Proceedings of Machine Learning Research vol 247:1–22, 2024 37th Annual Conference on Learning Theory

Inherent limitations of dimensions for characterizing learnability of distribution classes

Tosca Lechner TOSCA.LECHNER@UWATERLOO.CA and **Shai Ben-David** SHAI@UWATERLOO.CA

Editors: Shipra Agrawal and Aaron Roth

Abstract

We consider the long-standing question of finding a parameter of a class of probability distributions that characterizes its PAC learnability. While for many learning tasks (such as binary classification and online learning) there is a notion of dimension whose finiteness is equivalent to learnability within any level of accuracy, we show, rather surprisingly, that such parameter does not exist for distribution learning.

Concretely, our results apply for several general notions of characterizing learnability and for several learning tasks. We show that there is no notion of dimension that characterizes the *sample complexity* of learning distribution classes. We then consider the weaker requirement of only characterizing learnability (rather than the quantitative sample complexity function). We propose some natural requirements for such a characterization and go on to show that there exists no characterization of learnability that satisfies these requirements for classes of distributions. Furthermore, we show that our results hold for various other learning problems. In particular, we show that there is no notion of dimension characterizing PAC-learnability for any of the tasks: *classification learning* w.r.t. a restricted set of marginal distributions and learnability of classes of real-valued functions with continuous losses.

Keywords: List of keywords

1. Introduction

The celebrated 'fundamental theorem of statistical learning' provides a clean characterization of PAC learnability of binary classification in terms of the combinatorial Vapnik Chervonenkis dimension (VC-dimension) (Blumer et al., 1989). Furthermore, the learning rates for any class H of binary valued functions are fully determined (up to constants) by the VC-dimension of that class.

That result sparked a quest for notions of dimension that similarly characterize the learnability of other learning tasks. For some tasks, such as online learning of binary classifiers such dimensions have indeed been established. For other statistical learning tasks, some parameters have been proposed but not proven to provide the required characterizations.

In contrast, the results of (Ben-David et al., 2017) showed for the first time that for some type of problems, such as EMX learnability[1] no such characterization can be proved to exist by the common axioms of mathematics (the ZFC set theory).

In this paper, we investigate the existence of characterizing dimensions for several statistical learning problems for which that question remained open (most notably the learnability of classes of discrete probability distributions). We show that, quite surprisingly, no such characterizations

1. EMX, or General Statistical Learning, is the problem in which given a class of random variables and an i.i.d. sample generated by some unknown probability distribution P, the learner aims to find a member of that class that has close to maximal expectation w.r.t. P

exist. More concretely, we show that learnability of these problem cannot be characterized in a *scale-invariant* way (Like the VC-dimension characterization of binary classification). Our results do not rule out the existence of scale-sensitive characterization that depend on fixing the accuracy parameter.

Our results answer some long-standing open questions; The survey paper by (Diakonikolas, 2016) asks (Open Problem 1.5.1): *"Is there a "complexity measure" of a distribution class C that characterizes the sample complexity of learning C?"*

(Hopkins et al., 2023) state *"Unlike the standard model, very little is known about distribution-family learnability. While a number of works have made some progress on this front, a characterization of learnability remains elusive despite some 30 years of effort"* (end of Section 4 there).

(Benedek and Itai, 1991) ask about the characterization of PAC learnability of binary-valued classifiers w.r.t. a given class of probability distributions. They conjectured a characterization that was refuted by (Dudley et al., 1994). The latter repeats the question of finding a characterization for that task. Similar open questions are later stated by (Kulkarni and Vidyasagar, 1997) and (Vidyasagar et al., 2001). It is worth noting that the earlier proposed characterizations were *scale sensitive*. Our results show that there can be no scale invariant characterization of such learnability.

Of course, results showing the non-existence of some characterization, rely on a precise definition of such "characterizations". The notions we discuss in this paper are generalizations of a family of well studied learnability characterizations that refer to learnability to an arbitrarily small inaccuracy parameter. These include the fundamental Vapnik-Chervonenkis dimension, the Littlestone dimension, and the notions discussed for EMX learning ((Ben-David et al., 2017)). However, our definitions do not capture the so called 'scale sensitive' dimensions such as γ - fat shattering (Alon et al. (1997)). We discuss this further in subsection 3.3 below.

1.1. Our contributions

The motivation for this research was to address the question of characterizing learnability, hopefully via some appropriate parameter of a class (a.k.a. dimension) of various tasks for which no such characterization have been shown.

- The main finding of our work is that for the tasks listed, there can be no characterization (or learnability indicating dimension) that meets some natural requirements.
- As a secondary contribution, we propose formal requirements on what learnability characterizations should satisfy. All the notions that are known to characterize PAC learnability (rather than some form of weak learnability, or learnability up to some fixed accuracy) meet our proposed requirements.
- We also present two novel techniques for proving such non-existence results that we hope would find further applications.

1.2. Paper outline

In Section 2 we give a general definition for the kind of statistical learning models we will consider (of which distribution learning is a special case) and review some general definitions of ordered sets which we will use later in the paper. In Section 3 we introduce our notions of characterization of learnability. In Subsection 3.1 we introduce quantitative notions of characterization for learnability, which aim to characterize the sample complexity. In Section 3.2 we define a qualitative notion characterization of learnability (Definition 5), which is only required to distinguish learnable from non-learnable classes.

We present our main general results in Section 4: We first state a combinatorial condition of learning tasks that implies that no quantitative characterization (characterising the sample complexity of learning classes) exists. This result is shown in Theorem 10.

We then show that there are some general conditions which imply un-characterizability of a learning task in the qualitative sense (Theorem 11).

In Section 5, we use the theorems of Section 4 to show the impossibility of characterizing distribution learning, for both quantitative (Theorem 14) and qualitative (Theorem 11) notions of characterization. We also show an impossibility result for characterizing classes of distributions which are learnable with polynomial sample complexity for a slightly more restrictive notion of qualitative characterization (Theorem 21). In Section 6 we show un-characterizability by scale-invariant dimension for other learning tasks using the results from Section 4 and following the similar construction ideas as in Section 5. In particular, we show impossibility of quantitative and qualitative characterizations of classification learning of distribution classes (Theorem 23 and Theorem 24) and learning of real-valued functions with continuous losses (Theorem 27 and Theorem 28). Lastly, we discuss some implications of our results and perspectives for future research in Section 7.

2. Setup

2.1. Learning model

We consider a general notion of statistical learning tasks. These consist of the following elements

- a domain $\mathcal{Z}$ from which the input-instances/training-instances are sampled
- A class of benchmark models H (in some cases we denote it by $\mathcal{Q}$).
- A class of permissible data generating distributions $\mathcal{P} \subset \Delta(\mathcal{Z})$, where $\Delta(\mathcal{Z})$ denotes all distributions over the domain $\mathcal{Z}$.
- A set of possible outputs of a learner $\mathcal{F}$. Usually, $H \subseteq \mathcal{F}$.
- A loss/ approximation measure $L : \mathcal{F} \times \Delta(\mathcal{Z}) \to \mathbb{R}_0^+$ (where $\mathbb{R}_0^+$ denotes the set of non-negative real numbers).

The approximation error of a class H w.r.t. some data generating distribution P is defined as $\mathrm{opt}(H, P) = \inf_{h \in H} L(h, P)$.

Definition 1 (PAC learnability) *The setup that we address here is more general than the common PAC learning setup in that we consider a class of permissible data generating distributions on top of the commonly discussed class of models. In the definition below we refer to a fixed set of learner outputs $\mathcal{F}$ and a fixed loss function $L : \mathcal{F} \times \Delta\mathcal{Z} \to \mathbb{R}_0^+$.*

- *A pair of classes $H \times \mathcal{P}$ is α-agnostic PAC learnable w.r.t. to $L : \mathcal{F} \times \Delta\mathcal{Z} \to \mathbb{R}_0^+$, if there is a learner $\mathcal{A} : \bigcup_{m \in \mathbb{N}} \mathcal{Z}^m \to \mathcal{F}$ and a sample complexity function $m_H^\alpha : (0,1)^2 \to \mathbb{N}$, such that for every $\epsilon, \delta > 0$, every $P \in \mathcal{P}$ and every $m \geq m_H^\alpha(\epsilon, \delta)$, we have*

$$L(\mathcal{A}(S), P) \leq \alpha \cdot \mathrm{opt}(H, P) + \epsilon$$

with probability $1-\delta$ *over* $S \sim P^m$.

We omit the parameter α in the definition when we refer to $\alpha = 1$.

- *We denote by* $m_H(\epsilon, \delta)$ *the minimum number* m *that satisfies the above requirement for* $\alpha = 1$.
- *We say a class of models* H *is learnable if* $H \times \Delta(\mathcal{Z})$ *is learnable.*
- *We say* H *is PAC-learnable in the* realizable *case w.r.t.* L *if* $H \times \{P \in \Delta(\mathcal{Z}) : opt(H, P) = 0\}$ *is PAC learnable with respect to* L. *We will sometimes refer to the sample complexity of realizable learning by* m^{rlzb} *to distinguish it from the sample complexity of agnostic learning.*
- *We say a class of distributions* $\mathcal{P}$ *is PAC-learnable with respect to* $L : \mathcal{F} \times \Delta(\mathcal{Z}) \to \mathbb{R}_0^+$ *if* $\mathcal{F} \times \mathcal{P}$ *is PAC learnable with respect to* L.

We note that the standard definition PAC learning of a distribution class $\mathcal{P}$ with respect to total variation(TV)-distance is captured by the above definition. In this case, we have $H = \mathcal{P}$, $\mathcal{F} = \Delta(Z)$ and $L = d_{\mathrm{TV}} : \Delta(Z) \times \Delta(Z) \to \mathbb{R}_0^+$.

Definition 2 *For a given learning task, we say that a class of outputs* $H' \subset \mathcal{F}$ *is an* ϵ*-approximation for* $H \times \mathcal{P}$ *w.r.t. to* L*, if for every* $(h, p) \in H \times \mathcal{P}$*, there is a* $h' \in H'$ *such that* $L(h', p) \leq L(h, p) + \epsilon$. *Using the same slight abuse of notation as in the definition of PAC learning, we will say that:*

- H' *is an* ϵ*-approximation for* H *if it is an* ϵ*-approximation for* $H \times \Delta(\mathcal{Z})$
- H' *is an* ϵ*-approximation for* $\mathcal{P}$ *if it is an* ϵ*-approximation for* $\mathcal{F} \times \mathcal{P}$

3. Notions of characterization of learning tasks

Towards showing the "characterization" of some learning tasks is impossible, we need clear definitions of what such characterizations are.

We consider two common types of characterization of learning:

1. Quantitative notions that reflect the sample complexity of the learning task (the way the fundamental theorem of statistical learning shows that the *Vapnik-Chervonenkis dimension* characterizes the learning rates of leaning w.r.t. a given hypothesis class).
2. Qualitative notions that distinguish between learnable and non-learnable classes of models.

Below, we propose formal requirements for both types of characterizations.

All the notions that are known to characterize PAC learnability (rather than some form of weak learnability, or learnability up to some fixed accuracy) meet our proposed requirements. These include VC-dimension, Littlestone dimension, Natarajan dimension and Graph dimension.

Our main results show that no such characterization is possible for a variety of learning tasks, including the task of learning discrete distributions ((Kearns et al., 1994; Devroye and Lugosi, 2001; Silverman, 1986), learning of binary classifications w.r.t. a restricted sets of marginal distributions(Benedek and Itai (1991), Dudley et al. (1994)), and learnability of classes of real-valued functions with continuous losses.

We now elaborate our definitions of such notions.

3.1. Notions of quantitative characterization of statistical learning

Let $\mathcal{C}$ denote the family of all learnable classes w.r.t. some learning task.

Definition 3 *A* strong sample complexity dimension *is a mapping from $d : \mathcal{C} \to \mathbb{N} \cup \{\infty\}$, such that a class $\mathcal{Q}$ of models is PAC learnable if and only if $d(\mathcal{Q}) \neq \infty$ and there are functions $f : \mathbb{N} \to \mathbb{N}$ and $g : (0,1)^2 \to \mathbb{N}$ such that for every PAC learnable class of distributions $\mathcal{Q}$, $m_{\mathcal{Q}}(\epsilon, \delta) \leq f(d(\mathcal{Q}))g(\epsilon, \delta)$ for all $(\epsilon, \delta) \in (0, 1)$. In other words, there is a sample complexity upper bound function that factorizes into a factor depending only on the dimension of a class and a factor depending only on the accuracy and confidence parameters.*

Note that the fundamental theorem of statistical learning Blumer et al. (1989) shows that the *VC-dimension* is a strong sample complexity dimension for binary classification.

Definition 4 *A* weak sample complexity dimension *is a mapping from $d : \mathcal{C} \to \mathbb{N} \cup \{\infty\}$, such that a class $\mathcal{Q}$ of models is PAC learnable if and only if $d(\mathcal{Q}) \neq \infty$ and there are functions $f : \mathbb{N} \times (0,1)^2 \to \mathbb{N}$ such that for every PAC learnable class of distributions $\mathcal{Q}$, and every $(\epsilon, \delta) \in (0, 1)$, $m_{\mathcal{Q}}(\epsilon, \delta) \leq f(d(\mathcal{Q}), \epsilon, \delta)$. In other words, all the information needed about a class of distributions $\mathcal{Q}$ to determine (or upper bound) its sample complexity function $m_{\mathcal{Q}}(\epsilon, \delta)$ is captured in its dimension $d(\mathcal{Q})$.*

We will also sometimes refer to monotone real mappings $d : \mathcal{C} \to \mathbb{R} \cup \{\infty\}$ as ”weak sample complexity dimensions” if there exists a corresponding f satisfying the condition in Definition 4.

Clearly, every strong sample complexity dimension is also a weak one. We also note that, for a satisfying characterization, one might also require a lower bound of the sample complexity in terms of f and d. However, as we are only presenting impossibility results, it suffices to show that even this less ambitious goal is not achievable.

3.2. Notions of qualitative characterization of statistical learning

The notions of dimension that we have discussed above were *quantitative* - aiming to capture the sample complexity functions of learning classes. We showed that such dimensions do not exist for problems like distribution learning.

Lacking quantitative notions one can still seek *qualitative* characterizations of learnability. Namely, conditions that distinguish learnable classes from non-learnable ones. In the case of binary classification tasks, the distinction between finite and infinite VC-dimension serves as such a characterization.

Definition 5 *A* finitary characterization *of learnability for a learning task is a countable set of formulas*[2] *W such that:*

1. *A class H is not learnable if and only if it satisfies all the formulas in W.*

2. formally, these are properties of the learning problem expressed as first order formulas in many-sorted logic that has types/sorts for elements of the class of models $\mathcal{F}$, generating distributions (members of $\mathcal{P}$), domain elements and rational numbers for values of the loss function L. For brevity we keep it be clarified by the examples below.

2. *For every $\alpha \in W$ and every H that satisfies α there is a finite subset $H_\alpha \subseteq H$ such that for every H', if $H_\alpha \subseteq H'$ then H' satisfies α.*

   We say that a finitary characterization W is uniformly bounded *if for every $\alpha \in W$ there is a finite number n_α such that for every H satisfying α there is a subset $H_\alpha \subseteq H$ as above of size at most n_α.*

Note that most (if not all) of the known characterizations of learnability of learning tasks are finitary.

Examples:

1. The characterization of binary classification learning by VC-dimension; The characterizing W can be any set that contains the formulas stating *"H shatters a set of size d"* for infinitely many d's.

2. The characterization of online learnability by the Littlestone dimension; The characterizing W can be any set that contains the formulas stating *"H L-shatters a tree of depth d"* for infinitely many d's. dimension characterizing robust learning

3. The characterization of multi-class learnability by the finiteness of the Natarajan dimension, of by the finiteness of the graph dimension.

4. The characterizations by a combinatorial dimension based on the one-inclusion graph. In these characterizations, the complexity of a problem is demonstrated by a finite graph (the 1-inclusion graph). However, there is no a-priori bound to the sizes of these graphs. The resulting complexity parameters are reflected by either the maximum degree or maximum out-degree of some orientation of that graph. Such a characterizations were shown for multi-class learning Daniely and Shalev-Shwartz (2014), Brukhim et al. (2022), partial concept classes Alon et al. (2021), and adversarially robust learning Montasser et al. (2022).

There are also several conjectured characterizations that fall into this category. For example, characterizing the learnability of a class of probability discrete distributions by the finiteness of the VC-dimension of the Yatracos sets induced by that class. For a class F of functions from X to $\mathbb{R}$, their Yatracos class is the family of subsets of X defined as

$$Y(F) := \{\{x \in X : f_1(x) \geq f_2(x)\} \text{ for some } f_1, f_2 \in F\}$$

.

3.3. Comparing the different notions of characterization

The two types of characterizations introduced above are related but none of them implies the other. Sample complexity dimensions do not restrict the format (or syntax) of a characterization - the dimension function d can be *any* function (from classes to reals or natural numbers). In this respect, the notion of finitary characterization is more restrictive - it restricts the format of the characterization.

On the other hand, finitary characterizations are weaker, in the sense that they do not provide any information about learning rates. They are only required to distinguish learnable from non-learnable classes.

In many cases, however, there are characterizations that meet both definitions. Every notion of sample complexity dimension where the statements of the form $d(H) \geq k$ have finite size 'evidences' (like a set of members of H shattering a domain subset). In such case the finitary characterization set W is just the set of statements $\{d(H) \geq k : k \in \mathbb{N}\}$.

The notion of *finite character dimension* of Ben-David et al. (2017) has two definitions there. The first one requires that there is a class parameter, $D(F)$ such that a class F is PAC learnable if and only if $D(F)$ is finite. They also require *finite character*. Namely, that the parameter (or dimension) $D(F)$ has a "finite" character if for every $d \in \mathbb{N}$ and a class F, the statement $D(F) \geq d$ can be demonstrated by a finite set of domain points and a finite set of members of F". This definition is clearly equivalent (for classes of functions) to our notion of finitary characterization.

The second definition there, requiring that the statements $D(F) \geq d$ can be expressed by certain first-order formulas, is more syntactically restrictive than our definition, but on the other hand, more lenient as it does not require the existence of finite evidence sets.

Finally, it is worth mentioning again that all the types of characterizations that we discuss in the paper refer to PAC Learnability. Namely, learnability to arbitrarily small accuracy parameter. In setups where no such a characterization is known to exist, there is also a discussion of *scale-sensitive* learnability dimensions (e.g., Alon et al. (1997)). Scale-sensitive characterizations allow setting a separate condition for each level of accuracy. Our results do not rule out scale-sensitive characterizations. One should note that in setups satisfying Boosting a scale-sensitive characterization is also a PAC learnability characterization (and therefore ruled out by our results).

4. General un-charachterizability results

Our main tool for showing the impossibility of having sample complexity dimensions that provide a quantitative characterization for learning tasks is the basic notion of cofinality of an ordered set. We start by recalling its definition and some basic properties:

4.1. Some notions of ordered sets

Definition 6 (Cofinality)
Let $(X, \leq)$ be an ordered set.

- *For subsets $A, B \subseteq X$, we say that A is* cofinal *in B if for every $b \in B$ there exists some $a \in A$ such that $b \leq a$.*
- *The* cofinality of *an ordered set $(X, \leq)$ is the minimal cardinality of a subset A that is cofinal in X.*

Note that for subsets A, B, C of X, if A is cofinal in B and B is cofinal in C then A is cofinal in C.

Definition 7 (Dominance ordering of functions) *Let $(X, \leq_X)$, $(Y, \leq_Y)$ be linearly ordered sets where X has no maximal element. For functions $f, g : X \to Y$, we say that f* eventually dominates *g if there exists some $x \in X$ such that for every $x' \in X$, if $x \leq_X x'$ then $g(x') \leq_Y f(x')$. We denote this relation by $g \leq_{ed} f$.*

Claim 1 *Consider $\mathbb{N}^{\mathbb{N}}$ (the set of all functions from the natural numbers to natural numbers). The cofinality of $(\mathbb{N}^{\mathbb{N}}, \leq_{ed})$ is uncountable.*

Proof Consider any countable $A \subseteq \mathbb{N}^{\mathbb{N}}$, let $\{g_n : n \in \mathbb{N}\}$ be an enumeration of the members of A. Define $f : \mathbb{N} \to \mathbb{N}$ by $f(n) = \max\{g_i(n) : i \leq n\} + 1$. Clearly f dominates every member of A (and no member of A dominates f) showing that A is not cofinal in $(\mathbb{N}^{\mathbb{N}}, \leq_{ed})$. ■

4.2. A condition implying no qualitative characterization

The next lemma applies this notion to the existence of characterizing sample complexity (quantitative) dimensions.

Lemma 8 *Let $\mathcal{C}$ denote the family of all learnable classes w.r.t. some learning task. For any given learnable class $\mathcal{Q}$ consider the function $m_{\mathcal{Q}}(\frac{1}{\cdot}, \frac{1}{7}) : \mathbb{N} \to \mathbb{N}$ that maps a natural number k to $m_{\mathcal{Q}}(\frac{1}{k}, \frac{1}{7})$.*

If the set $\{m_{\mathcal{Q}}(\frac{1}{\cdot}, \frac{1}{7}) : \mathcal{Q} \in \mathcal{C}\}$ is cofinal in $\mathbb{N}^{\mathbb{N}}$ (under the eventual dominance ordering) then there exists no weak sample complexity dimension for that task.

Proof Let $d : \mathcal{C} \to \mathbb{N}$ be a weak sample complexity dimension. For any $d \in \mathbb{N}$, let $f'_d(k) = f(d, 1/k, 1/7)$ for all k. $\{f'_d : \mathcal{Q} \in \mathcal{C}\}$ is cofinal in the set of sample complexity functions $\{m_{\mathcal{Q}}(1/\cdot, 1/7) \in \mathbb{N}^{\mathbb{N}} : \mathcal{Q} \in \mathcal{C}\}$ (under the $\leq_{ed}$ ordering of functions). Thus the cofinality of $(\{m_{\mathcal{Q}}(1/\cdot, 1/7) : \mathbb{N} \to \mathbb{N} : \mathcal{Q} \in \mathcal{C}\}, \leq_{ed})$ is at most countable. Since we assume that the set $\{m_{\mathcal{Q}}(\frac{1}{\cdot}, \frac{1}{7}) : \mathcal{Q} \in \mathcal{C}\}$ is cofinal in $\mathbb{N}^{\mathbb{N}}$, we get a contradiction to the uncountable cofinality of $\mathbb{N}^{\mathbb{N}}$. ■

Definition 9 *For an ordered set $(\mathcal{X}, \leq)$, we say that a notion of dimension $d : \mathcal{C} \to \mathcal{X}$ is* monotonic *if for every pair of classes $\mathcal{Q}_1, \mathcal{Q}_2$, the implied sample complexity functions $f(d(\mathcal{Q}), 1/k, 1/7)$ are monotonically increasing. Namely,*

$$d(\mathcal{Q}_1) \geq d(\mathcal{Q}_2) \text{ implies that } f(d(\mathcal{Q}_1), 1/k, 1/7) \geq_{ed} f(d(\mathcal{Q}_2), 1/k, 1/7).$$

Theorem 10 *If the set $\{m_{\mathcal{Q}}(\frac{1}{\cdot}, \frac{1}{7}) : \mathcal{Q} \in \mathcal{C}\}$ is cofinal in $\mathbb{N}^{\mathbb{N}}$ (under the eventual dominance ordering) then there exists no monotonic real-valued function that is a weak sample complexity dimension for that task.*

Proof Noting that the real numbers have countable cofinality, the proof of the natural-valued dimension applies to the monotonic real-valued dimension as well. ■

4.3. A condition implying no sample complexity characterization

We now state and prove our main general result concerning our notion of qualitative characterizations of learnability.

Theorem 11 *The learnability of any learning task that satisfies the following two properties cannot be characterized by a finitary characterization.*

1. *Every finite union of learnable classes of hypotheses is learnable.*

2. *There exists a learnable class H_0 and non-learnable classes $\{H_k : k \in \mathbb{N}\}$ such that for every $k \in \mathbb{N}$, H_0 is an ϵ_k approximation of H_k and $\lim_{k\to\infty} \epsilon_k = 0$.*

Proof Assume, b.w.o.c. that W is a finitary characterization of the learning task. Let $W = \{\alpha_k : k \in \mathbb{N}\}$ be any enumeration of W. For each H_k let $\hat{H}_k$ be a finite subset of H_k be such that every $H \supset \hat{H}_k$ satisfies α_k and let $\hat{H} = \bigcup_{k\in\mathbb{N}} \hat{H}_k$. On one hand, since for every k, $\hat{H}_k \subseteq \hat{H}$, $\hat{H}$ satisfies every α_k and is therefore not learnable (by the first requirement from a characterizing W). Towards a contradiction, let us show that $\hat{H}$ is learnable. This holds because given any $\epsilon > 0$, the set $K_\epsilon = \{k : \epsilon_k \geq \epsilon/2\}$ is a finite set. Therefore by our assumptions on the learning task, the class $H_0 \cup \bigcup_{k\in K_\epsilon} \hat{H}_k$ is learnable. Therefore, for any given δ there is a learner $\mathcal{A}$ and some $m(\epsilon/2, \delta)$ so that training samples of larger size guarantee $(\epsilon/2, \delta)$ success for $\mathcal{A}$ on such samples. Since H_0 is an $\epsilon/2$ approximation to each H_k for which $k \notin K_\epsilon$ (w.r.t. the given learning instance) so being $\epsilon/2$ off the minimum loss minimizer on H_0 implies being within ϵ of the loss minimizer in $\hat{H}$. ■

5. Impossibility of Characterizing Distribution Learning

For now, we consider learning over the domain $\mathcal{X} = \mathbb{N}$. Thus all subsets of our domain are measurable. We consider learning of distribution classes with respect to total variation distance, i.e. our distance measure over distributions is given by $TV(p_1, p_2) = \sup_{A\subset\mathcal{X}} |p_1(A)-p_2(A)|$. Concretely, we consider the following PAC learning task.

Definition 12 ((realizable) PAC learning of a distribution class) Silverman (1986); Devroye and Lugosi (2001) *We say that a class $\mathcal{Q}$ of probability distributions over some domain set $\mathcal{X}$ is* PAC learnable *if there exists a function $\mathcal{A} : \bigcup_{m\in\mathbb{N}} \mathcal{X}^m \to \mathcal{P}$ and a function $m^{rlzb}_{\mathcal{Q}} : (0, 1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0, 1)^2$ and every $Q \in \mathcal{Q}$, if $m \geq m^{rlzb}_{\mathcal{Q}}(\epsilon, \delta)$ then*

$$\mathbb{P}_{S\sim Q^m}[TV(\mathcal{A}(S), Q) > \epsilon] \leq \delta.$$

Definition 13 (3-agnostic PAC learning of distribution class) Silverman (1986); Devroye and Lugosi (2001) *We say that a class $\mathcal{Q}$ of probability distributions over some domain set $\mathcal{X}$ is* PAC learnable *if there exists a function $\mathcal{A} : \bigcup_{m\in\mathbb{N}} \mathcal{X}^m \to \mathcal{P}$ and a function $m_{\mathcal{Q}} : (0, 1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0, 1)^2$ and every $Q \in \mathcal{Q}$, if $m \geq m_{\mathcal{Q}}(\epsilon, \delta)$ then*

$$\mathbb{P}_{S\sim Q^m}[TV(\mathcal{A}(S), Q) > 3 \cdot \inf_{Q'\in\mathcal{Q}}(TV(Q, Q')) + \epsilon] \leq \delta.$$

We note that these definitions are special cases of the PAC learning definition in the Setup Section (Definition 1), where $\mathcal{Z} = \mathcal{X} = \mathbb{N}$, $\mathcal{F} = \Delta(\mathbb{N})$, $\mathcal{Q} = H$ and $L = TV : \Delta(\mathbb{N}) \to \Delta(\mathbb{N}) \to \mathbb{R}^+_0$.

We now state the two main theorems of this section, showing the impossibility of both quantitative as well as quantitative characterizations of distribution learning.

Theorem 14 *There is no weak sample complexity dimension for distribution learning (neither in the realizable nor in the 3-agnostic case of distribution learning).*

Corollary 15 *There exist no monotonic real-valued function that is a weak sample complexity dimension for distribution learning.*

Corollary 15 follows directly from Theorem 14 and Theorem 10.

Theorem 16 *The learnability of classes of discrete distributions cannot be characterized by a finitary characterization. This statement holds both for realizable PAC learnability and for 3-agnostic PAC learnability.*

We note, that while we only consider constructions of discrete distributions in this section, the corresponding results on uncharacterizability for *general distribution learning* follow directly from these results.

We will show these theorems using Lemma 8 and Theorem 11 respectively. In order to do so, we need to construct classes of distributions that meet the requirements of these more general results. We will first describe a construction that can be used for both theorems. We then show some properties of this construction which will be needed for both theorems, namely an upper bound (Lemma 18) and a lower bound (Lemma 19) on its sample complexity. We will then state the proofs of Theorem 14 and Theorem 16. Lastly, we will end the section with discussing an extension of the qualitative impossibility result to classes of distributions with polynomial sample complexity.

Throughout this section we will also need the fact that finite classes are learnable. We state Theorem 3.4 from Ashtiani et al. (2018), which is a slight rephrasing of Theorem 6.3 from Devroye and Lugosi (2001).

Theorem 17 *Ashtiani et al. (2018),Devroye and Lugosi (2001) For any finite class of distributions $\mathcal{Q} = \{q_1, \ldots, q_m\}$, there exists a deterministic algorithm which 3-agnostic PAC learns Q with sample complexity $m_{\mathcal{Q}}(4\epsilon, \delta) \leq \frac{\log(3m^2)+\log(1/\delta)}{2\epsilon^2}$.*

We will now describe our construction. For a natural number $n \in \mathbb{N}$ and a (usually small) mixture parameter $\gamma \in (0, 1)$, we define the finite class,

$$P_{\gamma,n} = \{(1-\gamma)\delta_0 + \gamma U_A : A \subset \{1, \ldots, n\}\},$$

where δ_0 denotes the distribution with all its mass on point 0 and U_A denotes a uniform distribution over the set A. Intuitively, this class thus consists of a heavy non-flexible part ($(1-\gamma)\delta_0$) and a highly flexible part with low weight (γU_A). For any distribution p the TV-distance to an element of $P_{\gamma,n}$ only depends in a small part on the low-weight component. However, the low-weight flexible part, will make this class hard to learn for small ϵ. We now take union over these classes $P_{\gamma,n}$ for different combinations of γ and n, which allows us to control the behaviour of the sample complexity and fulfill the requirements for both. For sequences $\bar{\gamma} : \mathbb{N} \to [0, 1]$ and $\bar{n} : \mathbb{N} \to \mathbb{N}$, we define

$$\mathcal{Q}_{\bar{\gamma},\bar{n}} = \bigcup_{i=1}^{\infty} P_{\bar{\gamma}(i),\bar{n}(i)}. \tag{1}$$

We furthermore define $\bar{\gamma}^{-1}(\epsilon) = \arg\min\{i \in \mathbb{N} : \text{for every } j \geq i, \bar{\gamma}(j) \leq \epsilon\}$ and $n_{max}(i) = \max_{j\in\{1,\ldots i\}} \bar{n}(j)$. We will now show that infinite classes of this kind can be learnable, even as n grows to infinity (and thus making the class in some sense "infinitely flexible"), by controlling the mixture parameter γ.

Lemma 18 *Let $\mathcal{Q} = \mathcal{Q}_{\bar{\gamma},\bar{n}}$ with $\lim_{i\to\infty} \bar{\gamma}(i) = 0$. Then $\mathcal{Q}$ is 3-agnostic PAC learnable with sample complexity $m_{\mathcal{Q}}(\epsilon,\delta) \leq (128(\log(3(\bar{\gamma}^{-1}(\frac{\epsilon}{4})n_{max})\bar{\gamma}^{-1}(\frac{\epsilon}{4})+1)^n) + \log(\frac{1}{\delta})))/(\epsilon^2)$.*

Proof Assume $\lim_{i\to\infty} \bar{\gamma}(i) = 0$. Let $\epsilon > 0$. Then for every $\frac{\epsilon}{4} > 0$ there is an N, such that for every $N' \geq N$, $\bar{\gamma}(N') < \frac{\epsilon}{4}$. We can now focus on 3-agnostic learning the finite class $\mathcal{Q}' = \{\delta_0\} \cup (\bigcup_{i=0}^{N} P_{\bar{\gamma}(i),\bar{n}(i)})$, as learning as for any $q \in \mathcal{Q}_{\bar{\gamma},\bar{n}}$, there is $p \in \mathcal{Q}'$ with $TV(p,q) < \frac{\epsilon}{4}$. Thus by triangle inequality for any $q' \in \Delta(\mathcal{Z})$, we get $\inf_{p\in\mathcal{Q}'} TV(p,q') \leq \inf_{q\in\mathcal{Q}} TV(q,q') + \frac{\epsilon}{4}$. Thus if we have a 3-agnostic PAC learner $\mathcal{A}$ for $\mathcal{Q}'$ with sample complexity $m_{\mathcal{Q}'}(\frac{\epsilon}{4},\delta)$, we can use it as a PAC learner for $\mathcal{Q}$ with sample complexity $m_{\mathcal{Q}}(\epsilon,\delta)$. Now using Theorem 17, we can conclude that $\mathcal{Q}$ is learnable with sample complexity $m_{\mathcal{Q}}(\epsilon,\delta) \leq 128\frac{\log(3(\bar{\gamma}^{-1}(\frac{\epsilon}{4})n_{max}(\bar{\gamma}^{-1}(\frac{\epsilon}{4})))^2)+\log(1/\delta)}{\epsilon^2}$. ■

However, we can also lower bound the sample complexity of these kinds of classes in the following way.

Lemma 19 *For $\mathcal{Q} = P_{\gamma,4n}$, we have $m_{\mathcal{Q}}(\frac{\gamma}{8},\frac{1}{7}) \geq m_{\mathcal{Q}}^{rlzb}(\frac{\gamma}{8},\frac{1}{7}) \geq n$.*

The construction and argument follow from a common no-free lunch style argument. For details we refer the reader to Section B in the Appendices (in the supplementary file).

We can now use this construction and bounds to prove Theorem 14.

Proof [Proof of Theorem 14] Based on the confinality considerations described above, it suffices to show that the set $\{m_{\mathcal{Q}}(1/\cdot, 1/7) \in \mathbb{N}^{\mathbb{N}} : \mathcal{Q} \in \mathcal{C}\}$ is cofinal in $\mathbb{N}^{\mathbb{N}}$. Let $g \in \mathbb{N}^{\mathbb{N}}$ be arbitrary. Now consider the class $\mathcal{Q} = \mathcal{Q}_{\bar{\gamma},\bar{n}}$ as constructed in the previous section, where $\bar{\gamma}(k) = \frac{8}{k}$ and $\bar{n}(k) = 8(g(k)+1)$. Then according to Lemma 18, $\mathcal{Q}$ is learnable, as $\lim_{k\to\infty} \bar{\gamma}(k) = 0$. Furthermore, we know that for every $k \in \mathbb{N}$ we have $P_{\frac{8}{k},\bar{n}(k)} \subset \mathcal{Q}$. Thus, by Lemma 19, for every $k \in \mathbb{N}$: $g(k) < \frac{\bar{n}(k)}{4} \leq m_{\mathcal{Q}}(\frac{1}{k},\frac{1}{7})$. Thus, $g \leq_{ed} m_{\mathcal{Q}}(\frac{1}{\cdot},\frac{1}{7})$. Therefore $\{m_{\mathcal{Q}}(1/\cdot,1/7) : \mathcal{Q} \in m_{\mathcal{C}}\}$ is indeed cofinal in $\mathbb{N}^{\mathbb{N}}$. As the bounds of Lemma 18 and Lemma 19 both hold for the realizable distribution learning, we can prove that $\{m_{\mathcal{Q}}^{rlzb}(1/\cdot,1/7) : \mathcal{Q} \in m_{\mathcal{C}}\}$ is cofinal in $\mathbb{N}^{\mathbb{N}}$ by the same construction and argument. Thus, we have proved our claim.

■

We can now focus our attention the impossibility of qualitative characterization of distribution learning and finally prove Theorem 16.

Proof [Proof of Theorem 16] We only need to show that the two conditions from Theorem 11 are fulfilled by the problem of distribution learning. Condition 1 holds, as according to Theorem 17 every finite class of distributions is learnable. This means we can define a learner for any finite union of learnable sets, by running the learners for each of the learnable sets on the input to create a finite set of candidates and then learn the candidate set. Condition 2 holds by the following construction: $H_0 = \{\delta_0\}$ and $H_k = \mathcal{Q}_{\bar{\gamma}_n,\bar{n}}$, as defined in equation (1), where $\bar{\gamma}_k(i) = 1/k$ and $n(i) = i$. It is clear that, H_0 is an ϵ_k-approximation of H_k for $\epsilon_k = 1/k$ as all elements of H_k have $\frac{k-1}{k}$ mass on the point 0. Furthermore we have $\lim_{k\to\infty} \epsilon_k = 0$. Lastly, we have to argue that every H_k is not learnable. We now note that for every $n \in \mathbb{N}$ the class $P_{\epsilon_k,4n} \subset H_k$ (as defined in Section 3). Thus we can apply Lemma 19 to obtain, for every $n \in \mathbb{N}$, $m_{H_k}(\frac{\epsilon_k}{8},\frac{1}{7}) \geq n$. We note that any instance of the word "learnable" in this proof can either mean "realizeale PAC learnable" or "3-agnostic PAC learnable". The proof is correct in both cases. ■

5.1. Polynomial complexity distribution learning

Another, more restricted definition of learning, is one that requires specific bounds on the sample complexity.

Definition 20 *We say a class H is polynomially PAC learnable, if $m_H(\epsilon, \delta) \in poly(1/\epsilon, 1/\delta)$.*

We note, that for many learning tasks, like binary classification, classes are polynomially PAC learnable, if and only if they are PAC learnable. However, we have seen in this section that for the task of distribution learning, there are PAC learnable classes which are not polynomially PAC learnable. Arguably, in many scenarios one is more interested in polynomially learnable classes. We therefore pose the question, whether it is possible to give a *qualitative* characterization of *polynomially learnable* classes. In the case of distribution learning, we can give a partial answer, showing that there is no *uniformly-bounded* finitary characterization of polynomial distribution learning.

Theorem 21 *There is no uniformly-bounded finitary characterization of polynomial distribution learning (w.r.t TV-distance). This result holds for both the realizable and the 3-agnostic case of distribution learning.*

6. Impossibility of Characterizing Other Learning Tasks

In this section we consider two learning problems different from distribution learning. In Section 6.1 we consider classification learning for fixed distribution classes, as proposed by Benedek and Itai (1991), and in Section 6.2 we consider learning with real-valued functions and real-valued losses. For both of theses settings we show that there is no scale-invariant dimension which characterizes these learning tasks. Our results do not rule out the existence of scale-sensitive characterization that depend on any fixed accuracy parameter.

6.1. Classification Learning with respect to a restricted class of distributions

The standard PAC learning definitions for classification learning are 'distribution free'. Namely, learnability requires success with respect to any data-generating probability distributions. Benedek and Itai (1991) proposed a variant of that notion that requires success only w.r.t. some fixed probability distribution. They then raise the idea of extending the definitions by considering a *class* of distributions. This idea is then picked up in Dudley et al. (1994) with the following definition:

Definition 22 *A hypothesis class H is PAC-learnable with respect to a class of marginal distributions $\mathcal{D}$, if there is a learner $\mathcal{A}$ and a sample complexity function $m_{H,\mathcal{D}} : (0,1)^2 \to \mathbb{N}$, if for every $\epsilon, \delta > 0$, every $D \in \mathcal{D}$ and every labeling rule $h \in H$ and every $m \geq m^{0/1}_{H,\mathcal{D}}(\epsilon, \delta)$, with probability $1 - \delta$ over $S \sim (D, h)^m$, we have*

$$L^{0/1}_P(\mathcal{A}(S)) < \epsilon.$$

(where $L^{0/1}_P$ is the usual $0/1$ loss).

We will say a class $\mathcal{P} = \{(D, h) : D \in \mathcal{D}, h \in H\}$ of distribution classes is classification learnable, if the class of labelling rules H is learnable with respect to $\mathcal{D}$.

We will denote $\mathcal{D}^{0/1} = \{(D,h) : D \in D, h \in H_{all}\}$, where $H_{all} = 2^X$ is the class of all labelling rules. We note, that while in this definition, the marginal and the labelling rules are independent, an impossibility for characterizing this learning problem yields an impossibility for the more general setting in which the marginal contains information about the labelling rule.

We will now construct classes and show uncharacterizability of both of these learning problems in the sense of the previous chapters. Thus we show that the learning problem proposed by Benedek-Itai(Benedek and Itai (1991)) cannot be characterized in the sense of our notions of scale-invariant characterizing dimension. Using our results from the previous sections, we can get the following uncharacterizability results for this learning task.

Theorem 23 *There is no weak sample complexity dimension for classification learning of distribution classes.*

Theorem 24 *There is no finitary characterization of classification learning for distribution classes.*

Theorem 23 follows from Lemma 8 and Theorem 24 follows as a corollary of Theorem 11. Both theorems use the same class $\mathcal{Q}_{\bar{\gamma},\bar{n}}$ from Section 5. That is we consider learning the class of all labelling functions $H_{all} = 2^{\mathcal{X}}$ with respect to the class of marginal distributions $\mathcal{Q}_{\bar{\gamma},\bar{n}}$. In other words we consider the classification learnability of the class $\mathcal{Q}^{0/1}_{\bar{\gamma},\bar{n}} = \{(D,f) : D \in \mathcal{Q}_{\bar{\gamma},\bar{n}}, f \in H_{\text{all}}\}$. For the proof of Theorem 16 we use the classes $H_0 = \{h_a : a \in \{0,1\}$, where $h_a(x) = a$ for all $a \in \mathcal{X}\}$ and $H_k = \mathcal{Q}^{0/1}_{\bar{\gamma}_k,\bar{n}} = \{(D,h), D \in \mathcal{Q}_{\bar{\gamma}_k,\bar{n}}, h \in H_{\text{all}}\}$, where for all $i \in \mathbb{N}$, we have $\bar{n}(i) = i$ and $\bar{\gamma}_k(i) = \frac{1}{k}$. We note that H_0 is an ϵ-appoximation for every H_k in the sense of the second bullet-point of Definition 2.

Lemma 25 *If $\lim_{i\to\infty} \gamma(\bar{i}) = 0$, then the class $H_{all} = 2^{\mathcal{X}}$ of all labelling rules is PAC learnable with respect to the class $\mathcal{Q} = \mathcal{Q}_{\bar{\gamma},\bar{n}}$ is classification PAC learnable.*

The proof follows the same idea as the proof of Lemma 18. The proof can be found in Section B.

Furthermore, we can show a lower bound on the sample complexity for a given $P_{\gamma,2n}$.

Lemma 26 *For $\mathcal{Q} = P_{\gamma,2n}$ and $H = H_{all}$, we have $m^{0/1}_{H,\mathcal{Q}}(\frac{\gamma}{8}, \frac{1}{7}) \geq n$.*

This Lemma follows directly from the proof of the no-free-lunch theorem in Shalev-Shwartz and Ben-David (2014).

These two lemmas can now be used to show the theorems of this section. For more details on the proof on this section we refer the reader to Section B.

6.2. Learning Real-Valued Functions with Real-Valued Losses

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We will now PAC learning of real-valued functions with continuous losses. Let $\ell^g : 2^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+_0$ be a (point-wise) loss, such there is a continuous function $g : \mathbb{R} \to \mathbb{R}$ with

- for all $x \in \mathcal{X}, y \in Y, h \in 2^{\mathcal{X}}$: $\ell^g(h,x,y) = g(|h(x) - y|)$.
- $g(0) = 0$, i.e. perfect prediction incurs loss 0.
- There is $a > 0$, such that $g(a) > 0$, i.e. some level of miss-estimation will incur positive loss.

We now analyse PAC-learnability of a class $H \subset \mathcal{Y}^{\mathcal{X}}$ with respect to $L^g(h, P) = \mathbb{E}_{(x,y)\sim P}[\ell^g(h, x, y)]$.

We now state the main theorems of this subsection.

Theorem 27 *There is no weak sample complexity dimension for PAC learning real-valued function classes with respect to L^g (in neither the realizable nor the 1-agnostic case).*

Theorem 28 *There is no finitary characterization of PAC learning real-valued function classes with respect to L^g (in neither the realizable nor the 1-agnostic case).*

We note that these results do not stand in contradition to the positive result on characterizing learnability for real-valued functions given by Alon et al. (1997), as this result gives a characterization for ϵ-weak learnability for every ϵ, rather than a characterization for PAC-learnability. To show our theorems, we need a similar construction as before, which we then use to apply Lemma 8 and Theorem 11 respectively.

The construction for this result is similar in spirit to our previous constructions and can be found in Section A.

7. Discussion

We showed the uncharacterizability of learnability a variety of learning tasks, with respect to notions of characterization that meet some intuitive requirements. We discussed both the quantitative and quantitative characterizations and proposed some general properties of learning tasks that imply that their learnability is not captured by such characterizations.

Our work inspired by the work of Ben-David et al. (2017) which was the first to show the existence of a learning task that cannot be characterized. While their work laid the groundwork and gave a first formal definition of general dimensions for statistical learnability, we extended those definitions and also proposed a definition for quantitative characterizability.

Our work expands the understanding of uncharacterizability of learning problems in crucial ways. The results from Ben-David et al. (2017) applied to a newly defined learning task - Expectation Maximization - and relied on the existence of classes whose learnability is undecidable in ZFC. In contrast, our results apply to several natural learning tasks whose characterizability have thus far eluded the community (Diakonikolas (2016)) and are 'absolute' in not referring to notions of provability.

Another distinction to the EMX learning task in Ben-David et al. (2017) is the fact that the definition of EMX learning requires learning to be *proper*, i.e. the output of a successful learner needs to be element of the class that is being learned. Without this requirement the EMX setting becomes trivial, as any class can be learned by the constant learner that outputs the whole domain set for any input. In contrast, our results address the more general case of learning (and can also be easily extended to the proper case as well).

We note that all of our results rely on the construction of a sequence ϵ-weakly learnable classes for decreasing ϵ, which are not fully PAC learnable. For most tasks with known characterizations, there is an equivalence between weak learning and PAC learning. It might be interesting to further explore the connection between that equivalence and the characterizability of a learning problem.

We believe that we have not exhausted the implications of our approach and that our definitions of characterizations and our techniques are also applicable to more learning tasks.

Acknowledgments

Tosca Lechner was supported by a Vector Research Grant and a Waterloo Apple PhD Fellowship. Resources used in preparing this research were provided, in part, by the Government of Canada through CIFAR, and the Vector Institute.

References

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997. URL https://doi.org/10.1145/263867.263927.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 658–671. IEEE, 2021. doi: 10.1109/FOCS52979.2021.00070. URL https://doi.org/10.1109/FOCS52979.2021.00070.

Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3416–3425, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/70ece1e1e0931919438fcfc6bd5f199c-Abstract.html.

Shai Ben-David, Pavel Hrubes, Shay Moran, Amir Shpilka, and Amir Yehudayoff. A learning problem that is independent of the set theory ZFC axioms. *CoRR*, abs/1711.05195, 2017. URL http://arxiv.org/abs/1711.05195.

Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theor. Comput. Sci.*, 86(2):377–390, 1991. doi: 10.1016/0304-3975(91)90026-X. URL https://doi.org/10.1016/0304-3975(91)90026-X.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989. URL https://doi.org/10.1145/76359.76371.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pages 943–955. IEEE, 2022. URL https://doi.org/10.1109/FOCS54457.2022.00093.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 287–316. JMLR.org, 2014. URL http://proceedings.mlr.press/v35/daniely14b.html.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer series in statistics. Springer, 2001. ISBN 978-0-387-95117-1.

Ilias Diakonikolas. Learning structured distributions. In Peter Bühlmann, Petros Drineas, Michael J. Kane, and Mark J. van der Laan, editors, *Handbook of Big Data*, pages 267–283. Chapman and Hall/CRC, 2016. doi: 10.1201/b19567-21. URL http://www.crcnetbase.com/doi/abs/10.1201/b19567-21.

R. M. Dudley, Sanjeev R. Kulkarni, T. J. Richardson, and Ofer Zeitouni. A metric entropy bound is not sufficient for learnability. *IEEE Trans. Inf. Theory*, 40(3):883–885, 1994. doi: 10.1109/18.335898. URL https://doi.org/10.1109/18.335898.

Max Hopkins, Daniel M. Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. 2023. URL https://arxiv.org/pdf/2111.04746.pdf.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '94, page 273–282, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916638. doi: 10.1145/195058.195155. URL https://doi.org/10.1145/195058.195155.

Sanjeev R. Kulkarni and M. Vidyasagar. Learning decision rules for pattern classification under a family of probability measures. *IEEE Trans. Inf. Theory*, 43(1):154–166, 1997. URL https://doi.org/10.1109/18.567668.

Omar Montasser, Steve Hanneke, and Nati Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 37458–37470. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f392c6bbb14548df50092f10c9db440f-Paper-Conference.pdf.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5. URL http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Springer, 1986. ISBN 978-1-4899-3324-9. doi: 10.1007/978-1-4899-3324-9. URL https://doi.org/10.1007/978-1-4899-3324-9.

M Vidyasagar, S Balaji, and Barbara Hammer. Closure properties of uniform convergence of empirical means and pac learnability under a family of probability measures. *Systems I& amp; control letters*, 2001.

Appendix A. Learning Real-Valued Functions with Real-Valued Losses (Extended)

This section is an extended version of Section 6.2. We will restate all definitions and theorems of that section and elaborate on the construction needed to prove the theorems. We will now PAC learning of real-valued functions with continuous losses. Let $\ell^g : 2^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_0^+$ be a (point-wise) loss, such there is a continuous function $g : \mathbb{R} \to \mathbb{R}$ with

- for all $x \in \mathcal{X}, y \in Y, h \in 2^{\mathcal{X}}$: $\ell^g(h, x, y) = g(|h(x) - y|)$.
- $g(0) = 0$, i.e. perfect prediction incurs loss 0.
- There is $a > 0$, such that $g(a) > 0$, i.e. some level of miss-estimation will incur positive loss.

We now analyse PAC-learnability of a class $H \subset \mathcal{Y}^{\mathcal{X}}$ with respect to $L^g(h, P) = \mathbb{E}_{(x,y)\sim P}[\ell^g(h, x, y)]$.

Definition 29 *We say a class $H \subset \mathcal{Y}^{\mathcal{X}}$ is 1-agnostic PAC-learnable w.r.t. L^g, if there exists a learner $\mathcal{A}$ and a sample complexity function $m_H : (0,1)^2 \to \mathbb{N}$, such that for every $\epsilon, \delta > 0$ and every distribution P over $\mathcal{X} \times \mathcal{Y}$, we have for every $m \geq m_H(\epsilon, \delta)$,*

$$Pr_{S\sim P^m}[L_P^g(\mathcal{A}(S)) \leq \inf_{h\in H} L_P^g(h) + \epsilon] \leq 1 - \delta.$$

We say a class H is PAC-learnable w.r.t. L^g in the realizable case if it is learnable with respect to all distributions P with $\inf_{h\in H} L^g(h) = 0$. The sample complexity in the realizable case will be denoted by m_H^{rlzb}.

We now state the main theorems of this subsection.

Theorem 27 *There is no weak sample complexity dimension for PAC learning real-valued function classes with respect to L^g (in neither the realizable nor the 1-agnostic case).*

Theorem 28 *There is no finitary characterization of PAC learning real-valued function classes with respect to L^g (in neither the realizable nor the 1-agnostic case).*

We note that these results do not stand in contradition to the positive result on characterizing learnability for real-valued functions given by Alon et al. (1997), as this result gives a characterization for ϵ-weak learnability for every ϵ, rather than a characterization for PAC-learnability. To show our theorems, we need a similar construction as before, which we then use to apply Lemma 8 and Theorem 11 respectively.

We will now state the needed construction and then prove learnability as well as a lower bound on the sample complexity needed for the theorems. Let $g_{max} = \min\{\max_{a>0} g(a), 1\}$ and $g^{-1} : [0, g_{max}] \to \mathbb{R}_0^+$.

For every $\gamma \in [0, g_{max}]$,$n \in \mathbb{N}$ and $A \subset \{1, \dots, n\}$, let

$$f_{\gamma,n}^A(x) = \begin{cases} g^{-1}(\gamma) & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Now for a fixed γ and a fixed n, we define

$$F_{\gamma,n} = \{f_{\gamma,n}^A : A \subset \{1, \dots, n\}\}.$$

We then define $H_{\bar{\gamma},\bar{n}}$ for sequences $\bar{\gamma} : \mathbb{N} \to [0, g_{max}]$ and $\bar{n} : \mathbb{N} \to \mathbb{N}$, as

$$H_{\bar{\gamma},\bar{n}} = \bigcup_{i=1}^{\infty} \{f^A_{\gamma,n} : A \subset \{1, \dots, \bar{n}(i)\}\}.$$

Lemma 30 *If $\lim_{i\to\infty} \gamma\bar{(i)} = 0$, then $H_{\bar{\gamma},\bar{n}}$ is classification is 1-agnostic PAC learnable with respect to L^g.*

Lemma 31 *For $F = F_{\gamma,2n}$, we have $m_F(\frac{\gamma}{8}, \frac{1}{7}) \geq m^{rlzb}_F(\frac{\gamma}{8}, \frac{1}{7}) \geq n$.*

Now following the same proof strategy as for our previous results, this construction can be used fulfill the requirements needed to prove the theorems of this section. In particular we use Theorem 11 with classes $H_0 = \{h_0 : h_0(x) = 0 \text{ for all } x \in \mathcal{X}\}$ and $H_k = H_{\bar{\gamma}_k,\bar{n}}$ with $\bar{n}(i) = i$ and $\bar{\gamma}_k = \frac{1}{k}$. We refer the reader to Section B at the end of the paper.

Appendix B. Proofs

Proof [Proof of Lemma 19] Our proof follows a typical no-free-lunch-style argument. Consider $\mathcal{Q}' = \{(1-\gamma)\delta_0 + \gamma U_A : A \subset \{1, \dots, 4n\} \text{ and } |A| = n\}$. We will show a lower bound of learning this class of distributions and then conclude that this lower bound also holds for $\mathcal{Q}$, as $\mathcal{Q}' \subset \mathcal{Q}$. Now let $\mathcal{A}$ be any learner. Furthermore, let $S_1, \dots S_k$ be the set of all sequences of size n with elements in the set $\{0, \dots, 4n\}$. We have

$$\mathbb{E}_{S\sim q_i}[TV(q_i, \mathcal{A}(S))] = \sum_{j=1}^{k} q_i^n(S_j) TV(q_i, \mathcal{A}(S_j))$$

Now for every S_j and every $q_{i_1}, q_{i_2} \in \mathcal{Q}'$, we have that if $S_j \in \text{supp}(q^n_{i_1})$, then

$$q^n_{i_2}(S_j) = \begin{cases} q^n_{i_1}(S_j) & \text{if } S_j \subset supp(q^n_{i_2}) \\ 0 & \text{otherwise} \end{cases}$$

Let us denote $C_j = \{x \in \{1, \dots, 4n\} : x \in S_j\}$. Furthermore for a set $A = \{x_1, \dots, x_p\}$ with $C_j \subset A \subset \{1, \dots, 4n\}$ and $|A| \leq 2n$, let $\bar{A} = \{x'_1, \dots, x'_{p'}\}$, be such that $x_{l_1} < x_{l_2}$ and $x'_{l_1} < x'_{l_2}$ for $l_1 < l_2$. Now let us define $g_j(A) = C_j \cup \{x'_l \in \bar{A} : x_{p-l} \in A \setminus C_j\}$. We note that $|g_j(A)| = |A|$, $A \cap g_j(A) = C_j$ and $g_j(g_j(A)) = A$.

Now for any $q_i \in \mathcal{Q}'$, let us denote $A_i = \text{supp}(q_i) \setminus \{0\}$. Now let us define

$$f_j(q_i) = \begin{cases} (1-\gamma)\delta_0 + \gamma U_{g_j(A_i)} & \text{if } C_j \in A_i \\ \delta_{4n+1} & \text{otherwise.} \end{cases}$$

We note that if $q_i^n(S_j) > 0$, then $C_j \in A_i$ and $f_j(f_j(q_i)) = q_i$ and $f_j(q_i) \in \mathcal{Q}'$. In this case we furthermore have $TV(q_i, f_j(q_i)) \geq \frac{\gamma}{2}$ Furthermore if $q_i^n(S_j) = 0$, then $f_j(q_i)^n(S_j) = 0$. Taking both of these cases together we have for all i and for all j: $q_i^n(S_j) = f_j(q_i)^n(S_j)$.

Now we can put everything together into a no-free-lunch style argument.

$$
\begin{aligned}
\max_{q_i \in \mathcal{Q}'} \mathbb{E}_{S \sim q_i}[TV(q_i, \mathcal{A}(S))] &= \max_{q_i \in \mathcal{Q}'} \sum_{j=1}^{k} q_i^n(S_j) TV(q_i, \mathcal{A}(S_j)) \\
&\geq \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{k} q_i^n(S_j) TV(q_i, \mathcal{A}(S_j)) \\
&= \frac{1}{2T} \sum_{i=1}^{T} \sum_{j=1}^{k} q_i^n(S_j) TV(q_i, \mathcal{A}(S_j)) + \frac{1}{2T} \sum_{i=1}^{T} \sum_{j=1}^{k} f_j(q_i)^n(S_j) TV(f_j(q_i), \mathcal{A}(S_j)) \\
&= \frac{1}{2T} \sum_{i=1}^{T} \sum_{j=1}^{k} q_i^n(S_j) TV(q_i, \mathcal{A}(S_j)) + f(q_i)^n(S_j) TV(f_j(q_i), \mathcal{A}(S_j)) \\
&= \frac{1}{2T} \sum_{i=1}^{T} \sum_{j=1}^{k} q_i^n(S_j)(TV(q_i, \mathcal{A}(S_j)) + TV(f_j(q_i), \mathcal{A}(S_j))) \\
&\geq \frac{1}{2T} \sum_{i=1}^{T} \sum_{j=1}^{k} q_i^n(S_j) TV(q_i, f_j(q_i)) \\
&\geq \frac{1}{2T} \sum_{i=1}^{T} \sum_{j=1}^{k} q_i^n(S_j) \frac{\gamma}{2} = \frac{\gamma}{4}.
\end{aligned}
$$

Now, by Lemma B.1 of Shalev-Shwartz and Ben-David (2014), we get

$$
\begin{aligned}
\max_{q_i \in \mathcal{Q}'} \mathbb{P}_{S \sim q_i^n}[TV(q_i, \mathcal{A}(S)) \geq \frac{\gamma}{8}] &= \max_{q_i \in \mathcal{Q}'} \mathbb{P}_{S \sim q_i^n}[TV(q_i, \mathcal{A}(S)) \geq 1 - \frac{7\gamma}{8}] \\
&\geq \max_{q_i \in \mathcal{Q}'} \frac{\mathbb{E}_{S \sim q_i^n}[TV(q_i, \mathcal{A}(S))] - \frac{1}{8}}{\frac{7}{8}} \geq \frac{1}{7}.
\end{aligned}
$$

Thus, $m_{\mathcal{Q}'}(\frac{\gamma}{8}, \frac{1}{7}) \geq n$. Therefore $m_{\mathcal{Q}}^{rlzb}(\frac{\gamma}{8}, \frac{1}{7}) \geq n$. ■

Proof [Proof of Theorem 21] Assume, b.w.o.c. that W is a uniformly bounded finitary characterization of polynomial distribution learning. From W being uniformly bounded, we know that for every α there is n, such that for every class H, there is a subset H_α with $|H_\alpha| \leq n$ and such that for every H' if $H' \subset H_\alpha$, then H' satisfies α. Let $W = \{\alpha_k : k \in \mathbb{N}\}$ an enumeration of W, that is ordered by the size of the corresponding n_k, i.e. such that for every $k \leq k'$, we have $n_k \leq n_{k'}$. We define $f(k) = k \cdot n_k$ Consider $H_0 = \{\delta_0\}$ and $H_k = \mathcal{Q}_{\bar{\gamma}_k, \bar{n}}$ with $\bar{n}(i) = i$ and $\bar{\gamma}_k(i) = \max\{1/f(i), 1/f(k)\}$. Now H_k is not learnable as for every $m \in \mathbb{N}$, $P_{1/f(k),4m} \subset H_k$, meaning that by Lemma 19 for every m, $m_{H_k}(1/(8f(k)), 1/7) \geq m$. From the uniformly bounded finitary characterization, we know that there is $\hat{H}_k$, with $|\hat{H}_k| = n_k$ and every H' if $H' \subset \hat{H}_k$ then H' satisfies α_k. Let $\hat{H} = H_0 \cup (\bigcup_{k \in \mathbb{N}} \hat{H}_k)$. By construction, we have that $\hat{H}$ satisfies W. Furthermore, when aiming for (ϵ, δ)-success, it is sufficient to restrict our attention to learning the $\epsilon/4$ approximation $H_{\epsilon/4} = H_0 \cup (\bigcup_{k=1}^{\bar{\gamma}^{-1}(\epsilon/4)} \hat{H}_k)$ of $\hat{H}$, where $\bar{\gamma}^{-1}(\epsilon) = \min\{i : \text{ for all } j > i, \bar{\gamma}(j) < \epsilon\} = \min\{k : kn_k \geq \frac{1}{\epsilon}\}$. Thus, $|H_{\frac{\epsilon}{4}}| \in poly(\frac{1}{\epsilon})$. From

Theorem 17 we thus get that $H_{\epsilon/4}$ is polynomially learnable, which implies that $\hat{H}$ is polynomially PAC learnable w.r.t. to TV distance. Learning here can either mean 3-agnostic or realizable learnability. The result holds for both cases.

■

Proof [Proof of Lemma 25] The proof is equivalent to the proof of Lemma 18. Let $\epsilon > 0$. $\mathcal{Q}^{0/1}_{\frac{\epsilon}{2}} = \bigcup_{i:\bar{\gamma}(i)>\frac{\epsilon}{2}} P^{0/1}_{\bar{\gamma}(i),\bar{n}(i)}$. From $\lim_{i\to\infty}\bar{\gamma}(i) = 0$, we know that this class is finite. For a class Q, define the hypothesis class $H(Q) = \{h \in \{0,1\}^{\mathcal{X}} : \exists q \in Q \text{ with } q(x,1) \geq q(x,0) \text{ if and only if } h(x) = 1\}$. Now let us consider $H = H(\mathcal{Q}^{0/1}_{\frac{\epsilon}{2}})$. By construction, this class is finite and can therefore be PAC learned (in the binary classification sense). Furthermore, we have constructed H in such a way that for every $q \in \mathcal{Q}^{0/1}$, we can bound $\inf_{h\in H} L^{0/1}_q(h) \leq \inf_{p\in\mathcal{Q}_{\frac{\epsilon}{2}}}(TV(p,q) + \inf_{h\in H} L^{0/1}_p(h)) \leq \frac{\epsilon}{2}$. Thus, if we have a learner that $(\epsilon/2,\delta)$-successfully learns H, we can use it to successfully (ϵ,δ) learned $\mathcal{Q}^{0/1}$ w.r.t. to $L^{0/1}$. ■

Proof [Proof of Lemma 26]

Let us consider $P \in P^{0/1}_{\gamma,4n} = \{(D,h) : D = (1-\gamma)\delta_0 + \gamma U\}$. For any $h \in 2^{\mathbb{N}}$, we can decompose the loss $L^{0/1}_{(D,h')}(h) = (1-\gamma)\mathbf{1}[h(0) \neq h'(0)] + \gamma L^{0/1}_U(h) \geq \gamma L^{0/1}_U(h)$. Now for every learner $\mathcal{A}$, we can derive the lower bound $\max_{P\in P_{\gamma,4n}} \mathbb{E}_{S\sim P}[L^{0/1}_{U,h}(A(S))] \geq \frac{1}{4}$, according to the same argument as in the no-free-lunch theorem in Shalev-Shwartz and Ben-David (2014). Thus we have $\max_{P\in P_{\gamma,4n}} \mathbb{E}_{S\sim P}[L^{0/1}_P(A(S))] \geq \frac{\gamma}{4}$. Thus by Lemma B.1 of Shalev-Shwartz and Ben-David (2014), we get the claimed result.

■

Proof [Proof of Theorem 23] Let $\mathcal{C}$ be the collection of all classification-learnable distribution classes. According to Lemma 8 it is sufficient to show that $\{m_Q(\frac{1}{\cdot},\frac{1}{7}) : \mathcal{Q} \in \mathcal{C}\}$ is cofinal in $\mathbb{N}^{\mathbb{N}}$. Let $g \in \mathbb{N}^{\mathbb{N}}$ be arbitrary. Now consider the class $\mathcal{Q} = \mathcal{Q}_{\bar{\gamma},\bar{n}}$ with $\bar{\gamma}(k) = \frac{8}{k}$ and $\bar{n}(k) = 8(g(k)+1)$. Then, according to Lemma 25 is learnable. Furthermore, since for every $k \in \mathbb{N}$, we have $P_{\frac{8}{k},\bar{n}(k)} \subset \mathcal{Q}$, by Lemma 26 we have $g(k) \leq \frac{\bar{n}(k)}{4} < m_Q(1/k, 1/7)$. This shows that $\{m_Q(\frac{1}{\cdot},\frac{1}{7}) : \mathcal{Q} \in \mathcal{C}\}$ is indeed cofinal in $\mathbb{N}^{\mathbb{N}}$, which concludes the proof of our claim. ■

Proof [Proof of Theorem 24] We will again use Theorem 11 to show this claim. Thus we only need to show the conditions for Theorem 11 hold.

- *Any union of finitely many learnable classes is learnable*: Let us denote this class by $Q = \bigcup_{i=1} Q^i$. For every Q^i there is a learner A_i with sample complexity m_i. Let $\epsilon > 0, \delta > 0$ and let $m_{\max} = \max(\{m_i(\epsilon/2,\delta/2) : 1 \leq i \leq k\} \cup \{\frac{4(\log(k)+\frac{2}{\delta})}{\epsilon^2}\})$. Now for some $q \in Q$, let $S \sim q^{m_{max}}$. For every $i \in \{1,\ldots,k\}$, we run A_i on S and denote the output hypothesis by h_i. We know that there is $j \in \{1,\ldots,k\}$, such that $q \in Q^j$. By the success-guarantee of A_j, we know that with probability $1-\frac{\delta}{2}$, $L^{0/1}_q(h_j) \leq \frac{\epsilon}{2} + L^{0/1}_q(f_q)$. We can now use a PAC-learner for the finite class $H = \{h_i : 1 \leq i \leq k\}$ (which we know to exist from PAC learnability of binary classification of hypothesis classes (see Shalev-Shwartz and Ben-David (2014)). We know that a sample complexity of m_{max} guarantees a $(\epsilon/2,\delta/2)$-learning success for learning H. Taking everything together, we have constructed a learner that guarantees (ϵ,δ)-success for learning Q.

- We let $H_0 = \{h_a : a \in \{0,1\}$ where for all $x \in \mathbb{N} : h_a(x) = a\}$. Furthermore, we let H_k be $H_k = \mathcal{Q}_{\bar{\gamma}_k,\bar{n}}$, where $\bar{\gamma}_k(i) = 1/k$ and $n(i) = i$ for all $i \in \mathbb{N}$. By Lemma 26 all H_k are not learnable. Furthermore, by construction, we have that for every $\epsilon_k = \frac{1}{k}$, H_0 is an ϵ_k approximation for H_k. Since $\lim_{k\to\infty} \epsilon_k = 0$, the second condition of Theorem 11 is fulfilled. This concludes our proof.

■

Proof [Proof of Lemma 30] Let h_0 be the all-zero function, i.e. $h_0(x) = 0$ for all $x \in \mathbb{N}$. We start by noting that for every γ, every n and every $A \subset \{1,\dots,n\}$ if $P = (D, f^A_{\gamma,n})$ for any marginal D over $\mathbb{N}$, then $L^g_P(h_0) \leq g(g^{-1}(\gamma)) = \gamma$. Furthermore, we know that finite hypothesis classes of hypotheses with finite range are learnable due to uniform convergence (which we get from first using Hoeffding on each of the elements of the class and then using a union bound). We can now use the same proof idea as in Lemma 18. Let $\epsilon > 0$. Since $\lim_{i\to\infty} \bar{\gamma}(i) = 0$, there is N, such that for every $M \geq N$, $\bar{\gamma}(M) \leq \frac{\epsilon}{2}$. Thus we know that for every P and every $h \in \mathcal{F}_{\bar{\gamma},\bar{n}}$, there is $h' \in \bigcup_{i=1}^N \{f^A_{\bar{\gamma}(i),\bar{n}(i)} : A \subset \{1,\dots,n\}\}$, such that $L^g_P(h') \leq L^g_P(h) + \frac{\epsilon}{2}$. We now use the fact that $H' = \bigcup_{i=1}^N \{f^A_{\bar{\gamma}(i),\bar{n}(i)} : A \subset \{1,\dots,n\}\}$ is a finite class of hypotheses with finite range and can therefore be successfully PAC-learned w.r.t. L^g. Now we can use the learner for H' with sample complexity $m_{H'}$ on an i.i.d. sample of size $m \geq m_{H'}(\epsilon/2, \delta)$, to guarantee (ϵ,δ)-success for learning $H_{\bar{\gamma},\bar{n}}$. This ■

Proof [Proof of Lemma 31] For this lower bound, let us only consider the realizable case. The agnostic case follows directly from it. We note that for a fixed γ and n, for every $h_1, h_2 \in F_{\gamma,2n}$, every $x \in \mathbb{N}$, we have either $\ell^g(h_1, x, h_2(x)) = \gamma$ or $\ell^g(h_1, x, h_2(x)) = 0$. We can thus treat l^g as a binary loss. We see that the statement now becomes equivalent to the No-Free-Lunch Theorem for binary classification (see Theorem 5.1 in Shalev-Shwartz and Ben-David (2014)) with the ϵ-parameter in the sample complexity being scaled by γ. Thus we can conclude $m^{rlzb}_F(\frac{\gamma}{8}, \frac{1}{7}) \geq n$. This concludes our proof.

■

Proof [Proof of Theorem 27] Let $\mathcal{C}$ be the class of all PAC-learnable function classes with respect to L^g. We know from Lemma 8 that it is sufficient to show that $\{m_H(1/\cdot, 1/7) : H \in \mathcal{C}\}$ is cofinal in $\mathbb{N}^{\mathbb{N}}$. Now let $g \in \mathbb{N}^{\mathbb{N}}$ be arbitrary. We now construct a class H, such that $m_H(\frac{1}{\cdot}, \frac{1}{7})$ eventually dominates g. Let $H = H_{\bar{\gamma},\bar{n}}$, with $\bar{\gamma}(k) = 8/k$ and $\bar{n} = 4(g(k)+1)$ for every $k \in \mathbb{N}$. From Lemma 25 we know that the class is learnable as $\lim_{k\to\infty} \frac{8}{k} = 0$. From Lemma 31 we get $g(k) < \frac{\bar{n}(k)}{4} < m_H(1/k, 1/7)$. We note that the notion of "learnability" used here can either refer to the realizable case and sample complexity function m^{rlzb}_H or the 1-agnostic case and sample complexity function m_H. In either case, the statements are true, giving us both claims. ■

Proof [Proof of Theorem 28] We use Theorem 11 to prove the claim. We thus only need to show that the two conditions of the theorem are fulfilled. We will focus on the sub-task of learning function classes with range bounded by 1. Since we only use function classes of the form $H_{\bar{\gamma},\bar{n}}$ and those all consist of functions with range in $[0,1]$, this does not cause any issue.

- *Every finite union of learnable classes of hypotheses is learnable*:Let $H = \bigcup_{i=1}^k H^i$ be a union of learnable classes. Let A_i denote the learner and m_i denote the sample complexity for learning H_i. Now define $m_{max}(\epsilon, \delta) = \max(\{m_i(\epsilon/2, \delta/2)\} \cup \{\frac{4k+4\log 2/\delta}{\epsilon^2}\})$. Let A be the learner that first runs every A_i on an input sample to create a finite hypothesis class $H'(S) = \{A_i(S) : 1 \leq i \leq k\}$ of candidates of size k and then runs ERM on the finite hypothesis class. Now P be some distribution over $\mathcal{X} \times \mathcal{Y}$ and let $m \geq m_{max}(\epsilon, \delta)$. Furthermore let $S \sim P^m$. We know that there is j, such that $\inf_{h \in H^j} L_P^g(h) = \inf_{h \in H} L_P^g(h)$. Now A_j guarantees that $\inf_{h \in H'(S)} L_P^g \leq \frac{\epsilon}{2}$. Furthermore, from Hoeffding's inequality and union bound, we get that a sample size of $m_{max}(\epsilon, \delta)$ is sufficient for an ERM to guarantee $(\frac{\epsilon}{2}, \frac{\delta}{2})$-success, when learning any finite hypothesis class of size k with functions with range $[0, 1]$. Thus the learner A successfully PAC learns H with sample complexity m_{max}.

- We define $H_0 = \{h_0\}$, with $h_0(x) = 0$ for all $x \in \mathbb{N}$. Furthermore we let $H_k = H_{\bar{\gamma}_k, \bar{n}}$, where $\bar{\gamma}_k(i) = 1/k$ and $\bar{n}(i) = i$. From Lemma 31 we know that none of the classes H_k are learnable. Furthermore for every $k \in \mathbb{N}$, H_0 is an $\epsilon_k = \frac{1}{k}$ approximation of H_k, as argued in the proof of Lemma 30. Lastly $\lim_{k \to \infty} \epsilon_k = 0$. This concludes our proof.

■