
Logical perspectives on learning statistical objects

Aaron Anderson *University of Pennsylvania*

Michael Benedikt *University of Oxford*

Abstract

We consider the relationship between learnability of a “base class” of functions on a set X , and learnability of a class of statistical functions derived from the base class. For example, we refine results showing that learnability of a family $h_p : p \in Y$ of functions implies learnability of the family of functions $h_\mu = \lambda p : Y.E_\mu(h_p)$, where E_μ is the expectation with respect to μ , and μ ranges over probability distributions on X . We will look at both Probably Approximately Correct (PAC) learning, where example inputs and outputs are chosen at random, and online learning, where the examples are chosen adversarially. We establish improved bounds on the sample complexity of learning for statistical classes, stated in terms of combinatorial dimensions of the base class. We do this by adapting techniques introduced in model theory for “randomizing a structure”. We give particular attention to classes derived from logical formulas, and relate learnability of the statistical classes to properties of the formula. Finally, we provide bounds on the complexity of learning the statistical classes built on top of a logic-based hypothesis class.

1 Introduction

Much of classical learning theory deals with learning a function into the reals based on training examples consisting of input-output pairs. In the special case where the output space is $\{0, 1\}$ we refer to a *concept class*. There are many variations of the set-up. Training examples can be random – as in “Probably Approximately Correct” (PAC) learning – or they can be adversarial, as in “online learning”. We may assume that the examples match one of the hypothesis functions (the *realizable case*), or not (the *agnostic case*).

Here we will consider *learning statistical objects*, where the function we are learning is itself a distribution, and we are given not individual examples about it, but statistical information. One motivation is from database query processing, where we have queries to evaluate on a massive dataset, and we have stored some statistics about the dataset, for inputs of certain shape. For example, the dataset might be a graph with vertices having numerical identifiers, and we have computed histograms giving information about the average number of vertices connected to elements in certain intervals. In order to better estimate future queries, we may extrapolate the statistics to other unseen intervals. One formalization of this problem, focused specifically on learning an unknown probability distribution from statistics, was given in (Hu et al., 2022). In this setting, we have a set of points X , and a collection of subsets of X , referred to as ranges. The random object we are trying to learn is a distribution on X , and we try to learn it via samples of the probabilities of ranges: that is, each distribution can be considered as a function mapping a range to its probability. The main result of (Hu et al., 2022) is that if the set to subsets is itself learnable, then the set of functions induced by distributions is learnable.

We generalize the setting of (Hu et al., 2022) in several directions. We start with a hypothesis class which can consist of either Boolean-valued or real-valued functions on some set X , indexed by a parameter space Y . We use such a “base hypothesis class” to form several new “statistical hypothesis classes”, which will be real-valued functions over random objects. Two such classes are indexed by distributions μ , where the corresponding functions map an input to an expectation against μ . In one class, μ represents randomization over the *parameters*, and the functions will be on the input space of the original class; in the second class, μ represents randomization over *range elements*. We show that learnability of the base class allows us to derive learnability of the corresponding statistical classes, and establish new bounds on sample complexity of learning in terms of dimensions of the base class.

We analyze these two statistical classes by embedding in the *randomization of a hypothesis class*, inspired by work in model theory (Keisler, 1999; Ben Yaacov & Keisler, 2009; Ben Yaacov, 2009). In those works, we move from a base structure to another structure, its randomization. In the PAC learning scenario, we can apply prior work in model theory (Ben Yaacov, 2009) to conclude that when the base class is PAC learnable, the randomization class is PAC learnable. We show that this implies preservation of learnability of both distribution classes. We refine these arguments in several directions: to apply to new statistical classes, to deal with real-valued functions in the base class, and to get bounds on the number of samples needed to learn. We also examine whether the same phenomenon applies to other learning scenarios: e.g. realizable PAC learning, online learning.

We will pay particular attention to the setting where classes come from varying parameters within a logical formula over an infinite structure: *definable families*. Standard examples of learnable classes – e.g. rectangles, families of regions defined by polynomials of fixed degree – fit into this framework. We show that for many common structures, all the statistical classes built on top of definable families are learnable.

Contributions: preservation, sample complexity, and decidability. Our first results concern whether learnability is preserved when moving from a base hypothesis class to the corresponding statistical class. For agnostic learning, we provide positive results on preservation, accompanied by sample complexity bounds for the statistical class, stated in terms of combinatorial dimensions of the base class. For realizable learning, we show negative results. A high-level overview is in Table 1. The formal definitions are in the next sections.

The results above concern sample complexity – the number of samples needed to learn a statistical object within a given tolerance and a given confidence. We also examine computational complexity of learning, focusing exclusively on classes defined by first-order logic formulas. We show that under some decidability-

Learning	Agnostic	Realizable
Online	Bounded Online FatSh Rakhlin et al. (2015b) Preserved (Cor. 2)	Uniform Regret or Finite Online Dim. Not Preserved for Dual Dist. Class (Prop. 7)
PAC	Finite FatSh Bartlett & Long (1995) Preserved (Thm 1)	Finite OIG dimension Attias et al. (2023) Not Preserved for Distr. or Dual Distr. Class (Prop. 5)

Table 1: Dimensions governing learning, and preservation moving from a base class to its statistical class

related hypothesis on the logical structure, we can effectively estimate the value of a statistical function on a new input point, within a given tolerance.

Organization. Section 2 reviews different flavors of learnability that we study here. Section 3 reviews how learnability and hypothesis classes arise from logic. Section 4 defines the notion of “statistical class built over a base class” that will be our central object of study.

Section 5 studies preservation of PAC learnability for statistical classes, in two variations of the learning set up: agnostic and realizable. Section 6 performs the same study for online learning. The bounds in the bulk of this work are on the *sample complexity of learning*: how many samples do we need to make a prediction with a certain confidence. Section 7 briefly discusses the computational complexity of learning. We discuss related work in Section 8. We close with conclusions in Section 9.

We defer the proofs of a few propositions and lemmas to the appendix.

2 Preliminaries

In our preliminaries, and elsewhere in the paper, we use *fact* environments to denote results quoted from prior work.

Hypothesis classes and their duals. Our notions of learnability will be properties of a *hypothesis class*, a class of functions \mathcal{H} from some (in our case, usually infinite) set X , the *range space of the class*, to some interval in the reals, usually $[0, 1]$. A *concept class* \mathcal{C} is a family of subsets of X , which can be considered as a hypothesis class with range $\{0, 1\}$.

Functions in a hypothesis class \mathcal{H} are expressed as h_p where p ranges over the *parameter space of the class* Y . A family of functions on X indexed by a set Y can be considered as a single function from $X \times Y$ to $[0, 1]$. This corresponds naturally to a function $Y \times X \rightarrow [0, 1]$, and thus also defines a class of functions on Y indexed by X , the *dual class* of \mathcal{H} .

PAC learning. We recall the standard notion of a function class being learnable (Kearns & Schapire, 1994) from random supervision: *Probably Approximately Correct* or PAC learnable below. Fixing X , let $Z = X \times [0, 1]$. We call the elements of Z *samples*. For each hypothesis $h \in \mathcal{H}$ and sample $z = (x, y)$, let $l_h(z) = (h(x) - y)^2$: this is the *loss* of using this hypothesis at sample z . For a distribution P over Z , we let $\text{ExpLoss}_P(h)$ be the expected loss of h with respect to P . We let $\text{BestExpLoss}(P)$ be the infimum of $\text{ExpLoss}_P(h)$ over every $h \in \mathcal{H}$.

A *PAC learning procedure* is a mapping A from finite sequences in Z to \mathcal{H} . For parameters $\delta, \epsilon > 0$, we say that \mathcal{H} is δ, ϵ *agnostic PAC learnable* if there exists a learning procedure A and number $n_{\delta, \epsilon}$ such that for $n \geq n_{\delta, \epsilon}$, for every distribution P over Z ,

$$P^n(\vec{z} \mid \text{ExpLoss}_P(A(\vec{z})) \leq \text{BestExpLoss}_P + \epsilon) \geq 1 - \delta.$$

Here P^n denotes the n -fold product of P .

If a specific number $n_{\delta, \epsilon}$ suffices, we say that \mathcal{H} is δ, ϵ agnostic PAC learnable with *sample complexity* $n_{\delta, \epsilon}$. Alternately, if we just refer to bounding the δ, ϵ sample complexity of PAC learning \mathcal{H} , we mean the smallest number $n_{\delta, \epsilon}$ that suffices.

We say \mathcal{H} is *agnostic PAC learnable* if and only if it is δ, ϵ agnostic PAC learnable for all $\delta, \epsilon > 0$. Given a function $S(\epsilon, \delta)$ from $\epsilon, \delta \in [0, 1]$ to integers, we say that \mathcal{H} is *agnostic PAC learnable with sample complexity S* if for all $\epsilon, \delta \in [0, 1]$, \mathcal{H} is δ, ϵ agnostic PAC learnable with sample complexity $S(\delta, \epsilon)$.

The qualification “agnostic” above refers to the fact that we do not assume that there is an unknown true hypothesis that lies in our hypothesis class. Instead, we just try to find the best approximation in our class. This contrasts with *realizable PAC learning*, where we consider a true hypothesis $h_0 \in \mathcal{H}$, and randomly choose only *inputs*, with the outputs taken from h_0 . Realizable PAC learning requires the learning procedure to work as above, without knowledge of $h_0 \in \mathcal{H}$ or the distribution, but only for samples $(x, h_0(x))$ produced by applying h_0 to the randomly chosen x .

For us, the distinction between agnostic and realizable PAC learning will be important only in the case of real-valued functions:

Fact 1 (See, e.g. Shalev-Shwartz & Ben-David (2014)). *For a concept class, realizable PAC learnability coincides with agnostic PAC learnability. But the sample complexity can be lower in the realizable case.*

Fact 2. *Attias et al. (2023) For real-valued classes, agnostic learnability is strictly more restrictive than realizable learnability.*

Online learning. Another framework we consider is online learning of a hypothesis class \mathcal{H} (Ben-David et al., 2009). Here the learner receives examples with supervision, but not picked randomly, but “adversarially”: that is, arbitrarily, so the learner must consider the worst case. A (probabilistic) online learning algorithm A receives a finite sequence $s = (x_1, y_1) \dots (x_n, y_n)$ of pairs from $X \times [0, 1]$ along with an input x from X and returns a probability distribution over $y \in [0, 1]$. An *adversary* is an algorithm that receives a sequence of triples $s = (x_1, y_1, y'_1) \dots (x_n, y_n, y'_n)$ and returns a new pair (x, y) . Informally the first pair of each triple represents an input and real-valued output, while the last component is the value predicted by the learner. A *run* of a learning algorithm against an adversary for T rounds is a sequence $s = (x_1, y_1, y'_1) \dots (x_T, y_T, y'_T)$, where for each $i < T$, x_{i+1}, y_{i+1} is chosen by running the adversary on the prefix up to i , and y'_{i+1} is chosen by running the learning algorithm on the concatenation of the prefix up to i , projected to be a sequence of pairs, and the new adversary-generated example (x_{i+1}, y_{i+1}) . The *loss* of the algorithm on such a run of length T is, by default, $\sum_{i \leq T} |y'_i - y_i|$.¹ The *regret* for the run is the difference between the loss of the algorithm and the infimum of the loss obtained by an algorithm that uses a fixed $h \in \mathcal{H}$ to predict at each step. Note that if the run is highly inconsistent with all $h \in \mathcal{H}$, it will be much more difficult to predict; but each individual h will also fail to predict well. If we fix a strategy for the adversary, along with a probabilistic learner, we get a distribution on runs, and hence an *expected regret*. The *minimax regret* is the infimum over all learning algorithms of the supremum over all adversaries, of the expected regret. Following (Rakhlin et al., 2015b, Definition 2) we call a hypothesis class *online learnable in the agnostic case* if there is a learning algorithm whose minimax regret against any adversary in T rounds is dominated by T as T goes to ∞ .

As with PAC learning, “agnostic” here is contrasted to *online learnability in the realizable case*. This refers to the restriction on adversaries: they must be realizable, in that each should come from applying some hypothesis $h_0 \in \mathcal{H}$. We say that \mathcal{H} is *online learnable in the realizable case* if there is an algorithm that has *bounded loss*, uniform in T , for each realizable adversary.

Although the definitions of learnability are very different, for concept classes the dividing line for learnability is the same in the realizable case as in the agnostic case:

Fact 3. *Ben-David et al. (2009): A concept class is online learnable in the realizable case if and only if it is online learnable in the agnostic case.*

In fact, both of these are the same as having “bounded Littlestone dimension” defined below.

As with PAC learning, the dividing line for learnability diverges in the case of real-valued functions.

¹Here we use absolute value in online learning, but for our purposes square distance, as in PAC learning, would yield the same results.

Dimensions of real-valued families. Each of our notions of learnability corresponds to a combinatorial dimension of the class. For agnostic PAC learning, the characterization involves the fat-shattering definition originating in (Kearns & Schapire, 1994).

Definition 1 (Fat shattering). *For $\gamma \in (0, 1]$, we say \mathcal{H} γ fat-shatters a set $A \subseteq X$ if there exists a function $s : A \rightarrow [0, 1]$ such that, for every $E \subseteq A$, there exists some $h_E \in \mathcal{H}$ satisfying: For every $x \in A \setminus E$, $h_E(x) \leq s(x) - \gamma$, and for every $x \in E$, $h_E(x) \geq s(x) + \gamma$. The γ fat-shattering dimension of \mathcal{H} , denoted $\text{FatSHDim}_\gamma(\mathcal{H})$, is the supremum of the cardinalities of a γ fat-shattered subset of X .*

These dimensions give bounds on sampling:

Fact 4 (Bartlett & Long (1995, Thm 14)). *FatSHDim $_{\frac{\epsilon}{8}}(\mathcal{H})$ is finite for all γ if and only if \mathcal{H} is agnostic PAC learnable. In that case, there is an agnostic δ, ϵ PAC learning algorithm with sample complexity*

$$O\left(\frac{1}{\epsilon^2} \cdot \left(\text{FatSHDim}_{\frac{\epsilon}{8}}(\mathcal{H}) \cdot \log^2\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).$$

The notion of dimension simplifies for a concept class. A concept class is said to *shatter* a subset A of X if for every $E \subseteq A$ there is $c_A \in \mathcal{C}$ with c_A containing E and disjoint from $A \setminus E$. The *Vapnik-Chervonenkis (VC)-dimension* of \mathcal{C} is the supremum of the cardinalities of shattered subsets.

Agnostic online learning is linked to a sequential version of the fat-shattering dimension, defined in (Rakhlin et al., 2015b).

Definition 2 (Sequential fat-shattering dimension and Littlestone dimension). *Let $\{-1, 1\}^{<d}$ denote $\bigcup_{t=0}^{d-1} \{-1, 1\}^t$.*

A binary tree of depth d in X is a function $z : \{-1, 1\}^{<d} \rightarrow X$. Given such a binary tree and $t < d$, let z_t be the restriction of z to inputs in $\{-1, 1\}^t$. If $B \in \{-1, 1\}^d$ is a branch, we write $z_t(B)$ to denote z_t applied to the restriction of B to the first t entries.

Branches $B \in \{-1, 1\}^d$ can also be viewed as functions $B : \{0, \dots, d-1\} \rightarrow \{-1, 1\}$, with $B(t)$ the t th entry of B .

For $\gamma \in (0, 1]$ say \mathcal{H} γ fat-shatters a binary tree z in X , where z is of depth d , if there exists a binary tree s of depth d in \mathbb{R} and a labelling of each $B \in \{-1, 1\}^d$ with some $h_B \in \mathcal{H}$ satisfying: For every $0 \leq t < d$, if $B(t) = -1$, then $h_B(z_t(B)) \leq s_t(B) - \frac{\gamma}{2}$ and, if $B(t) = 1$, then $h_B(z_t(B)) \geq s_t(B) + \frac{\gamma}{2}$. The γ sequential fat-shattering dimension of \mathcal{H} , denoted $\text{FatSHDim}_\gamma^{\text{Seq}}(\mathcal{H})$, is the supremum of the depths of γ fat-shattered binary trees in X .

In the discrete case, a concept class shatters a binary tree $T : \{-1, 1\}^{<n} \rightarrow X$ when for every branch $B \in \{-1, 1\}^n$ of the tree, there is $c_B \in \mathcal{C}$ such that for all $k < n$, $T(B|_{\{-1, 1\}^k})$ is in c_B if and only if $B(k) = 1$. The Littlestone dimension of \mathcal{C} is the supremum of the depths of shattered binary trees.

Littlestone dimension characterizes online learnability (realizable or agnostic), in the case of concept classes, analogously to the way VC dimension characterizes PAC learnability (realizable or agnostic):

Fact 5 (Alon et al. (2021, Theorem 12.1)). *If \mathcal{H} is a $\{0, 1\}$ -valued concept class with Littlestone dimension at most d , then the minimax regret of a T -round online learner is bounded by*

$$O(\sqrt{dT}).$$

Conversely, if the dimension is infinite, then the minimax regret is infinite.

In the real-valued case, the bound is more complicated, but finiteness of all sequential fat-shattering dimensions is equivalent to agnostic online learnability:

Fact 6 (Rakhlin et al. (2015b, Part of Proposition 9)). *If \mathcal{H} is a hypothesis class taking values in $[0, 1]$, then the minimax regret of a T -round online learner is bounded below by*

$$\frac{1}{4 \cdot \sqrt{2}} \sup_{\gamma} \min \left(\sqrt{\text{FatSHDim}_\gamma^{\text{Seq}}(\mathcal{H}) \cdot T}, T \right)$$

and above by

$$\inf_{\gamma} \left(4 \cdot \gamma \cdot T + 12 \cdot \sqrt{T} \cdot \int_{\gamma}^1 \sqrt{\text{FatSHDim}_{\beta}^{\text{Seq}}(\mathcal{H}) \log \left(\frac{2 \cdot e \cdot T}{\beta} \right) d\beta} \right).$$

In particular, the minimax regret is sublinear if and only if $\text{FatSHDim}_{\gamma}^{\text{Seq}}(\mathcal{H})$ is finite for every γ .

Duality and agnostic learnability. It is well-known that agnostic PAC learnability is closed under moving to the dual:

Fact 7. *Kleer & Simon (2023)* A function class is agnostic PAC learnable exactly when its dual class is.

Using the combinatorial characterizations, one can show that same for online learning (see Appendix A):

Proposition 1. *A function class is agnostic online learnable exactly when its dual class is.*

3 Learnable classes from logic

We will now review a way to get hypothesis classes using logic, by looking at the hypothesis class associated to a formula and a partition of its free variables.

We deal with the standard definition of first-order logic, parameterized by a vocabulary L , consisting of relation and function symbols of different arities, including zero (e.g. for constant symbols). The first-order *terms* are built up from variables and constant symbols by applying function symbols: e.g. $f(g(c), v)$. The first-order logic *atomic formulas* are of the form $R(\tau_1 \dots \tau_n)$ where R has arity n and $\tau_1 \dots \tau_n$ are terms, along with $\tau_1 = \tau_2$ where τ_i are terms. The first-order logic formulas are built up from atomic formulas via Boolean operations \wedge, \vee, \neg along with quantifications $\exists x \phi, \forall x \phi$. The semantics are standard (Hodges, 1993).

A structure \mathfrak{M} for language L comes with a domain or universe, which we denote as M , along with an interpretation of each symbol in L : for example, a unary predicate will be interpreted by a subset of M . We fix a structure \mathfrak{M} with vocabulary L and domain M . We consider any first-order logic formula $\phi(x_1 \dots x_j; y_1 \dots y_k)$ where in the notation we also fix a partition of the free variables into \vec{x} and \vec{y} .

Definition 3 (Concept class of a (variable-partitioned) first-order formula). *A partitioned formula ϕ defines in the obvious way a function \mathcal{H}_{ϕ} on $M^j \times M^k$ to $\{0, 1\}$, which can be considered as a family of functions on range space M^j indexed by parameters in M^k . By convention, we write the variables that are considered parameters second in a partition. Thus the canonical concept class associated with $\phi(x_1 \dots x_j; y_1 \dots y_k)$, denoted \mathcal{C}_{ϕ} , has $y_1 \dots y_k$ as parameters, with each parameter value giving a concept over M^j .*

Definition 4 (NIP partitioned formulas and structures). *Fixing \mathfrak{M} we say that a partitioned first-order logic formula $\phi(\vec{x}, \vec{y})$ in the language of \mathfrak{M} is NIP if the corresponding family \mathcal{C}_{ϕ} has finite VC dimension, and say that \mathfrak{M} is NIP if every partitioned first-order formula has this property.*

NIP stands for “Not the Independence Property”, referring to an *independent family of sets*, one where every Boolean combination of family members is non-empty. So NIP partitioned formulas are those where the corresponding family does not have arbitrarily large independent subfamilies.

When we have established that a structure is NIP, we know how to use any partitioned formula to get a PAC learnable class. There are many tools available to verify that a structure is NIP, and examples include (Simon, 2015):

- The real ordered group $(\mathcal{R}, +, <)$ and the real ordered field $(\mathcal{R}, +, \cdot, <)$
- The real exponential field $(\mathcal{R}, +, \cdot, \lambda x.e^x, <)$
- The complex field $(\mathbb{C}, +, \cdot)$
- The infinite complete binary tree with predicates for the two successor relations.
- Integer additive arithmetic $(\mathbb{N}, +, <)$

Thus any definable family in each of these structures is learnable. Logic-based concept classes include many of the familiar examples of finite VC-dimension, such as rectangles. Concrete bounds on the corresponding

dimensions, and thus the sample complexity of PAC learning, have been established in many cases (Karpinski & Macintyre, 1997; Chistikov et al., 2022).

3.1 Real-valued functions from continuous logic

We now discuss how a structure can provide a collection of learnable real-valued functions. For this we utilize *continuous logic* (CL for short), defined over structures where the atomic predicates are real-valued: that is, a k -ary real-valued predicate is just a function from k -tuples in the model into the reals. We may also have k -ary function symbols $f(x_1 \dots x_k)$, which are associated to functions on elements in the model, as in first-order logic. We write \mathfrak{C} for such a structure.

We will be able to make do with a simplified presentation of the syntax and semantics of CL, since many aspects in the logic literature – e.g. metrics, variations of the compactness theorem, Lowenheim-Skolem theorem, ultraproducts – will not be needed for the results we state. We form the set of *basic formulas* over \mathfrak{C} , following (Hart, 2023). For any valuation of the free variables, the semantics will associate these formulas with real numbers.

- Terms are built up from variables using function symbols as in first-order logic. For any j -ary predicate P in the signature and terms $\tau_1 \dots \tau_j$ we have a basic formula $P(\tau_1 \dots \tau_j)$. The semantics under a valuation for the variables just evaluates each τ_i to get a vector of elements in the model, and then applies the real-valued function associated to P .
- If $f : [0, 1]^n \rightarrow [0, 1]$ is continuous, and $\phi_1 \dots \phi_n$ are basic formulas, then $f(\phi_1, \dots, \phi_n)$ is a basic formula, with the obvious semantics.
- If $\phi(x, \vec{y})$ is a basic formula with x a variable, then $\sup_x \phi$ and $\inf_x \phi$ are basic formulas.

Notice that if we start with a classical first-order structure, the range of each basic formula is a finite set, and we are not really gaining expressiveness over first-order logic: a formula will just partition tuples in the model into first-order definable sets and then associate each partition with a real number.

A *formula* of continuous logic extends the basic formulas above by closing under an infinite summation construct: if we have formulas $\phi_i : i \in \mathbb{N}$, then we can form a new formula $\sum_i \frac{\phi_i}{2^i}$. The semantics are again obvious: inductively, we can show that the range of each formula is bounded, which guarantees that the sum converges.

Definition 5 (Hypothesis class of a real-valued formula). *Consider a CL formula $\phi(x_1 \dots x_j; y_1 \dots y_k)$ with a partition of the free variables into \vec{x} and \vec{y} . From this we get a function from $M^j \times M^k \rightarrow [0, 1]$, which we can consider as a hypothesis class \mathcal{H}_ϕ over M^j indexed by M^k .*

We say that a CL structure \mathfrak{C} is NIP if for every partitioned real valued formula $\phi(\vec{x}; \vec{y})$, the corresponding family of real valued functions (as we vary \vec{y} over the structure), has bounded fat-shattering dimension – or equivalently, by Fact 4, the family is agnostic PAC learnable.

Any first-order structure can be viewed as a continuous logic structure: we have $\{0, 1\}$ valued basic predicates, and we can also consider equality as a basic predicate, but we have many more formulas. We write $\mathfrak{C}(\mathfrak{M})$ for this continuous logic structure. It is easy to prove that if \mathfrak{C} is NIP, then so is $\mathfrak{C}(\mathfrak{M})$. That is, throwing in these additional formulas does not give us function classes that are not agnostic PAC learnable: see Appendix B.

Proposition 2. *If a first-order structure \mathfrak{M} is NIP, then for any general continuous logic partitioned formula ϕ , \mathcal{H}_ϕ has finite fat-shattering dimension and is thus agnostic PAC learnable. Thus $\mathfrak{C}(\mathfrak{M})$ is also NIP.*

Proposition 2 gives a way of using first-order structures to get real-valued hypothesis classes that are well-behaved in terms of learning. It is particularly useful when applying continuous logic to first-order structures over the reals. Let \mathfrak{M} be a classical first-order structure on \mathcal{R} in a signature including the ordering $<$.

Lemma 1. *Let $f : \mathcal{R}^n \rightarrow [0, 1]$ be a definable function in \mathfrak{M} : one whose graph is defined by a formula $\phi(x_1 \dots x_n, y)$. Then there is a formula in continuous logic over $\mathfrak{C}(\mathfrak{M})$ that has f as its semantic function.*

See Appendix C for the simple proof. Combining the previous proposition and lemma, we can get many NIP CL-definable function classes since there are multiple NIP structures over the reals (van den Dries, 1998).

Example 3.1. Consider the hypothesis class $\mathcal{H}_{\vec{p}}$ consisting of rational functions $\frac{P(x)}{Q(x)}$, where $P(x)$ and $Q(x)$ are real polynomials restricted to an interval where Q has no zeros (with the function defined to be zero off this interval), the degrees of P and Q are at most 5, and the value of the quotient function on the interval is in $[0, 1]$. This is a class definable over the real field by a formula $\phi(x, y; \vec{p})$ where \vec{p} are parameters representing the coefficients of P and Q , x represents the function input and y the output. Thus since the real field is known to be an NIP expansion of $(\mathcal{R}, <)$ (van den Dries, 1998), Lemma 1 implies that this class of real-valued functions indexed by \vec{p} is agnostic PAC learnable.

We emphasize that *every concept class can be trivially seen as the concept class associated with a logical formula in a structure*. We just take a two-sorted structure, with one sort being the parameter space of the concept, the other being the range space. Similarly, *every real-valued hypothesis class can be trivially seen as being given by some continuous logic formula*. We do not formalize these comments here, since most of our proofs can be translated directly from logic-based classes to general ones.

The logical perspective allows us to generate many families of learnable examples. It also gives an elegant way of seeing duality: if our hypothesis class is given by a partitioned formula $\phi(\vec{x}; \vec{y})$, then moving to the dual just swaps the role of \vec{y} and \vec{x} .

4 Hypothesis classes from statistical objects

We now turn to the basic object of study in our paper:

Definition 6 (Distribution class and Dual Distribution class). *Given a concept class \mathcal{C} on X , parameterized by Y the distribution function class of \mathcal{C} , denoted $\text{Distr}_{\mathcal{C}}$, is a real-valued hypothesis class on the same range space X , consisting of the hypotheses h_{μ} indexed by the distributions μ on Y such that for each x , the set of $p \in Y$ such that $h_p(x) = 1$ is measurable. The hypothesis h_{μ} is defined as the function mapping x to the probability, with respect to μ , that $h_p(x) = 1$.*

The dual distribution function class of \mathcal{C} , denoted $\text{DualDistr}_{\mathcal{C}}$, is defined as above, but starting with the dual class of \mathcal{C} . That is, a hypothesis is parameterized by a distribution μ on X , such that each $C_p \in \mathcal{C}$ is measurable. The function h_{μ} maps $p \in Y$ to $\mu(C_p)$.

We extend these definitions to work on top of a real-valued function class \mathcal{H} , replacing probability under μ with expectation. For example, for the distribution function class, the functions are indexed by distributions μ on Y again, but now we restrict to distributions μ such that each $h \in \mathcal{H}$ is a μ measurable function, and h_{μ} maps $h \in \mathcal{H}$ to $E_{\mu}(h)$.

Example 4.1. Let \mathcal{H} be the concept class of rectangles over the reals. The range space is thus the collection of points in the plane – pairs of reals – while the parameter space consists of 4-tuples of reals, representing the lower left corner and upper right corners of the rectangle.

The dual distribution class of \mathcal{H} will consist of “random elements of the plane”: functions parameterized by distributions ν over the plane. Each such ν induces a function h_{ν} on rectangles, mapping each rectangle to its ν -probability. We can equivalently consider h_{ν} as a real-valued function on 4-tuples. In learning such a function h_{ν} , we will be given supervision in the form of a sequence of pairs, each pair consisting of a rectangle (or 4-tuple) and the ν -probability that a point is in the rectangle.

In contrast, the distribution class of \mathcal{H} will consist of “random rectangles”: functions parameterized by distributions μ over 4-tuples. Given such a μ , we have a function h_{μ} that maps a real pair \vec{x} to the μ probability that \vec{x} is in rectangle h_{μ} . In learning such a function, our supervision will consist of a point in the real plane and the probability that the point is in random rectangle μ .

Only the dual distribution class has been studied in the past literature, and exclusively for the case of a concept class. The following result is our starting point:

Fact 8. *Hu et al. (2022) Suppose a concept class \mathcal{C} is (Agnostic or Realizable) PAC learnable, and the VC-dimension of the class is λ . Then the dual distribution function class of \mathcal{C} is Agnostic PAC learnable*

with sample complexity $\tilde{O}(\frac{1}{\epsilon^{\lambda+1}})$ where \tilde{O} indicates that we drop terms that are polylogarithmic in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ for constant λ .

Randomization and random variable-based classes. We can embed the distribution class and the dual distribution class in a more general family of hypothesis classes, where parameters and range elements are treated symmetrically. We present the construction for logic-based classes, as in (Ben Yaacov, 2009; 2013). But the construction can be adapted for any hypothesis class.

Definition 7 (Well behaved collection of M -valued random variables). *Given a CL structure M , an M -valued random variables or functions from some atomless probability space $\Omega^* = (\Omega, \Sigma, \mu_0)$ to M . A collection of M -valued random variables is well-behaved if for each CL formula $\phi(\vec{x})$ over \mathfrak{C} , and each tuple \vec{X} of functions in the set, $\phi(\vec{X})$ is Ω^* -measurable.*

For example, the set of functions with countable image, such that the pre-image of each point is in Σ , is well-behaved. Thus, given $F_1 \dots F_k$ from well-behaved S , the composition of $\phi(F_1 \dots F_k)$ is a real-valued random variable on Ω .

Definition 8 (Randomization of a structure). *Consider a classical first-order structure or a continuous-valued structure \mathfrak{C} with domain M . We can form a new structure, the randomization of \mathfrak{C} , $\text{RAND}_{\mathfrak{C}}$ (Keisler, 1999; Ben Yaacov & Keisler, 2009; Ben Yaacov, 2013). For each (classical first-order or real-valued) formula $\phi(x_1 \dots x_k)$ over the signature of \mathfrak{C} , there is a new real-valued predicate $E\phi(X_1 \dots X_k)$, where $X_1 \dots X_k$ are new variables.*

This completes the description of the signature. We now describe the structure. Its domain is a well-behaved subset S of the M -valued random variables, based on atomless probability space $\Omega^ = (\Omega, \Sigma, \mu_0)$. Given $F_1 \dots F_k$ from S , the composition of $\phi(F_1 \dots F_k)$ is a real-valued random variable on Ω , and we set $E\phi$ evaluated at $F_1 \dots F_k$ to be the expected value of this random variable under μ_0 .*

The randomization is not unique. However, all the results presented will hold for *any* randomization, and thus we will sometimes abuse notation by referring to it as if it were unique.

Note that we consider the randomization of \mathfrak{C} as a CL structure, and thus we can apply CL connectives and quantification to form new formulas based on the primitive formulas $E\phi$ above. Our requirement on well-behavedness says that our random variables “play well” with the expectation of formulas from the base structure. The following result states that if we have this, then these random variables also play well with CL formulas over the randomization:

Fact 9. *Ben Yaacov & Keisler (2009); Ben Yaacov (2013). For any real-valued formula $f(x)$ in the language of $\text{RAND}_{\mathfrak{C}}$, any probability space $(\Omega, \mathcal{B}, \mu)$, and any $a \in (\text{RAND}_{\mathfrak{C}})^x$, the function $\omega \mapsto f(a(\omega))$ is measurable.*

Now fix a variable-partitioned formula $\phi(\vec{x}; \vec{y})$ in the language of a structure \mathfrak{M} , and consider the real-valued formula $E\phi$ in the randomization. It has free variables \vec{X} and \vec{Y} , both ranging over random variables. In (Ben Yaacov, 2009) (for concept classes) and (Ben Yaacov, 2013) (for real-valued based classes) it is shown that moving from ϕ in \mathfrak{C} to $E\phi$ in the randomization preserves agnostic PAC learnability:

Fact 10. *(Ben Yaacov, 2013, Theorem 4.10) If the hypothesis class induced by the formula $\phi(\vec{x}; \vec{y})$ over a continuous logic structure \mathfrak{C} is agnostic PAC learnable, then so is the class corresponding to the real-valued formula $E\phi(\vec{X}; \vec{Y})$ in the randomization. Further, if the structure \mathfrak{C} is NIP – that is, every definable family is PAC learnable – then the same holds for the real-valued structure $\text{RAND}_{\mathfrak{C}}$.*

Fact 10 will also be a corollary of our Theorem 1, with the latter providing more precise bounds.

We now want to apply this to the distribution and dual distribution class. We need to address the fact that in these classes, we consider distributions on the parameter space or the range space, not on some external sample space. Given two measure spaces (Ω, Σ, μ_0) and (X, Σ', μ') , and a function F from Ω to X , we say that F induces μ_0 from μ' if for every $S' \in \Sigma'$, $F^{-1}(S') \in \Sigma$ and $\mu_0(F^{-1}(S')) = \mu'(S')$. The following is a basic result in measure theory (Fremlin, 2002), see Appendix D:

Proposition 3. *For any set X , we can choose (Ω, Σ, μ) and well-behaved set of random variables RV_X such that for each probability measure μ', Σ' on X there is $F \in RV_X$ inducing μ' from μ .*

From now on for a model \mathfrak{C} we assume a suitably rich probability space and a well-behaved set of random variables, and refer to the randomization structure based on them as $\mathcal{RAN}_{\mathfrak{C}}$. We use analogous terminology for a partitioned formula $\phi(\vec{x}; \vec{y})$ over \mathfrak{C} – by the randomization class of ϕ we refer to the hypothesis class corresponding to the CL formula $E\phi$. We could similarly refer to the randomization class for a hypothesis class \mathcal{H} over range space X and parameter space P , referring to the hypothesis class with random variables into X as range elements and random variables into P as parameters.

We start with the observation that agnostic learnability of the randomization classes is sufficient to get agnostic learnability for the distribution and dual distribution classes:

Proposition 4. *For a formula $\phi(x_1 \dots x_j; y_1 \dots y_k)$ over a model \mathfrak{C} , if the randomization class corresponding to the formula $E\phi(\vec{X}; \vec{Y})$ is agnostic PAC learnable, then so are the distribution class and dual distribution class for \mathcal{H}_{ϕ} . Similarly for a hypothesis class \mathcal{H} , if its randomization class is agnostic PAC learnable, then so are its distribution and dual distribution classes. The same holds for agnostic online learnability.*

Proof. We consider only the PAC case for simplicity.

Consider the functions in the randomization class $E\phi(\vec{X}; \vec{Y})$, varying the random parameters \vec{Y} but restricting the function arguments \vec{X} to be deterministic. These are functions that map a vector \vec{x} to the expectation of $\phi(\vec{x}, \vec{Y})$ with respect to a given random variable \vec{Y} . By Proposition 3, the distributions induced by \vec{Y} cover all distributions over the parameter space M^k . Thus this class of restrictions give exactly the distribution class over $\mathcal{H}_{\phi(\vec{x}; \vec{y})}$. But if a class of functions is agnostic PAC learnable, then so is the class obtained by restricting its domain to a subset.

For the dual distribution class, we can apply Fact 7 and then the argument above. \square

Example 4.2. Consider the base model to be additive arithmetic $(\mathbb{N}, +, <)$. We can write a partitioned formula $\phi(x; y_1, y_2)$ that states that x is an even integer between y_1 and y_2 . Thus as we vary the parameters y_1, y_2 we get the set of intervals intersected with the even numbers.

Fix a probability space $\Omega^* = (\Omega, \Sigma, \mu_0)$ and consider the hypothesis space where the parameter space Y consists of “pairs of random integers”: pairs of Ω^* measurable functions from Ω into \mathbb{N} with countable range. Given a parameter $Y = \langle Y_1, Y_2 \rangle$, we can consider the function h_{Y_1, Y_2} mapping random integers X (Ω^* -measurable functions to \mathbb{N}) to the probability that the integer produced by X is even and between Y_1, Y_2 : that is, we get the function class for $E\phi(X; Y_1, Y_2)$. If we restrict each of the functions in the class to the deterministic integers X , we get the distribution function class corresponding to $\mathcal{C}_{\phi(x; y_1, y_2)}$.

If we reverse the role of parameters and range values, we get a function class indexed by random integers, mapping each random pair of integers to the probability. If we restrict this function class to deterministic pairs, we get the dual distribution function class for $\phi(x; y_1, y_2)$.

5 PAC learning of statistical classes

This section will be devoted to a more fine-grained investigation of how PAC learnability for these statistical classes follow from PAC learnability of the base class. Our first main result is a new proof that agnostic PAC learnability is preserved by moving to any of these classes, along with a new bound on sample complexity in terms of dimensions of the base class:

Theorem 1. *For a function class \mathcal{H} having $\frac{\epsilon}{50}$ fat-shattering dimension at most d , one can perform agnostic PAC learning on the randomization class of \mathcal{H} with sample complexity:*

$$O\left(\frac{d}{\epsilon^4} \cdot \log^2 \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \cdot \log \frac{1}{\delta}\right).$$

If \mathcal{H} is $\{0, 1\}$ -valued and has VC-dimension at most d , one can perform agnostic PAC learning on the randomization class of \mathcal{H} with sample complexity:

$$O\left(\frac{1}{\epsilon^2} \left(d \log \frac{d}{\epsilon} + \log \frac{1}{\delta}\right)\right).$$

Thus we get the same bounds for the distribution class and (moving to the dimension of the dual class) for the dual distribution class.

Example 5.1. Recall Example 3.1, where we consider a family \mathcal{H} of rational functions definable by a partitioned formula $\phi(x, y; \vec{p})$ in the real field, which can be considered as real-valued formula $\tau(x; \vec{p})$. The distribution function class of \mathcal{H} is parameterized by random \vec{p} . It is agnostic PAC learnable with sample complexity as in Theorem 1. Thus given supervision based on the expectations for various x_i , we can learn the hypothesis in the distribution class that has the best expected fit in terms of sum of differences.

Note that the base class is defined by *equalities* over the reals, so it is a restriction of a function class defined by equalities over the complex field. Thus our later results (Section 6) will imply that the statistical classes are also learnable in an online setting.

5.1 Combinatorial and statistical tools

We will follow the methods of (Ben Yaacov, 2009), which relate upper and lower bound on the combinatorial fat-shattering dimensions to upper and lower bounds on the *Rademacher mean width*.

Definition 9. [Height of a set] Let $A \subseteq \mathcal{R}^n$ be bounded. We define $h_A : \mathcal{R}^n \rightarrow \mathcal{R}$, the height of A in a particular direction, to be

$$h_A(\vec{b}) = \sup_{\vec{a} \in A} \vec{a} \cdot \vec{b}.$$

Lemma 2. For any bounded $A \subseteq \mathcal{R}^n$, the height function $h_A : \mathcal{R}^n \rightarrow \mathcal{R}$ is Borel measurable.

Proof. If A is countable, then h_A is the supremum of a countable family of continuous functions, and is thus Borel measurable.

If $D \subseteq A$, then clearly $h_D \leq h_A$. If D is dense in A , then for every $\vec{a} \in A$, $\vec{v} \in \mathcal{R}^n$, and $\epsilon > 0$, there is some $\vec{d} \in D$ such that $|\vec{v} \cdot (\vec{a} - \vec{d})| \leq \epsilon$, so we can conclude that $h_D(\vec{v}) \geq h_A(\vec{v}) - \epsilon$, so in fact, $h_D = h_A$.

Every subspace of \mathcal{R}^n is separable, so for every A there is a countable dense D , and $h_A = h_D$ is measurable. \square

Definition 10. [Mean width] Let β be a Borel probability measure on \mathcal{R}^n . Define the mean width of A w.r.t. β , $w(A, \beta)$, as $\mathbb{E}_\beta [h_A(\vec{b})]$, where \vec{b} is a random variable with distribution β . By the measurability result mentioned above, this is always defined.

If β is the distribution that samples uniformly from $\{+1, -1\}$ n times independently, we define the Rademacher mean width, denoted $w_{\mathcal{R}}(A)$, as $w(A, \beta)$.

Let $f(x, y)$ be a function on $X \times Y$, we let $f(\vec{x}, b)$ for $\vec{x} = (x_1, \dots, x_n) \in X^n$ denote the vector $(f(x_1; b), \dots, f(x_n; b))$, and let $f(\vec{x}, Y)$ be the set of vectors $\{f(\vec{x}, b) : b \in Y\}$.

Then we extend the definition of Rademacher mean width to be a function of an integer n , given the function f ; $\mathcal{R}_f(n) = \sup_{\vec{x} \in X^n} w_{\mathcal{R}}(f(\vec{x}, Y))$. We will refer to this function as the Rademacher mean width of the function f , but this only differs from the “Rademacher complexity” of f in other literature by a factor of n .

The reason for looking at Rademacher mean width will be that we can show it behaves well under averaging with respect to an arbitrary measure: see Theorem 2 to follow.

From a bound on the Rademacher width of a function class, we can infer a bound on the Rademacher width of its expectation. Using this and some relationships between Rademacher width bounds and fat-shattering, we are able to bootstrap from a class to its randomization, proving Theorem 1.

Glivenko-Cantelli Dimension. Our arguments for PAC learnability will go through the following dimension:

Definition 11 (Glivenko-Cantelli dimension). *The Glivenko-Cantelli dimension of a hypothesis class, denoted $GC_{\mathcal{H}}(\epsilon, \delta)$ is parameterized by $\delta, \epsilon > 0$:*

$$GC_{\mathcal{H}}(\epsilon, \delta) = \min \{n : \forall m \geq n, \forall D \text{ Distribution on } X \\ D^m \left\{ (x_1, \dots, x) \mid \forall h \in \mathcal{H}, \left| \frac{1}{m} \cdot (\sum_{i=1}^m h(x_i)) - \int h(u) dD(u) \right| > \epsilon \right\} \leq \delta \}$$

Recall that the law of large numbers implies that if we fix any bounded measurable function f and are given a δ and ϵ , then we can find an n so that, for any distribution D , for all but δ of the n -samples from the distribution, the sample mean of f is within ϵ of the the mean of f . Most proofs that a class is agnostic PAC learnable go through showing that for each $\epsilon, \delta > 0$, the dimension $GC_{\mathcal{H}}(\epsilon, \delta)$ is finite. From the Glivenko-Cantelli dimension for ϵ and δ , we can easily obtain bounds on the number of samples needed to learn to given tolerances δ and ϵ .

Glivenko-Cantelli bounds are used to derive learnability bounds in (Alon et al., 1997) and (Bartlett & Long, 1995, Theorem 14). The proof of (Anthony & Bartlett, 2009, Theorem 19.1) gives the following version of the connection between these concepts:

Fact 11 (Anthony & Bartlett (2009)). *The sample size needed to learn \mathcal{H} with error at most ϵ and error probability at most δ is at most $GC_{\mathcal{H}}\left(\frac{\epsilon}{2}, \frac{\delta}{2}\right)$.*

Because the bounds we will derive for $GC_{\mathcal{H}}(\epsilon, \delta)$ will always be polynomial in ϵ and δ , the factors of 2 in this fact will only change the bound by a constant multiple, which will only change the constant of asymptotic notation. While we will need to prove Glivenko-Cantelli bounds for most classes we study, it will be helpful to refer to the following general bound, which, combined with a version of Fact 11, comprised the proof of (Bartlett & Long, 1995, Theorem 14)(our Fact 4):

Fact 12 (Bartlett & Long (1995, Theorem 9)). *The Glivenko-Cantelli dimension is bounded by*

$$GC_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2} \cdot \left(\text{FatSHDim}_{\frac{\delta}{2}}(\mathcal{H}) \cdot \log^2\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).$$

5.2 Bounding the mean width for a derived class in terms of a base class

In this subsection, we will establish connections between combinatorial dimensions of a class of sets or functions and dimensions of its randomization class. Following the approach in (Ben Yaacov, 2009), we will first establish this connection for notions of mean width.

We will show that Rademacher mean width does not increase under averaging. With that in hand, if we are able to bound learnability of f through mean width, the same bound will apply to $\mathbb{E}[f]$. This strategy stems from (Ben Yaacov, 2009, Theorem 4.1), where it was applied to Gaussian mean width.

Assume that (Ω, Σ, μ) is a probability space, and consider a family $(f_{\omega} : \omega \in \Omega)$ of functions $X \times Y \rightarrow [0, 1]$ such that for each $x \in X^n$ and $v \in \mathcal{R}^n$, the function $\omega \mapsto h_{f_{\omega}(\bar{x}, Y)}(v)$ is measurable. This measurability follows trivially when Y is countable, and will also hold when these functions derive from the randomization class.

We now note a relationship between suprema of expectations and expectations of suprema.

Lemma 3. *Let (Ω, Σ, μ) be a probability space, and let $(f_{\omega} : \omega \in \Omega)$ be such that $\omega \mapsto h_{f_{\omega}(\bar{x}, Y)}(v)$ is measurable. Fix $\bar{x} \in X^n$ and $v \in \mathcal{R}^n$. Then*

$$h_{\mathbb{E}_{\mu}[f_{\omega}](\bar{x}, Y)}(v) \leq \mathbb{E}_{\mu}[h_{f_{\omega}(\bar{x}, Y)}(v)].$$

Proof. The supremum of the expectations of a family of functions is at most the expectation of their suprema, so we have

$$\begin{aligned}
h_{\mathbb{E}_\mu[f_\omega](\bar{x}, Y)}(v) &= \sup_{y \in Y} v \cdot (\mathbb{E}_\mu[f_\omega](\bar{x}, y)) \\
&= \sup_{y \in Y} \mathbb{E}_\mu[v \cdot (f_\omega(\bar{x}, y))] \\
&\leq \mathbb{E}_\mu \left[\sup_{y \in Y} v \cdot f_\omega(\bar{x}, y) \right] \\
&= \mathbb{E}_\mu[h_{f_\omega(\bar{x}, Y)}(v)].
\end{aligned}$$

which proves the lemma. ² □

And we can now make a conclusion about how mean width behaves under expectation.

Lemma 4. *Let (Ω, Σ, μ) be a probability space, and let $(f_\omega : \omega \in \Omega)$ be a family of real-valued functions $f : X \times Y \rightarrow [0, 1]$ such that for each $(x, y) \in X \times Y$ and $v \in \mathcal{R}^n$, the functions $\omega \mapsto f_\omega(x, y)$ and $\omega \mapsto h_{f_\omega(\bar{x}, Y)}(v)$ are measurable.*

Fix n , $\bar{x} \in X^n$, and a Borel probability measure β on \mathcal{R}^n . Then

$$w(\mathbb{E}_\mu[f_\omega](\bar{x}, Y), \beta) \leq \mathbb{E}_\mu[w(f_\omega(\bar{x}, Y), \beta)].$$

Proof. To prove this, we only need to unfold the definition of $w(A, \beta)$ and apply Lemma 3 and Fubini's Theorem.

$$\begin{aligned}
w(\mathbb{E}_\mu[f_\omega](\bar{x}, Y), \beta) &= \mathbb{E}_\beta \left[h_{\mathbb{E}_\mu[f_\omega](\bar{x}, Y)}(\vec{b}) \right] \\
&\leq \mathbb{E}_\beta \mathbb{E}_\mu [h_{f_\omega(\bar{x}, Y)}(\vec{b})] \\
&= \mathbb{E}_\mu \mathbb{E}_\beta [h_{f_\omega(\bar{x}, Y)}(\vec{b})] \\
&= \mathbb{E}_\mu [w(h_{f_\omega(\bar{x}, Y)}, \beta)]
\end{aligned}$$

□

We are now ready to bound the Rademacher mean width of an expectation using the Rademacher mean width of the underlying class, as mentioned in the body of the paper:

Theorem 2 (Pushing a Mean Width Bound through an Expectation). *Let (Ω, Σ, μ) be a probability space, and let $(f_\omega : \omega \in \Omega)$ be a family of real-valued functions $f : X \times Y \rightarrow [0, 1]$ such that for each $(x, y) \in X \times Y$ and $v \in \mathcal{R}^n$, the functions $\omega \mapsto f_\omega(x, y)$ and $\omega \mapsto h_{f_\omega(\bar{x}, Y)}(v)$ are measurable.*

Then

$$\mathcal{R}_{\mathbb{E}[f]}(n) \leq \sup_{\omega} \mathcal{R}_{f_\omega}(n).$$

Proof. For each n , where β is uniformly distributed on $\{-1, 1\}^n$, it suffices to show that

$$\sup_{\bar{x} \in X^n} w(\mathbb{E}_\mu[f_\omega](\bar{x}, Y), \beta) \leq \sup_{\omega} \sup_{\bar{x} \in X^n} w(f_\omega(\bar{x}, Y), \beta),$$

which, as the suprema commute, amounts to showing that for each $\bar{x} \in X^n$,

$$w(\mathbb{E}_\mu[f_\omega](\bar{x}, Y), \beta) \leq \sup_{\omega} w(f_\omega(\bar{x}, Y), \beta),$$

²This result is implicit in the proof of (Ben Yaacov, 2009, Theorem 4.1). For a fixed \bar{x} , it is stated there that $\mathbb{E}_\mu[f_\omega](\bar{x}, Y) \subseteq \mathbb{E}_\mu[\overline{\text{Conv}}(f_\omega(\bar{x}, Y))]$, where the latter expectation is an expectation of convex compact sets. This amounts to saying that for every $\vec{v} \in \mathcal{R}^n$,

$$h_{\mathbb{E}_\mu[f_\omega](\bar{x}, Y)}(\vec{v}) \leq \mathbb{E}_\mu[h_{f_\omega(\bar{x}, Y)}(\vec{v})].$$

which follows from Lemma 4 as

$$\mathbb{E}_\mu[w(f_\omega(\bar{x}, Y), \beta)] \leq \sup_\omega w(f_\omega(\bar{x}, Y), \beta).$$

□

5.3 Glivenko-Cantelli Bounds through Mean Width

Above we have seen how to estimate how moving to an expectation impacts means width. We will now look at the impact on Glivenko-Cantelli dimension. We can bound the Glivenko-Cantelli dimension with Rademacher mean width. The following is a restatement of (Wainwright, 2019, Theorem 4.10) in terms of GC-dimension, using the fact that for any probability measure μ on X and any f , the *Rademacher complexity* $\frac{1}{n}\mathbb{E}_{\mu^n}[w_{\mathcal{R}}(f(\bar{x}, Y))]$ is at most $\frac{1}{n}\mathcal{R}_f(n)$.

Fact 13. *Let $f : X \times Y \rightarrow [0, 1]$ be a real-valued function, which we view as a hypothesis class on X parametrized by Y . For any $\delta > 0$ and n , then*

$$GC_f\left(\frac{2 \cdot \mathcal{R}_f(n)}{n} + \delta, \exp\left(-\frac{n\delta^2}{2}\right)\right) \leq n.$$

We can rephrase this fact in a form that makes it easier to calculate the Glivenko-Cantelli dimension.

Lemma 5 (From Rademacher Width of a Base Class to GC of the Expectation class). *Let (Ω, Σ, μ) be a probability space, and let $(f_\omega : \omega \in \Omega)$ be a family of real-valued functions $f : X \times Y \rightarrow [0, 1]$ such that for each $(x, y) \in X \times Y$ and $v \in \mathcal{R}^n$, the functions $\omega \mapsto f_\omega(x, y)$ and $\omega \mapsto h_{f_\omega(\bar{x}, Y)}(v)$ are measurable.*

For any $\epsilon, \delta > 0$, if N is such that for all $n \geq N$, $\frac{\mathcal{R}_{f_\omega}(n)}{n} \leq \frac{\epsilon}{4}$ for each $\omega \in \Omega$, then

$$GC_{\mathbb{E}[f]}(\epsilon, \delta) \leq N + \frac{8}{\epsilon^2} \log \frac{1}{\delta}.$$

Roughly speaking, the lemma says that, when fixing ϵ , if we can find a linear bound on the Rademacher mean width, then we can bound the ϵ, δ GC dimension, which will allow us to get a bound on the ϵ, δ sample complexity.

Proof. Suppose that for all $n \geq N$ and $\omega \in \Omega$, $\frac{\mathcal{R}_{f_\omega}(n)}{n} \leq \frac{\epsilon}{4}$. Now fix $n \geq N + \frac{8}{\epsilon^2} \log \frac{1}{\delta}$, and observe that $\frac{\mathcal{R}_{f_\omega}(n)}{n} \leq \frac{\epsilon}{4}$ still holds for all $\omega \in \Omega$.

Then by Theorem 2, we see that n is also large enough that $\frac{\mathcal{R}_{\mathbb{E}[f]}(n)}{n} \leq \frac{\epsilon}{4}$. Then setting $\gamma = \epsilon - \frac{2\mathcal{R}_{\mathbb{E}[f]}(n)}{n}$, our assumption implies that $\gamma \geq \frac{\epsilon}{2}$. Plugging in γ for δ in Fact 13 we have $GC_{\mathbb{E}[f]}(\epsilon, \exp(-\frac{n\gamma^2}{2})) \leq n$. As $\gamma \geq \frac{\epsilon}{2}$ and $n \geq \frac{8}{\epsilon^2} \log \frac{1}{\delta}$, we have

$$\exp\left(-\frac{n\gamma^2}{2}\right) \leq \exp\left(-\frac{n\epsilon^2}{8}\right) \leq \delta$$

Thus the conclusion holds. □

5.4 Proof of Theorem 1 in the concept class case

We now apply the lemma on pushing Rademacher mean width through an expectation in the context of a concept class.

In this setting, we can estimate $\mathcal{R}_f(n)$ in terms of VC-dimension. For a concept class over X , indexed by Y , the characteristic function of the class is a function $X \times Y \rightarrow \{0, 1\}$.

Fact 14 ((Wainwright, 2019, Lemma 4.14 and Equation 4.24)). *Assume that f is the characteristic function of a concept class with VC-dimension at most d_f . Then for $n \geq 1$,*

$$\mathcal{R}_f(n) \leq 2 \cdot \sqrt{d_f \cdot n \cdot \log(n+1)}.$$

This fact will give us the bound on Rademacher mean width that we can plug in to the “pushing through expectation lemma”, Lemma 5.

Recall that \mathcal{H} is a $\{0, 1\}$ -valued class (that is, a concept class) with VC-dimension at most d . Putting Fact 14 together with Lemma 5, we see that to bound the GC-dimension of the randomization class of \mathcal{H} , it suffices to find N large enough that for $n \geq N$, $2\sqrt{\frac{d \log(n+1)}{n}} \leq \frac{\epsilon}{4}$, and add $\frac{1}{\epsilon^2} \log \frac{1}{\delta}$. This inequality is equivalent to

$$\frac{\log(n+1)}{n} \leq \frac{\epsilon^2}{64d},$$

which is guaranteed by

$$\frac{\log n}{n} \leq \frac{\epsilon^2}{64d \log 2},$$

where we call the constant on the right γ , noting that we can assume $\epsilon \leq 1$ and thus $\gamma < e^{-1}$. Because the function on the left is decreasing for $n > e$, it suffices to find some N for which this inequality holds. We try $N = C\gamma^{-1} \log \gamma^{-1}$, and see that

$$\frac{\log N}{N} = \frac{\log C + \log \gamma + \log \log \gamma^{-1}}{C\gamma \log \gamma^{-1}} \leq \gamma \left(\frac{\log C + 2 \log \gamma^{-1}}{C \log \gamma^{-1}} \right) \leq \gamma \left(\frac{\log C + 2}{C} \right),$$

using the fact that $\log \gamma^{-1} \geq 1$. For sufficiently large C (independent of γ), this is at most γ as desired. Thus

$$N = O(\gamma^{-1} \log \gamma^{-1}) = O\left(\frac{d}{\epsilon^2} \log \frac{d}{\epsilon^2}\right) = O\left(\frac{d}{\epsilon^2} \log \frac{d}{\epsilon}\right),$$

and

$$GC_{\mathbb{E}[f]}(\epsilon, \delta) \leq N + \frac{8}{\epsilon^2} \log \frac{1}{\delta} = O\left(\frac{1}{\epsilon^2} \left(d \log \frac{d}{\epsilon} + \log \frac{1}{\delta}\right)\right).$$

By Fact 11, this completes the proof of Theorem 1 in the case of concept classes.

5.5 Extension to the real-valued case

We now extend to get sample complexity bounds for random objects, but where we start with a class of real-valued functions. Our aim will be:

Theorem 3. *For any $\epsilon, \delta > 0$, if d is the $\frac{\epsilon}{50}$ fat-shattering dimension of f , then the sample complexity of agnostic PAC learning is bounded by:*

$$O\left(\frac{d}{\epsilon^4} \log^2 \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$$

The challenge will be in getting the required linear bound on Rademacher mean widths, so that we can apply the lemma on pushing mean width through an expectation, Lemma 5.

We will use covering numbers. We first recall unnormalized ℓ_p norms: the ℓ_p norm on \mathcal{R}^n is $(\sum_{i=1}^n x_i^p)^{1/p}$, while ℓ_∞ is $\max_{i=1}^n x_i$.

Definition 12. *For $A \subseteq \mathcal{R}^n$, $\gamma > 0$, and $1 \leq p \leq \infty$, we let $\mathcal{N}_p(\gamma, A)$ denote the minimum number of γ -balls in the ℓ_p -metric that cover A .*

We also let $\mathcal{N}_p(\gamma, f, n)$ denote $\sup_{\bar{x} \in X^n} \mathcal{N}_p(\gamma, f(\bar{x}, Y))$, with $f(\bar{x}, Y)$ defined as in Definition 10.

We have defined these for arbitrary p , but from now we will use only $p = 2$ and $p = \infty$. The relevant relation between them is that for all $x \in \mathcal{R}^n$, we have $|x|_2 \leq \sqrt{n}|x|_\infty$, so for any $A \subseteq \mathcal{R}^n$, $\mathcal{N}_2(\gamma\sqrt{n}, A) \leq \mathcal{N}_\infty(\gamma, A)$, and for any f and n ,

$$\mathcal{N}_2(\gamma\sqrt{n}, f, n) \leq \mathcal{N}_\infty(\gamma, f, n).$$

We can bound covering numbers using fat-shattering:

Fact 15 (From the proof of (Alon et al., 1997, Lemma 3.5)). *Let $f : X \times Y \rightarrow [0, 1]$ be a real-valued function, which we view as a hypothesis class on X parametrized by Y . Let d be the $\frac{\gamma}{4}$ fat-shattering dimension of f . Then*

$$\mathcal{N}_\infty(\gamma, f, n) \leq 2 \left(\frac{4n}{\gamma^2} \right)^{d \log(2en/d\gamma)}.$$

Here e is the base of the natural logarithm.

To connect covering numbers to Rademacher mean width, we pass through another width notion, *Gaussian mean width*.

Definition 13. *Let $\beta = (\beta_1, \dots, \beta_n)$, where the σ_i s are independent Gaussian variables with distribution $N(0, 1)$. We define the Gaussian mean width, denoted $w_G(A)$, as $w(A, \beta)$, where*

$$w(A, \beta) = \mathbb{E}_\beta \left[h_A(\vec{b}) \right]$$

as in the definition of Rademacher mean width.

We can easily relate Gaussian to Rademacher mean width, using the following fact:

Fact 16 ((Wainwright, 2019, Exercise 5.5)). *For any $A \subseteq [0, 1]^n$,*

$$w_{\mathcal{R}}(A) \leq \sqrt{\frac{\pi}{2}} w_G(A) \leq 2\sqrt{\log n} w_{\mathcal{R}}(A).$$

We will only use the first of these two inequalities, but together they show that Gaussian and Rademacher mean widths are closely connected. Covering numbers allow us to estimate how Gaussian mean width, and thus also Rademacher mean width, grows with dimension:

Fact 17 ((Wainwright, 2019, Equation 5.36)). *For $A \subseteq \mathcal{R}^n$ with ℓ_2 -diameter at most D , and $0 \leq \gamma \leq D$,*

$$w_G(A) \leq \gamma\sqrt{n} + 2D\sqrt{\log \mathcal{N}_2(\gamma, A)}.$$

We will concern ourselves with $A \subseteq [0, 1]^n$, so $D \leq \sqrt{n}$. Thus for $0 \leq \gamma \leq 1$, we may plug in $\gamma\sqrt{n} \leq D$, and get

$$w_G(A) \leq \gamma n + 2\sqrt{n \log \mathcal{N}_2(\gamma\sqrt{n}, A)}.$$

Combining the previous two facts gives us a straightforward way to relate covering numbers to Rademacher mean width:

Corollary 1. *Let $A \subseteq [0, 1]^n$ and let $\gamma \in [0, 1]$. Then*

$$w_{\mathcal{R}}(A) \leq \sqrt{\frac{\pi}{2}} \left(\gamma n + 2\sqrt{n \log \mathcal{N}_2(\gamma n, A)} \right) \leq \sqrt{\frac{\pi}{2}} \left(\gamma n + 2\sqrt{n \log \mathcal{N}_\infty(\gamma, A)} \right).$$

We now prove the remainder of Theorem 1, using the following bound on Glivenko-Cantelli dimension:

Theorem 4. *Let $f : X \times Y \rightarrow [0, 1]$, thus f can be considered as a hypothesis class of real valued functions on X , indexed by Y . For any $\epsilon, \delta > 0$, if d is the $\frac{\epsilon}{50}$ fat-shattering dimension of f , then*

$$GC_{\mathbb{E}[f]}(\epsilon, \delta) = O \left(\frac{d}{\epsilon^4} \log^2 \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \log \frac{1}{\delta} \right)$$

.

Proof. By Lemma 5, it suffices to show that there is a constant $C > 0$ such that if

$$n \geq C \frac{d}{\epsilon^4} \log^2 \frac{d}{\epsilon},$$

then

$$\frac{\mathcal{R}_f(n)}{n} \leq \frac{\epsilon}{4}.$$

As an intermediate bound, we can use Corollary 1 to bound the Rademacher complexity in terms of covering numbers, using $\gamma = \frac{\epsilon}{\sqrt{32\pi}}$:

$$\begin{aligned} \frac{\mathcal{R}_f(n)}{n} &= \sup_{\bar{x} \in X^n} \frac{w_{\mathcal{R}}(f(\bar{x}, Y))}{n} \\ &\leq \sup_{\bar{x} \in X^n} \sqrt{\frac{\pi}{2}} \left(\gamma + 2\sqrt{\frac{\log \mathcal{N}_{\infty}(\gamma, f(\bar{x}, Y))}{n}} \right) \\ &\leq \sqrt{\frac{\pi}{2}} \left(\gamma + 2\sqrt{\frac{\log \mathcal{N}_{\infty}(\gamma, f, n)}{n}} \right) \\ &= \frac{\epsilon}{8} + \sqrt{\frac{2\pi \log \mathcal{N}_{\infty}(\gamma, f, n)}{n}}. \end{aligned}$$

Thus it suffices to show that for suitably large n ,

$$\sqrt{\frac{2\pi \log \mathcal{N}_{\infty}(\gamma, f, n)}{n}} \leq \frac{\epsilon}{8},$$

or equivalently,

$$\frac{\log \mathcal{N}_{\infty}(\gamma, f, n)}{n} \leq \frac{\epsilon^2}{128\pi}.$$

By Fact 15, we see that for all n ,

$$\log \mathcal{N}_{\infty}(\gamma, f, n) = O\left(d \log\left(\frac{4n}{\gamma^2}\right) \log\left(\frac{2en}{d\gamma}\right)\right) = O\left(d \log^2\left(\frac{n}{\gamma^2}\right)\right) = O\left(d \log^2\left(\frac{32\pi n}{\epsilon^2}\right)\right).$$

Now let D be the constant of this inequality, so that for all n ,

$$\log \mathcal{N}_{\infty}(\gamma, f, n) \leq Dd \log^2\left(\frac{32\pi n}{\epsilon^2}\right).$$

It now suffices to show for suitably large n that

$$\frac{Dd \log^2\left(\frac{32\pi n}{\epsilon^2}\right)}{n} \leq \frac{\epsilon^2}{128\pi}.$$

Setting $a = \frac{32\pi}{\epsilon^2}$ and $b = \frac{\epsilon^2}{128\pi Dd}$, we may assume that $a \geq 1$ and $\log(ab^{-1}) > 1$. We can restate our desired inequality as

$$\frac{\log^2(an)}{bn} \leq 1,$$

and for some $C > 0$, we see that if $n \geq Cab^{-1} \log^2(ab^{-1})$, then

$$\frac{\log^2(an)}{bn} \leq \frac{(\log C + \log(a^2b^{-1}) + \log \log^2(ab^{-1}))^2}{bC(ab^{-1}) \log^2(ab^{-1})} \leq \frac{(\log C + 4 \log(ab^{-1}))^2}{aC \log^2(ab^{-1})} \leq \frac{(\log C / \log(ab^{-1}) + 4)^2}{C} \leq \frac{(\log C + 4)^2}{C},$$

and it is clear that this bound is at most 1 for large C .

We then see that

$$Cab^{-1} \log^2(ab^{-1}) = O\left(\frac{d}{\epsilon^4} \log^2 \frac{d}{\epsilon^4}\right) = O\left(\frac{d}{\epsilon^4} \log^2 \frac{d}{\epsilon}\right),$$

which completes the proof. \square

Theorem 1 follows from Theorem 4 using Fact 11.

5.6 Simpler and better VC bounds for the distribution function class

A much simpler argument is available that provides bounds for the distribution class or dual distribution class. We explain the idea for the dual distribution class formed over a concept class. We know that if a concept class is agnostic PAC learnable, then so is its dual class, where the concepts are given by elements x of the range space. We can then conclude by routine calculation with dimensions, that for each k , the functions on the parameter space given by normalized sums of elements $\frac{x_1 + \dots + x_k}{k}$ is agnostic PAC learnable. But by a basic result in statistical learning theory, the fact that our original concept class is PAC learnable means we can approximate each function in the dual distribution class arbitrarily closely by normalized sums.

Recall that in the special case of a concept class, the result for the dual distribution class had been proven in prior work (Hu et al., 2022). The simpler proof we present here, like the analytic proof that goes via the randomization class, improves on the bound given in prior work. We now explain the idea of this alternative approach.

Fix a concept class \mathcal{C} on X indexed by parameter set Y such that \mathcal{C} has finite VC dimension $d_{\mathcal{C}}$ and dual VC dimension $d_{\mathcal{C}}^*$.

We let $\chi_x^{\mathcal{C}}$ be the dual family of characteristic functions: the family of functions on Y , indexed by elements of X , given by $\chi_x^{\mathcal{C}}(c) = 1$ if $x \in \mathcal{C}_c$, 0 otherwise.

Thus for any γ , the γ fat-shattering dimension of this family is just the dual VC dimension $d_{\mathcal{C}}^*$. This is for every $1 > \gamma > 0$, independent of γ .

We let $\text{Avg}_m(v_1 \dots v_m)$ be the average of $v_1 \dots v_m$, thus Avg_m is a function from \mathcal{R}^m to \mathcal{R} .

Let $\chi_m^{\mathcal{C}}$ be the composed class, indexed by $x_1 \dots x_m \in X$, with each function taking $c \in Y$ to

$$\text{Avg}_m(\chi_{x_1}^{\mathcal{C}}(c) \dots \chi_{x_m}^{\mathcal{C}}(c))$$

This fits the composition framework in Theorem 1 of (Attias & Kontorovich, 2024).

Fact 18. *Attias & Kontorovich (2024) The γ fat-shattering dimension of the composed class $\chi_m^{\mathcal{C}}$ is bounded by:*

$$25 \cdot D_{\gamma} \log^2(90 \cdot D_{\gamma})$$

where D_{γ} here is the sum from 1 to m of the γ fat-shattering dimension of the constituent class, which in this case is just $m \cdot d_{\mathcal{C}}^*$.

Thus the γ fat-shattering dimension of the class $\chi_m^{\mathcal{C}}$ is bounded by:

$$25 \cdot m \cdot d_{\mathcal{C}}^* \cdot [\log(90) + \log m + \log d_{\mathcal{C}}^*]^2$$

Call this $J(m, d_{\mathcal{C}}^*)$.

We now want to control the relationship between arbitrary measures and averages.

Fix $\gamma > 0$. Recall that $\mathcal{DualDistr}_{\mathcal{C}}$ is the dual distribution function class for the concept class \mathcal{C} . Suppose the γ fat-shattering dimension of $\mathcal{DualDistr}_{\mathcal{C}}$ were greater than or equal to k' . Let n_k be large enough that every measure has a $\frac{\gamma}{2}$ -approximation of size n_k : that is, there is a set of size n_k elements of X such that for any \tilde{c} , the percentage of elements in the set satisfying $\mathcal{C}_{\tilde{c}}$ is within $\frac{\gamma}{2}$ of the measure of the set. Then the $\frac{\gamma}{2}$ fat-shattering dimension of the class $\chi_{n_k}^{\mathcal{C}}$ would be above k' .

Thus $k' \leq J(n_k, d_{\mathcal{C}}^*)$, or restated, the γ fat-shattering dimension of $\mathcal{DualDistr}_{\mathcal{C}}$ is bounded by $J(n_k, d_{\mathcal{C}}^*)$.

We can use the following fact, which can be found in standard learning theory texts: see, e.g. (Li et al., 2001) or for an exposition (Raban, 2023) Theorem 4.2.

Fact 19. *There is a constant L , such that if we let n_k be such that $L \cdot \frac{\sqrt{d_{\mathcal{C}}}}{\sqrt{n_k}} \leq \frac{\gamma}{2}$, then every measure has a $\frac{\gamma}{2}$ -approximation of size n_k .*

Thus a bound for n_k can be taken to be:

$$\frac{L' \cdot d_{\mathcal{C}}}{\gamma^2}$$

for another universal constant L' .

Plugging into the bound for $J(n_k, d_{\mathcal{C}}^*)$, and ignoring log factors and universal constants, we get a bound on the γ fat-shattering dimension for $DualDistr_{\mathcal{C}}$ on the order of:

$$\frac{(d_{\mathcal{C}}) \cdot d_{\mathcal{C}}^*}{\gamma^2}$$

Plugging into the sample complexity bound of Fact 4 we get the sample complexity of agnostic PAC learning $Distr_{\mathcal{C}}$ is bounded by:

$$O\left(\frac{1}{\epsilon^2} \cdot \left[\frac{d_{\mathcal{C}} \cdot d_{\mathcal{C}}^*}{\left(\frac{\epsilon}{9}\right)^2} \log^2\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

Recall from the body of the paper that in (Hu et al., 2022) the dual distribution class over a concept class was considered, and a sample complexity bound of $\tilde{O}\left(\frac{1}{\epsilon^{\lambda+1}}\right)$ was obtained, where \tilde{O} indicates that we drop terms that are polylogarithmic in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ for constant λ . In contrast, in our bound above, we have a constant in the exponent of ϵ in the denominator, rather than the VC-dimension.

We now show that this approach generalizes easily from concept classes to function classes. That is, we give an alternative proof of the preservation of agnostic PAC learnability, without going through randomization. We do not compute sample complexity bounds explicitly for this alternative proof, but they are similar to those given above.

Theorem 5. *Let \mathcal{H} be a class of functions over X , indexed by Y . Suppose \mathcal{H} is agnostic PAC learnable (equivalently has finite γ fat-shattering dimension for each γ) then the same holds for the dual distribution function class (and the distribution function class) of \mathcal{H} .*

We let \mathcal{H}^* be the dual family. Now this is a class of functions over Y , indexed by X . Since agnostic PAC learning for real-valued functions is closed under dualization by Fact 7, for any γ , the γ fat-shattering dimension of this family is also finite.

As before, let $\text{Avg}_m(v_1 \dots v_m)$ be the average of $v_1 \dots v_m$. Let $\text{Avg}_m(\mathcal{H}^*)$ be the composed class, indexed by $x_1 \dots x_m$ in X^m taking c to $\text{Avg}_m(h_{x_1}(c) \dots h_{x_m}(c))$.

This again fits the composition framework in Theorem 1 of (Attias & Kontorovich, 2024). From the theorem we have the γ fat-shattering dimension of the composed class is bounded by:

$$25 \cdot m \cdot D_{\gamma} \cdot \log^2(90 \cdot D_{\gamma})$$

where D_{γ} here is the sum from 1 to m of the γ fat-shattering dimension of the constituent class, which in this case is just $m \cdot \text{FatSHDim}_{\gamma}(\mathcal{H}^*)$.

Thus the γ fat-shattering dimension of the class $\text{Avg}_m(\mathcal{H}^*)$ is bounded by:

$$25 \cdot (m^2) \cdot \text{FatSHDim}_{\gamma}(\mathcal{H}^*) [\log 90 + \log m + \log \text{FatSHDim}_{\gamma}(\mathcal{H}^*)]^2$$

Call this $J(m, \text{FatSHDim}_{\gamma}(\mathcal{H}^*))$.

Fix γ , and again let $DualDistr_{\mathcal{H}}$ be the dual distribution function class for \mathcal{H} . Suppose the γ fat-shattering dimension of $DualDistr_{\mathcal{H}}$ were greater than or equal to k' . This time let n_k be large enough that every

distribution has an n_k sized $\frac{\gamma}{2}$ -approximation, in the sense that there is an n_k sized tuple (x_1, \dots, x_{n_k}) in X such that for any $h \in \mathcal{H}$, the average of h over the elements of the tuple is within $\frac{\gamma}{2}$ of the mean of h - that is,

$$\left| \frac{1}{n_k} \sum_{i=1}^{n_k} h(x_i) - \mathbb{E}[h] \right| \leq \frac{\gamma}{2}.$$

Then the $\frac{\gamma}{2}$ fat-shattering dimension of the class $\chi_{n_k}^\phi$ would be above k' .

Thus $k' \leq J(n_k, \text{FatSHDim}_\gamma(\mathcal{H}^*))$. Restated: the γ fat-shattering dimension of $\mathcal{DualDistr}_\mathcal{H}$ is bounded by $J(n_k, \text{FatSHDim}_\gamma(\mathcal{H}^*))$.

It suffices to take n_k large enough that, for some fixed $0 < \delta < 1$, $GC_\mathcal{H}(\frac{\gamma}{2}, \delta) \leq n_k$, because if this is the case, the probability that a randomly-selected tuple is a $\frac{\gamma}{2}$ -approximation is at least $1 - \delta$. Thus, fixing δ , we may use

$$n_k = O\left(\frac{1}{\epsilon^2} \cdot \text{FatSHDim}_\epsilon(\mathcal{H}) \cdot \log^2\left(\frac{1}{\epsilon}\right)\right).$$

5.7 The realizable case for PAC learning

Randomization, and the related distribution class constructions, do not preserve PAC learnability in the realizable case.

Proposition 5. *There is a hypothesis class \mathcal{H} that is realizable PAC learnable, but the distribution function class, dual distribution function class, and random variable class based on \mathcal{H} are not realizable PAC learnable.*

In the proof, we let \mathcal{H}_0 be the following slight modification of a class from (Attias et al., 2023, Example 1). Let X be a set, partitioned into nonempty pieces X_0, X_1, \dots , with characteristic functions χ_{X_i} . Let $B \subseteq \{0, 1\}^\mathbb{N}$ consist of all sequences of bits with only finitely many ones, and let $\mathcal{H}_0 = \{h_b : b \in B\}$, where if $b = (b_0, b_1, \dots)$, then

$$h_b(x) = \frac{3}{4} \sum_{i=0}^{\infty} b_i \cdot \chi_{X_i}(x) + \frac{1}{8} \sum_{i=0}^{\infty} b_i \cdot 2^{-i}.$$

The class \mathcal{H}_0 has infinite γ fat-shattering dimension for all $\gamma < \frac{1}{4}$, so is not agnostic PAC learnable. But it is realizable learnable – in fact, it is learnable from one sample, as for any x and distinct $h, h' \in \mathcal{H}_0$, $h(x) \neq h'(x)$.

We can easily see that the dual distribution class of this class is not realizable PAC learnable:

Lemma 6. *The dual class of \mathcal{H}_0 is not PAC learnable in the realizable case. Hence the dual distribution class is not PAC learnable in the realizable case.*

Proof. A dual class element is given by an x_0 in the range space, with any other element in the same partition X_i as x_0 inducing the same function. If we see samples $\vec{b}_1 \dots \vec{b}_n$, with value at most $\frac{1}{8}$, we will be able to exclude some partition elements X_i , but we will have infinitely many X_i possible. \square

We now show the same thing for the distribution class.

Lemma 7. *The distribution class and (hence) the random variable class of \mathcal{H}_0 are not PAC learnable in the realizable case. In fact, we may simply look at the class on X of “two choice distributions”, consisting of all hypotheses $\lambda \cdot h_b + (1 - \lambda) \cdot h_{b'}$ for rational $\lambda \in [0, 1]$ with $b, b' \in B$.*

Proof. We prove this by showing that this new class has infinite $\frac{1}{8}$ -graph dimension, as defined in (Attias et al., 2023), which we now review. Fix a natural number n . For our purposes, we only need to know when a class has $\frac{1}{4}$ -graph dimension at least n . The definition of graph dimension states that this holds when we have $x_0, \dots, x_{n-1} \in X$, $f_1, \dots, f_{n-1} \in [0, 1]$, and for each $\beta \in \{0, 1\}^n$, a hypothesis h'_β such that

- $h'_\beta(x_i) = f_i$ when $\beta_i = 0$
- $|h'_\beta(x_i) - f_i| > \frac{1}{8}$ when $\beta_i = 1$.

Note that, unlike with fat-shattering, we have an equality for the zero values of a branch and an inequality for the one value. Theorem 1 of (Attias et al., 2023) shows that an infinite $\frac{1}{4}$ -graph dimension — that is, having such witnesses for each n — implies that the class is not realizable PAC learnable. To find the required witnesses, let $x_i \in X_i$ for $i < n$, let $f_i = \frac{1}{2}$ for each i , and let $1_n = (1, 1, \dots, 1, 0, 0, \dots) \in \{0, 1\}^{\mathbb{N}}$ be such that $(1_n)_i = 1$ exactly when $i < n$. For each $\beta \in \{0, 1\}^n$, let $\beta' \in \{0, 1\}^{\mathbb{N}}$ be the sequence extending β with zeroes, that is, $\beta'_i = 0$ for $i \geq n$. For one more piece of notation, we let

$$c_{\beta'} = \frac{1}{8} \sum_{i=0}^n \beta'_i \cdot 2^{-i} \text{ and}$$

$$c_{1_n} = \frac{1}{8} \sum_{i=0}^n (1_n)_i \cdot 2^{-i},$$

noting that $0 \leq c_{\beta'}, c_{1_n} < \frac{1}{4}$ are both rational.

We claim that there is some $\lambda \in [0, 1] \cap \mathbb{Q}$ such that letting $h'_\beta = \lambda \cdot h_{\beta'} + (1 - \lambda) \cdot h_{1_n}$ for each β , we will obtain the two required properties above. We find that for x_i with $i < n$, we have, for each β

$$\begin{aligned} h'_\beta(x_i) &= \lambda h_{\beta'}(x_i) + (1 - \lambda) h_{1_n}(x_i) \\ &= \lambda \left(\frac{3}{4} \beta_i + c_{\beta'} \right) + (1 - \lambda) \left(\frac{3}{4} + c_{1_n} \right) \end{aligned}$$

Again fixing β , we observe that $c_{\beta'} < \frac{1}{2} < \frac{3}{4} + c_{1_n}$ and both $c_{\beta'}$ and c_{1_n} are rational. Thus we can choose $\lambda \in [0, 1] \cap \mathbb{Q}$ with

$$\lambda(c_{\beta'}) + (1 - \lambda) \left(\frac{3}{4} + c_{1_n} \right) = \frac{1}{2}.$$

Then for each i , if $\beta_i = 0$, we have

$$h'_\beta(x_i) = \lambda(c_{\beta'}) + (1 - \lambda) \left(\frac{3}{4} + c_{1_n} \right) = \frac{1}{2} = f_i,$$

and if $\beta_i = 1$, we have both $h_{\beta'}(x_i), h_{1_n}(x_i) \geq \frac{3}{4}$, so $h'_\beta(x_i) = \lambda \cdot h_{\beta'}(x_i) + (1 - \lambda) \cdot h_{1_n}(x_i) \geq \frac{3}{4}$. \square

We now show that the logic perspective in some sense “fixes” these anomalies. If we consider “global learnability of a structure” – realizable PAC learnability for *all* definable families in a model – then there is no difference between the realizable and agnostic case, and we do have preservation under moving to statistical structures.

Proposition 6. *For any CL structure \mathfrak{C} , and definable predicate $\phi(x; y)$ with infinite γ fat-shattering dimension for some $\gamma > 0$, there is another definable predicate which defines a class that is not realizable PAC learnable. Thus in particular if every definable family for \mathfrak{C} is realizable PAC learnable, then every definable family is agnostic PAC learnable, and thus every definable family in the randomized structure (and thus every distribution class or dual distribution class family) is agnostic PAC learnable and realizable PAC learnable.*

We have stated the proposition for simplicity in the case where the partitioned formula has two free variables, but it extends to the general case of a partitioned formula.

Before beginning the proof of the proposition, we recall a basic result on the fat shattering dimension. If a class has infinite γ fat-shattering dimension for some $\gamma > 0$, this means we get arbitrary large powersets that we can capture with a γ -gap. The following result states that realize these counterexamples with a uniform choice of number $r < s$ that are at least γ -apart:

Fact 20 ((Alon et al., 1997, Thm. 4.1)). *Consider a real-valued hypothesis class \mathcal{H} consisting of function h_y for y in some parameter set. Suppose that \mathcal{H} has infinite γ fat-shattering dimensions. Then for every natural number d , there are also $0 \leq r < s \leq 1$ such that $s - r \geq \gamma$ and there are $x_1, \dots, x_d \in X$ and $(y_b : b \in \{0, 1\}^d)$ such that for each b and i , if $b(i) = 0$, then $h(x_i; y_b) \leq r$, and if $b(i) = 1$, then $h(x_i; y_b) \geq s$.*

In the literature, this is sometimes phrased by saying that \mathcal{H} has “infinite V_γ -dimension” (where V is for Vapnik).

We now begin the proof of Proposition 6:

Proof. Fix \mathfrak{C} and $\phi(x; y)$. Assume that $\phi(x; y)$ has infinite γ fat-shattering dimension for some $\gamma > 0$. Thus we get arbitrary large powersets that we can capture with a γ -gap. Then by Fact 20 for every natural number d , there are also $0 \leq r < s \leq 1$ such that $s - r \geq \gamma$ and there are $x_1, \dots, x_d \in X$ and $(y_d : b \in \{0, 1\}^d)$ such that for each b and i , if $b(i) = 0$, then $\phi(x_i; y_b) \leq r$, and if $b(i) = 1$, then $\phi(x_i; y_b) \geq s$.

Let $f : [0, 1] \rightarrow [0, 1]$ be a continuous, monotone function such that $f(r) = 0$ and $f(s) = 1$. Then consider the real-valued function $\tau(x, y)$ defined as $f(\phi(x; y))$, which is also a definable predicate.

We now see that for every d , there are x_1, \dots, x_d in the domain of \mathfrak{C} and $(y_d : b \in \{0, 1\}^d)$ such that for each b and i , if $b(i) = 0$, then $\tau(x_i; y_b) = 0$, and if $b(i) = 1$, then $\tau(x_i; y_b) = 1$, so the 1-graph dimension is at least d , so the class defined by $\tau(x; y)$ is not PAC learnable in the realizable case. \square

6 Online learnability of statistical classes

We will now be interested in getting an analog of Theorem 1 for online learning, deriving bounds on learnability for statistical classes based on dimensions of the base class. We will calculate the following regret bound in terms of the sequential fat-shattering dimension:

Theorem 6. *For a function class \mathcal{H} having $\frac{\epsilon}{50}$ sequential fat-shattering dimension at most d , the minimax regret of online learning for the randomization class of \mathcal{H} over runs of length n is at most*

$$4 \cdot \gamma \cdot n + 12 \cdot (1 - \gamma) \cdot \sqrt{d \cdot n \cdot \log \left(\frac{2 \cdot e \cdot n}{\gamma} \right)}.$$

If \mathcal{H} is $\{0, 1\}$ -valued and has Littlestone dimension at most d , the minimax regret of online learning for the randomization class of \mathcal{H} over runs of length n is at most $O(\sqrt{d \cdot n})$.

As finiteness of γ sequential fat-shattering dimension for all $\gamma > 0$ is equivalent to sublinear minimax regret by Fact 6, we have the following corollary:

Corollary 2. *If function class \mathcal{H} is agnostic online learnable, in the sense of having sublinear minimax regret, then so is its randomization class, and thus so is its distribution class and the dual distribution class.*

For the corollary, we use that the restrictions of the randomization class to deterministic range elements give the dual distribution class. We also use the fact that agnostic online learnability is closed under dualization by Proposition 1.

6.1 Proof of the main theorems on agnostic online learnability of statistical classes, with quantitative bounds

We now present the proof of Theorem 6.

Recall that to push agnostic PAC learning from a base class into a statistical class, we went through notions of width. We will do something similar here.

To bound regret for online learning for the randomization class, we will adapt the framework of sequential Rademacher mean width developed in (Rakhlin et al., 2015a;b). We will define it here, slightly amending the definition to better match our conventions for mean width.

Definition 14 (Sequential Rademacher mean width). *For any n , let $\{1, -1\}^{<n} = \bigcup_{t=0}^{n-1} \{1, -1\}^t$. For each sequence $s = (s_1, \dots, s_n) \in \{1, -1\}^n$, let $v_s \in \mathcal{R}^{\{1, -1\}^{<n}}$ be the vector such that for $1 \leq t \leq n$, $(v_s)_{(s_1, \dots, s_{t-1})} = s_{t+1}$, while all other entries are 0. Then let $T_n = \{v_s : s \in \{1, -1\}^n\}$. Recall the definition of mean width from Definition 10. Define the sequential Rademacher mean width of a set $A \subseteq \mathcal{R}^{\{1, -1\}^{<n}}$, $w_{\mathcal{R}}^S(A)$, to be $w(A, \beta)$ where β is the uniform distribution on the 2^n elements of T_n .*

For a function $f(x, y)$, we then define the sequential Rademacher mean width, $\mathcal{R}^{Seq}_f(n)$, to be $\sup_{\bar{x} \in X^{\{1, -1\}^{<n}}} w_{\mathcal{R}^{Seq}}(f(\bar{x}, Y))$. This definition coincides with the one from (Rakhlin et al., 2015a) except for the factor of $\frac{1}{n}$, although we call it a mean width instead of a complexity.

Our argument will rely on the following bound, extracted from (Rakhlin et al., 2015b, Proposition 9). The differences between this fact and the statement in (Rakhlin et al., 2015b) are due to our slightly different definition and the fact that our function classes take values in $[0, 1]$ instead of $[-1, 1]$.

Fact 21 (See (Rakhlin et al., 2015b, Proposition 9)). *Let \mathcal{H} be a function class on X taking values in $[0, 1]$. The minimax regret of online learning for \mathcal{H} on a run of length n is at most $\mathcal{R}^{Seq}_{\mathcal{H}}(n)$, which is in turn at most*

$$\inf_{\gamma} 4 \cdot \gamma \cdot n + 12 \cdot \sqrt{n} \cdot \int_{\gamma}^1 \sqrt{\text{FatSHDim}_{\beta}^{Seq}(\mathcal{H}) \cdot \log\left(\frac{2 \cdot e \cdot n}{\beta}\right)} d\beta.$$

As with Rademacher mean width, the advantage of this dimension is that we can push it through expectations.

Theorem 7. *Let (Ω, Σ, μ) be a probability space, and let $(f_{\omega} : \omega \in \Omega)$ be a family of real-valued functions $f : X \times Y \rightarrow [0, 1]$ such that for each $(x, y) \in X \times Y$ and $v \in \mathcal{R}^n$, the functions $\omega \mapsto f_{\omega}(x, y)$ and $\omega \mapsto h_{f_{\omega}(\bar{x}, Y)}(v)$ are measurable. Then*

$$\mathcal{R}^{Seq}_{\mathbb{E}[f]}(n) \leq \sup_{\omega} \mathcal{R}^{Seq}_{f_{\omega}}(n).$$

Proof. Recall that the definition of sequential Rademacher mean width is the same as Rademacher mean width, except with a different probability distribution.

For any n , the distribution β we use on $\mathcal{R}^{\{1, -1\}^{<n}}$ is the uniform distribution on a particular finite set T_n . Then the sequential Rademacher mean width of a set $A \subseteq \mathcal{R}^{\{1, -1\}^{<n}}$, $w_{\mathcal{R}}^S(A)$ is $w(A, \beta)$. For a function $f(x, y)$, the sequential Rademacher mean width was defined by

$$\mathcal{R}^{Seq}_f(n) = \sup_{\bar{x} \in X^{\{1, -1\}^{<n}}} w_{\mathcal{R}^{Seq}}(f(\bar{x}, Y)).$$

This allows us to restate this theorem statement as showing that for each n ,

$$\sup_{\bar{x} \in X^{\{1, -1\}^{<n}}} w(\mathbb{E}_{\mu}[f_{\omega}](\bar{x}, Y), \beta) \leq \sup_{\omega} \sup_{\bar{x} \in X^{\{1, -1\}^{<n}}} w(f_{\omega}(\bar{x}, Y), \beta),$$

and the proof of this is essentially identical to the proof of Theorem 2, but with a new distribution β . \square

From this, we can prove a bound on regret for the randomization class in terms of a single sequential fat-shattering dimension, proving the first part of Theorem 6.

Theorem 8. *The minimax regret of online learning for the randomization class of \mathcal{H} with γ sequential fat-shattering dimension at most d on a run of length n is at most*

$$4 \cdot \gamma \cdot n + 12 \cdot (1 - \gamma) \cdot \sqrt{d \cdot n \cdot \log\left(\frac{2 \cdot e \cdot n}{\gamma}\right)}.$$

Proof. By Theorem 7, any bound on $\mathcal{R}^{Seq}_{\mathcal{H}}(n)$ also applies to the randomization class, so regret for the randomization class is bounded by

$$4 \cdot \gamma \cdot n + 12 \cdot \sqrt{n} \cdot \int_{\gamma}^1 \sqrt{\text{FatSHDim}_{\beta}^S(\mathcal{H}) \cdot \log\left(\frac{2 \cdot e \cdot n}{\beta}\right)} d\beta.$$

Because the function in the integral is decreasing in β , we may bound it naïvely by the value at γ . \square

In case \mathcal{H} is a $\{0, 1\}$ -valued concept class, all sequential fat-shattering dimensions coincide with the Littlestone dimension, and the following improved bound holds:

Fact 22 ((Alon et al., 2021, See Lemma 6.4 and Theorem 12.2)). *If \mathcal{H} is a $\{0, 1\}$ -valued concept class with Littlestone dimension at most d , then*

$$\mathcal{R}^{Seq}_{\mathcal{H}}(n) = O(\sqrt{d \cdot n}).$$

From this, we are able to conclude that the minimax regret for the randomization class of \mathcal{H} is also at most $O(\sqrt{d \cdot n})$, which proves the second part of Theorem 6.

6.2 Preservation of online learnability in moving to statistical classes, via logic

Above we showed that online learnability is preserved in moving to statistical classes. Without the quantitative bounds, this can also be derived from prior results in model theory. We can show that agnostic online learnability is equivalent to the notion of *stability* — a strengthening of NIP — for real-valued logic. We can then use prior results (Ben Yaacov, 2009; 2013) on the preservation of stability in real-valued logic under randomization. We now explain how to analyze online learnability via notions from logic, beginning with online learnable concept classes coming from first-order formulas over classical structures.

Consider a first-order structure \mathfrak{M} and a formula with partitioned variables $\phi(\vec{x}, \vec{y})$. We say ϕ is *stable* in \mathfrak{M} if there do not exist arbitrarily large sequences $\vec{a}_i, \vec{b}_i : 1 \leq n$ such that $\forall i, j \leq n \phi(\vec{a}_i, \vec{b}_j)$ if and only if $i < j$. Roughly speaking stability says that ϕ does not define arbitrarily large linear orders.

Recall that for PAC learning of concept classes the critical dimension is VC dimension or NIP: not having arbitrarily large shattered sets. Stability is the analogous dividing line for online learning, agnostic or realizable, in the case of concept classes:

Fact 23. (Chase & Freitag, 2019) *A formula $\phi(\vec{x}, \vec{y})$ is stable if and only if the concept class \mathcal{C}_{ϕ} is online learnable.*

Here we refer to learnability either in the realizable case or the agnostic case, which are equivalent for a concept class, as noted in the preliminaries. Both are equivalent to \mathcal{C}_{ϕ} having finite Littlestone dimension by Fact 6.

A first-order structure is said to be stable when every partitioned formula is stable. As with NIP, many classical structures are already known to be stable, which implies that all definable families are online learnable. These include the complex field, vector spaces, and commonly studied equivalence relations (Simon, 2015, Example 2.61).

Thus far we are reviewing a connection between the definability property stability and online learnability, which is already known. We now turn to real-valued structures and continuous logic, where to the best of our knowledge, there was no characterization of which structures were online learnable. Recall that in the case of continuous logic structures partitioned real-valued formulas define classes of real-valued functions. The notion of a stable formula generalizes to this setting, but now requires a real parameter:

Fact 24 ((Ben Yaacov & Usvyatsov, 2010, Section 7)). *If $\phi(x; y)$ is a continuous logic formula well-defined over a structure \mathfrak{C} and $\gamma > 0$, the following are equivalent:*

- *There is some d such that in \mathfrak{C} there are no $a_1, \dots, a_d, b_1, \dots, b_d$ such that for all $i < j$,*

$$|\phi(a_i; b_j) - \phi(a_j; b_i)| \geq \gamma$$

- There is some d such that in \mathfrak{C} there are no $a_1, \dots, a_d, b_1, \dots, b_d$ and r, s with $r + \gamma \leq s$ such that for all $i < j$, $\phi(a_i; b_j) \leq r$ and $\phi(a_j; b_i) \geq s$.

We call such a formula γ -stable. If $\phi(x; y)$ is γ -stable for all $\gamma > 0$, then it is called stable. Roughly speaking stability says that we cannot use gaps in function values, discretized up to some γ , to define arbitrarily large linear orders.

The main result in this subsection shows that the connection between stability and online learnability extends to real-valued classes:

Theorem 9. *A partitioned formula $\phi(x; y)$ of continuous logic defined over a structure \mathfrak{C} is stable if and only if for every $\gamma > 0$, the sequential fat-shattering dimension $\text{FatSHDim}_\gamma^{\text{Seq}}(\mathcal{H}_\phi)$ is finite.*

Specifically, if $\phi(x; y)$ is δ -stable for $0 < \delta < \frac{\gamma}{2}$, then $\text{FatSHDim}_\gamma^{\text{Seq}}(\mathcal{H}_\phi)$ is finite, and if $\text{FatSHDim}_\gamma^{\text{Seq}}(\mathcal{H}_\phi)$ is finite, then $\phi(x; y)$ is γ -stable.

The interest in the above theorem is that stability has already been shown to be preserved in moving to from a base class to a statistical class:

Fact 25. *(Ben Yaacov & Keisler, 2009; Ben Yaacov, 2013) If a partitioned CL formula $\phi(\vec{x}, \vec{y})$ is stable over a real-valued model \mathfrak{C} , then the corresponding formula $E\phi(\vec{X}; \vec{Y})$ over the randomization is also stable in the randomization \mathfrak{C} .*

Thus we can combine Theorem 9, Fact 25, and the characterization of agnostic online learnability via sequential fat-shattering in Fact 6 to give another proof that preservation of agnostic online learning in moving to statistical classes:

Corollary 3. *For any partitioned $\phi(\vec{x}; \vec{y})$ over a classical or real-valued model \mathfrak{C} , the hypothesis class of ϕ is agnostic online learnable if and only if the randomized version $E\phi(\vec{X}; \vec{Y})$ is agnostic online learnable. In particular, if the hypothesis class associated to the formula is agnostic online learnable, so are the distribution class and the dual distribution class.*

Again, we emphasize, that have stated the result here for classes defined by formulas, since this matches the prior development in model theory. But the result extends to general real-valued classes.

We now turn to proving Theorem 9:

Proof. Assume that $\phi(x; y)$ is not γ -stable. We will show that \mathcal{H}_ϕ γ fat-shatters a binary tree of depth d . First, we linearly order the set $\{-1, 1\}^{\leq d}$ in a variation of lexicographical fashion, so that for any string s of length k with $k < d$, and any string t strictly extending s , if $t_k = -1$ then $t < s$, and if $t_k = 1$ then $t > s$. Thus by instability, there are $(a_s b_s : s \in \{-1, 1\}^{\leq d})$ such that for all $s < t$ in this order, $\phi(a_s; b_t) \leq r$ and $\phi(a_t; b_s) \geq s$.

We claim we can shatter the tree T defined by sending each sequence $E \in \{-1, 1\}^{< d}$ to a_E . For every $E \in \{-1, 1\}^d$, let $E_{< t} = (E(0), \dots, E(t-1))$. We consider b_E , and see that for all $0 \leq t < d$, if $E(t) = -1$, then $E < E_{< t}$, so $\phi(a_{E_{< t}}; b_E) \leq r$, while if $E(t) = 1$, we have $E > E_{< t}$ and $\phi(a_{E_{< t}}; b_E) \geq s$. As $s - r \geq \gamma$, this tree is γ -shattered. Thus we have finished the proof of one direction of Theorem 9.

To prove the other direction, we fix $0 < \delta < \frac{\gamma}{2}$ and assume that for every d , \mathcal{H}_ϕ γ -shatters a binary tree T in X of depth d , and prove that $\phi(x; y)$ is not δ -stable.

To do this, we show by induction on d that if $k > (\gamma - 2\delta)^{-1}$ and \mathcal{H} is a class on X that γ -shatters a binary tree T of depth $\frac{k^{d+1}-1}{k-1}$, then there are $x_1, \dots, x_d \in X$ and $h_1, \dots, h_d \in \mathcal{H}$ such that for all $i < j$, $|h_i(x_j) - h_j(x_i)| \geq \delta$.

The base case is $d = 1$. We just need that X and \mathcal{H} are nonempty, which is satisfied by the existence of any shattered tree.

For the induction step, we will need a Ramsey-theoretic fact about partitions on the nodes of a binary tree, and to state it, we need to define a *subtree*:

Definition 15 (Subtree). Let $T : \{-1, 1\}^{<d} \rightarrow X$ be a binary tree in X of depth d . If x, y are both in the image of T , say that y is a left descendant of x when $x = T(E)$ and $y = T(E')$, with E' extending $(E_0, \dots, E_t, -1)$. If instead E' extends $(E_0, \dots, E_t, 1)$, we call y a right descendant of x .

Then a subtree of T of depth $d' \leq d$ is a map $T' : \{-1, 1\}^{<d'} \rightarrow X$ such that if y is a left/right descendant of x in the image of T' , it is also a left/right descendant of x in the image of T .

The Ramsey-theoretic fact is:

Fact 26 (Alon et al. (2019, Lemma 16)). If p, q are positive integers, $T : \{-1, 1\}^{<p+q-1} \rightarrow X$ is a binary tree in X of depth $p + q - 1$, and the elements of X are partitioned into two sets, blue and red, then either there is a blue subtree of T with depth p , or a red subtree of depth q .

By applying this repeatedly, we find a more useful version for our purposes:

Corollary 4 (Tree Ramsey). If d_1, \dots, d_k are positive integers, $T : \{-1, 1\}^{<d_1+\dots+d_k-k+1} \rightarrow X$ is a binary tree in X of depth $d_1 + \dots + d_k - k + 1$, and the elements of X are partitioned into k sets X_1, \dots, X_k , then there is some i such that the set X_i contains a subtree of T of depth d_i .

Returning to the inductive step, assume that d is such that we have the inductive invariant for d : if \mathcal{H} is a class on X that γ -shatters a binary tree of depth $\frac{k^{d+1}-1}{k-1}$, then there are $x_1, \dots, x_d \in X$ and $h_1, \dots, h_d \in \mathcal{H}$ such that for all $i < j$, $|h_i(x_j) - h_j(x_i)| \geq \delta$.

Now assume the hypothesis of the invariant for $d + 1$: \mathcal{H} is a class on X that γ -shatters a binary tree T of depth $\frac{k^{d+2}-1}{k-1}$. In the text below, by the width of a real interval with endpoints $a < b$, we mean $b - a$. We pick some $h \in \mathcal{H}$, partition $[0, 1]$ into k intervals I_1, \dots, I_k each of width at most $\gamma - 2\delta$, and then partition X into sets X_1, \dots, X_k where if $x \in X_i$, then $h(x) \in I_i$. Then by Corollary 4, as $k \left(\frac{k^{d+1}-1}{k-1} + 1 \right) - k + 1 = \frac{k^{d+2}-1}{k-1}$, the depth of T , some X_a contains the set of values decorating a subtree T' of T of depth $\frac{k^{d+1}-1}{k-1} + 1$. Let x' be the root of T' . By the shattering hypothesis, there are r and s with $r + \gamma \leq s$ such that the tree of left descendants of x' in T' is γ -shattered by the set of h' with $h'(x') \leq r$, and the tree of right descendants is γ -shattered by all h' with $h'(x') \geq s$. Because I_a has width at most $\gamma - 2\delta$, either the interval $[0, r]$ or the interval $[s, 1]$ has distance to I_a at least δ . Assume without loss of generality that it is $[0, r]$. The tree of left descendants of x' in T' has depth $\frac{k^{d+1}-1}{k-1}$, and is γ -shattered by the set of h' with $h'(x') \leq r$. Thus by the inductive hypothesis, there are left descendants x_1, \dots, x_d of x' in T' and $h_1, \dots, h_d \in \mathcal{H}$ with $h_i(x') \leq r$ for each i such that for each $i < j \leq d$, $|h_i(x_j) - h_j(x_i)| \geq \delta$. We now let $x_{d+1} = x'$ and $h_{d+1} = h$, and observe that for $i \leq d$, $h_i(x') \leq r$ while $h(x_i) \in I_a$, so $|h_i(x') - h(x_i)| \geq \delta$.

Thus we have completed the proof of the other direction of Theorem 9. \square

6.3 Realizable Online Learning for statistical classes

We now turn to preservation of realizable online learning for statistical classes.

We start by reviewing the relationship between realizable and agnostic online learning. Recall that for PAC learning, agnostic learnability is weaker than realizable learnability, and strictly weaker for real-valued function classes: realizable learning is a special case where the optimal hypothesis gives zero error. For realizable learning, the situation is different, since we have a stronger hypothesis on the target concept, but also a stronger requirement for our learning algorithm: a uniform bound on regret. As with PAC learnability, there is no difference in the boundary line for learnability between realizable and agnostic for concept classes. For real-valued classes, there is a difference between agnostic and realizable learning, just as in the PAC case.

In terms of dimensions, while agnostic online learning is characterized via the sequential fat-shattering dimension mentioned previously, realizable online learning has recently been characterized using *online dimension* (Attias et al., 2023).

Definition 16. [Online dimension] A hypothesis class \mathcal{H} on a set X has online dimension greater than D when there is some d , some X -valued binary tree $T : \{-1, 1\}^{<d} \rightarrow X$, a real-valued binary tree $\tau : \{-1, 1\}^{<d} \rightarrow [0, 1]$, and a \mathcal{H} -labelling of each branch in such a tree, $\Lambda : \{-1, 1\}^d \rightarrow \mathcal{H}$, such that

- for every two branches $b_0, b_1 \in \{-1, 1\}^d$, if the last node at which they agree is t , then

$$|\Lambda(b_0)(T(t)) - \Lambda(b_1)(T(t))| \geq \tau(t)$$

- for every branch $b \in \{-1, 1\}^d$, whose restrictions to previous levels are t_0, \dots, t_{d-1} , we have $\sum_{i=0}^{d-1} \tau(t_i) > D$.

Finiteness of online dimensions characterizes realizable online learnability.

Fact 27 (Attias et al. (2023, Theorem 4)). Let \mathcal{H} be a hypothesis class on a set X . Then \mathcal{H} has bounded regret for realizable online learning if and only if its online dimension is finite. If D is greater than the online dimension of \mathcal{H} , there is an algorithm for realizable online learning with regret at most D .

Online dimension has a (one way) relationship to fat-shattering of binary trees:

Lemma 8. Let \mathcal{H} be a hypothesis class on a set X , let $\gamma > 0$, and $d \in \mathbb{N}$. If \mathcal{H} γ -fat-shatters a tree of depth d , then the online dimension of \mathcal{H} is at least $\gamma \cdot d$.

Proof. Suppose T is the γ -fat-shattered tree. Then fix a binary tree s of depth d in \mathcal{R} , and label each branch $b \in \{-1, 1\}^d$ with some $h_b \in \mathcal{H}$ such that for all nodes t of the tree $\{-1, 1\}^d$, b_{-1}, b_1 are branches extending t , and b_i extends t concatenated with i for $i = \pm 1$, then

$$h_{b_{-1}}(T(t)) \leq s(t) - \frac{\gamma}{2},$$

while

$$h_{b_1}(T(t)) \geq s(t) + \frac{\gamma}{2}.$$

We now show that for any $\epsilon > 0$, the online dimension of \mathcal{H} is greater than $d(\gamma - \epsilon)$. We let $\tau : \{-1, 1\}^{<d} \rightarrow [0, 1]$ be the real-valued labelled binary tree with constant value $\gamma - \epsilon$, and let Λ label each branch b with h_b . Then for any two branches, we may without loss of generality call the branches b_{-1}, b_1 , let t be the last node at which they agree, and assume that b_i extends t concatenated with i for $i = \pm 1$. Then

$$\begin{aligned} & |h_{b_1}(T(t)) - h_{b_{-1}}(T(t))| \geq \\ & \left| \left(s(t) + \frac{\gamma}{2} \right) - \left(s(t) - \frac{\gamma}{2} \right) \right| = \gamma > \gamma - \epsilon. \end{aligned}$$

Thus T, τ, Λ satisfy the requirements to show that the online dimension of \mathcal{H} is greater than

$$\min_{b \in \{-1, 1\}^d} \sum_{t=0}^{d-1} \tau(t_i),$$

where t_i is the restriction of b to level i . As τ takes a constant value $\gamma - \epsilon$, the online dimension is greater than $d(\gamma - \epsilon)$. \square

From the lemma we infer an important consequence, saying that the containment between realizable and agnostic goes the opposite way in online learning as compared to PAC learning:

Corollary 5. If \mathcal{H} is realizable online learnable, it has some finite online dimension D , and thus for any $\gamma > 0$, \mathcal{H} has sequential γ -fat-shattering dimension at most $\frac{D}{\gamma}$.

Because this gives a finite bound for all γ , realizable online learnability implies agnostic online learnability, even for real-valued function classes.

We are now ready to show that in the case of real-valued functions, moving from a base class to statistical classes does not preserve realizable online learnability. In fact, this will follow from lack of closure under dualization:

Proposition 7. *There is a real-valued hypothesis class \mathcal{H} that is realizable online learnable, but its dual class is not realizable online learnable. Thus the dual distribution class based on \mathcal{H} is not online learnable.*

Before proving the proposition, we note a pathology for realizable online learning which will introduce the example classes that are relevant to the proof of the proposition. We show that the notion of learnability is less robust, in the sense that it is not preserved under composition with increasing homeomorphisms of $[0, 1]$.

Theorem 10. *There is a hypothesis class \mathcal{H} on a set X that has finite γ sequential fat-shattering dimension for all $\gamma > 0$, but has infinite online dimension. In fact, there are classes $\mathcal{H}, \mathcal{H}'$ on the same set X , both indexed by a set Y , such that both have finite γ sequential fat-shattering dimension for all $\gamma > 0$, \mathcal{H} has infinite online dimension, \mathcal{H}' has finite online dimension, and there is an increasing homeomorphism $f : [0, 1] \rightarrow [0, 1]$ such that $f \circ \mathcal{H} = \mathcal{H}'$, as functions $X \times Y \rightarrow [0, 1]$.*

In terms of learning, this means that both classes are agnostic online learnable, but only \mathcal{H}' is realizable online learnable. In particular, this shows that the dividing lines for these notions of learnability are different.

We proceed to the proof of the theorem.

Proof. Let X be the infinite binary tree $X = \bigcup_{t=0}^{\infty} \{0, 1\}^t$. Fix a decreasing sequence $\Gamma = \gamma_0, \gamma_1, \dots$ of positive reals to be determined, with $\lim_d \gamma_d = 0$. We define \mathcal{H}^Γ , a hypothesis class indexed by the infinite branches $\{0, 1\}^\mathbb{N}$ of that infinite binary tree.

Given an infinite branch b , and $x \in X$, the hypothesis $h_b(x) = 0$ when x is not an initial segment of b . If x is an initial segment of b , let d be its length. Then we let $h_b(x) = b_d \cdot \gamma_d$. Thus for any $x \in \{0, 1\}^d$ and branches b, b' , the difference $|h_b(x) - h_{b'}(x)|$ is either 0 or γ_d , with the latter only occurring when at least one of b, b' extends x . In particular, if $b_1, b_2, b_3 \in \{0, 1\}^\mathbb{N}$ are pairwise distinct, then there must be some pair $i \neq j$ with $i, j \in \{1, 2, 3\}$ with $|h_{b_i}(x) - h_{b_j}(x)| = 0$, as otherwise, we must have $(b_i)_d \neq (b_j)_d$ for each $i \neq j$, but $(b_1)_d, (b_2)_d, (b_3)_d \in \{0, 1\}$, so all three bits cannot be pairwise distinct.

For any $\gamma > 0$, we will calculate that \mathcal{H}^Γ has finite γ sequential fat-shattering dimension. Specifically, if \mathcal{H}^Γ γ fat-shatters a binary tree, it is clear that the nodes of this tree must all be nodes of X of length at most d , where d is the largest number such that $\gamma_d \geq \gamma$. Thus the depth of the γ fat-shattered tree must be at most d , so the γ sequential fat-shattering dimension is at most d . In particular, regardless of the rate at which γ_i goes to zero, the resulting class is agnostic online learnable.

We now characterize when the class is realizable online learnable, which will depend on the rate at which the parameters γ_i go to 0.

Note that the tree $\{0, 1\}^{<d} \subseteq X$ is γ_d fat-shattered: for each maximal branch b of the tree, we label the branch with a hypothesis corresponding to any infinite branch extending b . Thus by Lemma 8, \mathcal{H}^Γ will have online dimension at least $d \cdot \gamma_d$. If the sequence $(d \cdot \gamma_d : d \in \mathbb{N})$ is unbounded, then the online dimension is infinite, and there is no uniform bound for regret for realizable online learning.

Now we will show that if Γ is chosen such that the sum $\sum_{i=0}^{\infty} 2^i \gamma_i$ converges, then the online dimension of \mathcal{H}^Γ is at most $\sum_{i=0}^{\infty} 2^i \gamma_i$, and in particular, \mathcal{H}^Γ has finite online dimension. Assume for contradiction that the online dimension is greater than $\sum_{i=0}^{\infty} 2^i \gamma_i$. For this to be true, it must be witnessed by some d , an X -valued tree T of depth d , a tree τ of real-valued errors, and an assignment of elements from \mathcal{H} to each branch of the tree.

We claim that if $s, t \in \{-1, 1\}^{<d}$ are such that s is a strict initial substring of t and $T(s) = T(t)$, then either $\tau(s) = 0$ or $\tau(t) = 0$. Let $b_1, b_2 \in \{-1, 1\}^d$ be two branches extending t , while b_3 extends s in such a way that its last common node with b_1, b_2 is s . As noted earlier, there must be two of these three branches such that $i \neq j$ but $\Lambda(b_i)(T(s)) = \Lambda(b_j)(T(s))$. If $\Lambda(b_1)(T(s)) = \Lambda(b_2)(T(s))$, then as $T(s) = T(t)$,

$$\tau(t) \leq |\Lambda(b_1)(T(t)) - \Lambda(b_2)(T(t))| = 0.$$

Otherwise, for some $i \in \{1, 2\}$, we have $\Lambda(b_i)(T(s)) = \Lambda(b_3)(T(s))$, so

$$\tau(s) \leq |\Lambda(b_i)(T(s)) - \Lambda(b_3)(T(s))| = 0.$$

Now consider a branch of $\{-1, 1\}^{<d}$ consisting of nodes t_0, \dots, t_{d-1} . We can bound the sum of the weights of that branch by grouping the indices i by the value of $T(t_i)$:

$$\sum_{i=0}^{d-1} \tau(t_i) \leq \sum_{x \in X} \left(\sum_{0 \leq i < d-1: T(t_i)=x} \tau(t_i) \right).$$

For each $x \in X$, there is at most one i such that $T(t_i) = x$ and $\tau(t_i) > 0$, so only one i can contribute to the sum $\sum_{x \in X} \left(\sum_{0 \leq i < d-1: T(t_i)=x} \tau(t_i) \right)$. If x has length ℓ and i is such that $T(t_i) = x$, then $\tau(t_i) \leq \gamma_\ell$, so

$$\sum_{x \in X} \left(\sum_{0 \leq i < d-1: T(t_i)=x} \tau(t_i) \right) \leq \gamma_\ell.$$

As there are only 2^ℓ elements of X with length ℓ , the online dimension D of \mathcal{H} is bounded by

$$D < \sum_{i=0}^{d-1} \tau(t_i) \leq \sum_{x \in X} \left(\sum_{0 \leq i < d-1: T(t_i)=x} \tau(t_i) \right) \leq \sum_{\ell=0}^{\infty} 2^\ell \gamma_\ell.$$

By assumption, this latter sum converges, so the online dimension is finite.

We now claim that if \mathcal{H}^Γ is constructed from the sequence $\Gamma = \gamma_1, \gamma_2, \dots$ while $\mathcal{H}^{\Gamma'}$ is constructed in the same way from the sequence $\Gamma' = \gamma'_1, \gamma'_2, \dots$, then there is an increasing homeomorphism $f : [0, 1] \rightarrow [0, 1]$ such that $f \circ \mathcal{H}^\Gamma = \mathcal{H}^{\Gamma'}$. To do this, define $f(1) = 1$, and for each d , define $f(\gamma_d) = \gamma'_d$. We can then extend this to a piecewise linear definition, with countably many pieces, on $(0, 1]$, where $\lim_{x \rightarrow 0} f(x) = 0$, so defining $f(0) = 0$ will maintain continuity.

We now see that by choosing $\Gamma = \gamma_1, \gamma_2, \dots$ so that $\lim_d d \cdot \gamma_d = \infty$ and choosing $\Gamma' = \gamma'_1, \gamma'_2, \dots$ so that $\sum_{i=1}^{\infty} \gamma'_i$ converges, we find \mathcal{H}^Γ with infinite online dimension and $\mathcal{H}^{\Gamma'}$ with finite online dimension such that $f \circ \mathcal{H}^\Gamma = \mathcal{H}^{\Gamma'}$. \square

Using the same family of examples, we now prove Proposition 7:

Proof. Let $\Gamma = \gamma_i : i > 0$ be a sequence such that the $i \cdot \gamma_i$ is unbounded. Let \mathcal{H}^Γ be the class from Theorem 10. Recall that range points are prefixes p (finite sequences). Hypotheses are parameterized by infinite sequences s and the value of a hypothesis h_s on a prefix p is either zero or $\frac{1}{\gamma_n}$ for n the length of p . Then, as proven in Theorem 10 H_γ is not online learnable in the realizable case.

Let D_γ be the dual class. So points are now ω -sequences s , hypotheses are prefixes p , and the value of a hypothesis h_p at a sequence s is 0 if s does not extend p and is $\frac{1}{\gamma_n}$ if s does extend p , where again n is the length of prefix p . We claim that D_γ is realizable online learnable. Consider the definition of online learnability in terms of a game between learner and adversary. Adversary is playing range points for D_γ – that is, infinite sequences s . And at a move for learner with previous adversary range points $s_1 \dots s_k$, learner knows the values $v_1 \dots v_{k-1}$ of a D_γ -consistent hypothesis for $s_1 \dots s_{k-1}$. Note that if adversary ever reveals a value v_i that is non-zero, learner will know the hypothesis, since there is a unique prefix of s_i that would give such a value. Thus adversary should always reveal value zero. Thus learner has a strategy that will achieve bounded loss: play zero until a non-zero value appears.

To finish the proof, we note that H_γ is the dual of D_γ . \square

6.4 An Alternate Approach to Realizable Online Learning

Thus far we have shown that realizable online learning is not preserved under moving to statistical classes. We have indicated that this is related to another pathology, that online learnability is not preserved under applying continuous mappings. We will now look at using this intuition to “fix” the pathologies of realizable online learning. We will do this by changing the definition, applying alternate loss functions which do not sum the cumulative losses, but rather discretize them. The change is motivated by similar discretizations that appear in continuous logic.

Recall that a run of a learning algorithm for T rounds yields a sequence $(z_1, y_1, y'_1), \dots, (z_T, y_T, y'_T)$, where (z_i, y_i) are the moves by the adversary, and y'_i are the moves by the learner. Earlier, the loss was defined as $\sum_{i \leq T} |y'_i - y_i|$. Here we let the loss be $\sum_{i \leq T} \ell(|y'_i - y_i|)$, for certain $\ell : [0, 1] \rightarrow [0, \infty)$ nondecreasing. We then define regret in terms of this new loss function ℓ , and the remainder of the setup remains unchanged, including the restriction on the adversary in the realizable case.

Definition 17 (Online learnability for a general loss function). *Given a loss function $\ell : [0, 1] \rightarrow [0, \infty)$, say that a hypothesis class \mathcal{H} is ℓ -online learnable in the agnostic case when there is a learning algorithm whose minimax regret with loss function ℓ against any adversary is sublinear in T .*

Say that a hypothesis class \mathcal{H} is ℓ -online learnable in the realizable case when there is a learning algorithm whose minimax regret with loss function ℓ against any realizable adversary is bounded, uniform in T .

The lower the loss function, the easier it is to learn:

Lemma 9. *Let $C > 0$, and suppose $\ell_1, \ell_2 : [0, 1] \rightarrow [0, 1]$ are loss functions with $C\ell_1(x) \leq \ell_2(x)$ for all $x \in [0, 1]$.*

Then in either the agnostic or realizable case, if a hypothesis class \mathcal{H} is ℓ_2 -online learnable, then it is ℓ_1 -online learnable.

Proof. The same learning algorithm will suffice. On any given run of the algorithm, the regret with ℓ_1 as the loss function will be at most the regret with ℓ_2 as the loss function:

$$C \sum_{i \leq T} \ell_1(|y'_i - y_i|) \leq \sum_{i \leq T} \ell_2(|y'_i - y_i|),$$

so regret with loss function ℓ_1 will satisfy the same upper bound required of regret with loss function ℓ_2 , up to the fixed factor C . \square

We now define the loss functions we will focus on:

Definition 18 (ϵ -truncated loss functions). *Given $\epsilon > 0$, define $\ell_\epsilon, L_\epsilon : [0, 1] \rightarrow [0, \infty)$ by*

$$\begin{aligned} \ell_\epsilon(x) &= \max(x - \epsilon, 0) \\ L_\epsilon(x) &= \begin{cases} 0 & \text{if } x < \epsilon \\ 1 & \text{if } x \geq \epsilon. \end{cases} \end{aligned}$$

The usual loss function we just denote by ℓ_{id} (this is the identity, so $\ell_{\text{id}}(x) = x$). Using these other loss functions make it easier for a class to be online learnable:

Lemma 10. *In either the agnostic or realizable case, online learnability implies L_ϵ -online learnability which implies ℓ_ϵ -online learnability.*

Proof. Online learnability is equivalent to the standard notion ℓ_{id} -online learnability, and we see that for any $x \in [0, 1]$,

$$\begin{aligned} \epsilon L_\epsilon(x) &\leq \ell_{\text{id}}(x) \\ \ell_\epsilon(x) &\leq L_\epsilon(x), \end{aligned}$$

so the result follows by Lemma 9. \square

The reason for studying these loss functions is that when we require online learnability with respect to any ℓ_ϵ , the gap between agnostic and realizable online learnability vanishes. The proof below will apply prior characterizations of agnostic online learnability, which are given in terms of notions from model theory.

Theorem 11. *For a hypothesis class \mathcal{H} , the following are equivalent:*

- \mathcal{H} is online learnable in the agnostic case
- $\forall \epsilon > 0$, \mathcal{H} is L_ϵ -online learnable in the agnostic case
- $\forall \epsilon > 0$, \mathcal{H} is ℓ_ϵ -online learnable in the agnostic case
- $\forall \epsilon > 0$, \mathcal{H} is L_ϵ -online learnable in the realizable case
- $\forall \epsilon > 0$, \mathcal{H} is ℓ_ϵ -online learnable in the realizable case

As a corollary, we see that these properties of \mathcal{H} , being equivalent to online learnability in the agnostic case, are also preserved under moving to statistical classes:

Corollary 6. *If for every $\epsilon > 0$, \mathcal{H} is ℓ_ϵ -online learnable in the realizable case, or for every $\epsilon > 0$, \mathcal{H} is L_ϵ -online learnable in the realizable case, then the same holds of the randomization class of \mathcal{H} , and thus for both the distribution class and the dual distribution class.*

We now work towards the proof of Theorem 11.

To handle realizable online learnability with a loss function, we need to expand online dimension to include a loss function. In fact, (Attias et al., 2023) defined it for an even more broad notion of a loss function, but this version will suffice for our purposes here:

Definition 19. *[Online dimension for a general loss function] In the context of a loss function $\ell : [0, 1] \rightarrow [0, 1]$, a hypothesis class \mathcal{H} on a set X has online dimension greater than D when there is some d , some X -valued binary tree $T : \{-1, 1\}^{<d} \rightarrow X$, a real-valued binary tree $\tau : \{-1, 1\}^{<d} \rightarrow [0, 1]$, and a \mathcal{H} -labelling of each branch in such a tree, $\Lambda : \{-1, 1\}^d \rightarrow \mathcal{H}$, such that*

- for every two branches $b_0, b_1 \in \{-1, 1\}^d$, if the last node at which they agree is t , then

$$\ell(|\Lambda(b_0)(T(t)) - \Lambda(b_1)(T(t))|) \geq \tau(t)$$

- for every branch $b \in \{-1, 1\}^d$, whose restrictions to previous levels are t_0, \dots, t_{d-1} , we have $\sum_{i=0}^{d-1} \tau(t_i) > D$.

This dimension still characterizes realizable online learnability, as in the full version of (Attias et al., 2023, Theorem 4):

Fact 28 (Attias et al. (2023, Theorem 4)). *Let \mathcal{H} be a hypothesis class on a set X . Then when $\ell : [0, 1] \rightarrow [0, 1]$ is a loss function, \mathcal{H} has bounded regret for realizable online learning if and only if $\text{Onl}_\ell(\mathcal{H}) < \infty$.*

Specifically, if $\text{Onl}_\ell(\mathcal{H}) < D$, then there is an algorithm for realizable online learning with regret at most D with respect to loss function ℓ . Conversely, if $\text{Onl}_\ell(\mathcal{H}) > D$, then for any algorithm for realizable online learning, the minimax regret with respect to loss function ℓ is at least $\frac{D}{2}$.

Lemma 11. *For any non-decreasing loss function $\ell : [0, 1] \rightarrow [0, 1]$, if a hypothesis class \mathcal{H} on X γ sequentially fat-shatters a binary tree in X of depth d , then $\text{Onl}_\ell(\mathcal{H}) \geq d \cdot \ell(\gamma)$, and also the minimax regret of a d -round online learner in the agnostic case with loss function ℓ is at least $\frac{1}{3}d \cdot \ell(\gamma)$.*

Proof. We first show that $\text{Onl}_\ell(\mathcal{H}) \geq d \cdot \ell(\gamma)$. This is only a slight modification of Lemma 8.

As in that proof, suppose T is the γ -fat-shattered tree. Let the binary tree $s : \{-1, 1\}^{<d} \rightarrow \mathcal{R}$ and the branch labelling Λ which labels each branch $b \in \{-1, 1\}^d$ with h_b be as in that proof. Then as in that proof, for any two branches, we may refer to those branches without loss of generality as b_{-1}, b_1 , where t is the last node at which they agree, and assume that b_i extends t concatenated with i for $i = \pm 1$. The essential property of sequential fat-shattering is that then

$$|h_{b_1}(T(t)) - h_{b_{-1}}(T(t))| \geq \gamma,$$

so we may choose the real labelling $\tau : \{-1, 1\}^{<d} \rightarrow [0, 1]$ given by $\tau(t) = \ell(\gamma) - \epsilon$, and then T, τ, Λ satisfy the requirements to show that the online dimension of \mathcal{H} is greater than

$$\min_{b \in \{-1, 1\}^d} \sum_{t=0}^{d-1} \tau(t_i),$$

where t_i is the restriction of b to level i . As τ takes a constant value $\ell(\gamma) - \epsilon$, the online dimension is greater than $d(\ell(\gamma) - \epsilon)$. \square

The loss functions L_ϵ were chosen so that online dimension would capture sequential fat-shattering dimension:

Lemma 12. *For any hypothesis class \mathcal{H} and any $\epsilon > 0$, if $\text{Onl}_{L_\epsilon}(\mathcal{H})$ is infinite, then \mathcal{H} is not agnostic online learnable.*

Proof. Here we will use the connection of agnostic online learnability with stability, which was utilized earlier in the appendix.

Specifically we will show that for $0 < \delta < \frac{\gamma}{2}$, \mathcal{H} is not δ -stable, which suffices by Theorem 9.

Because L_ϵ is $\{0, 1\}$ -valued, if D is an integer and $\text{Onl}_{L_\epsilon} \geq D$, then there is a tree $T : \{-1, 1\}^{<d} \rightarrow X$, a $\{0, 1\}$ -valued binary tree $\tau : \{-1, 1\}^{<d} \rightarrow \{0, 1\}$, and an \mathcal{H} -labelling of each branch in the tree, $\Lambda : \{-1, 1\}^d \rightarrow \mathcal{H}$, such that

- for every two branches $b_0, b_1 \in \{-1, 1\}^d$, if the last node at which they agree is t , and $|\Lambda(b_0)(T(t)) - \Lambda(b_1)(T(t))| \geq \epsilon$, then $\tau(t) = 1$.
- for every branch $b \in \{-1, 1\}^d$ whose restrictions to previous levels are t_0, \dots, t_{d-1} , at least D of these nodes have $\tau(t_i) = 1$.

We will show that there must also be a binary tree of depth D that is ϵ fat-shattered by \mathcal{H} . To do this recursively. It will helpful below for the reader recall the definitions of descendants and subtrees from Definition 15.

We claim that for any integer D , if every branch of a finite-depth $\{0, 1\}$ -valued binary tree $t : \{-1, 1\}^{<d} \rightarrow \{0, 1\}$ has at least D nodes labelled 1, then there is a depth- D subtree of this binary tree with all nodes labelled 1. That is, there is a function $\iota : \{-1, 1\}^{<D} \rightarrow \{-1, 1\}^{<d}$, increasing in the tree order, such that $t \circ \iota(v) = 1$ for all nodes v .

We construct the subtree recursively, inducting on D . This is trivial for $D = 0$. Suppose that this is true for D , and we now consider a finite-depth $\{0, 1\}$ -valued binary tree $t : \{-1, 1\}^{<d} \rightarrow \{0, 1\}$ where every branch has at least $D + 1$ nodes labelled 1. Let v be a node of minimal length with $\tau(v) = 1$. We now consider the tree of all left descendants of v . Suppose this tree has depth d' . Then the labelling t on this subtree induces a labelling $t' : \{-1, 1\}^{<d'} \rightarrow \{0, 1\}$, given by concatenating each node $w \in \{-1, 1\}^{<d'}$ with v and -1 to form a left descendant w' of v , and then setting $t'(w) = t(w')$. Every branch $b \in \{-1, 1\}^{d'}$ of this smaller tree corresponds uniquely to a branch $b' \in \{-1, 1\}^d$ of the original tree which extends v to the left. To see this, let $v = (v_0, \dots, v_\ell)$, and let $b = (b_0, \dots, b_{d'-1})$ - we then define this new branch to be $b' = (v_0, \dots, v_\ell, -1, b_0, \dots, b_{d'-1})$. We then see that of the nodes leading up to b' , at least $D + 1$ are labelled 1 by t . Let S be the set of such nodes. Because the elements of S lie on the same branch, they are comparable. As this branch contains v , they are thus either substrings of v , or descendants of v . By the minimality assumption, the only substring of v which can be labelled 1 by t is v , so the remaining $\geq D$ elements of S are descendants of v . As these lie on the branch b' , they are left descendants. These correspond to nodes of the smaller tree $\{-1, 1\}^{<d'}$, which lie on the branch b , and are labelled 1 by t' . Thus every branch b' contains at least D elements labelled 1 by t' , so the induction hypothesis applies. There is thus a subtree of $\{-1, 1\}^{<d'}$ of depth D where all nodes are labelled 1 by t' - this is given by a map $\iota_{-1} : \{-1, 1\}^{<D} \rightarrow \{-1, 1\}^{<d'}$ such that for all nodes $w \in \{-1, 1\}^{<D}$, concatenating v with -1 and $\iota_{-1}(w)$ gives a node w' with $t(w') = 1$. The same must be true, by symmetry, of the right descendants, and these two trees, together with v , form a subtree of depth $D + 1$, with all nodes labelled 1.

We now return to our trees T, τ , and our branch labelling Λ . By our claim, there is a subtree of $\{-1, 1\}^{<D}$ of depth D where every element is labelled 1 by τ . We can thus choose an increasing (in the tree partial order) function $\iota : \{-1, 1\}^{<D} \rightarrow \{-1, 1\}^{<d}$ with $\tau \circ \iota(v) = 1$ for all v . For every branch $b \in \{-1, 1\}^D$ of $\{-1, 1\}^{<D}$, choose a branch $b' \in \{-1, 1\}^d$ extending $\iota(v)$ for all nodes v comprising b' . Label b with $\Lambda'(b) = \Lambda(b')$.

Now for any two branches b_{-1}, b_1 of $\{-1, 1\}^D$, if t is the last node at which they agree, we can assume that b_i extends t concatenated with i for $i = \pm 1$. As above, for $i = \pm 1$, let $b'_i \in \{-1, 1\}^d$ be the chosen branch corresponding to b_i in the larger tree. Because we have only chosen nodes labelled with 1, we have

$$\begin{aligned} L_\epsilon(|\Lambda'(b_{-1})(T \circ \iota(t)) - \Lambda'(b_1)(T \circ \iota(t))|) &= L_\epsilon(|\Lambda(b'_{-1})(T \circ \iota(t)) - \Lambda(b'_1)(T \circ \iota(t))|) \\ &\geq \tau \circ \iota(t) \\ &= 1, \end{aligned}$$

so in particular, $|\Lambda'(b_{-1})(T \circ \iota(t)) - \Lambda'(b_1)(T \circ \iota(t))| \geq \epsilon$.

We will connect this to stability using an argument that is a slight modification of the proof of one direction of Theorem 9. Fix $0 < \delta < \frac{\epsilon}{2}$ and $k > (\epsilon - 2\delta)^{-1}$. We wish to show that for all d , there are $x_1, \dots, x_d \in X$ and $h_1, \dots, h_d \in \mathcal{H}$ such that for all $i < j$, $|h_i(x_j) - h_j(x_i)| \geq \delta$.

First, we observe that in the first part of this proof, we have shown that for all d , there exists an X -valued binary tree T of depth $\frac{k^{d+1}-1}{k-1}$, and a labelling Λ of the branches of T with elements of \mathcal{H} such that for any branches b_{-1}, b_1 , if t is the last node at which they agree, then $|\Lambda(b_{-1})(T(t)) - \Lambda(b_1)(T(t))| \geq \epsilon$. We say that such a tree T is ϵ spread-shattered by \mathcal{H} , and that Λ witnesses spread-shattering of T .

To finish the proof, we show by induction on d that if \mathcal{H}' is a class on X that ϵ spread-shatters an X -valued binary tree T of depth $\frac{k^{d+1}-1}{k-1}$, then there are $x_1, \dots, x_d \in X$ and $h_1, \dots, h_d \in \mathcal{H}'$ such that for all $i < j$, $|h_i(x_j) - h_j(x_i)| \geq \delta$. As this holds for \mathcal{H} and any d , the result follows.

For a base case, let $d = 1$. We simply need that X and \mathcal{H}' are nonempty, which is satisfied by the existence of any shattered tree.

Now assume that d is such that we have the inductive invariant for d : if \mathcal{H}' is a class on X that ϵ -shatters a binary tree of depth $\frac{k^{d+1}-1}{k-1}$, then there are $x_1, \dots, x_d \in X$ and $h_1, \dots, h_d \in \mathcal{H}'$ such that for all $i < j$, $|h_i(x_j) - h_j(x_i)| \geq \delta$.

Also assume the hypothesis of the invariant for $d+1$: \mathcal{H}' is a class on X that ϵ -shatters a binary tree T of depth $\frac{k^{d+2}-1}{k-1}$. As before, by the width of a real interval with endpoints $a < b$, we mean $b - a$. We pick some $h \in \mathcal{H}'$, partition $[0, 1]$ into k intervals I_1, \dots, I_k each of width at most $\epsilon - 2\delta$, and then partition X into sets X_1, \dots, X_k where if $x \in X_i$, then $h(x) \in I_i$. Then by our Ramsey result for trees, Corollary 4, as $k \left(\frac{k^{d+1}-1}{k-1} + 1 \right) - k + 1 = \frac{k^{d+2}-1}{k-1}$, the depth of T , some X_a contains the set of values decorating a subtree T' of T of depth $\frac{k^{d+1}-1}{k-1} + 1$. Let x' be the root of T' .

Let $\mathcal{H}_L \subseteq \mathcal{H}'$ consist of all hypotheses $\Lambda(b)$ where b is a branch of T extending x' to the left, and let $\mathcal{H}_R \subseteq \mathcal{H}'$ consist of all hypotheses $\Lambda(b)$ where b is a branch of T extending x' to the right. By the spread-shattering hypothesis, if $h_L \in \mathcal{H}_L$ and $h_R \in \mathcal{H}_R$, then $|h_L(x') - h_R(x')| \geq \epsilon$. Because I_a has width at most $\epsilon - 2\delta$, either the set $\{h_L(x') : h_L \in \mathcal{H}_L\}$ or the set $\{h_L(x') : h_L \in \mathcal{H}_L\}$ has distance to I_a at least δ . Assume without loss of generality that it is $\{h_L(x') : h_L \in \mathcal{H}_L\}$. The tree of left descendants of x' in T' has depth $\frac{k^{d+1}-1}{k-1}$, and is ϵ -shattered by \mathcal{H}_L . Thus by the inductive hypothesis, there are left descendants x_1, \dots, x_d of x' in T' and $h_1, \dots, h_d \in \mathcal{H}_L$ for each i such that for each $i < j \leq d$, $|h_i(x_j) - h_j(x_i)| \geq \delta$. We now let $x_{d+1} = x'$ and $h_{d+1} = h$, and observe that for $i \leq d$, $h_i \in \mathcal{H}$ while $h(x_i) \in I_a$, so $|h_i(x') - h(x_i)| \geq \delta$. \square

We are now ready to prove Theorem 11:

Proof. Many of these implications follow from Lemma 10. To show all of the agnostic case conditions are equivalent, it suffices to show that if for every $\epsilon > 0$, \mathcal{H} is ℓ_ϵ -online learnable, \mathcal{H} is online learnable.

We show this through the contrapositive. If \mathcal{H} is not online learnable in the agnostic case, then there is some $\epsilon > 0$ such that \mathcal{H} sequentially 2ϵ fat-shatters a tree of depth d for every d . By Lemma 11, the regret of agnostic ℓ_ϵ -online learning in d rounds is at least $d\ell_\epsilon(2\epsilon) = d\epsilon$, so this is not sublinear.

To show that the realizable case conditions are equivalent to agnostic online learnability, we first observe that by Lemma 12, if \mathcal{H} is online learnable in the agnostic case, then for any $\epsilon > 0$, \mathcal{H} is L_ϵ -online learnable in the realizable case.

Using Lemma 10 once more, it suffices to show that if for every $\epsilon > 0$, \mathcal{H} is ℓ_ϵ -online learnable in the realizable case, \mathcal{H} is online learnable in the agnostic case. The contrapositive of this follows from the other part of Lemma 11. \square

7 Complexity of prediction for statistical classes

Thus far we considered the sample complexity needed to learn. To learn a real-valued function we need to perform empirical risk minimization: find the $h \in \mathcal{H}$ that minimizes the empirical risk, or is within a given ϵ of the minimum.

In our case, the function h we are learning usually has an infinite domain: it maps parameters to the corresponding measures. Thus it is not clear how to represent it. Further, there can be many h 's that minimize empirical risk on a sample set, and they may provide different predictions on other inputs. At test time, what we actually need is not h itself, but the predictions generated by h . Thus, rather than pick some representation of an empirical risk minimizer and study how to generate it, in this section we will look at decision problems associated with the possible predictions of an empirical risk minimizer. We formalize this below.

Fix a real-valued function class \mathcal{H} over a range space whose elements can be effectively represented – like the integers or rationals. Given a finite collection of pairs $(x_1, y_1) \dots (x_n, y_n)$, a hypothesis in \mathcal{H} is ϵ -optimal if its error, using the standard mean squared loss, is within ϵ of the infimum of the error for any hypothesis in the class. It is ϵ -fitting if its error is at most ϵ . We can also drop ϵ and talk about optimal or fitting hypotheses for a training set, noting that such a hypothesis might not exist. We can consider *prediction over ϵ -optimal hypotheses*:

Given a sequence of input-output pairs $x_1, y_1 \dots x_n, y_n$, where y_i are rational, along with a new input x , a rational q , and rational $\epsilon \geq 0$, determine whether for each $h \in \mathcal{H}$ that is ϵ -optimal, $h(x) \geq q$.

We can likewise consider ϵ -fitting, optimal, fitting. Notice that these definitions do not require us to represent the functions: instead we quantify over their predictions. We will be interested in the complexity of this problem for a *fixed* \mathcal{H} , ϵ , and q : thus the input is only $x_1, y_1 \dots x_n, y_n$.

When \mathcal{H} is the randomization class of a concept class or function class, it is not obvious how to represent elements of the range space – the inputs x_i – since they are random variables. We focus on the case of the dual distribution function class formed over a hypothesis class \mathcal{H}_0 . Recall that the parameter space of this class consists of measures on the range space of \mathcal{H}_0 , and each such distribution μ specifies a function that mapping $h_p \in \mathcal{H}_0$ (C_p for a concept class) to its μ mean (resp. its μ probability). For a concept class, our training set consists of elements of the parameter space and the corresponding μ -probability, while in the real-valued case we have the corresponding μ -mean.

We look first at the case where \mathcal{H}_0 comes from a definable family over a first-order structure. This will allow us to relate the *decidability of the prediction problem* to *decidability of satisfaction for logics over the structure*. We always assume that the probabilities in our training set, as well as the tolerance parameter ϵ , are presented as rationals, coded in binary. We likewise restrict training inputs – which are elements of the model – to be given by constants of the model: in the case of models over the reals, they would be rational.

It is easy to show that for decidable structures – those where we can decide whether a first-order sentence is true – any of these prediction problems are decidable, not just for the definable families themselves, but for the distribution class families. For some common decidable structures, we can even get reasonable bounds in terms of the size of the training set.

Theorem 12. *If \mathfrak{M} is a decidable first-order structure, then the optimal prediction problems for the dual distribution or distribution class are decidable as well. Let \mathfrak{M} be any of Presburger arithmetic, the real closed field, the real closed order group. The ϵ -fitting prediction problem can be solved in NP for the distribution class (or the dual distribution class) formed over any fixed \mathcal{C}_ϕ , and in PTIME when fixing $\epsilon = 0$.*

Note that since we are fixing the partitioned formula $\phi(\vec{x}; \vec{y})$, we are fixing the number of parameters that need to be learned. This contrasts with most prior work on fitting problems (Abrahamsen et al., 2021; Bertschinger et al., 2023; Goel et al., 2021), which provide hardness results in settings where the number of weights is part of the input.

Before turning to statistical classes, we show that the optimal prediction problem is decidable as long as the underlying structure is decidable:

Proposition 8. *If \mathcal{H}_0 is a concept class given by partitioned formula $\phi(\vec{x}; \vec{y})$ in a decidable model, then all the prediction problems (fitting, optimal, ϵ -optimal) are decidable for \mathcal{H}_0 .*

Here we assume decidability of arbitrary sentences for simplicity, but it will be clear from the proof that we only need to decide certain formulas related to ϕ .

Proof. We assume the realizable case – the optimal hypothesis gives error zero – for brevity.

If the defining formula is $\phi(\vec{x}, \vec{y})$, our sequence of inputs will be $\vec{x}_1 \dots \vec{x}_n$, and our outputs y_i will be 0 or 1. It suffices to show that the problem is decidable when the loss is 0 – since we can enumerate all subsets S of $1 \dots n$ and look for hypotheses that are correct on this subset. For simplicity assume that our sample data asserts that $h(\vec{x}_i)$ is true on each i . Since the output of h on the new element x_{n+1} is binary, our loss on x_{n+1} is discrete, and we can consider the two possible outputs separately. To determine if the outcome true holds on input x_{n+1} , we need to decide if:

$$\forall \vec{y} \left[\bigwedge_{i \in S} \phi(\vec{x}_i, \vec{y}) \rightarrow \phi(\vec{x}_{n+1}, \vec{y}) \right]$$

We can do this appealing to the decidability of the structure. □

Of course, we are not interested in prediction problems about the original hypothesis class \mathcal{H}_ϕ . Instead we are interested in prediction problems for statistical classes based on \mathcal{H}_ϕ . We next turn to the dual distribution class, and show decidability of the optimal prediction problem:

Proposition 9. *If \mathcal{H}_ϕ is the dual distribution class built over a concept class given by a partitioned first-order formula $\phi(\vec{x}; \vec{y})$ in a decidable model, then the optimal prediction problems are decidable.*

Here we deal with the dual distribution class – where the measures are defined over the range space \vec{x} . But the same comment holds for the distribution class, just swapping the roles of \vec{x} and \vec{y} .

Proof. For simplicity, we only give the argument for the ϵ -fitting version. In this case our decision problem reduces to:

Given $\epsilon, \vec{y}_i, r_i : i \leq n, \vec{y}_{n+1}$ and q , is there a measure μ such that $\sum_{i \leq n} |\mu(\vec{x} | \phi(\vec{x}, \vec{y}_i)) - r_i| < \epsilon$ and $\mu(\vec{x} | \phi(\vec{x}, \vec{y}_{n+1})) > q$.

All that is relevant from μ is its values on the atomic measure algebra generated by $\phi(\vec{x}, \vec{y}_i)$ for $i \leq n+1$. Let A be the generators of this algebra, which have size at most 2^{n+1} . Our algorithm will have two stages, the *decision procedure stage*, and the *linear arithmetic stage*. In the decision procedure stage, we make 2^{n+1} calls to the decision procedure for the theory to determine the subset A_\emptyset of Boolean combinations that are empty. Let A_i be the elements of the algebra that include a conjunct $\phi(\vec{x}, \vec{y}_i)$. We are thus asking whether there are numbers $r_a \in [0, 1]$ for $a \in A$ such that:

- $\sum_{a \in A \setminus A_\emptyset} r_a = 1$

- For $i \leq n$, $\sum_{a \in A_i \setminus A_\emptyset} |r_a - r_i| < \epsilon$
- $\sum_{a \in A_{n+1} \setminus A_\emptyset} r_a > q$.

In the *linear arithmetic stage* we can solve this by appealing to the decidability of the reals with addition and order (Weispfenning, 1988). In the above case, it suffices to invoke decidability of solving systems of linear inequations. \square

Above we did not make any assumption on the base hypothesis class in terms of learnability/VC dimension/fat-shattering. Note that if the model is NIP, then the size of $A - A_\emptyset$ will be polynomial in n , say $O(n^k)$. We can compute $A - A_\emptyset$ with $O(n^k)$ many calls to the decision procedure. We iteratively find elements of $A - A_\emptyset$ represented by Boolean combinations of at most i elements. At round $i + 1$ we take all the combinations of size i and make calls to the decision procedure to see if the intersection with $\phi(\vec{x}, \vec{y}_{i+1})$ and its complement are both non-empty, adding any non-empty element to the working set. Our working set never grows past $O(n^k)$, and thus at each of the n rounds we make at most $O(n^k)$ calls.

Thus the decision procedure stage consists of polynomial many calls to a decision procedure for the model. However, each of these calls concerns a formula of the form $\Gamma(U_1, U_2) = \exists \vec{x} \bigwedge_{\vec{a}_i \in U_1} \phi(\vec{x}, \vec{a}_i) \wedge \bigwedge_{\vec{a}_i \in U_2} \neg \phi(\vec{x}, \vec{a}_i)$, where U_1, U_2 will be finite sets whose size will not grow in the process.

The comments above apply to the dual distribution class or the distribution class equally: since one has finite VC-dimension iff the other does, and since ϕ is fixed in our complexity analysis any blow up in moving between a class and its dual is irrelevant.

To complete the proof of the complexity assertions in Theorem 12, we show that for common decidable models like those listed in the theorem, the linear programming phase can be handled in polynomial time in the sizes of U_1 and U_2 , for fixed ϕ . This follows easily from:

Theorem 13. *Let M be any of Presburger arithmetic, the real closed field, the real closed order group. For any fixed Γ , there is an algorithm that decides statements of the form $\Gamma(U_1, U_2)$ in polynomial time in the size of U_1, U_2 .*

Proof. We rely on two properties of the model. One property is that it has quantifier elimination in a language where atomic formulas can be implement in polynomial time: this property is well-known for the models in the theorem. The second property is *restricted quantifier collapse*, which is also known for these models: we explain this property next.

Let U be a finite relational vocabulary – for the application, $U = \{U_1, U_2\}$ suffices. A *restricted quantifier sentence* is a sentence in the language of $L(M) \cup U$ built up from atomic formulas via Boolean operators and quantifications over predicates in U . In this case $\exists x \in U_i \phi$. It is known (see e.g. (Benedikt, 2006)) that in each of the models above, we can convert an arbitrary $L(M) \cup U$ sentence into a restricted-quantifier sentence. Clearly each such sentence can be evaluated in time polynomial in the sizes of predicates in U . \square

Thus the only source of intractability is in the linear programming part. In the case $\epsilon = 0$ we are simply determining feasibility of a set of linear equations. Thus using standard algorithms for feasibility of linear systems (Schrijver, 1986) (e.g. ellipsoid), we conclude that the problem is in polynomial time. In the general case it is in NP.

Extensions to real-valued classes. We now discuss optimal prediction problems for distribution classes where the base class consists of real-valued functions.

Proposition 10. *Let \mathcal{H}_ϕ be a real-valued class given by a bounded family of real-valued functions definable by partitioned formula $\phi(\vec{x}, y; \vec{p})$ over the real ordered field.*

Then the ϵ -optimal prediction problem for the distribution class or dual distribution class is decidable.

Proof. We deal with the dual distribution class for simplicity. Consider parameters $\vec{p}_1 \dots \vec{p}_n$, let $\vec{F} = F_1 \dots F_n$ be the corresponding functions where F_i is formed by fixing \vec{p}_i in ϕ . For a probability distribution μ such

that the \vec{F} are measurable, let $\mathbf{Mean}_\mu(\vec{F})$ be the vector of μ -means of \vec{F} : a vector in \mathcal{R}^n . Let $\mathbf{Means}(\vec{F})$ be the collection of such vectors.

We claim that $\mathbf{Means}(\vec{F})$ is definable over the real ordered field. For a fixed j , let $\mathbf{Means}_j(\vec{F})$ be all the vectors that arise from measures that are convex combinations of j point masses. Clearly this set is definable, since these are restricted j -sums of combinations of the ranges of the F_i . However, $\mathbf{Means}(\vec{F})$ is the same as $\mathbf{Means}_n(\vec{F})$: for discretely-supported distributions, this follows from Caratheodory's theorem (Cook & Webster, 1972). While for general distributions we will show below that it can be reduced to the finitely-supported case. Note that this argument shows that in the concept class case, without assuming NIP, we can always restrict to probability distributions that give non-zero mass to only polynomially many Boolean combinations of the concepts in the training set. However, in the NIP case, we can compute a superset of these combinations without considering the training output values.

All of our prediction problems reduce to statements about $\mathbf{Means}(\vec{F})$, and since M is decidable, we can decide these.

We now fill in the details about why $\mathbf{Means}(\vec{F})$ is the same as $\mathbf{Means}_n(\vec{F})$ for general distributions. First consider measures μ is an arbitrary measure on some arbitrary domain X . Let $F_1 \dots F_k$ be arbitrary bounded real-valued functions on X .

First, note that by the law of large numbers, for every number n , there is a finitely supported probability measure μ^n such that for each $i \leq k$, the μ^n -mean of each F_i is within $\frac{1}{n}$ of the μ mean of F_i . By Caratheodory's theorem, we can take each μ^n to have support of size k , thus given by k -tuple of domain elements \vec{x}^n and a k -tuple of reals \vec{r}^n in $(0, 1]$, with \vec{r}^n summing to one. By moving to a subsequence, we can assume \vec{r}^n converges in the product topology over the reals. Let r_j be the limit of the j^{th} component.

Now suppose that the domain X of our functions, and of our measure μ is a product of the reals. For $j \leq k$ let \vec{x}_j^n denote the j^{th} component of \vec{x}^n .

Let D be the set of j such that \vec{x}_j^n does not eventually stay in some compact subset of the reals: thus D is the set of “divergent indices”. Since μ is a probability distribution, the mass of μ outside of compact set must go to zero as the compacts expand towards infinity. From this we see that for $j \in D$, the weight r_j^n corresponding to \vec{x}_j^n must go to zero. Thus $r_j = 0$ for $j \in D$, which means that the sum of r_j for j outside of D must be 1.

For j outside of D , \vec{x}_j^n is eventually contained in a compact set, so by moving to a subsequence we can assume it converges to some real vector \vec{x}_j .

Up until this point, we have not used any additional properties of the F_i . Let us now assume that F_i are continuous. Then for each j outside of D $F_i(\vec{x}_j^n)$ converges to $F_i(\vec{x}_j)$.

It is now clear that, again assuming continuity of each F_i , if we take the measure μ^∞ that assigns \vec{x}_j to r_j for $j \leq k$ outside of D , we have the finitely-supported probability measure we need.

We note that this argument for continuous F_i does not require the measure to be a probability distribution, but applies also to a subprobability distribution.

Finally, we will make use of the fact that our functions F_i are not arbitrary function. Recall that in the theorem we have a formula $\phi(\vec{x}, y; \vec{p})$ first-order over the real field, such that for each \vec{p}_i , $\phi(\vec{x}, y, \vec{p}_i)$ defines a function, which we can denote by $F_{\vec{p}_i}^\phi$. Our functions F_i are all $F_{\vec{p}_i}^\phi$ for some \vec{p}_i .

We utilize the following fact about definable functions, which holds for every \mathcal{o} -minimal expansion of the real field:

Fact 29. *van den Dries (1998) For every $\phi(x_1 \dots x_d, y; \vec{p})$ there is a number j such that for each \vec{p}_0 , there is a partition of \mathcal{R}^d into definable sets $S_1 \dots S_j$ such that $F_{\vec{p}_0}^\phi$ is continuous on each S_i .*

By taking a common refinement, we can assume a single partition for each of our functions $F_{\vec{p}_i}^\phi$. Thus, by the argument for continuous functions above, we can get finitely-supported measures that correctly produce

the mean of each function on each partition. By taking the weighted sum of these measures, weighting the measure for S_i by the μ measure of S_i we can get a single finitely-supported measure.

Note that the process above gives a measure with support larger than the number of functions. But applying Caratheodory’s theorem again, we can reduce the support to equal the number of functions. \square

8 Related Work

As noted earlier, what we refer to as the “dual distribution class” is from (Hu et al., 2022), while the distribution class is new to our work.

The notions of NIP formulas and structures have been developed over many decades, beginning with work of Shelah, see for example (Shelah, 1990). Many common structures, such as the real field and real exponential field were shown to be NIP (van den Dries, 1998). A fundamental result in (M.C.Laskowski, 1992) is that NIP first order theories are precisely those where definable families are PAC learnable: we use this as a definition of NIP in this work.

The notion of randomization of a structure was introduced in (Keisler, 1999). It was later reformulated in terms of continuous logic by Keisler and Ben Yaacov, see (Ben Yaacov, 2009). (Keisler, 1999; Ben Yaacov & Keisler, 2009) present an extensive study of the randomization transformation, presenting in particular axioms for the randomization. In this paper we deal with a simpler context, and we do not reason across models. Thus our presentation of randomization is simplified.

NIP and other model-theoretic properties were shown to be preserved in moving from a class to its randomization in (Ben Yaacov, 2009; 2013): our bounds refine the analysis given in these works. Characterizations of which partitioned first-order formulas lead to the corresponding concept class being online learnable can be found in (Chase & Freitag, 2019). Model-theoretic characterizations of which formulas are learnable in other learning models (e.g. Private PAC learning) are provided in (Alon et al., 2019).

In the last part of the paper, we deal with prediction and fitting problems for statistical classes: finding parameters that fit training data, or quantifying over these. Such problems have been extensively studied for standard neural architectures (Abrahamsen et al., 2021; Bertschinger et al., 2023; Goel et al., 2021), and also in the presence of SoftMax and other exponential activation functions (Hankala et al., 2023). Our results deal with a different set of function classes, and in the problem we will deal with a fixed hypothesis class, thus fixing the number of parameters that can be set, in contrast to (Abrahamsen et al., 2021; Bertschinger et al., 2023; Goel et al., 2021; Hankala et al., 2023).

9 Conclusions

We investigated a mapping that takes a “base” hypothesis class, consisting of either Boolean or real-valued functions, to other classes based on probability distributions over either the range space or the parameter space of the class. We connected this to the theory of randomizations in model theory: there we map a classical or continuous-valued structure to another continuous-valued structure based on random variables. We have proved that these transformations preserve agnostic PAC learnability and agnostic online learnability, refining results from both learning of database queries (Hu et al., 2022) and the model theory (Ben Yaacov & Keisler, 2009). In addition to providing a linkage between these communities, our results provide improved bounds. For realizable learning, we have provided counterexamples to preservation. Finally, we provide an initial exploration of the computational complexity of decision problems related to learning distribution classes.

Our motivation concerns distribution classes, but we obtain our positive results by embedding into a more general class, the randomization of a class or structure. This class is strictly more general than the distribution and dual distribution class, since it allows correlation between range elements and parameters. We are currently exploring how to exploit this generality.

We leave open the question of whether realizable online learnability of a base class implies realizable online learnability of the distribution class. For the dual distribution class, we showed failure of preservation in Proposition 7.

Our results on fitting/prediction problems are far from comprehensive, showing only that logical techniques, and more generally model-theoretic properties of the base structure, can be relevant to these problems.

References

- Mikkel Abrahamsen, Linda Kleist, and Tillmann Miltzow. Training neural networks is \exists -complete. In *NeurIPS*, 2021.
- 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.
- Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *STOC*, 2019.
- Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *STOC*, 2021.
- Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.
- Idan Attias and Aryeh Kontorovich. Fat-shattering dimension of k -fold aggregations. *Journal of Machine Learning Research*, 25:1–29, 2024.
- Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velezgas. Optimal learners for realizable regression: Pac learning and online learning. In *NeurIPS*, 2023.
- Peter L. Bartlett and Philip M. Long. More theorems about scale-sensitive dimensions and learning. In *COLT*, 1995.
- Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.
- Itai Ben Yaacov. Continuous and random Vapnik-Chervonenkis classes. *Israel Journal of Mathematics*, 173(1):309–333, 2009.
- Itai Ben Yaacov. On theories of random variables. *Israel Journal of Mathematics*, 194:957–1012, 2013.
- Itai Ben Yaacov and H. Jerome Keisler. Randomizations of models as metric structures. *Confluentes Mathematici*, 1(2):197–223, 2009.
- Itai Ben Yaacov and Alexander Usvyatsov. Continuous first order logic and local stability. *Transactions of the American Mathematical Society*, 362(10):5213–5259, 2010.
- Itai Ben Yaacov, Alexander Berenstein, C. W. Henson, and Alex Usvyatsov. *Model theory for metric structures*, pp. 315–427. 2008.
- Michael Benedikt. *Generalizing Finite Model Theory*, volume 24. Cambridge University Press, 2006.
- Daniel Bertschinger, Christoph Hertrich, Paul Jungeblut, Tillmann Miltzow, and Simon Weber. Training fully connected neural networks is \exists -complete. In *NeurIPS*, 2023.
- Hunter Chase and James Freitag. Model theory and machine learning. *The Bulletin of Symbolic Logic*, 25(3):319–332, 2019.
- Dmitry Chistikov, Christoph Haase, and Alessio Mansutti. Geometric decision procedures and the VC dimension of linear arithmetic theories. In *LICS*, 2022.
- W. D. Cook and R. J. Webster. Carathéodory’s theorem. *Canadian Math. Bull.*, 15:293, 1972.

-
- D.H. Fremlin. Measure theory, 2002. URL <http://www.essex.ac.uk/maths/staff/fremlin/int.htm>.
- Surbhi Goel, Adam R. Klivans, Pasin Manurangsi, and Daniel Reichman. Tight hardness results for training depth-2 relu networks. In *ITCS*, 2021.
- Teemu Hankala, Miika Hannula, Juha Kontinen, and Jonni Virtama. Complexity of neural network training and ETR: extensions with effectively continuous functions. In *AAAI*, 2023.
- Bradd Hart. Introduction to continuous model theory, 2023. <https://arxiv.org/abs/2303.03969v1>.
- Wilfrid Hodges. *Model Theory*. Cambridge university press, 1993.
- Xiao Hu, Yuxi Liu, Haibo Xiu, Pankaj K. Agarwal, Debmalaya Panigrahi, Sudeepa Roy, and Jun Yang. Selectivity functions of range queries are learnable. In *SIGMOD*, 2022.
- Marek Karpinski and Angus Macintyre. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *JCSS*, 54(1):169–176, 1997.
- Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *JCSS*, 48(3):464–497, 1994.
- H.Jerome Keisler. Randomizing a model. *Advances in Mathematics*, 143(1):124–158, 1999.
- Pieter Klerer and Hans Simon. Primal and dual combinatorial dimensions. *Discrete Applied Mathematics*, 327:185–196, 2023.
- Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.
- M.C.Laskowski. Vapnik - Chervonenkis classes of definable sets. *J. London Math. Soc.*, 45:377–384, 1992.
- Daniel Raban. The Glivenko-Cantelli Theorem and Introduction to VC Dimension, 2023. <https://pillowmath.github.io>.
- Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015a.
- Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015b.
- Alexander Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1986.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.
- Saharon Shelah. *Classification theory and the number of non-isomorphic models*. Elsevier, 1990.
- Pierre Simon. *A Guide to NIP Theories*. Cambridge University Press, 2015.
- L. P. D. van den Dries. *Tame Topology and O-minimal Structures*. Cambridge University Press, 1998.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Volker Weispfenning. The complexity of linear problems in fields. *Journal of Symbolic Computation*, 5(1):3–27, 1988.

A Agnostic online learnability and duality: Proof of Proposition 1

Recall Proposition 1:

A function class is agnostic online learnable exactly when its dual class is.

Although this is surely well known, we include a proof for completeness. This can be done directly using the sequential fat-shattering dimension, Fact 6. However, we can also use the characterization in terms of stability of a CL formula in Theorem 9: although this is stated in the context of a logical formula, any function class can be considered a formula. The theorem states that a function class is online learnable if and only if:

There is some d such that in \mathfrak{C} there are no $a_1, \dots, a_d, b_1, \dots, b_d$ such that for all $i < j$,

$$|\phi(a_i; b_j) - \phi(a_j; b_i)| \geq \gamma.$$

Clearly, if the roles of parameters are swapped, the same thing holds.

B Continuous logic over classical first-order structures: proof of Proposition 2

Recall that we allow continuous logic to be defined over a classical first-order structure \mathfrak{M} . We now have more formulas, including infinite convergent sums and applying continuous functions. For example, if our structure has infinitely many unary predicates $U_i(x)$, we can form a new formula

$$\chi(x) = \sum_i \frac{U_i(x)}{2^i}.$$

If we have only a single binary predicate $G(x, y)$, we could similarly start by letting $\phi_i(x)$ state that there is a G path of length i originating at x , and then set

$$\chi(x) = \sum_i \frac{\phi_i(x)}{2^i}.$$

Recall Proposition 2 from the body of the paper:

If a first-order structure \mathfrak{M} is NIP, then for any general continuous logic partitioned formula ϕ , \mathcal{H}_ϕ has finite fat-shattering dimension and is thus agnostic PAC learnable. Thus $\mathfrak{C}(\mathfrak{M})$ is also NIP.

That is, if \mathfrak{M} is NIP (equivalently, all partitioned formulas are agnostic PAC learnable) as a classical first-order structure, then it is still NIP as a CL structure: even though we have more formulas, hence more hypothesis classes, the new ones are still learnable.

This is implicit in Ben Yaacov (2009), but we spell out a few more details here. Recall the distinction between basic formulas and general formulas of CL: the latter are formed by closing under convergent sum. We first note that general formulas can be uniformly approximated by basic formulas. From this it follows that to show general partitioned formulas are NIP (i.e. induce function classes with finite γ fat-shattering dimension), it suffices to show the same for basic formulas.

Call a CL formula *essentially FO* if it is of the form $\sum_{i \leq n} \chi_i(\vec{x}) \cdot r_i$ where χ_i is the characteristic function of a first-order formula. It is easy to see (and see (Ben Yaacov et al., 2008, Remark 9.21) for a proof) that basic formulas are essentially FO.

If a classical structure is NIP then each χ_i has finite VC dimension, hence trivially each one has finite γ fat-shattering dimension. It is easy to see that finite fat-shattering dimension is preserved under scalar multiples and summing.

C Lemmas on continuous logic, NIP, and definable functions: proof of Lemma 1

Our goal will be to prove Lemma 1 from the body. Recall that we deal there with a first-order structure \mathfrak{M} over the reals that has at least the ordering relation. We first note the following:

Lemma 13. *The function $f_{[0,1]} : \mathcal{R} \rightarrow [0, 1]$ given by*

$$f(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 \leq x \leq 1 \\ 1 & 1 \leq x \end{cases}$$

is a formula in $\mathfrak{C}(\mathfrak{M})$.

Proof. For each n , let ϕ_n be the indicator function of the set $\bigcup_{i=1}^{2^{n-1}} \left(\frac{2i-1}{2^n}, \frac{2i}{2^n}\right]$. Each of these is the indicator function of a definable set in \mathfrak{M} , so it is a formula in $\mathfrak{C}(\mathfrak{M})$. Also let ψ be the indicator function for the definable set $(1, \infty)$. Then we can take an infinite weighted sum, and see that $\psi + \sum_{n=1}^{\infty} \frac{\phi_n}{2^n}$ is also a formula. At any $x \in [0, 1]$, this sum will evaluate to x , as $\phi_n(x)$ will give the n^{th} bit in a binary expansion of x . This will also evaluate to 0 for $x < 0$, and to $\psi(x) = 1$ for $x > 1$, so this sum formula is $f_{[0,1]}$. \square

Once we have this formula, we are able to construct all other bounded definable functions.

Recall the statement of Lemma 1:

Let $f : \mathcal{R}^n \rightarrow [0, 1]$ be a definable function in \mathfrak{M} . Then it is also a formula in continuous logic over $\mathfrak{C}(\mathfrak{M})$.

Proof. We may apply any formula in one variable to a definable function and get another formula, and in this case, we find that $f_{[0,1]} \circ f = f$, so f is a formula. \square

D Proof of Proposition 3: inducing measures

Recall the statement of the proposition from the body of the paper:

For any range X , we can choose (Ω, Σ, μ) and well-behaved RV_X such that for each measure μ', Σ' on X there is an $F \in RV_X$ that induces μ' from μ .

Proof. We choose \mathcal{P} to be a product of all measure spaces on X . Then every measure is induced as a projection, and we take RV_X to be all such projections. \square