Proceedings of Machine Learning Research vol 313:1–27, 2026 37th International Conference on Algorithmic Learning Theory

No Scale Sensitive Dimension for Distribution Learning

Tosca Lechner TOSCA.LECHNER@VECTORINSTITUTE.AI
Vector Institute, Ontario, Canada

Shai Ben-david SHAI@UWATERLOO.CA
University of Waterloo and Vector Institute, Ontario, Canada

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Learning probability distributions is one of the most basic statistical learning tasks. While for many learning tasks learnability of a class can be characterized by a combinatorial dimension (like the VC-dimension for binary classification prediction), no such characterization is known for classes of probability distributions. A leap toward resolving this long-standing problem was made recently by (Lechner and Ben-David, 2024) who showed that there can be no *scale invariant* characterization of PAC style learnability of such classes. The question of *scale sensitive* characterization remained open. In this paper we fully resolve the question by showing that there can be no *scale sensitive* combinatorial characterization of PAC style learnability of classes of probability distributions.

Keywords: PAC learnability, classes of probability distributions, scale-sensitive characterization of learnability.

1. Introduction

We investigate the fundamental problem of learning classes of probability distributions from finite randomly generated samples (PAC style learning) with respect to the total variation variation distance. Characterizing learnability of classes of models is a hallmark of machine learning theory. Starting from the characterization of classes of binary classification by the Vapnik-Chervonenkis dimension (Blumer et al., 1989), through the characterization of online learning by the Littlestone dimension (Littlestone, 1987; Ben-David et al., 2009), and on to charactrization of multi-class learnability (Daniely and Shalev-Shwartz, 2014; Brukhim et al., 2022), of robust learnability (Montasser et al., 2022) and more.

However, for some basic leaning tasks, most notably the learnability of classes of probability distributions, such characterizations were not found despite considerable research effort over more than 30 years. A leap towards a solution of this problem was made recently by (Lechner and Ben-David (2024)), showing that there can be no *scale invariant* characterization for distribution learning, as well several other not-yet-characterized problems.

Given a learning setup a *scale-invariant* characterization of learnability in that setup determines, for every class of models, whether it can be learned for any level of accuracy. The prototypical example of such a characterization is the Vapnik-Chervonenkis dimension, a combinatorial dimension that depends only on the class of models (hypotheses) and determines whether it is learnable for any level of accuracy. In contrast, a *scale-sensitive* characterization of learnability characterizes for any given level of accuracy whether the class is learnable to such accuracy. For a given class of models, the statistical complexity of learning it to a given accuracy depends on the class 'richness' (or expressive power) at that level of accuracy. A typical example of scale-sensitive characterization is the characterization of real valued learnability of classes of functions by the fat-shattering dimension (Alon et al., 1997). A class may have a different γ- fat shattering dimension for different values of γ. Interestingly, Lechner and Ben-David (2024) show that for that task, no scale-invariant dimension can characterize learnability.

The main result of this work is proving that there can also be no *scale-sensitive* characterization of the learnability of classes of probability distributions.

On our way towards this result we discuss some other fundamental issues in learnability, such as different notions of learnability characterizing dimension and two compactness properties, i.e. properties that allow us to assess the learning behaviour of a class by via the learning behaviour of its finite or countable subsets. We prove a novel non-compactness result for distribution learning as a tool for our main non-characterizability theorem and a novel compactness result as a tool for showing a positive characterization for countable classes over countable domains. We also develop a new general tool for proving the non-learnability of classes of probability distributions. We believe that the techniques we introduce in this work and the questions we raise will be of independent interest for future research.

1.1. Related Work

As mentioned above, the characterization of learnability in different learning setups by notions of combinatorial dimensions is a key theme in the theory of machine learning. Until recently, all of the results along this line were positive - coming up with definitions of dimensions and proving that they indeed characterize learnability.

The first negative results appeared in Ben-David et al. (2017). They discuss a general statistical learning problem (expectation maximization, EMX), and use set theoretic arguments to show that there can be no combinatorial dimension characterizing such learnability. A similar result for proper learnability in the multiclass setting is shown in Asilis et al. (2025) (by reducing that problem to the EMX learnability problem). Naturally, such non-existence results require a precise definition of what a combinatorial characterization of learnability is. Lechner and Ben-David (2024) discuss some relaxations of the characterization definitions of Ben-David et al. (2017) and prove the non-exists of such characterizations for scale invariant learning of several learning setups, including the setup of distribution learning which is the focus of this work. We extend the definitions of Lechner and Ben-David (2024) to the task of scale-sensitive characterizations. It is worthwhile noting that a scale sensitive characterization is a weaker notion than scale invariant characterization, and therefore non-existence results for scale-sensitive dimensions are stronger than similar scale invariant results.

Distribution learning (a.k.a. density estimation) is a central task in statistics and machine learning. Discussions of learnability of classes of distributions consider two facets of the problem; the statistical (information complexity, or sample size) aspects and the computational complexity of such tasks. Here we focus on the statistical aspect. There are quite a few results showing the statistical complexity of learning specific classes (e.g., mixtures of Gaussian distributions Ashtiani et al. (2018)) and some sufficient conditions for learnability of such classes (such as Yatracos (1985)). However, the question of *characterizing* learnability of such classes has been recognized as a significant open problem for a long time (for example, it is listed as Open Problem 1.5.1. in Diakonikolas (2016) and as a problem that ”remains elusive despite some 30 years of effort” in Hopkins et al. (2023)).

Another aspect of this work is the use of compactness properties of sample complexity and of learnability of classes. Earlier, a (different) learning compactness result was shown in Asilis et al. (2024) in the context of transductive learning.

1.2. Outline of our results

Our main result states, as described above, that there can be no scale sensitive dimension that characterizes the learnability of classes of probability distributions (Theorem 10, Section 3.2).
Our first step toward showing that result is to provide a formal definition (Definition 8) of such a notion (including a discussion of some alternative variations) in Section 3.
We discuss two compactness properties, *countable compactness of learnability* and *pseudo-compactness for sample complexity of learning*.
We show that the existence of scale sensitive learnability characterizing dimension implies a countable compactness property of learnability (Lemma 9).
Our main results now follows from constructing a class $\mathcal{C}_{\text{counter}}$ of distributions that strongly violates that countable compactness of learnability property. Namely, every countable subclass of that class is learnable but the full class is not (Theorem 11).
We give a high level description of the proof and the class $\mathcal{C}_{\text{counter}}$ in Section 3.2 and a formal description and full proofs in Section 6. The proof of non-learnability of the full class is carried out by introducing a novel (as far as we are aware) technique for proving non-learnability of classes of probability distributions (Theorem 22) in Section 5. This tool may be of independent interest.
Finally, we discuss c-pseudo-compactness of sample-complexity (Definition 14) in Section 4, a property that states, for every class $\mathcal{C}$, if a sample size m suffices to guarantee (ε, δ)-success for all finite subclasses of $\mathcal{C}$, then a sample of size m suffices to guarantee $(c \cdot \varepsilon, \delta)$-success. We then show that learning tasks that satisfy this pseudo-compactness, are characterized by an implied notion of scale sensitive dimension. We conclude our investigation by showing that countable classes of distributions over countable domains, do satisfy the pseudo-compactness requirement, and we deduce that in such cases a scale sensitive learnability characterizing dimension does exist.

2. Setup

2.1. Distribution Learning

Let $\mathcal{X}$ be a domain set and $\Sigma(\mathcal{X})$ be a σ-algebra over it. We consider learnability of classes of distributions $\mathcal{C}$ corresponding to probability measures defined over measure space $(\mathcal{X}, \Sigma)$. We will denote the set of all probability measures over $(\mathcal{X}, \Sigma)$ by $\Delta((\mathcal{X}, \Sigma))$. Thus, we consider learnability of classes $\mathcal{C} \subset \Delta((\mathcal{X}, \Sigma))$.

As distance measure between two distributions $p, q \in \Delta(\mathcal{X}, \Sigma)$, we consider the *total variation distance* defined by:

$$d_{\mathrm{TV}}(p, q) = \sup_{B \in \Sigma} |p(B) - q(B)|.$$

We will denote $\mathcal{X}^* = \bigcup_{m=1}^{\infty} \mathcal{X}^m$. When learning a class $\mathcal{C}$, we are interested in the following success criteria.

Definition 1 (realizable (ε, δ)-success) *Let $\mathcal{C}$ be a class of distributions. We say a sample size m* guarantees (realizable) (ε, δ)-success for $\mathcal{C}$, *if there exists a learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$, such that for* *every $p \in \mathcal{C}$, with probability $1 - \delta$ over $S \sim p^m$, we have*

$$d_{\mathrm{TV}}(\mathcal{A}(S), p) \leq \varepsilon.$$

In the above definition, we are only considering distributions p from the distribution class $\mathcal{C}$. However, one may also be interested in the more general agnostic learning success.

Definition 2 (α-agnostic (ε, δ)-success) *Let $\mathcal{C}$ be a class of distributions. We say a sample size m* guarantees α-agnostic (ε, δ)-success for $\mathcal{C}$, *if there exists a learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$, such that for every $p \in \Delta(\mathcal{X})$, with probability $1 - \delta$ over $S \sim p^m$, we have*

$$d_{\mathrm{TV}}(\mathcal{A}(S), p) \leq \alpha \cdot \inf_{q \in \mathcal{C}} d_{\mathrm{TV}}(p, q) + \varepsilon.$$

Definition 3 (sample complexity) *Let $\mathcal{C}$ be a distribution class. Let*

$$M_{\mathcal{C}}(\varepsilon, \delta) = \{m \in \mathbb{N} : m \textit{ guarantees } (\varepsilon, \delta)\textit{-success for } \mathcal{C}.\}$$

The realizable sample complexity $m_{\mathcal{C}} : [0,1]^2 \to \mathbb{N} \cup \{\infty\}$ *is defined by*

$$m_{\mathcal{C}}(\varepsilon, \delta) = \begin{cases} \min\{m : m \in M_{\mathcal{C}}(\varepsilon, \delta)\} & \textit{, if } M_{\mathcal{C}}(\varepsilon, \delta) \neq \emptyset \\ \infty & \textit{, otherwise.} \end{cases}$$

Similarly, for $\alpha > 1$, we define the set

$$M_{\mathcal{C}}(\varepsilon, \delta)^{\alpha} = \{m \in \mathbb{N} : m \textit{ guarantees } \alpha\textit{-agnostic } (\varepsilon, \delta)\textit{-success for } \mathcal{C}.\}$$

We then define the α-agnostic sample complexity $m^{\alpha}_{\mathcal{C}} : [0,1]^2 \to \mathbb{N} \cup \{\infty\}$ *by*

$$m^{\alpha}_{\mathcal{C}}(\varepsilon, \delta) = \begin{cases} \min\{m : m \in M^{\alpha}_{\mathcal{C}}(\varepsilon, \delta)\} & \textit{, if } M^{\alpha}_{\mathcal{C}}(\varepsilon, \delta) \neq \emptyset \\ \infty & \textit{, otherwise.} \end{cases}$$

We now define weak learnability and PAC learnability

Definition 4 (ε-weak learnability) *A class $\mathcal{C}$ is* ε-weakly learnable, *if for every $\delta \in (0,1)$, $m_{\mathcal{C}}(\varepsilon, \delta) < \infty$. A class is* α-agnostically, *ε-weakly learnable, if for every $\delta \in (0,1)$, $m^{\alpha}_{\mathcal{C}}(\varepsilon, \delta) < \infty$.*

Definition 5 (PAC learnability) *A class $\mathcal{C}$ is* PAC learnable in the realizable case, *if for every $\varepsilon, \delta \in (0,1)$, $m_{\mathcal{C}}(\varepsilon, \delta) < \infty$. A class is* α-agnostically PAC learnable, *if for every $\varepsilon, \delta \in (0,1)$, $m^{\alpha}_{\mathcal{C}}(\varepsilon, \delta) < \infty$.*

General Learning Tasks While we generally focus on distribution learning in this work, some of our results, also hold for a more general notion of statistical learning tasks. We adapt the notion of statistical learning tasks described in Lechner and Ben-David (2024). The definitions for (ε, δ)-success, sample-complexity, weak learnability and PAC learnability are analogous to the definitions for distribution learning stated above. For a detailed description of what these learning tasks consist of and these definitions, we refer the reader to Appendix A.

3. Scale sensitive dimension

In this section we define the notion of scale sensitive dimension in line with the definition of characterizing dimension proposed in Lechner and Ben-David (2024). Towards proving the main result of this paper, we first show that the existence of such a dimension implies a learnability compactness theorem. Namely, that every class that is not ε-weakly learnable must contain a countable subclass that is also not ε-weakly learnable (Lemma 9). Our main theorem, stating that the task of distribution learning cannot be characterized by a scale sensitive dimension, follows by constructing a class of distributions $\mathcal{C}_{\text{counter}}$ that violates this compactness property. We conclude the section by describing the class $\mathcal{C}_{\text{counter}}$ and sketching the proof of the non-characterizability result.
We start by recalling the definition of *scale invariant* characterizing dimension given in Lechner and Ben-David (2024).

Definition 6 (Scale invariant characterizing dimension (Definition 5 in Lechner and Ben-David (2024))) *A* scale-invariant characterization *(also called* finitary characterization of learnability *in (Lechner and Ben-David, 2024)) of a learning task is a countable set of formulas $W = \{\beta_d : d \in \mathbb{N}\}$ such that:*

1. *A class $\mathcal{C}$ is* not *learnable if and only if it satisfies all formulas in W.*

2. *For every $\beta \in W$ and every class $\mathcal{C}$ that satisfies β there exists a finite subset $\mathcal{C}_\beta \subset \mathcal{C}$, such that every $\mathcal{C}'$, if $\mathcal{C}' \supset \mathcal{C}_\beta$, then $\mathcal{C}'$ satisfies β.*

Many characterizations of common learning tasks are characterized by a dimension satisfying the definition above. For example, PAC learning of binary hypothesis classes is characterized by the VC-dimension (Blumer et al., 1989) with the formulas β_d corresponding to the statement ”The class $\mathcal{C}$ shatters a set of size d”.

In contrast, Lechner and Ben-David (2024) showed that there are many learning tasks — including regression learning and distribution learning with respect to TV-distance — that cannot be characterized by a scale invariant dimension as described in Definition 6. However, Alon et al. (1997) showed that agnostic regression learning is characterized by a scale-sensitive dimension, the fat-shattering dimension.

Definition 7 (fat-shattering dimension Kearns et al. (1994)) *A class of real-valued functions $\mathcal{F}$ is said to γ-*shatter *a set $S = \{x_1, \ldots, x_d\} \subset \mathcal{X}$ if there exists a function $s : \mathcal{X} \to \mathbb{R}$, such that for every subset $B \subset S$, there exists $f_B \in \mathcal{F}$, such that:*

- *If $x \in B$, then $f_B(x) \geq s(x) + \gamma$.*

- *If $x \in S \setminus B$, then $f_B(x) \leq s(x) - \gamma$.*

The γ-fat-shattering dimension of $\mathcal{F}$ is the largest number d, such that there exists a set S of size d that is γ-shattered by $\mathcal{F}$.

The characterization of regression learning by Alon et al. (1997) now shows that there exists $c \leq 24$, such that every class $[0, 1]$-valued class is:

- $c \cdot \gamma$-learnable, if $\mathcal{F}$ has finite γ-fat-shattering dimension.

- Not γ-learnable, if $\mathcal{F}$ does not have finite γ-fat-shattering dimension.

We see that both the notion of "shattering" as well as the type of learnability are sensitive to the "scale" γ. We now give a definition which aims to capture a general notion of scale sensitive characterizations with a dimension similar to Definition 6.

Definition 8 (Scale-sensitive dimension with slackfactor c) . *A set of formulas*[1] $W = \bigcup_{\varepsilon \in (0,1)} W_\varepsilon$, *where every* $W_\varepsilon = \{\beta_{d,\varepsilon} : d \in \mathbb{N}\}$ *is a* scale-sensitive dimension with slackfactor c *for a learning task if the following properties hold:*

1. *If a class $\mathcal{C}$ satisfies (all formulas in) W_ε, then $\mathcal{C}$ is not ε-weakly learnable.*

2. *If there exists a formula in W_ε that $\mathcal{C}$ does not satisfy, then $\mathcal{C}$ is $c \cdot \varepsilon$-weakly learnable.*

3. *Every formula in W_ε has finite evidence property, i.e., for every $\beta \in W_\varepsilon$ and every class $\mathcal{C}$ that satisfies β, there exists a finite evidence set $\mathcal{C}_\beta \subset \mathcal{C}$, such that every superset $\mathcal{C}' \supset \mathcal{C}_\beta$ satisfies β.*

We can easily verify that the fat-shattering dimension is a scale-sensitive dimension with slack factor $c \leq 24$. For the fat-shattering dimension, the formula $\beta_{\varepsilon,d}$ corresponds to the statement "There exists a set of size d, which is ε-shattered by the class $\mathcal{C}$". Thus, similar to the formulas in Definition 6, before the formulas $\beta_{\varepsilon,d}$ in Definition 8 correspond to a (scale sensitive) notion of shattering or hardness condition.

Typically, if a class C satisfies $\beta_{d,\varepsilon}$ then it also satisfies $\beta_{d',\varepsilon}$ and $\beta_{d,\varepsilon'}$ for $d' \leq d$ and $\varepsilon' \geq \varepsilon$. However, while this commonly holds for scale sensitive dimensions, we do not need this property to show our negative results. For the results in which we show the existence of a scale sensitive dimension, this monotonicity property does hold.

While a perfect scale sensitive characterization would have a slackfactor of 1 and thus characterize ε-weak learnability exactly, many known scale sensitive dimensions from the literature, such as fat shattering dimension, are only known to be scale sensitive dimensions with slack factor $c > 1$. To our knowledge every notion of scale sensitive characterizing dimensions proposed in the literature satisfies Definition 8. These include γ-Natarayan-dimension, γ-graph-dimension, γ-OIG-dimension (shown to characterize realizable regression learning (Attias et al., 2023)), k-fat-shattering dimension (characterizing agnostic k-list regression (Pabbaraju and Sarmasarkar, 2025)) and k-OIG-dimension (characterizing realizable k-list regression (Pabbaraju and Sarmasarkar, 2025)).

3.1. Countable compactness of learnability

The following lemma states that any tasks that has a scale sensitive dimension characterizing its learnability, must satisfy a "countable compactness" property of learnability.

Lemma 9 *If a learning task has a scale sensitive dimension (with any slack factor $c \geq 1$), then for every class $\mathcal{C}$ that is not $c \cdot \varepsilon$-weakly learnable, there exists a countable subclass $\mathcal{C}' \subset \mathcal{C}$ that is not ε-weakly learnable.*

[1] The term 'formula' here does not refer to a specific logical language. It stands for a property of a class. Any predicate that when applied to a class can be True or False.

Proof Let W be a scale sensitive characterization for the learning task and let $\mathcal{C}$ be a class that is not $c \cdot \varepsilon$-weakly learnable. Then $\mathcal{C}$ satisfies all the formulas $\beta_{d,\varepsilon}$ in the countable set W_ε. Now for every such $\beta_{d,\varepsilon}$ there exists a finite $\mathcal{C}_{\beta_{d,\varepsilon}} \subset \mathcal{C}$, such that every superset of $C_{\beta_{d,\varepsilon}}$ satisfies $\beta_{d,\varepsilon}$. Now consider the countable class

$$\mathcal{C}' = \bigcup_{d=1}^{\infty} C_{\beta_{d,\varepsilon}}.$$

Clearly, for every $d \in \mathbb{N}$ the class $\mathcal{C}'$ is a superset of $C_{\beta_{d,\varepsilon}}$ and therefore satisfies $\beta_{d,\varepsilon}$. Thus $\mathcal{C}'$ satisfies every formula in W_ε. By the definition of scale-sensitive dimension it follows that $\mathcal{C}'$ is not ε-weakly learnable. The same statement holds in the α-agnostic case. ■

3.2. Distribution Learning is not characterized by a scale sensitive dimension

In this subsection, we state the main result of this work, Theorem 10 and outline its proof.

Theorem 10 *The task of distribution learning, cannot be characterized by a scale sensitive dimension (for any possible slack factor $c \geq 1$).*

To show this result, we construct a class of distributions, $\mathcal{C}_{\text{counter}}$, is *not* $c \cdot \varepsilon$-weakly learnable for every c, but for which all of its countable subsets are ε-weakly learnable.

Theorem 11 *There exists a class $\mathcal{C}_{\text{counter}}$ of distributions for which the following two statements are true:*

1. *$\mathcal{C}_{\text{counter}}$ is not PAC learnable. In particular, for every learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$ and every $m \in \mathbb{N}$, there is $p \in \mathcal{C}_{\text{counter}}$ such that*

$$\mathbb{P}_{S \sim p^m}[d_{\text{TV}}(\mathcal{A}(S), p) = 1] = 1.$$

2. *Every countable subset $\mathcal{C}' \subset \mathcal{C}_{\text{counter}}$ is (realizably) learnable with sample complexity $m_{\mathcal{C}'}(\varepsilon, \delta) = 1$ for every (ε, δ). The class is also 2-agnostic PAC learnable and thus 2-agnostic ε-weakly learnable for every $\varepsilon \in (0, 1)$*

We will describe the class $\mathcal{C}_{\text{counter}}$ and outline a proof of this theorem below. Wishing to convey the main ideas concisely, we defer a full proof of this theorem to Section 6. Theorem 11 readily implies Theorem 10 as follows:

Proof of Theorem 10 Fix any constant $c \geq 1$. From Theorem 11, we know that there exists a class $\mathcal{C}_{\text{counter}}$, such that for every $\varepsilon \in (0, 1/c)$ the class $\mathcal{C}_{\text{counter}}$ is not $c \cdot \varepsilon$-weakly learnable. Furthermore, every countable subset of the class $\mathcal{C}_{\text{counter}}$ is PAC learnable and therefore ε-weakly learnable for every $\varepsilon \in (0, 1)$. Lemma 9 now implies that the task of distribution learning cannot be characterized by a scale-sensitive dimension. The result also holds for 2-agnostic ε-weak learning for every $\varepsilon \in (0, 1)$. ■

Next, we describe the class $\mathcal{C}_{\text{counter}}$.

Proof sketch of Theorem 11 We construct $\mathcal{C}_{\text{counter}}$ to be a class which satisfies the following two criteria:

(i.) For every $\varepsilon < 1$, $\mathcal{C}_{\text{counter}}$ is not ε-weakly learnable.

(ii.) For every two $p, q \in \mathcal{C}_{\text{counter}}$ with $p \neq q$, we have $d_{\text{TV}}(p, q) = 1$.

The first condition (i.) clearly implies (1.) of Theorem 11. In order to show that the second condition (ii.) implies (2.) of Theorem 11, we require the following lemma.

Lemma 12 *For a class $\mathcal{C}$, if for every $p, q \in \mathcal{C}$ with $p \neq q$, we have $d_{\text{TV}}(p, q) = 1$, then every countable subclass $\mathcal{C}' \subset \mathcal{C}$ is PAC learnable with $m_{\mathcal{C}'}(0, 0) = 1$ in the realizable case, and also 2-agnostic PAC learnable.*

The proof of this lemma can be found in Section 6.1. Thus the only remaining thing to show is that there is indeed a class $\mathcal{C}_{\text{counter}}$ which satisfies conditions (i.) and (ii.). We will now describe a class meetings those criteria.

Description of counterexample $\mathcal{C}_{\text{counter}}$ Let $\mathcal{X}_d$ be a ball of dimension d around a point $(3d, 0, 0, \ldots, 0) \in \mathbb{R}^d$ of radius 1. We consider the domain $\mathcal{X} = \bigcup_{n=2}^{\infty} \mathcal{X}_d$. We define the class $\mathcal{C}_{\text{counter}}$ as a countable union $\bigcup_{d=2}^{\infty} \mathcal{D}_d$, where every class $\mathcal{D}_d$ is defined over a subdomain $\mathcal{X}_d$. The class $\mathcal{D}_d$ consists of uniform distributions over sets $B \in \mathcal{B}_d$, where $\mathcal{B}_d$ is the set of all intersections of $\mathcal{X}_d$ with $d-1$-dimensional hyperplanes.

From this construction it follows that for every $p, q \in \mathcal{D}_d$, with $p \neq q$ we have $d_{\text{TV}}(p, q) = 1$. First, if we take the intersection of any two distinct subsets of $\mathcal{B}_d$, they can be described as an intersection of $\mathcal{X}_d$ and a $d-2$-dimensional hyperplane and thus is a set of measure 0 with respect to both p and q. Furthermore, if $p, q \in \mathcal{C}_{\text{counter}}$ come from distinct sets $\mathcal{D}_d$, $\mathcal{D}_{d'}$ with $d' \neq d$ their supports are disjoint. Thus, condition (ii.) is satisfied and therefore, via Lemma 12, the class $\mathcal{C}_{\text{counter}}$ satisfies condition (2.) of Theorem 11.

Now to show the remaining condition (1) in Theorem 11, we argue, that for every sample size m, the subclass $\mathcal{D}_{m+2} \subset \mathcal{C}_{\text{counter}}$ is not learnable with sample complexity m. Intuitively, this hardness follows from the observation, that every sample $S \sim p^m$ with $p \in \mathcal{D}_d$ and $m \leq d-2$, lies in the intersection of $\mathcal{X}_d$ and a $d-2$ dimensional hyperplane. Thus the sample S is consistent with an uncountable subset of $\mathcal{B}_d$ and thus it is consistent with an uncountable subset $\mathcal{D}_S \subset \mathcal{D}_d$. Since every two elements of $\mathcal{D}_S \subset \mathcal{D}_d$ have total variation distance 1, for any choice of $\mathcal{A}(S)$ the learner will pick a bad distribution for a majority of consistent distributions. To make this intuition precise and show hardness for learning $\mathcal{C}_{counter}$ formally we will use a lower bound from Section 5, which requires the definition of a meta-distribution over elements of $\mathcal{C}_{\text{counter}}$. The formal description of the class $\mathcal{C}_{counter}$ as well as the full proof of Theorem 11 will be shown in Section 6.

4. Scale-sensitive characterization in the countable regime via sample complexity compactness

In this section, we discuss sample complexity compactness and a weaker version, pseudo-compactness. We show that for many learning tasks — including distribution learning — satisfying (pseudo) sample complexity compactness, implies that the task *can* be characterized by a scale-sensitive dimension. We finish this section by a discussion of sample complexity compactness of distribution learning. We first show that the class $\mathcal{C}_{\text{counter}}$ defined in the previous section shows that the task of distribution learning does not satisfy (pseudo)-compactness. However, if we restrict our attention to distribution learning of countable classes over countable domains, we find that this learning problem satisfies 2-pseudo compactness. Our other results in this section then imply that we can define a scale-sensitive characterizing dimension for distribution learning of countable classes over countable domains in line with Definition 8. Furthermore, we also discuss the possibility of boosting δ and its implications for a quantitative scale sensitive dimension.

We will now state the definition of sample-complexity compactness.

Definition 13 (Sample complexity compactness) *A learning task is said to satisfy* sample complexity compactness *if for every class $\mathcal{C}$ and every $m \in \mathbb{N}$ and every $\varepsilon, \delta \in \mathbb{N}$, the following two statements are equivalent:*

- $m_{\mathcal{C}}(\varepsilon, \delta) \leq m$.
- *For every finite subset $\mathcal{C}' \in \mathcal{C}$ we have $m_{\mathcal{C}'}(\varepsilon, \delta) \leq m$.*

We can also define a weaker version, where a uniform bound on the sample complexity for all finite classes does imply a certain level of learning success for the whole class for samples of the same size, that does not match the accuracy parameter exactly.

Definition 14 (c-pseudo sample complexity compactness) *A learning task is said to satisfy c-*pseudo sample complexity compactness *if for every class $\mathcal{C}$ and every $m \in \mathbb{N}$ and every $\varepsilon, \delta \in \mathbb{N}$, if for every finite subset $\mathcal{C}' \in \mathcal{C}$ we have $m_{\mathcal{C}'}(\varepsilon, \delta) \leq m$, then $m_{\mathcal{C}}(c\varepsilon, \delta) \leq m$. :*

These notions are in line with Asilis et al. (2024) who defined sample compexity compactness for classification settings and showed both compactness and 2-pseudocompactness results for transductive learning.

It is easy to verify that the sample complexity compactness defined before corresponds to 1-pseudo sample complexity compactness.

Monotonicity of sample complexity We say a learning task satisfies *sample complexity monotonicity* if for every $\mathcal{C}' \subset \mathcal{C}$ and every ε, δ, we have $m_{\mathcal{C}'}(\varepsilon, \delta) \leq m_{\mathcal{C}}(\varepsilon, \delta)$. Most (improper) statistical learning tasks, including distribution learning, satisfy sample complexity monotonicity. This follows directly from the definition of PAC learning (when defined without additional constraints such as properness)

We will now show that scale-sensitive characterizing dimensions and (pseudo) sample complexity compactness are closely linked, via the following lemma.

Theorem 15 *Consider a learning task for which the following statements are true:*

- *Every finite class is α-agnostically learnable.*
- *The learning task satisfies c'-pseudo sample complexity compactness.*
- *The learning tasks satisfies sample-complexity monotonicity.*

Then this learning task is characterized by a scale-sensitive dimension.

In particular, consider $W = \bigcup_{\varepsilon \in (0,1)} W_\varepsilon$ with $W_\varepsilon = \{\beta_{d,\varepsilon} : d \in \mathbb{N}\}$,
where $\beta_{d,\varepsilon}$ expresses the statement

“$\exists$ finite $\mathcal{C}_d \subset \mathcal{C}$ with $m_{\mathcal{C}_d}(\varepsilon, 1/3) \geq d$.”

For every $c > \alpha \cdot c'$, W is a scale-sensitive dimension with slack factor c.

In particular, this implies the following result for distribution learning of countable classes over countable domains.

Theorem 16 *Let $c > 4$. The task of (realizable) learning of countable distribution classes over a countable domain is characterized by a scale sensitive dimension with slack factor c.*

This result follows directly from Theorem 20, Theorem 15 and Lemma 17 below.

Before proving Theorem 15 in Subsection 4.2, we will show a lemma about boosting δ in the next subsection.

4.1. δ-boosting

In this subsection, we will discuss how to boost a learning guarantee for a fixed pair (ε, δ) to learning guarantees for arbitrarily small δ'. More precisely, we will show that if a sample size m guaranees (ε, δ)-success for a given (ε, δ), we can show learning guarantees for pairs $(c\varepsilon, \delta')$ for arbitrarily small δ' and fixed c.

For the task of distribution learning, we get the following result.

Lemma 17 *Let $\mathcal{C}$ be a class of distributions. If there exists a learner $\mathcal{A}$ and parameters $\delta', \eta \in (0,1)$ and $m \in \mathbb{N}$, such that a sample size m guarantees $(\frac{\varepsilon - \eta}{2}, \delta')$-success, then $\mathcal{C}$ is ε-weakly learnable with sample complexity*

$$m_{\mathcal{C}}(\varepsilon, \delta) \leq mn + \tilde{O}\left(\frac{\log(n) \cdot \log(2/\delta)}{\eta^2}\right),$$

where $n = \log_{\delta'}(\frac{\delta}{2})$.

Furthermore, for general learning tasks which have a finite sample complexity bound for all finite classes of size k, we get the following result.

Lemma 18 (δ-boosting) *Consider a learning task for which there exists a function $m^{\alpha}_{\mathit{finite}} : \mathbb{N} \times (0,1)^2 \to \mathbb{N}$, such that every finite class $\mathcal{C}'$ with $|\mathcal{C}'| = k$ is α-agnostically learnable with sample complexity $m^{\alpha}_{\mathcal{C}'}(\varepsilon, \delta) \leq m^{\alpha}_{\mathrm{finite}}(k, \varepsilon, \delta)$ for every $\varepsilon, \delta \in (0,1)$.*

For any fixed ε, let $\mathcal{C}$ be a concept class for which there exists $\delta', \eta \in (0,1)$ such that a sample size m suffices for $\left(\frac{\varepsilon}{\alpha} - \eta, \delta'\right)$-success. Then, $\mathcal{C}$ is ε-weakly learnable with

$$m_C(\varepsilon, \delta) \leq mn + m_{\mathrm{finite}}\left(n, \eta, \frac{2}{\delta}\right),$$

where $n = \log_{\delta'}(\frac{\delta}{2})$.

The result follows from a standard argument for δ-boosting. A sample S is split into $n+1$ subsamples $S_{1,1},\dots,S_{1,n},S_2$. The samples $S_{1,i}$ are each of size m and are fed into a learner satisfying the $(\frac{\varepsilon-\eta}{2},\delta')$-success guarantee. The resulting outputs are then used to create a finite set $\mathcal{C}'$ of candidate hypothesis. Lastly, we use the last subsample S_2 to do hypothesis selection for $\mathcal{C}'$. The full proof can be found in Appendix B.

We note that most common statistical learning problems satisfy the condition that finite hypothesis classes are α-agnostically learnable for some $\alpha > 1$. In particular, this is the case for distribution learning and $\alpha = 2$ due to Bousquet et al. (2019, 2022). This directly implies Lemma 17.

4.2. Proof of Theorem 15

We will now prove Theorem 15.

Proof Let us assume that the learning task we consider satisfies the conditions listed above.

We first note, that the set W_ε is clearly countable.

We now want to show that if $\mathcal{C}$ satisfies W_ε, then $\mathcal{C}$ is not ε-weakly learnable. Clearly, if for every $d \in \mathbb{N}$, there exists a finite $\mathcal{C}_d \subset \mathcal{C}$, with $m_{\mathcal{C}_d}(\varepsilon, 1/3) \geq d$, then due to sample complexity monotonicity, we have $m_{\mathcal{C}}(\varepsilon, 1/3) \geq \sup_{d\in\mathbb{N}} m_{\mathcal{C}_d}(\varepsilon, 1/3) = \infty$. Thus $\mathcal{C}$ is not ε-weakly learnable.

For the other direction, we note, that if $\mathcal{C}$ does not satisfy W_ε, then there exists some $d \in \mathbb{N}$, such that $\mathcal{C}$ does not satisfy $\beta_{d,\varepsilon}$. Thus for every finite subset $\mathcal{C}' \subset \mathcal{C}$, we know $m_{\mathcal{C}}(\varepsilon, 1/3) < d$. Since the learning task satisfies c'-pseudo-compactness, we can infer that i.e., $m_{\mathcal{C}}(c' \cdot \varepsilon, 1/3) < d$. From Lemma 18 it now follows that $\mathcal{C}$ is $(\alpha \cdot c' \cdot \varepsilon + \eta')$-weakly learnable, for every $\eta' > 0$. Thus $\mathcal{C}$ is $c \cdot \varepsilon$-weakly learnable for every $c > \alpha \cdot c'$.

Lastly W clearly satisfies the finite evidence set property, as $\beta_{d,\varepsilon}$ is defined via the existence of a finite evidence set. ■

4.3. Pseudo-compactness and Distribution Learning

We now want to consider the implications of the above results to distribution learning and discuss whether distribution learning satisfies pseudo sample complexity compactness.

On the one hand the existence of the class $\mathcal{C}_{\text{counter}}$ and Theorem 11 imply the following corollary for pseudo-compactness

Corollary 19 *The task of (realizable) distribution learning does not satisfy c-pseudo sample complexity compactness for any $c \in \mathbb{R}$.*

Proof Fix any $c \geq 1$. Assume by way of contraction that the task of distribution learning would satisfy c-pseudo sample complexity compactness. Let $\varepsilon = \delta = \frac{1}{2c}$. From Theorem 11, we know that there is a class $\mathcal{C}_{\text{counter}}$, such that every countable subclass $\mathcal{C}' \in \mathcal{C}_{\text{counter}}$ we have $m_{\mathcal{C}'}(\varepsilon, \delta) = 1$. This guarantee also holds for every finite subset of $\mathcal{C}_{\text{counter}}$ by sample complexity monotonicity. Thus c-pseudo sample complexity robustness implies that $m_{\mathcal{C}_{\text{counter}}}(c{\cdot}\varepsilon, \delta) = m_{\mathcal{C}_{\text{counter}}}(1/2, 1/c) = 1$.

However from Theorem 11 we know for every $\varepsilon', \delta' \in (0,1)$, we have $m_{\mathcal{C}_{\text{counter}}}(\varepsilon', \delta') = \infty$. Thus we get a contradition, proving our result.

■

However, we note that the class $\mathcal{C}_{\text{counter}}$ is a continuous class over a continuous domain. It is natural to ask, whether pseudo-compactness holds when we restrict our attention the the countable regime. Indeed, we can show, that for countable domains and countable classes pseudo-compactness holds.

Theorem 20 *The task of (realizable) learning of countable distribution classes over a countable domain satisfies* 2*-pseudo sample complexity compactness.*

The proof of this result can be found in the appendix.

This result then implies the existence for a scale-sensitive dimension characterizing the learnability of countable classes over countable domains, as stated in Theorem 16.

4.4. Implications for quantitative scale-sensitive dimensions

The work of Lechner and Ben-David (2024) also considered quantitative characterizing dimensions. i.e., dimensions that yield a bound on the sample complexity of learning. In particular they considered the following definition.

Definition 21 A strong scale-invariant sample complexity dimension *is a mapping from* d *that takes as input a class* $\mathcal{C}$ *and outputs a value in* $\mathbb{N} \cup \{\infty\}$*, such that a class* $\mathcal{C}$ *of models is PAC learnable if and only if* $d(\mathcal{C}) \neq \infty$ *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* $\mathcal{C}$*,* $m_{\mathcal{C}}(\varepsilon, \delta) \leq f(d(\mathcal{C}))g(\varepsilon, \delta)$ *for all* $(\varepsilon, \delta) \in (0, 1)$*. 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.*

The most commonly discussed dimensions in the literature are both qualitative (learnability characterizing) and quantitative (sample-complexity characterizing) dimensions. The results of Lechner and Ben-David (2024) showed that there can not be a dimension characterizing sample-complexity of distribution learning (not even in a weaker version). This result relied on showing that the sample complexity for a pair (ε, δ) does not give any guarantees for the sample complexity for (ε', δ') with $\varepsilon' < \varepsilon$ and $\delta' \approx \delta$. This was then used to show that ε-dependent sample-complexity behaviours of distribution learning are essentially *too rich to be captured by the natural numbers*. We now want to address the potential of a *quantitative scale-sensitive dimension*, i.e. a scale-sensitive dimension, characterizing the sample complexity of learning. A natural scale-sensitive extension of Definition 21 would allow the mapping d to also depend on ε as a second argument. That is, a scale-sensitive sample-complexity dimension, would consist of a mapping d_ε and a mapping g, such that for every ε-weakly learnable class $\mathcal{C}$, we have

$$m_{\mathcal{C}}(\varepsilon, \delta) \leq d_\varepsilon(\mathcal{C}) \cdot g(\varepsilon, \delta).$$

We will now show that we get something close to such a dimension for learning tasks with pseudo-compactness that allow for δ-boosting: If a class $\mathcal{C}$ is ε-weakly learnable, then $\mathcal{C}$ is $c\varepsilon$-weakly learnable, with sample complexity

$$m_{\mathcal{C}}(c\varepsilon, \delta) \leq m_{\mathcal{C}}(\varepsilon, 1/2)\log_2(2/\delta) + m_{\text{finite}}\left(\log_2(1/\delta), \varepsilon, \frac{2}{\delta}\right)$$

for $c \geq 2\alpha$. Thus, for $d_{c\varepsilon}(\mathcal{C}) := m_{\mathcal{C}}(\varepsilon, 1/2)$ and $g(c\varepsilon, \delta) := \log_2(2/\delta) + m_{\text{finite}}(\log_2(1/\delta), \varepsilon, \frac{2}{\delta})$ we get

$$m_{\mathcal{C}}(c\varepsilon, \delta) \leq d_{c\varepsilon}(\mathcal{C}) \cdot g(c\varepsilon, \delta).$$

If $\mathcal{C}$ is a PAC learnable class and thus ε-weakly learnable for every $\varepsilon \in (0, 1)$, then this bound holds for all values of ε. If a class $\mathcal{C}$ is ε-weakly for $\varepsilon \in [\varepsilon_0, 1)$, then the above argument only gives an upper bound for $\varepsilon \in [c\varepsilon_0, 1)$. Thus, for $\varepsilon \in [\varepsilon_0, c\varepsilon_0)$ the class $\mathcal{C}$ could still be ε-learnable, while not showing easily characterizable sample-complexity behaviour captured by d_ε. We thus "almost" have a (trivial) quantitative scale-sensitive dimension, characterizing the sample complexity of ε-weak learning. The only "gap" in explaining the behaviour of sample complexity via these means exists for not PAC-learnable classes. For these classes the range of unexplained behaviour sits in a narrow interval around the transition from weak learnablity to non-learnablity. This is "gap" in the characterization of learning is in line with our notion of "slack" in Definition 8. For both scale-sensitive characterizations — qualitative and quantitative — the learning behaviour for most values of ε can be inferred from the characterization, however there remains a range of values for ε for which the corresponding learning behaviour is inaccessible to these notions.

5. Lower Bounds for Distribution Learning

In order to show Theorem 11 formally, we need to show that the class $\mathcal{C}_{\text{counter}}$ is not learnable. To this end, we will introduce a general lower bound technique (Theorem 22) for distribution learning. This technique is inspired by standard arguments of no-free-lunch theorems. The idea is to define a meta-distributions (denoted by Q) over a class of distributions. The meta-distribution is picked in such a way that all the support elements have high total variation distance from each other. Moreover, the meta-distribution causes a high level of indistinguishability between possible candidates. In particular, we consider the posterior distribution over possible generating distributions for a given sample S, with the meta-distribution Q as a prior. We pick Q in such a way that these posteriors cover a large (uncountable) space of potential candidates (all of which have large pairwise total variation distance). This inability to distinguish between candidates that no output-distribution could be simultaneously close to then implies the hardness results.

Let $\mathcal{C}$ be a class of distributions. Let Q be a meta-distribution over elements of $\mathcal{C}$. We define $|Q|^m$ to be the distribution over samples S of size m that results from first samples $q \sim Q$ and then $S = S_q \sim q^m$. Note that $|Q|^m$ can not be understood as an i.i.d distribution.

Let us denote the random variable $q_Q \sim Q$ and $S_Q \sim |Q|^m$. Formally, we consider a joint probability distribution $Q_{q,S}$ defined over $\mathcal{C} \times \mathcal{X}^m$, where for $q \subset \mathcal{C}$ and $S \subset \mathcal{X}^m$, with density

$$f_{q_Q, S_Q}(q, S) = f_Q(q) \cdot f_{q^m}(S).$$

We can now define the conditional probability distribution

$$\kappa_{q_Q|S_Q}(B_{\mathcal{C}}, S) = \frac{\int_{B_{\mathcal{C}}} f_{q_Q, S_Q}(q, S)}{\int f_{q_Q, S_Q}(q, S) dq}$$

Theorem 22 *Let $\mathcal{C}$ be a class of distributions. If for $m \in \mathbb{N}$, there exists a meta distribution Q over elements in $\mathcal{C}$, such that for every $q, p \in \operatorname{supp}(Q)$ with $p \neq q$, we have $d_{\mathrm{TV}}(p, q) = 1$ and for every $s \in \operatorname{supp}(|Q|^m) \subset \mathcal{X}^m$, for the conditional probability distribution over elements of $\mathcal{C}$ with Q as prior, conditioned on a sample $S_Q \sim |Q|^m$, we have for every finite subset $B_{\mathcal{C}} \subset \mathcal{C}$ and every $S \in \mathcal{X}^m$ we have*

$$\kappa_{q_Q|S_Q}(B, S) = 0,$$

then for every learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$, there exists $q \in \mathcal{C}$ with

$$\mathbb{P}_{S \sim q^m}[d_{\mathrm{TV}}(\mathcal{A}(S), q) = 1] = 1.$$

If for every $m \in \mathbb{N}$ such a meta-distribution exists, then $\mathcal{C}$ is not ε-weakly learnable for any $\varepsilon \in [0, 1]$.

In order for this bound to be applicable, we need to consider meta-distributions for which the above notions are well-defined. In this work, we will only consider meta-distributions that can be defined as pushforward measures of Lebesque measures via one-to-one mappings. This method bears some similarty to common lower bound techniques such as Fano's inequality, Le Cam's method and Assouad's inequality. In the appendix we will discuss the differences of our bound to these techniques. Crucially, these known techniques would all fail to show Theorem 11. The proof of this theorem can be found in Appendix C.

6. Learnability of distribution classes over continuous domains cannot be characterized by scale sensitive dimensions.

In this section we will go into more detail about the construction of the class $\mathcal{C}_{\text{counter}}$ proof of Theorem 11, which were already described in Section 3.2 and which were used there to prove that distribution learning cannot be characterized by a scale-sensitive dimension (Theorem 10).

We start this section by restating the main theorem of this section, Theorem 11.

Theorem 11 There exists a class $\mathcal{C}_{\text{counter}}$ of distributions for which the following two statements are true:

1. For every learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$ and every $m \in \mathbb{N}$, there is $p \in \mathcal{C}_{\text{counter}}$ such that

$$\mathbb{P}_{S \sim p^m}[d_{\mathrm{TV}}(\mathcal{A}(S), p) = 1] = 1.$$

2. Every countable subset $\mathcal{C}' \subset \mathcal{C}_{\text{counter}}$ is (realizably) learnable with sample complexity $m_{\mathcal{C}'}(\varepsilon, \delta) = 1$ for every (ε, δ). Every countable $\mathcal{C}' \subset \mathcal{C}_{\text{counter}}$ is also 2-agnostic learnable.

Before defining $\mathcal{C}_{\text{counter}}$ formally, we first prove Lemma 12, which will be useful, when showing that $\mathcal{C}_{\text{counter}}$ satisfies condition (2) of Theorem 11.

6.1. Proof of Lemma 12

Lemma 12 For a class $\mathcal{C}$, if for every $p, q \in \mathcal{C}$ with $p \neq q$, we have $d_{\mathrm{TV}}(p, q) = 1$, then every countable subclass $\mathcal{C}' \subset \mathcal{C}$ is realizably PAC learnable with $m_{\mathcal{C}}(0, 0) = 1$ and 2-agnostically PAC learnable, thus also 2-agnostically ε-weakly for every $\varepsilon \in (0, 1)$.
Proof Let $\mathcal{C}' \subset \mathcal{C}$ be countable. Fix an arbitrary enumeration $\mathcal{C}' = \{p_i : i \in \mathbb{N}\}$. Let $p_i, p_j \in \mathcal{C}'$ with $i \neq j$. From our assumption, we have

$$d_{\mathrm{TV}}(p_i, p_j) = 1.$$

For every p, q with $d_{\mathrm{TV}}(p, q) = 1$, there exists a set B such that $|p(B) - q(B)| = 1$. Thus, for the set B we have either $p(B) = 1$ and $q(B) = 0$ (and thus $q(\mathcal{X} \setminus B) = 1$) or $q(B) = 1$ and $p(B) = 0$.

For every $i, j \in \mathbb{N}$ with $i \neq j$, we can therefore choose sets $E_{i,j}$ and $E_{j,i}$ as sets meeting the following criteria:

- $p_i(E_{i,j}) = 1$,
- $p_j(E_{j,i}) = 1$,
- and $E_{i,j} \cap E_{j,i} = \emptyset$.

Define $E_i = \bigcap_{j=1}^{\infty} E_{i,j}$. Clearly $p_i(E_i) = 1$. Furthermore, for any two $i \neq j$, $E_i \cap E_j = \emptyset$. Now consider the learner $\mathcal{A} : \mathcal{X}^m \to \Delta(\mathcal{X})$, that is defined by

$$\mathcal{A}(S) = \begin{cases} p_i & \text{if for all either} j \in \mathbb{N} : |S \cap E_i| \geq |S \cap E_j| \text{ and } j > i \text{ or } |S \cap E_i| > |S \cap E_j| \\ p \text{ arbirary}, & \text{if } S \cap E_j = \emptyset \text{ for all} j \in \mathbb{N} \end{cases}$$

Now fix arbitrary $p_i \in \mathcal{C}'$. With probability 1 over $x \sim p_i$, $x \in E_i$. Thus for $m = 1$

$$\mathbb{P}_{S \sim p_i^m}[d_{\mathrm{TV}}(\mathcal{A}(S), p_i) = 0] = 1.$$

For the 2-agnostic case, we note that the required learning guarantee is only non-vacuuous, if the data-generating distribution p satisfies $d_{\mathrm{TV}}(p, p_i) < \frac{1}{2}$ for some $p_i \in \mathcal{C}'$. The probability that $\mathcal{A}$ outputs hypothesis p_i (which is the optimal hypothesis in $\mathcal{C}'$ in this instance) on a sample of size 1 is thus at least 1/2. Thus we have $m^2_{\mathcal{C}'}(0, 1/2)$. This now implies that $\mathcal{C}'$ is 2-agnostic PAC learnable and thus weak learnable for every $\varepsilon \in (0, 1)$ via δ-boosting (analogous to the proof of Lemma 17). ■

Counter example class We start this section by restating the definition of the class $\mathcal{C}_{\text{counter}}$. We consider the domain $\mathcal{X} = \bigcup_{d \in \mathbb{N}} \mathcal{X}_d$, where $\mathcal{X}_d$ is the ball with radius 1 in dimension d around the point $(3d, 0, \dots, 0)$, i.e., $\mathcal{X}_d = \{(x_1, \dots, x_d) \in \mathsf{R}^d : (x_1 - 3d)^2 + \sum_{i=2}^d x_i^2 \leq 1\}$. For every dimension d, we consider the collection $\mathcal{B}_d$ of sets B that are intersections of $\mathcal{X}_d$ with a $d-1$ dimensional hyperplane (note that each $B \in \mathcal{B}_d$ is a $d-1$ dimensional ball). We define the class of distributions $\mathcal{D}_d = \{q_B : B \in B_d\}$, where q_B is the uniform distribution on B with respect to Lebesgue measure. Let $\mathcal{C}_{\text{counter}} = \bigcup_{d=1}^{\infty} \mathcal{D}_d$. We note that any $B \in \mathcal{B}_d$ can be understood as a ball dimension $d-1$. The density function of q_B for a given $B \in \mathcal{B}_d$ is given by

$$f_{q_B} : x \mapsto \frac{1}{\mathrm{vol}_{d-1}(B)} 1[x \in B].$$

This class class is constructed in such a way that for any two distributions $q_{B_1}, q_{B_2} \in \mathcal{C}_{\text{counter}}$ their total variation distance is 1. This holds, since the intersection between any two distinct sets $B_1, B_2 \in \mathcal{B}_d$ has $d-1$-dimensional volume 0. Thus for both distributions $q_{B_1}(B_1 \cap B_2) = q_{B_2}(B_1 \cap B_2) = 0$. It immediately follows that $d_{\mathrm{TV}}(q_{B_1}, q_{B_2}) = 1$. With the use of Lemma 12 this shows that $\mathcal{C}_{\text{counter}}$ satisfies condition (2) of Theorem 11.

To satisfy (1), we will further show, that for every $m \in \mathbb{N}$, we can define a meta-distribution according to the conditions of Theorem 22. Given a sample size m, we will construct the meta-distribution Q as a uniform distribution over elements of $\mathcal{D}_{m+2}$. Since a sample S of size m is guaranteed to be contained in a m-dimensional hyperplane, there is an uncountable number of distributions in the support of Q that are consistent with S. This is sufficient to satisfy the conditions of Theorem 22.

We will now state the proof of Theorem 11.

6.2. Proof of Theorem 11

Proof We start by arguing that every two elements $p, q \in \mathcal{C}_{\text{counter}}, p \neq q$ have total variation distance $d_{\text{TV}}(p, q) = 1$. We note, that if $p \in \mathcal{D}_{d_1}$ and $p \in \mathcal{D}_{d_2}$ for $d_1 \neq d_2$, then p and q are defined over different subdomains. In this case, we clearly have $d_{\text{TV}}(p, q) = 1$. Thus, it remains to show that $d_{\text{TV}}(p, q) = 1$ for $p, q \in \mathcal{D}_d$ for some $d \in \mathbb{N}$. We note, that in this case these distributions are of the form $p = q_{B_1}$ and $q = q_{B_2}$ for distinct $B_1, B_2 \in \mathcal{B}_d$. Thus, either $B_1 \cap B_2 = \emptyset$ or $B_1 \cap B_2$ is the intersection of $\mathcal{X}_d$ with a $d-2$-dimensional hyperplane. In either case we have

$$\text{vol}_{d-1}(B_1 \cap B_2) = 0.$$

Thus, $q_{B_1}(B_1 \cap B_2) = 0$. Therefore,

$$d_{\text{TV}}(q_{B_1}, q_{B_2}) = 1.$$

Thus, $\mathcal{C}_{\text{counter}}$ consists of distributions with pairwise TV-distance 1. Using Lemma 12, we can thus conclude that countable subclasses of $\mathcal{C}_{\text{counter}}$ are learnable. Thus (2.) is satisfied.

It remains to show that condition (1.) is satisfied as well, namely, that for every learner $\mathcal{A}$ there exists a distribution p, on which the learner's output has $d_{\text{TV}}(\mathcal{A}(S), p) = 1$ with probability 1. We show this using Theorem 22. We need to show that for every $m \in \mathbb{N}$, there exists a meta-distribution Q such that for all $p, q \in \text{supp}(Q)$, we have $d_{\text{TV}}(p, q) = 1$ and for every finite subset $C \subset \text{supp}(Q)$ and every $S \in \text{supp}(|Q|^m)$, we the posterior distribution for q_Q given Q satisfies

$$\kappa_{q_Q|S_Q}(C, S) = 0.$$

We already know that pairwise TV-distances in $\mathcal{C}$ are 1, thus it satisfies, to show the condition on the posterior.

Given m, we will define a meta distribution Q over elements in $\mathcal{D}_d$ with $d = m+2$. Let $r(B)$ be the radius of a given set B. We note, that we can identify each set B with $0 < r(B) < 1$ uniquely by its center point. We can thus define $x \in \mathcal{X}^d \neq \{0_d\}$ $B(x) = \{B \in \mathcal{B}_d : x \text{ is the center point of} B\}$. We choose the distribution Q' over center points to be a uniform distribution over $\mathcal{X}_d$ with respect to Lebesgue measure normalized by the radius of the ball X_d in B. We then can pick q_B by randomly selecting a point in the d-dimensional ball $\mathcal{X}_d$.

Thus, for a given sample S we get the posterior distribution

$$f_{Q'} : x \mapsto \frac{1}{\text{vol}_d(\mathcal{X}_d)} 1[x \in \mathcal{X}^m]$$

We then define the meta-distribution Q as picking $q_Q = q_{B(x)}$ according to $x \sim Q'$.

Now, for a sample $S = (z_1, \ldots, z_m)$ with $m = d - 2$, we want to calculate the posterior $\kappa_{q_Q|S_Q}(\cdot, S)$. We first note that the density of S with respect to a given $q_{B(x)}$ is

$$f_{q^m_{B(x)}}(S) = \frac{1}{\text{vol}_d(\mathcal{X}_d)} 1[x \in \mathcal{X}^d] \cdot \prod_{i=1}^{m} \frac{1}{\text{vol}_{d-1}(B(x))} 1[z_i \in B(x)].$$

Thus,

$$f_{q_Q, S_Q}(q, S) = \frac{1}{\text{vol}_d(\mathcal{X}_d)} 1[q = q_{B(x)} \text{ and } x \in \mathcal{X}^d] \prod_{i=1}^{m} \frac{1}{\text{vol}_{d-1}(B(x))} 1[z_i \in B(x)].$$

Thus for the posterior, we have

$$\kappa_{q_Q|S_Q}(C,(z_1,\dots,z_m)) = \frac{\int_C f_{q_Q,S_Q}(q,(z_1,\dots,z_m))dq}{\int_{\mathcal{C}_{\text{counter}}} f_{q_Q,S_Q}(q,(z_1,\dots,z_m)dq}$$

$$\frac{\int_C \frac{1}{\text{vol}_d(\mathcal{X}_d)} 1[q = q_{B(x)} \wedge x \in \mathcal{X}^d] \prod_{i=1}^m \frac{1}{\text{vol}_{d-1}(B(x))} 1[z_i \in B(x)]dq}{\int_{\mathcal{C}_{\text{counter}}} \frac{1}{\text{vol}_d(\mathcal{X}_d)} 1[q = q_{B(x)} \wedge x \in \mathcal{X}^d] \prod_{i=1}^m \frac{1}{\text{vol}_{d-1}(B(x))} 1[z_i \in B(x)]dq}$$

We note, that the points $z_1, \dots, z_m$ must be contained in the intersection of $\mathcal{X}_d$ with a a m-dimensional hyperplane. Let us denote this intersection with L. For every such L, the set $\{B \in \mathcal{B}_d : L \subset B\}$ has uncountable cardinality, since $m < d - 1$. Furthermore, every distribution $q_B \in \mathcal{D}_d$ that is consistent with S is given positive density by the posterior distribution. And furthermore, the densities between two different elements in the support of the posterior differs by at most a real-valued factor. It follows that for every finite subset $C \subset \text{supp(Q)}$, we have $\kappa_{q_Q|S_Q}(C,(z_1,\dots,z_n)) = 0$.

Thus, by Theorem 22, for every learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$ and every $m \in \mathbb{N}$, there exists $p \in \mathcal{C}_{\text{counter}}$ with

$$\mathbb{P}_{S\sim p^m}[d_{\text{TV}}(\mathcal{A}(S), p) = 1] = 1.$$

■

Acknowledgments

Tosca Lechner was supported by a Vector Postdoctoral Fellowship. Shai Ben-David was supported by NSERC through a Discovery grant and by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring 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.

Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18, 2018.

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Transductive learning is compact. In *Advances in Neural Information Processing Systems 38, NeurIPS 2024*, 2024.

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Proper learnability and the role of unlabeled data. In *International Conference on Algorithmic Learning Theory, 2025*, volume 272 of *Proceedings of Machine Learning Research*, pages 112–133. PMLR, 2025.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: PAC learning and online learning. In *NeurIPS 2023*, 2023.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/032.pdf#page=1.

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.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Olivier Bousquet, Daniel M. Kane, and Shay Moran. The optimal approximation factor in density estimation. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 318–341, 2019.

Olivier Bousquet, Mark Braverman, Gillat Kol, Klim Efremenko, and Shay Moran. Statistically near-optimal hypothesis selection. In *Proceedings of the 62nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '21, pages 909–919. IEEE Computer Society, 2022.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *Foundations of Computer Science, FOCS*, 2022.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory, COLT*, 2014.

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.

Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing*, STOC '23, New York, NY, USA, 2023. ACM.

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Towards efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.

Tosca Lechner and Shai Ben-David. Inherent limitations of dimensions for characterizing learnability of distribution classes. In *The Thirty Seventh Annual Conference on Learning Theory, 2024*, volume 247 of *Proceedings of Machine Learning Research*, pages 3353–3374. PMLR, 2024.

Tosca Lechner and Shai Ben-David. Inherent limitations of dimensions for characterizing learnability of distribution classes. In *COLT*, 2024.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285–318, 1987.

Omar Montasser, Steve Hanneke, and Nati Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. In *Advances in Neural Information Processing Systems*, 2022.

Chirag Pabbaraju and Sahasrajit Sarmasarkar. A characterization of list regression. In Gautam Kamath and Po-Ling Loh, editors, *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 870–920. PMLR, 24–27 Feb 2025.

Yannis G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Appendix A. General Learning Tasks

While we generally focus on distribution learning in this work, some of our results, also hold for a more general notion of statistical learning tasks. We adapt the notion of statistical learning tasks described in Lechner and Ben-David (2024). A learning problem consists of the following elements:

- a domain $\mathcal{Z}$ from which the input-instances/training-instances are sampled
- A class of benchmark models $\mathcal{C}$
- 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, $\mathcal{C} \subseteq \mathcal{F}$.
- 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 $\mathcal{C}$ w.r.t. some data generating distribution p is defined as $\mathrm{opt}(\mathcal{C}, p) = \inf_{h \in \mathcal{C}} L(h, p)$.

A *learning task* is defined as a collection of learning problems. Usually they are grouped together by the loss function, e.g. "distribution learning with respect to total variation distance" only specifies L to be d_{TV} and $\mathcal{C} = \mathbb{P} \subset \mathcal{F} = \Delta(\mathcal{Z})$. However learning tasks can be restricted further, by assuming additional requirements, such as realizability, properness, ect. In this work, we also consider the learning task of "distribution learning of countable classes over countable domains (with respect to total variation distance)", which corresponds to including only countable $\mathcal{Z}$ and $\mathcal{C}$ in the collection of learning problems.

We define the notions of (ε, δ)-success, sample complexity, PAC learning and weak learning analogous to the definitions in the previous subsection.

Definition 23 (**(ε, δ)-success**) *Let $\mathcal{C} \times \mathcal{P}$ be a pair of classes and $L : \mathcal{F} \times \Delta(\mathcal{Z}) \to \mathbb{R}_0^+$ be a loss function. We say a* sample size m guarantees (ε, δ)-success *for $\mathcal{C} \times \mathcal{P}$ with respect to L if there exists a learner $\mathcal{A} : \mathcal{Z}^* \to \mathcal{F}$ such that for every $p \in \{p' \in \mathcal{P} : \mathrm{opt}(\mathcal{C}, p) = 0\}$, with probability $1 - \delta$ over $S \sim p^m$ we have*

$$L(\mathcal{A}(S), p) \leq \varepsilon.$$

We say a sample size m guarantees α-agnostic (ε, δ)-success for $\mathcal{C} \times \mathcal{P}$ with respect to L *if there exists a learner $\mathcal{A} : \mathcal{Z}^* \to \mathcal{F}$ such that for every $p \in \mathcal{P}$, with probability $1 - \delta$ over $S \sim p^m$ we have*

$$L(\mathcal{A}(S), p) \leq \alpha \cdot \mathrm{opt}(\mathcal{C}, p) + \varepsilon.$$

Definition 24 (sample complexity) *Let $\mathcal{C} \times \mathcal{P}$ be a pair of classes. Let*

$$M_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) = \{m \in \mathbb{N} : m \textit{ guarantees } (\varepsilon, \delta)\textit{-success for } \mathcal{C} \times \mathcal{P}.\}$$

The realizable sample complexity $m_{\mathcal{C}\times\mathcal{P}} : [0,1]^2 \to \mathbb{N} \cup \{\infty\}$ *is defined by*

$$m_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) = \begin{cases} \min\{m : m \in M_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta)\} & \textit{, if } M_{\mathcal{C}}(\varepsilon, \delta) \neq \emptyset \\ \infty & \textit{, otherwise.} \end{cases}$$

Similarly, for $\alpha > 1$, we define the set

$$M_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta)^{\alpha} = \{m \in \mathbb{N} : m \textit{ guarantees } \alpha\textit{-agnostic } (\varepsilon, \delta)\textit{-success for } \mathcal{C}.\}$$

We then define the α-agnostic sample complexity $m^{\alpha}_{\mathcal{C}\times\mathcal{P}} : [0,1]^2 \to \mathbb{N} \cup \{\infty\}$ *by*

$$m^{\alpha}_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) = \begin{cases} \min\{m : m \in M^{\alpha}_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta)\} & \textit{, if } M^{\alpha}_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) \neq \emptyset \\ \infty & \textit{, otherwise.} \end{cases}$$

We now define weak learnability and PAC learnability

Definition 25 (ε-weak learnability) *A pair of classes $\mathcal{C} \times \mathcal{P}$ is* ε-weakly learnable*, if for every $\delta \in (0,1)$, $m_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) < \infty$. A pair of classes $\mathcal{C} \times \mathbb{P}$ is* α-agnostically*, ε-weakly learnable, if for every $\delta \in (0,1)$, $m^{\alpha}_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) < \infty$.*

Definition 26 (PAC learnability) *A pair of classes $\mathcal{C} \times \mathcal{P}$ is* PAC learnable in the realizable case*, if for every $\varepsilon, \delta \in (0,1)$, $m_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) < \infty$. A pair of classes $\mathcal{C} \times \mathcal{P}$ is* α-agnostically PAC learnable*, if for every $\varepsilon, \delta \in (0,1)$, $m^{\alpha}_{\mathcal{C}\times\mathcal{P}}(\varepsilon, \delta) < \infty$.*

To simplify notation and phrasings we will often only refer to concept classes $\mathcal{C}$ and sample complexity function $m_{\mathcal{C}}$ in the main body of the work.

Appendix B. δ-boosting: Deferred proof

B.1. Proof of Lemma 18

Consider a learning task for which there exists a function $m^{\alpha}_{\text{finite}} : \mathbb{N} \times (0,1)^2 \to \mathbb{N}$, such that every finite class $\mathcal{C}'$ with $|\mathcal{C}'| = k$ is α-agnostically learnable with sample complexity $m^{\alpha}_{\mathcal{C}'}(\varepsilon, \delta) \leq m^{\alpha}_{\text{finite}}(k, \varepsilon, \delta)$ for every $\varepsilon, \delta \in (0,1)$.

For any fixed ε, let $\mathcal{C}$ be a concept class for which there exists $\delta', \eta \in (0,1)$ such that a sample size m suffices for $\left(\frac{\varepsilon}{\alpha} - \eta, \delta'\right)$-success. Then, $\mathcal{C}$ is ε-weakly learnable with

$$m_C(\varepsilon, \delta) \leq mn + m_{\text{finite}}\left(n, \eta, \frac{2}{\delta}\right),$$

where $n = \log_{\delta'}(\frac{\delta}{2})$.

Proof Given an sample S of size $mn + m_{\text{finite}}\left(n, \eta, \frac{2}{\delta}\right)$, we split S into $n+1$ subsamples $S_{1,1}, \dots, S_{1.n}, S_2$, where for every $i \in [n]$, we have $|S_{1,i}| = m$ and $|S_2| = m_{\text{finite}}\left(n, \eta, \frac{2}{\delta}\right)$. We know that there exists a learner $\mathcal{A}$ such that for every $p \in \{p' \in \mathcal{P} : \text{opt}(\mathcal{C}, p) = 0\}$ we have

$$\mathbb{P}_{S'^m \sim p}[d_{\text{TV}}(\mathcal{A}(S')) \leq \frac{\varepsilon}{\alpha} - \eta] \leq 1 - \delta'.$$

Thus, if S is i.i.d. sampled by some $p \in p \in \{p' \in \mathcal{P} : \text{opt}(\mathcal{C}, p) = 0\}$, then for every independent subsample $S_{1,i}$ the above theorem holds. Define a set of candidates

$$\mathcal{C}' = \{\mathcal{A}(S_{1,i}) : i \in [n]\}.$$

Since $|\mathcal{C}'| = n$, we can use an α-agnostic learner $\mathcal{A}_{\text{finite},\mathcal{C}'}$ for $\mathcal{C}'$ with sample complexity $m_{\text{finite}}(n, \cdot, \cdot)$. Since $|S_2| = m_{\text{finite}}\left(n, \eta, \frac{2}{\delta}\right)$ with probability $1 - \frac{\delta}{2}$

$$\mathcal{A}_{\text{finite},\mathcal{C}'}(S_2) \leq \alpha \cdot \text{opt}(p, \mathcal{C}) + \eta.$$

The probability that $\mathcal{C}'$ contains an element q with $d_{\text{TV}}(p, q) \leq \frac{\varepsilon}{\alpha} - \frac{\eta}{2}$ is $1 - (\delta')^n = 1 - \frac{\delta}{2}$. Thus, taking everything together, with probability $1 - \frac{\delta}{2} + \frac{\delta}{2} = 1 - \delta$, we have

$$\mathcal{A}_{\text{finite},\mathcal{C}'}(S_2) \leq \alpha \cdot \text{opt}(p, \mathcal{C}) - \eta \leq \alpha \cdot \frac{\varepsilon}{\alpha} - \eta + \eta = \varepsilon.$$

■

B.2. Proof of Lemma 17

Let $\mathcal{C}$ be a class of distributions. If there exists a learner $\mathcal{A}$ and parameters $\delta', \eta \in (0, 1)$ and $m \in \mathbb{N}$, such that a sample size m guarantees $(\frac{\varepsilon - \eta}{2}, \delta')$-success, then $\mathcal{C}$ is ε-weakly learnable with sample complexity

$$m_{\mathcal{C},\varepsilon}(\delta) \leq mn + \tilde{O}\left(\frac{\log(n) \cdot \log(2/\delta)}{\eta^2}\right),$$

where $n = \log_{\delta'}(\frac{\delta}{2})$.

Proof For this proof we will make use of the hypothesis selection algorithm that guarantees 2-agnostic learnability for finite hypothesis classes introduced in Bousquet et al. (2022). Their results states, that for a finite class of hypotheses $\mathcal{C}''$ with $|\mathcal{C}''| = n$ a learning success of

$$d_{\text{TV}}(\hat{p}, p) \leq 2 \cdot \text{opt} + \varepsilon'',$$

can be guaranteed with probabiltiy $1 - \delta''$, when trained on an i.i.d. sample of size $m'' \geq m_{\mathcal{C}''}(\varepsilon'', \delta'')$, where $m_{\mathcal{C}''}(\varepsilon'', \delta'') \in \tilde{O}\left(\frac{\log(n) \cdot \log(1/\delta'')}{\varepsilon''^2}\right)$. The result then follows as an immediate corollary of Lemma 18

■

Appendix C. Lower Bound for Distribution Learning

We start this subsection by giving a comparison of Theorem 22 and common lower bound techniques.

Comparisson to other lower bound techniques. Our result bears some similarity with common lower bound techniques from literature, such as Fano's inequality, Le Cam's method and Assouad's inequality, in the sense that all of these methods rely on creating some form of "indistinguishability". However, our result also differs in ways that are crucial when it comes to proving a lower bound for the class $\mathcal{C}_{\text{counter}}$.

Fano's inequality Fano's inequality requires distibutions to have pairwise small KL-divergence. Both of our methods yield positive results for distributions which have pairwise KL-divergence ∞. In particular, Fano's inequality could not yield a lower bound for $\mathcal{C}_{\text{counter}}$.

Assouad's inequality Assouad's inequality relies on a finite number of distributions. In our setting we require a lower bound that considers an infinite number of distributions in our class $\mathcal{C}_{\text{counter}}$, as this class is constructed in such a way that every finite subset is learnable for sample complexity $m(0,0) = 1$. Thus, any technique relying on only a finite subset could thus not yield a sufficient lower bound for learning $\mathcal{C}_{\text{counter}}$.

Le Cam's method Le Cam's two-point method would not obtain the required indistinguishability to obtain hardness (as an uncountable number of distributions is needed for that). Le Cam's mixture-vs-point method in contrast could handle the class $\mathcal{C}_{\text{counter}}$ to obtain some lower bound, as it also makes use of a meta-distribution. However, this method could bound the total variation distance of learning by at most $\frac{1}{2}$, which is worse than our result, which gives a lower bound of 1. Structurally, our bound creates indistinguishability in the posterior between different support elements, whereas Le Cam's mixture-vs-point method creates a meta-distribution Q such that $|Q|^m$ is hard to distinguish from some fixed $p \in \mathcal{C}$.

C.1. Proof of Theorem 22

C.1.1. THEOREM 22

Let $\mathcal{C}$ be a class of distributions. If for $m \in \mathbb{N}$, there exists a meta distribution Q over elements in $\mathcal{C}$, such that for every $q, p \in \text{supp}(Q)$ with $p \neq q$, we have $d_{\text{TV}}(p,q) = 1$ and for every $s \in \text{supp}(|Q|^m) \subset \mathcal{X}^m$, for the conditional probability distribution over elements of $\mathcal{C}$ with Q as prior, conditioned on a sample $S_Q \sim |Q|^m$, we have for every finite subset $B_{\mathcal{C}} \subset \mathcal{C}$ and every $S \in \mathcal{X}^m$ we have

$$\kappa_{q_Q|S_Q}(B,S) = 0,$$

then for every learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$, there exists $q \in \mathcal{C}$ with

$$\mathbb{P}_{S\sim q}[d_{\text{TV}}(\mathcal{A}(S), q) = 1] = 1.$$

If for every $m \in \mathbb{N}$ such a meta-distribution exists, then $\mathcal{C}$ is not ε-weakly learnable for any $\varepsilon \in [0,1]$.

Proof Let us consider n pairwise distinct distributions $p_1, \dots, p_n \in \text{supp}(Q)$. By our assumption, we know that for any $i, j \in [n]$ with $i \neq j$ we have $d_{\text{TV}}(p_i, p_j) = 1$. Then, for every pair (p_i, p_j) we can define a set $E_{i,j} \subset \Sigma_{\mathcal{X}}$ with $p_i(E_{i,j}) = 1$ and $p_i(E_{i,j}) = 0$, such that $E_{i,j}$ and $E_{j,i}$ are disjoint. Let $E_i := \bigcap_{j=1}^n E_{i,j}$. Clearly $p_i(E_i) = 1$ and for all $j \neq i$ we have $E_i \cap E_j = \emptyset$ and $p_j(E_i) = 0$. Let us consider any $\hat{q} \in \Delta(\mathcal{X})$. For every $i \in [n]$, we have

$$d_{\text{TV}}(\hat{q}, p_i) = \sup_{B \subset \Sigma_{\mathcal{X}}} |\hat{q}(B) - p_i(B)| \geq \hat{q}(E_i).$$

Thus, for every $\hat{q} \in \Delta(\mathcal{X})$, we have

$$\begin{aligned}\sum_{i=1}^{n} d_{\mathrm{TV}}(\hat{q}, p_i) &\geq \sum_{i=1}^{n}(1 - \hat{q}(E_i)) \\ &\geq n - \sum_{i=1}^{n} \hat{q}(E_i) \\ &\geq n - \hat{q}\left(\bigcup_{i=1}^{n} E_i\right) \\ &\geq n - 1.\end{aligned}$$

Let $B_{n,\hat{q}} \subset \mathrm{supp}(Q)$ be defined as $\{p \in \mathrm{supp}(Q) : d_{\mathrm{TV}}(\hat{q}, p) \leq \frac{n-1}{n}\}$. From the above a calculation, clearly $|B_n| \leq n$.

For a fixed $S \in \mathcal{X}^m$, consider the distribution defined by $\kappa_{q_Q|S_Q}(\cdot, S)$ and any learner $\mathcal{A} : \mathcal{X}^* \to \Delta(\mathcal{X})$:

$$\begin{aligned}\kappa_{q_Q|S_Q}(\{p \in \Delta(\mathcal{X}) : d_{\mathrm{TV}}(\mathcal{A}(S), p) < 1\}, S) &\leq \kappa_{q_Q|S_Q}\left(\{p \in \mathrm{supp}(Q) : d_{\mathrm{TV}}(\mathcal{A}(S), p) < 1\}, S\right) \\ &\leq \sum_{n=0}^{\infty} \kappa_{q_Q|S_Q}\left(\left\{p \in \mathrm{supp}(Q) : d_{\mathrm{TV}}(\mathcal{A}(S), p) < \frac{n-1}{n}\right\}, S\right) \\ &\leq \sum_{n=0}^{\infty} \kappa_{q_Q|S_Q}(B_n, S) \\ &\leq \sum_{n=0}^{\infty} 0 = 0.\end{aligned}$$

Thus $\kappa_{q_Q|S_Q}(\{p \in \Delta(\mathcal{X}) : d_{\mathrm{TV}}(\mathcal{A}(S), p) = 1\}, S) = 1$.

Now consider for an arbitrary learner $\mathcal{A}$

$$\begin{aligned}
\mathbb{P}_{(q_Q,S_Q)\sim Q_{\mathcal{C},m}}[d_{\mathrm{TV}}(\mathcal{A}(S_q),q_Q)=1] &= \int_{\mathcal{C}\times\mathcal{X}^m} f_{q_Q,S_Q}(q,S)1[d_{\mathrm{TV}}(\mathcal{A}(S),q)=1]d(q,S)\\
&= \int_{\mathcal{X}^m}\left(\int_{\mathcal{C}} f_{q_Q,S_Q}(q,S)1[d_{\mathrm{TV}}(\mathcal{A}(S),q)=1]dq\right)dS\\
&= \int_{\mathcal{X}^m}\left(\int_{\mathcal{C}} f_{q_Q,S_Q}(q,S)1[d_{\mathrm{TV}}(\mathcal{A}(S),q)=1]dq\right)dS\\
&= \int_{\mathcal{X}^m}\left(\int_{\{p\in\mathcal{C}:d_{\mathrm{TV}}(\mathcal{A}(S),p)=1\}} f_{q_Q,S_Q}(q,S)dq\right)dS\\
&= \int_{\mathcal{X}^m}\left(\frac{\int_{\{p\in\mathcal{C}:d_{\mathrm{TV}}(\mathcal{A}(S),p)=1\}} f_{q_Q,S_Q}(q,S)dq}{\int_{\mathcal{C}} f_{q_Q,S_Q}(q,S)dq}\cdot\int_{\mathcal{C}} f_{q_Q,S_Q}(q,S)dq\right)dS\\
&= \int_{\mathcal{X}^m}\left(\kappa_{q_Q|S_Q}(\{d_{\mathrm{TV}}(\mathcal{A}(S_q),q_Q)=1\},S)\cdot\left(\int_C f_{q_Q,S_Q}(q,S)dq\right)\right)dS\\
&= \int_{\mathcal{X}^m}\left(1\cdot\left(\int_C f_{q_Q,S_Q}(q,S)dq\right)\right)dSS\\
&= \int_{\mathcal{X}^m}\left(\int_C f_{q_Q,S_Q}(q,S)dq\right)dSS\\
&= \int_{\mathcal{C}\times\mathcal{X}^m}(\int_{\mathcal{C}} f_{q_Q,S_Q}(q,S)d(q,S)\\
&= 1.
\end{aligned}$$

Thus, for every learner $\mathcal{A}$:

$$\sup_{q\in\mathrm{supp}(Q)}\mathbb{E}_{S\sim q}[d_{\mathrm{TV}}(\mathcal{A}(S),q)]\geq\mathbb{E}_{q\sim Q,S\sim Q}[d_{\mathrm{TV}}(\mathcal{A}(S),q)]\geq 1.$$

Thus, for every learner $\mathcal{A}$, there exists $q\in\mathcal{C}$, with

$$\mathbb{P}_{S\sim q}[d_{\mathrm{TV}}(\mathcal{A}(S),q)=1]=1.$$

■

Appendix D. Pseudo-compactness for distribution learning for countable hypothesis classes over countable domains

D.1. Proof of Theorem 20

The learning task of learning countable distribution classes over a countable domain satisfies 2-pseudo sample complexity compactness.

Proof Let $\mathcal{C}$ be a countable class over a countable domain. Thus, we want to show that for every $\varepsilon,\delta\in(0,1)$ and every $m\in\mathbb{N}$, if for every finite subset $\mathcal{C}'\subset\mathcal{C}$, we have

$$m_{\mathcal{C}}(\varepsilon,\delta)\geq m$$

then

$$m_{\mathcal{C}}(2\varepsilon, \delta) \geq m.$$

Since $\mathcal{C}$ is countable, we can consider an arbitrary enumeration of $\mathcal{C} = \{p_i : i \in \mathbb{N}\}$. Let us assume for every finite subset $\mathcal{C}' \subset \mathcal{C}$, we have

$$m_{\mathcal{C}}(\varepsilon, \delta) \geq m$$

. In particular, if we consider the sets $\mathcal{C}_i = \{p_j : j \in [i]\}$, we get

$$m_{\mathcal{C}_i}(\varepsilon, \delta) \geq m$$

. Thus, there exists a learner $\mathcal{A}_i$ for $\mathcal{C}_i$, such that for every $p_j \in \mathcal{C}$, we have

$$\mathrm{P}_{S \sim p_j^m}[d_{\mathrm{TV}}(\mathcal{A}_i(S), p_i) \leq \varepsilon] \geq 1 - \delta.$$

Using this sequence of learner $(A_i)_{i \in \mathbb{N}}$, we now aim to define a learner $\mathcal{A}^*$ for $\mathcal{C}$ such that for every $p_j \in \mathcal{C} = \bigcup_{i=1} \mathcal{C}_i$, we have

$$\mathrm{P}_{S \sim p_j^m}[d_{\mathrm{TV}}(\mathcal{A}^*(S), p_i) \leq \varepsilon] \geq 1 - \delta.$$

We note that since $\mathcal{X}$ is countable, there is an enumeration of subsets S of size m, i.e. $\{S_k : k \in \mathbb{N}\} = \mathcal{X}^m$.

Now let $\beta(i, j, k)$ denote the following binary function:

$$\beta(i, j, k) = \begin{cases} 1 & \text{if } \Leftrightarrow d_{\mathrm{TV}}(\mathcal{A}_i(S_k), p_j) \leq \varepsilon \\ 0 & \text{otherwise.} \end{cases}$$

For every $i \geq j$, we know that

$$\sum_{k \in \mathbb{N}} \beta(i, j, k) p_j(S_k) \leq 1.$$

Our goal is now to define $\mathcal{A}^*(S_k)$ in such a way that preserves this learnability for every p_j. To do so we want to define a "limit behaviour" of the sequence $(\mathcal{A}_i)_{i\mathbb{N}}$.

For fixed j, k and an infinite indexset $I \in \mathbb{N}$, consider the sequence $(\beta(i, j, k))_{i \in I}$. We define

$$\beta^*((j, k), I) = \begin{cases} 1 & \text{if there are infinitely many } i \in I \text{ with } \beta(i, j, k) = 1 \\ 0 & \text{otherwise.} \end{cases}$$

And define the subsequence $I^*(j, k, I)$ by

$$i \in I^*((j, k), I) :\Leftrightarrow i \in I \text{ and } \beta^*(j, k, I) = \beta(i, j, k).$$

It is easy to verify that if I is an infinite sequence, then so is $I^*((j, k), I)$.

We can now travers the pairs $(j, k)_{j \in \mathbb{N}, k \in \mathbb{N}}$ "diagonally", by the following recursively defined sequence $(a_l)_{l\mathbb{N}}$:

- $a_1 = (1, 1)$

- If $a_l = (j,k)$ with $j \neq 1$,then $a_{l+1} = (j-1, k+1)$
- If $a_l = (1,k)$ then, $a_{l+1} = (k+1, 1)$.

Clearly $\{a_l : l \in \mathbb{N}\} = \mathbb{N}^2$.

We now define the sequence I_l^* recursively by $I_0^* = \mathbb{N}$ and

$$I_l^* = I^*(a_l, I_{l-1}^*).$$

We now define the index set $I^* = (i_l^*)_{l\mathbb{N}}$ by

$$i_l^* = I_l^*(l),$$

Where $I_l^*(n)$ refers to the n-th element of the sequence I_l^*.

Now if we consider the sequence $(\beta(i,j,k))_{i \in I^*} = (\beta(i_l^*, j, k))_{i_l^* : l \in \mathbb{N}}$, we get

$$\lim_{l \to \infty} \beta(i_l^*, j, k) = \beta(j, k, I^*).$$

We now want to argue that we can choose $\mathcal{A}^*$, such that

1. For every $k \in \mathbb{N}$ and every $j \in \mathbb{N}$: $\beta^*((j,k), I^*) = 1$ implies
$$d_{\mathrm{TV}}(\mathcal{A}^*(S_k), p_j) \leq 2 \cdot \varepsilon,$$
2. and for every $j \in \mathbb{N}$
$$\sum_{k=1}^{\infty} \beta^*((j,k), I^*) p_j(S_k) \geq 1 - \delta$$

We start by arguing (2.).

$$\begin{aligned}\sum_{k=1}^{\infty} b^*((j,k), I^*) p_j(S_k) &= 1 - \sum_{k=1}^{\infty} (1 - b^*((j,k), I^*)) p_j(S_k) \\ &= 1 - \lim_{t \to \infty} \sum_{k=1}^{t} (1 - \beta^*((j,k), I^*)) p_j(S_k)\end{aligned}$$

We now note that by the way we picked i_l^* and the way we enumerated pairs (j,k) via the sequence a_L, we know that for every $l \geq (j+k)^2$, we have $\beta(i_l^*, j, k) = \beta^*(i, j, I^*)$. Conversely, if for a fixed j and fixed t, we have $l \geq (j+t)^2$, then

$$\sum_{k=1}^{t} (1 - \beta(i_l^*, j, k)) p_j(S_k) = (1 - \beta^*((j,k), I^*)).$$

Furthermore, since I^* is a subsequence of $(i)_{i \in \mathbb{N}}$, we have

$$i_l \geq l \geq (j+t)^2 \geq j.$$

Thus, we get the following:

$$\begin{aligned}\sum_{k=1}^{\infty}\beta^*((j,k),I^*)p_j(S_k) &= 1-\sum_{k=1}^{\infty}(1-\beta^*((j,k),I^*))p_j(S_k)\\ &= 1-\lim_{t\to\infty}\sum_{k=1}^{t}(1-\beta^*((j,k),I^*))p_j(S_k)\\ &= 1-\lim_{t\to\infty}\sum_{k=1}^{t}(1-\beta(i^*_{(j+t)^2},j,k))p_j(S_k)\\ &\geq 1-\lim_{t\to\infty}\sum_{k=1}^{\infty}(1-\beta(i^*_{(j+t)^2},j,k))p_j(S_k)\\ &\geq 1-\lim_{t\to\infty}\delta\\ &= 1-\delta.\end{aligned}$$

Thus condition (2) is satisfied.

We now want to argue condition (1.). Given k, we define $\mathcal{Q}_k := \{p_j : j \in \mathbb{N} \text{ with } \beta^*(j,k,I^*) = 1\}$. If $\mathcal{Q}_k$ is empty then then statement (1) is trivially true. Thus, we will assume for the rest of the proof that $\mathcal{Q}_k$ contains at least one element. We want to choose $\mathcal{A}^*(S_k)$ in such a way that for every $p_j \in \mathcal{Q}_k$, we have

$$d_{\mathrm{TV}}(\mathcal{A}^*(S_k),p_j) \leq 2\cdot\varepsilon.$$

We note, that since for every $p_{j_1}, p_{j_2} \in \mathcal{Q}_k$, we can upper bound their total variation distance, by considering a learner $\mathcal{A}_{i^*_l}$ with $l > j_1, j_2$:

$$d_{\mathrm{TV}}(p_{j_1},p_{j_2}) \leq d_{\mathrm{TV}}(\mathcal{A}_{i^*_l}(S_k),p_{j_1}) + d_{\mathrm{TV}}(\mathcal{A}_{i^*_l}(S_k),p_{j_2}) \leq 2\varepsilon.$$

Thus we can define the learner $\mathcal{A}^*$ as:

$$\mathcal{A}^*(S_k) = \begin{cases} p_j & \text{where } j = \min\{j' : p_{j'} \in \mathcal{Q}_k\}\text{, if } \mathcal{Q}_k \text{ is not empty,}\\ \delta_1 & \text{otherwise.}\end{cases}$$

Clearly, this learner satisfies condition (1.). From (2.) we now have that $\mathcal{A}^*$ is a successful learner for $\mathcal{C}$.

■