. The cardinality of the natural

numbers is denoted by $\aleph_0$. Cantor's famous diagonalization argument shows that the cardinality of the continuum is strictly larger than $\aleph_0$. In particular, there are cardinalities that are larger than $\aleph_0$. The smallest cardinal that is larger than $\aleph_0$ is denoted by $\aleph_1$. The continuum hypothesis states that the cardinality of continuum is $\aleph_1$. Having defined $\aleph_0$ and $\aleph_1$, we can keep going and define $\aleph_{k+1}$ as the smallest cardinal that is larger than $\aleph_k$.

The key part that remains is to relate the cardinality of a set X to the size of a monotone compression of the following family of sets:

$$\mathcal{F}^X_{\text{fin}} = \{h \subseteq X : \text{h is a finite set}\}$$

Theorem 1. *For every integer $k \geq 0$ and every domain set X, the cardinality of X is at most $\aleph_k$ if and only if the class $\mathcal{F}^X_{\text{fin}}$ has a $(k+2) \to (k+1)$ monotone compression scheme.*

Before proving Theorem 1, let us explain how it implies our main theorem, that the EMX learnability of $\mathcal{F}^* = \mathcal{F}^{[0,1]}_{\text{fin}}$ is independent of the ZFC axioms. As discussed in the section 'Independence and learnability', it is known that the continuum hypothesis is independent of the ZFC axioms (see, for example, chapter 15 of ref. [14]). There are two models of set theory ('worlds') that are consistent with the axioms of ZFC:

1. In one model, the continuum hypothesis is true; the cardinality of [0, 1] is $\aleph_1$.
2. In the second model, the continuum hypothesis is far from being true; the cardinality of [0, 1] is larger than $\aleph_k$ for all integers k.

In the first model, $\mathcal{F}^*$ is learnable since it has a $3 \to 2$ monotone sample compression scheme. In the second model, $\mathcal{F}^*$ is not learnable since it has no monotone sample compression scheme. We see that the learnability of $\mathcal{F}^*$ cannot be proved nor refuted (unless the axioms lead to a contradiction).

Proof of Theorem 1. The monotone compression scheme for $\mathcal{F}^X_{\text{fin}}$ when the cardinality of X is small extends the strategy for the finite superset reconstruction game from the section 'Learning and compression'.

Let X be a set of cardinality $\aleph_k$. Given a set $S \subseteq X$ of size $k+2$, compress it as follows. Let $\prec_k$ be a well-ordering of X of order type ω_k. Let x_k be the $\prec_k$ maximal element of S. The key property of ω_k is that the cardinality of the initial segment $I_k := \{y \in X : y \prec_k x_k\}$ is at most $\aleph_{k-1}$ (see, for example, ref. [14]). Let $\prec_{k-1}$ be a well-ordering of I_k of order type ω_{k-1}. Let x_{k-1} be the $\prec_{k-1}$ maximal element of $S \backslash \{x_k\}$. The cardinality of the initial segment $I_{k-1} := \{y \in I_k : y \prec_{k-1} x_{k-1}\}$ is at most $\aleph_{k-2}$. Continue in a similar fashion to get $x_{k-2}, \ldots, x_1$ and the corresponding initial segments $I_{k-2}, \ldots, I_1$. The initial segment I_1 is countable. Let x_0 be the $\prec_1$ maximal element between the two elements of $S \backslash \{x_k, x_{k-1}, \ldots, x_1\}$. The initial segment $\{y \in I_1 : y \prec_1 x_0\}$ is finite, and it contains the only element in $S \backslash \{x_k, x_{k-1}, \ldots, x_0\}$. The final compression of S is to $S' = \{x_k, \ldots, x_0\}$. The decompression of S' reconstructs $x_k, \ldots, x_0$ and outputs S' union the finite set $\{y \in I_1 : y \prec_1 x_0\}$.

It remains to explain how to use a monotone compression for $\mathcal{F}^X_{\text{fin}}$ to deduce that the cardinality of X is small. This follows by induction using the following lemma.

Lemma 2. *Let k be a positive integer and $Y \subset X$ be infinite sets of cardinalities $|Y| < |X|$. If $\mathcal{F}^X_{\text{fin}}$ has a $(k+1) \to k$ monotone compression scheme then $\mathcal{F}^Y_{\text{fin}}$ has a $k \to (k-1)$ monotone compression scheme.*

The lemma implies the theorem as follows. Assume towards a contradiction that there is a $(k+2) \to (k+1)$ monotone compression scheme η_{k+1} for a set of cardinality $\aleph_{k+1}$. The lemma yields a $(k+1) \to k$ monotone compression scheme η_k for some set of cardinality $\aleph_k$. Repeating this $k+1$ times, we get a $1 \to 0$ monotone compression scheme η_0 for some set of cardinality $\aleph_0$. This is a contradiction; no infinite set has a $1 \to 0$ monotone compression scheme.

Proof of Lemma 2. Let η be a decompression function for $\mathcal{F}^X_{\text{fin}}$ such that for every $S \subset X$ of size $k+1$, there exists $S' \subset S$ of size $|S'| \leq k$ such that $\eta(S') \supseteq S$. The main observation is that the set

$$Z = \bigcup_{T \subset Y : |T| \leq k} \eta(T)$$

has the same cardinality as Y. This holds since Y is infinite, and since $\eta(T)$ is finite for each T. It follows that there is $x \in X$ that is not in Z, because $|X| > |Y|$. Therefore, for every $T \subset Y$ of size k, the compression S' of $S = T \cup \{x\}$ must contain x, since otherwise $x \notin \eta(S')$. So, $S' \backslash \{x\}$ is a subset of T of size $k-1$ such that $\eta(S') \supset T$.
We found a $k \to (k-1)$ compression scheme for Y. The compression is of the form $T \mapsto S' \setminus \{x\}$. The decompression is obtained by applying η and taking the intersection of the outcome with Y.

No general dimension for learning

Here we discuss the existence of a dimension-like quantity that captures learnability. All the notions of dimension described in section 'Dimensions for learning' can be abstracted as functions D that map a class of functions $\mathcal{F}$ to $\mathbb{N} \cup \{\infty\}$ and satisfy the following requirements:

1. *Characterizes learnability*: A class $\mathcal{F}$ is learnable if and only if $D(\mathcal{F})$ is finite.
2. *Of finite character*: For every $d \in \mathbb{N}$ and class $\mathcal{F}$, the statement $D(\mathcal{F}) \geq d$ can be demonstrated by a finite set of domain points and a finite collection of members of $\mathcal{F}$.

When it comes to EMX learnability, we have seen that both the size of a monotone compression for $\mathcal{F}$ and the sample size for (1/3, 1/3)-learnability of $\mathcal{F}$ satisfy the first requirement (Lemma 1). However, we now show that no such notion of dimension can both characterize EMX learnability and satisfy the finite character requirement. This can be interpreted as saying that there is no effective notion of dimension that characterizes learnability in full generality.

Let $\mathcal{X}, \mathcal{Y}$ be variables. A property is a formula $A(\mathcal{X}, \mathcal{Y})$ with the two free variables $\mathcal{X}, \mathcal{Y}$. A bounded formula ϕ is a first-order formula in which all the quantifiers are of the form $\exists x \in \mathcal{X}$, $\forall x \in \mathcal{X}$ or $\exists y \in \mathcal{Y}$, $\forall y \in \mathcal{Y}$.

Definition 4 (finite character). *A property $A(\mathcal{X}, \mathcal{Y})$ is a finite character property if there exists a bounded formula $\phi(\mathcal{X}, \mathcal{Y})$ so that ZFC proves that A and ϕ are equivalent.*

We think of $\mathcal{X}$ as corresponding to the domain X and $\mathcal{Y}$ as corresponding to the class $\mathcal{F}$. The intuition is that a finite character property can be checked by probing finitely many elements of $\mathcal{X}$ and $\mathcal{F}$.

For every integer d, the property 'VC dimension$(\mathcal{F}) \geq d$' is a finite character property. It can be expressed using only existential quantification into X and $\mathcal{F}$. Recall that PAC learnability is characterized by VC dimension.

Theorem. *For every integer d, there exists integers m, M so that for every set X and family $\mathcal{F}$ of subsets of X, the following holds*[2,4]:

- *If VC dimension $(\mathcal{F}) \leq d$ then the sample complexity of (1/3, 1/3)-PAC learning $\mathcal{F}$ is at most M.*
- *If VC dimension $(\mathcal{F}) > d$ then the sample complexity of (1/3, 1/3)-PAC learning $\mathcal{F}$ is at least m.*

The integers m, M tend to ∞ as d tends to ∞.

On the other hand, we have seen that EMX learnability is deeply related to cardinalities. As a corollary, we obtain the following theorem.

Theorem. *There is some constant $c>0$ so that the following holds. Assuming ZFC is consistent, there is no finite character property A so that for some integers m, $M>c$ for every set X and family of subsets $\mathcal{F}$ of X, the following holds:*

- *If $A(X,\mathcal{F})$ is true then the sample complexity of* (1/3, 1/3)*-EMX learning $\mathcal{F}$ is at most M.*
- *If $A(X,\mathcal{F})$ is false then the sample complexity of* (1/3, 1/3)*-EMX learning $\mathcal{F}$ is at least m.*

The theorem follows from the connection between EMX learnability and cardinalities. It is based on known constructions of two models as discussed in the section 'Monotone compressions and cardinalities'. In one model, $\mathcal{F}^*$ is EMX learnable. In the second world, $\mathcal{F}^*$ is not EMX learnable. The final crucial point is that the truth value of every finite character property must be the same in both models.

Conclusion

The main result of this work is that the learnability of the family of sets $\mathcal{F}^*$ over the class of probability distributions $\mathcal{P}^*$ is undecidable. While learning $\mathcal{F}^*$ over $\mathcal{P}^*$ may not be directly related to practical machine learning applications, the result demonstrates that the notion of learnability is vulnerable. In some general yet simple learning frameworks there is no effective characterization of learnability. In other words, when trying to understand learnability, it is important to pay close attention to the mathematical formalism we choose to use.

How come learnability can neither be proved nor refuted? A closer look reveals that the source of the problem is in defining learnability as the existence of a learning function rather than the existence of a learning algorithm. In contrast with the existence of algorithms, the existence of functions over infinite domains is a (logically) subtle issue.

The advantages of the current standard definitions (that use the language of functions) is that they separate the statistical or information-theoretic issues from any computational considerations. This choice plays a role in the fundamental characterization of PAC learnability by the VC dimension. Our work shows that this set-theoretic view of learnability has a high cost when it comes to more general types of learning.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Received: 19 August 2018; Accepted: 31 October 2018;
Published online: 7 January 2019

References

1. Valiant, L. G. A theory of the learnable. *Commun. ACM* **27**, 1134–1142 (1984).
2. Vapnik, V. N. & Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16**, 264–280 (1971).
3. Vapnik, V. N. & Chervonenkis, A. Y. *Theory of Pattern Recognition* [in Russian] (Nauka, Moscow, 1974).
4. Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. Learnability and the Vapnik–Chervonenkis dimension. *J. ACM* **36**, 929–965 (1989).
5. Vapnik, V. N. *Statistical Learning Theory* (Wiley, Hoboken, 1998).
6. Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **10**, 988–999 (1999).
7. Shalev-Shwartz, S., Shamir, O., Srebro, N. & Sridharan, K. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.* **11**, 2635–2670 (2010).
8. Ben-David, S., Cesa-Bianchi, N., Haussler, D. & Long, P. M. Characterizations of learnability for classes of {0, …, n}-valued functions. *J. Comput. Syst. Sci.* **50**, 74–86 (1995).
9. Daniely, A., Sabato, S., Ben-David, S. & Shalev-Shwartz, S. Multiclass learnability and the ERM principle. *J. Mach. Learn. Res.* **16**, 2377–2404 (2015).
10. Daniely, A., Sabato, S. & Shalev-Shwartz, S. Multiclass learning approaches: a theoretical comparison with implications. In *Proc. NIPS* 485–493 (ACM, 2012).
11. Daniely, A. & Shalev-Shwartz, S. Optimal learners for multiclass problems. In *Proc. COLT* 287–316 (2014).
12. Cohen, P. J. The independence of the continuum hypothesis. *Proc. Natl Acad. Sci. USA* **50**, 1143–1148 (1963).
13. Cohen, P. J. The independence of the continuum hypothesis, II. *Proc. Natl Acad. Sci. USA* **51**, 105–110 (1964).
14. Jech, T. J. *Set Theory: Third Millenium Edition, Revised and Expanded* (Springer, Berlin, 2003).
15. Kunen, K. *Set Theory: An Introduction to Independence Proofs* (Elsevier, Amsterdam, 1980).
16. Gödel, K. *The Consistency of the Continuum Hypothesis* (Princeton University Press, Princeton, 1940).
17. David, O., Moran, S. & Yehudayoff, A. Supervised learning through the lens of compression. In *Proc. NIPS* 2784–2792 (ACM, 2016).
18. Littlestone, N. & Warmuth, M. *Relating Data Compression and Learnability. Technical Report* (Univ. of California, 1986).
19. Moran, S. & Yehudayoff, A. Sample compression schemes for VC classes. *J. ACM* **63**, 1–21 (2016).
20. Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms* (Cambridge Univ. Press, New York, 2014).
21. Alon, N., Ben-David, S., Cesa-Bianchi, N. & Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM* **44**, 615–631 (1997).
22. Kearns, M. J. & Schapire, R. E. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.* **48**, 464–497 (1994).
23. Simon, H. U. Bounds on the number of examples needed for learning functions. *SIAM J. Comput.* **26**, 751–763 (1997).
24. Hanneke, S. The optimal sample complexity of PAC learning. *J. Mach. Learn. Res.* **15**, 1–38 (2016).
25. Ben-David, S. & Ben-David, S. Learning a classifier when the labeling is known. In *Proc. ALT 2011* (Lecture Notes in Computer Science Vol. 6925, 2011).

Acknowledgements

The authors thank D. Chodounský, S. Hanneke, R. Honzk and R. Livni for useful discussions. The authors also acknowledge the Simons Institute for the Theory of Computing for support. A.S.'s research has received funding from the Israel Science Foundation (ISF grant no. 552/16) and from the Len Blavatnik and the Blavatnik Family foundation. A.Y.'s research is supported by ISF grant 1162/15.

Competing interests

The authors declare no competing interests.

Additional information