Journal of Machine Learning Research 27 (2026) 1-55 Submitted 5/25; Published 6/26

From learnable objects to learnable random objects

Aaron Anderson AWANDERS@SAS.UPENN.EDU
Department of Mathematics
University of Pennsylvania
Philadelphia, PA

Michael Benedikt MICHAEL.BENEDIKT@CS.OX.AC.UK
Department of Computer Science
University of Oxford
Oxford, UK

Editor: Aryeh Kontorovich

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 $h_p : p \in \Theta$ of functions implies learnability of the family of functions $h_\mu(p) = \mathbb{E}_\mu[h_p]$, where $\mathbb{E}_\mu$ 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. For agnostic learning, we establish improved bounds on the sample complexity of learning for statistical classes, stated in terms of combinatorial dimensions of the base class. We connect these problems to techniques introduced in model theory for "randomizing a structure". We also provide counterexamples for realizable learning, in both the PAC and online settings.

Keywords: fat-shattering dimension, realizable PAC learning, sequential fat-shattering dimension, threshold dimension

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 Θ. 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 using a broader result about *random hypotheses classes* inspired by work in model theory (Keisler, 1999; Ben Yaacov and Keisler, 2009; Ben Yaacov, 2009). In the PAC learning scenario, we can apply prior work in model theory (Ben Yaacov, 2009) to conclude that when a random family of hypothesis classes is uniformly learnable, then the expectation of this class is also 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.

1.1 Contributions: Preservation and Sample Complexity

Our results concern whether various notions of learnability are preserved when moving from a base hypothesis class to the corresponding statistical class. For agnostic learning, both PAC and online, 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, which highlights the positive and negative preservation results we prove in moving from a base class to a statistical class, and the combinatorial dimension that characterizes learnability. In each of the positive results, we supply sample complexity bounds, not listed in the table. The formal definitions are in the next sections.

1.2 Organization

Section 2 reviews different flavors of learnability that we study here. Section 3 defines the notion of "statistical class built over a base class" that will be our central object of study. Section 4 studies preservation of PAC learnability for statistical classes, in two variations of

Learning	Agnostic	Realizable
Online	Bounded Online FatSh (Rakhlin et al., 2015b) Preserved (Cor. 36)	Uniform Regret or Finite Online Dim. Not preserved for dual dist. class (Prop. 52)
PAC	Finite FatSh (Bartlett and Long, 1995) Preserved (Thm 16)	Finite OIG dimension (Attias et al., 2023) Not preserved for (dual) dist. class (Prop. 31)

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

the learning set up: agnostic and realizable. Section 5 performs the same study for online learning. We discuss related work in Section 6. We close with conclusions in Section 7. 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.

2.1 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* Θ. We write $\mathcal{C}_p$ for a concept that is in a concept class $\mathcal{C}$.

A family of functions on X indexed by a set Θ can be considered as a single function from $X \times \Theta$ to $[0,1]$. This corresponds naturally to a function $\Theta \times X \to [0,1]$, and thus also defines a class of functions on Θ indexed by X, the *dual class* of $\mathcal{H}$.

2.2 PAC learning

We recall the standard notion of a function class being learnable (Kearns and 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 μ over Z, we let $\mathrm{ExpLoss}_\mu(h)$ be the expected loss of h with respect to μ. We let $\mathrm{BestExpLoss}(\mu)$ be the infimum of $\mathrm{ExpLoss}_\mu(h)$ 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 μ over Z,

$$\mu^n(\vec{z} \mid \mathrm{ExpLoss}_\mu(A(\vec{z})) \leq \mathrm{BestExpLoss}_\mu + \epsilon) \geq 1 - \delta.$$

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

If a specific number $n_{\delta,\epsilon}$ suffices, we say that $\mathcal{H}$ is δ, ϵ agnostic PAC learnable with *sample complexity* $n_{\delta,\varepsilon}$. 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 and 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.*

2.3 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, $\Sigma_{i \leq T} |y'_i - y_i|$. [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

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

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 Definition 2 (Rakhlin et al., 2015b), 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.

2.4 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 and 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 $s : A \to [0,1]$ 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* $\mathrm{FatSHDim}_\gamma(\mathcal{H})$*, is the supremum of the cardinalities of a γ fat-shattered subset of X.*

These dimensions give bounds on sampling:

Fact 4 (Bartlett and Long 1995, Theorem 14) $\mathrm{FatSHDim}_\gamma(\mathcal{H})$ *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*

$$O\left(\frac{1}{\epsilon^2} \cdot \left(\mathrm{FatSHDim}_{\frac{\epsilon}{9}}(\mathcal{H}) \cdot \log^2\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).$$

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

A binary tree of depth d in X *is a function* $z : \{1,-1\}^{<d} \to X$. *Given such a binary tree and* $t < d$, *let* z_t *be the restriction of* z *to inputs in* $\{1,-1\}^t$. *If* $\mathrm{B} \in \{1,-1\}^d$ *is a branch, we write* $z_t(\mathrm{B})$ *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* $B : \{0,\ldots,d-1\} \to \{-1,1\}$, *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* $\mathrm{B} \in \{1,-1\}^d$ *with some* $h_{\mathrm{B}} \in \mathcal{H}$ *satisfying: For every* $0 \leq t < d$, *if* $\mathrm{B}(t) = -1$, *then* $h_{\mathrm{B}}(z_t(\mathrm{B})) \leq s_t(\mathrm{B}) - \frac{\gamma}{2}$ *and, if* $\mathrm{B}(t) = 1$, *then* $h_{\mathrm{B}}(z_t(\mathrm{B})) \geq s_t(\mathrm{B}) + \frac{\gamma}{2}$. *The* γ sequential fat-shattering dimension *of* $\mathcal{H}$, *denoted* $\mathrm{FatSHDim}_\gamma^{Seq}(\mathcal{H})$, *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 $T : \{-1,1\}^{<n} \to X$ *when for every branch* $\mathrm{B} \in \{-1,1\}^n$ *of the tree, there is* $c_{\mathrm{B}} \in \mathcal{C}$ *such that for all* $k < n$, $T(\mathrm{B}|_{\{-1,1\}^k})$ *is in* c_A *if and only if* $\mathrm{B}(k) = 1$. *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*

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

and above by

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

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

2.5 Duality and Agnostic Learnability

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

Fact 7 (Kleer and 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 3 *A function class is agnostic online learnable exactly when its dual class is.*

3. Hypothesis Classes from Statistical Objects

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

Definition 4 (Distribution class and Dual Distribution class) *Given a concept class $\mathcal{C}$ on X, parameterized by Θ the* distribution function class of $\mathcal{C}$, *denoted $\mathcal{D}istr_{\mathcal{C}}$, is a real-valued hypothesis class on the same range space X, consisting of the hypotheses h_μ indexed by the distributions μ, with σ-algebra Σ_μ, on Θ such that for each x, the set of $p \in \Theta$ such that $c_p(x) = 1$ is measurable.*

The hypothesis h_μ is defined as the function mapping x to the probability, with respect to μ, that $c_p(x) = 1$. Note that, WLOG, we can assume that the Σ_μ is the Σ-algebra generated by $\{\{p \in \Theta : c_p(x) = 1\} : x \in X\}$, since h_μ is completely determined by the restriction of μ to this Σ-algebra.

The dual distribution function class of $\mathcal{C}$, *denoted $\mathcal{D}ualDistr_{\mathcal{C}}$, 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 $p \in \Theta$ to 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 the distribution function class, the functions are indexed by distributions μ on Θ 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 $\mathbb{E}_\mu[h]$. As with the concept class, we can WLOG assume that Σ_μ is the σ-algebra generated by the sets $\{p : h_p(x) \in I\}$ where I is an interval, $x \in X$.

Example 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 be parameterized by "random elements of the plane": functions induced by distributions μ over the plane. Each such μ induces a function h_ν on rectangles, mapping each rectangle to its ν-probability. We can equivalently consider h_μ as a real-valued function on 4-tuples. In learning such a function h_μ, we will be given supervision in the form of a sequence of pairs, each pair consisting of a rectangle (or 4-tuple) and the μ-probability that a point is in the rectangle.

In contrast, the distribution class of $\mathcal{H}$ is parameterized by "random rectangles": functions induced by distributions ν over 4-tuples. Given such a ν, we have a function h_ν that maps a real pair $\vec{x}$ to the ν probability that $\vec{x}$ is in rectangle h_ν. 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 ν.

Example 2 *Consider the hypothesis class $\mathcal{H}$ 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 with parameters $\vec{p}$ representing the coefficients of P and Q. We can easily show (see the appendix for a broader argument) that this class of real-valued functions indexed by $\vec{p}$ is agnostic PAC learnable.*

The dual distribution class is parameterized by "random arguments" to these functions: with each distribution inducing a function on parameters given by integrating against the distribution. The distribution class will be parameterized by "random coefficients": each distribution will induce a function on arguments x obtained by integrating the random function against the distribution on parameters.

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.*

The motivation in Hu et al. (2022) was from databases: we want to learn the selectivity of a table in a databases. Given some examples of the density of the table in some intervals, we want to predict the density in a new interval. Hu et al. (2022) deals with computational complexity as well as sample complexity, but here we will only be interested in sample complexity bounds.

3.1 Expectation of a Measurable Family of Hypothesis Classes

We will show that we can embed the distribution class and the dual distribution class in a more general construction for hypothesis classes, where parameters and range elements are treated symmetrically. This construction is fundamentally connected to the *randomization* construction for hypothesis classes coming from logical formulas, originating with Keisler (1999), refined later in Ben Yaacov and Keisler (2009); Ben Yaacov (2009).

We will start with a construction that works on something much more general than a hypothesis class.

Definition 5 (Measurable family of hypothesis classes) *Assume that (Ω, Σ, μ) is a probability space, and $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ is a family of hypothesis classes $\mathcal{H}_\omega : X \times \Theta \to [0,1]$ such that for each $x \in X, p \in \Theta$, the function $\omega \mapsto \mathcal{H}_\omega(x,p)$ is measurable. We call such a family a* measurable family of hypothesis classes on X parameterized by Θ.

Thus a measurable family can be thought of as a *randomized hypothesis class*, where for each $x \in X, p \in \Theta$ we randomly choose a real-valued output.

Definition 6 (Expectation class of a measurable family of hypothesis classes) *Let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes. Then the* expectation class *of $\mathcal{F}$ is the function $\mathbb{E}\mathcal{F} : X \times \Theta \to [0,1]$ defined by*

$$\mathbb{E}\mathcal{F}(x,p) = \mathbb{E}_\mu[\mathcal{H}_\omega(x,p)].$$

This is a single hypothesis class, with range space X and parameter space Θ.

3.2 Embedding Distribution Classes in Expectation Classes

Definition 7 (X-valued random variables) *If (Ω, Σ, μ) is a probability space and (X, Σ_X) is a measurable space, then let $\mathcal{RV}_X$ be the set of X-valued random variables over Ω: that is, measurable functions from $\Omega \to X$.*

Definition 8 (Randomized version of a hypothesis class) *Let $\mathcal{H}$ be a hypothesis class on range space X parameterized by Θ, and let $\Omega^* = (\Omega, \Sigma, \mu)$ be a probability space. A Θ-valued random variable P is said to be* compatible with $\mathcal{H}$ *if for each $x \in X$ the function mapping $\omega \in \Omega$ to $h_{P(\omega)}(x)$ is measurable in the usual sense.*

We write $\mathrm{CompatParamRV}(\mathcal{H})$ *for the set of Θ-valued random variables that are compatible with a given $\mathcal{H}$, omitting the dependence on Ω^* for brevity.*

*For a given $P \in \mathrm{CompatParamRV}(\mathcal{H})$ we refer to the real-valued function on X above as the P-*randomized version of $\mathcal{H}$.

Thus each compatible Θ-valued random variables thus give an alternative notion of "random parameter".

Definition 9 (Parameter randomized version of a hypothesis class) *Given a hypothesis class $\mathcal{H}$ on range space X parameterized by Θ and probability space (Ω, Σ, μ), we define a measurable family, which we will call the* parameter randomized family of $\mathcal{H}$, *denoted* $\mathrm{ParamRandom}(\mathcal{H}) = ((\mathrm{ParamRandom}(\mathcal{H}))_\omega : \omega \in \Omega)$. *Given $\omega \in \Omega$, let $(\mathrm{ParamRandom}(\mathcal{H}))_\omega$ be the hypothesis class on X consisting of the functions: For each $P \in \mathrm{CompatParamRV}(\mathcal{H})$, the function mapping $x \in X$ to $h_{P(\omega)}(x)$.*

Finally, we can define the expectation class that we want:

Definition 10 (Parameter randomized expectation class) *We consider the hypothesis class denoted $\mathcal{H}(\mathrm{CompatParamRV}(\mathcal{H}))$: it is on the same range space X, but parameterized by $\mathrm{CompatParamRV}(\mathcal{H})$. Given $P \in \mathrm{CompatParamRV}(\mathcal{H})$ h_P maps $x \in X$ to $\mathbb{E}_\mu$ of the P-randomized version of $\mathcal{H}$. That is, it is the expectation class of the parameter randomized family of $\mathcal{H}$.*

The parameter randomized expectation class of a hypothesis class $\mathcal{H}$ is very similar to the distribution class of $\mathcal{H}$. The difference is that the former considers Θ-valued functions from a fixed probability space, while the latter deals with the induced distributions on Θ, ignoring the underlying sample space. We now show that by choosing the probability space carefully, we can close the gap.

Let ν be a distribution on Θ, where the underlying Σ-algebra is generated by $S_{x,I} = \{p \in \Theta | h_p(x) \in I\}$ for I an interval and $x \in X$. We say ν is *induced* by a function f from Ω to Θ (with respect to (Ω, Σ, μ)) if $\nu(S_{x,I}) = \mu(\{\omega | f(\omega) \in S_{x,I}\}$.

The following simple result in measure theory states that by choosing (Ω, Σ, μ) appropriately, all the distributions over Θ are induced by functions from (Ω, Σ, μ):

Proposition 11 *For any Θ we can find (Ω, Σ, μ) such that for every distribution ν over the Σ-algebra on Θ generated by $S_{x,I}$, there is a function f such that ν is induced by f with respect to (Ω, Σ, ν).*

Although the result is well-known (see, e.g. Fremlin (2002)), we sketch the proof:
Proof We choose $\Omega^* = (\Omega, \Sigma, \mu)$ to be a product of all measure spaces on Θ. Then every measure is induced as a projection on one component. ■

From the proposition we see:

Proposition 12 *By choosing Ω^* as in Proposition 11, we have $\mathcal{H}(\text{CompatParamRV}(\mathcal{H}))$ contains all the functions in the distribution class of $\mathcal{H}$. Thus if the former is learnable, in any of the variants we have discussed (agnostic PAC, agnostic online etc.), then so is the latter, with the same sample complexity bounds.*

We can analogously talk about X-valued random variable X' being compatible with $\mathcal{H}$, and define the class $\mathcal{H}(\text{CompatRangeRV}(\mathcal{H}))$ of hypothesis parameterized by such random variables: these subsume the dual distribution class.

We will use these embeddings below to transfer results about learnability of the expectation class of a measurable family of hypothesis classes to results about the distribution and dual distribution classes. We emphasize that the notion of measurable family is much more general than the distribution class and dual distribution class. In the distribution class, only parameters are randomized, while range elements are deterministic, while in the dual distribution class it is the reverse. The measurable family allows us to model situations where the range elements and parameters are both randomized, in a correlated way.

Example 3 *Consider a concept class $\mathcal{C}$ containing characteristic functions $h(x)_{p_1,p_2}$ where both the range elements x and the parameters p_1, p_2 are integers. The concept contains any x that an even integer between p_1 and p_2. Thus as we vary the parameters p_1, p_2 we get the set of intervals intersected with the even numbers.*

Fix a probability space $\Omega^* = (\Omega, \Sigma, \mu_0)$. The parameterized randomized family of $\mathcal{H}$ is a measurable family of hypothesis class parameterized by a "random pair of integers": a measurable function taking $\omega \in \Omega$ to a pair p_1, p_2. The measurable family contains a class for each $\omega \in \Omega$ parameterized by measurable functions $P : \Omega \to \Theta$, where for each ω and P, the function maps x to $C(x)_{P(\omega)}$. The expectation class of this measurable family is the distribution class of the concept class $\mathcal{C}$.

If we reverse the role of parameters and range values, we get a measurable family of concept classes indexed by random integers X, where given ω and X we map (p_1, p_2) to $C(X(\omega))_{p_1,p_2}$. The expectation class of this measurable family is the dual distribution class of $\mathcal{C}$

We could also consider measurable families parameterized by functions from ω to triples (x, p_1, p_2), which represent a correlated pair X, P of random range elements and random parameters.

While the notation used in Ben Yaacov (2009) is different, it is proven there that a uniform bound on fat-shattering dimensions of the hypothesis classes in a measurable family implies finite fat-shattering dimension, and thus PAC learnability, of the expectation class.

Definition 13 *For $\gamma > 0$, say that a measurable family of hypothesis classes $\mathcal{F}$ has* uniformly bounded γ fat-shattering dimension *it there is some d such that each class $\mathcal{H} \in \mathcal{F}$ has γ fat-shattering dimension at most d.*

Fact 9 (Ben Yaacov 2009, Corollaries 4.2, 4.3) *If $\mathcal{F}$ is a measurable family of hypothesis classes, and for every $\gamma > 0$, the family has uniformly bounded γ fat-shattering dimension, then $\mathbb{E}\mathcal{F}$ has finite fat-shattering dimension.*

Now let us return to the distribution class, and recall that it can be embedded in the parameter randomized family of $\mathcal{H}$, which associates to $P \in \text{CompatParamRV}(\mathcal{H})$ the function mapping $x \in X$ to $h_{P(\omega)}(x)$. If $\mathcal{H}$ has finite γ fat-shattering dimension d, then for any ω, the corresponding class of functions above also has fat-shattering dimension d: thus the measurable family has uniformly bounded γ fat-shattering dimension. Similarly for the dual distribution class. Thus we get the following corollary of Fact 9:

Corollary 14 *If $\mathcal{H}$ is agnostic PAC learnable, so are the distribution class and dual distribution classes.*

In Section 4, we will adapt the proof from Ben Yaacov (2009) to provide concrete sample complexity bounds on learnability of the expectation class of a measurable family, which will provide improved bounds for the distribution and dual distribution class.

4. 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:

Remark 15 *A warning that* throughout this section and the next, we ignore several measurability issues that arise when uncountable hypothesis classes are considered – e.g. measurability of sets that involve intersecting over uncountably many objects. *These subtleties do not arise when the parameter set and range sets are countable, and all the results in the next sections are true without qualification in this case. Extensions to the uncountable case require further sanity conditions on the hypothesis classes. We defer a discussion of this to Appendix C.*

Theorem 16 *If $\mathcal{F}$ is a measurable family of hypothesis classes such that each class $\mathcal{H} \in \mathcal{F}$ has $\frac{\epsilon}{50}$ fat-shattering dimension at most d, one can perform agnostic PAC learning on the expectation class $\mathbb{E}\mathcal{F}$ with sample complexity:*

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

If $\mathcal{F}$ is a measurable family of $\{0,1\}$-valued concept classes with VC-dimension at most d, one can perform agnostic PAC learning on the expectation class $\mathbb{E}\mathcal{F}$ with sample complexity:

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

Thus, via Proposition 12, we get the same bounds for the distribution class and (moving to the dimension of the dual class) for the dual distribution class.

Example 4 *Recall Example 2, where we consider a family $\mathcal{H}$ of rational functions definable by real-coefficients $\vec{p}$. The distribution function class of $\mathcal{H}$ is parameterized by random $\vec{p}$. It is agnostic PAC learnable with sample complexity as in Theorem 16. 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.*

Thus the next few subsections will be devoted to the proof of Theorem 16. which will complete our sample complexity analysis of agnostic PAC learning for statistical classes.

4.1 Combinatorial and Statistical Tools

Our arguments for PAC learnability will go through the following dimension:

Definition 17 (Glivenko-Cantelli dimension) *The Glivenko-Cantelli dimension of a hypothesis class, denoted $GC_{\mathcal{H}}(\epsilon,\delta)$ is parameterized by $\delta,\epsilon>0$:*

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

Recall that the law of large numbers implies that if we fix any bounded measurable function f into the reals, 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 and Long (1995, Theorem 14). The proof of Theorem 19.1 in Anthony and Bartlett (2009) gives the following version of the connection between these concepts:

Fact 10 (Anthony and 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 $GC_{\mathcal{H}}(\varepsilon,\delta)$ 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 within asymptotic notation.

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*. We will be able to then move to bounds on the Rademacher width of the expectation class, and from there to bound on GC-dimension of the expection class.

Definition 18 *[Width of a set in a direction] Let $A \subseteq \mathcal{R}^n$ be bounded. Given $\vec{b} \in \mathcal{R}^n$, we define $w(A, \vec{b})$, the* width *of A in the particular direction $\vec{b}$, to be $w(A, \vec{b}) = \sup_{\vec{a} \in A} \vec{a} \cdot \vec{b}$.*

Lemma 19 *For any bounded $A \subseteq \mathcal{R}^n$, the width function $\vec{b} \mapsto w(A, \vec{b})$ is Borel measurable.*

Proof If A is countable, then $w(A, \vec{b})$ is the supremum of a countable family of continuous functions of $\vec{b}$, and is thus Borel measurable.

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

Every subspace of $\mathcal{R}^n$ is separable, so for every A there is a countable dense D, and $\vec{b} \mapsto w(D, \vec{b})$, and thus $\vec{b} \mapsto w(A, \vec{b})$, is measurable. ■

Definition 20 *[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 $\mathbb{E}_\beta \left[w(A, \vec{b}) \right]$, 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 $\{+1, -1\}^n$, we define the Rademacher mean width*, denoted $w_{\mathcal{R}}(A)$, as $w(A, \beta)$.*

For any function $h : X \to [0, 1]$ and $\bar{x} = (x_1, \ldots, x_n) \in X^n$, we let $h(\bar{x})$ denote the vector $(h(x_1), \ldots, h(x_n))$. In particular, if $\mathcal{H}$ is a hypothesis class on X parameterized by Θ, $\bar{x} = (x_1, \ldots, x_n) \in X^n$, and $p \in \Theta$, then $\mathcal{H}(\bar{x}, p) = (\mathcal{H}(x_1, p), \ldots, \mathcal{H}(x_n, p))$.

For any hypothesis class $\mathcal{H}$ on X parametrized by Θ and $\bar{x} = (x_1, \ldots, x_n) \in X^n$, let $\mathcal{H}(\bar{x}, \Theta)$ be the set of vectors

$$\{h(\bar{x}) : h \in \mathcal{H}\} = \{\mathcal{H}(\bar{x}, p) : p \in \Theta\} \subseteq [0, 1]^n.$$

Then we extend the definition of Rademacher mean width to be a function of an integer n, given the class $\mathcal{H}$ on X parametrized by Θ; $\mathcal{R}_{\mathcal{H}}(n) = \sup_{\bar{x} \in X^n} w_{\mathcal{R}}(\mathcal{H}(\bar{x}, \Theta))$. We will refer to this function as the Rademacher mean width of the class $\mathcal{H}$, but this only differs from the "Rademacher complexity" of $\mathcal{H}$ in other literature by a factor of n.

The reason for looking at Rademacher mean width will be that it behaves well under averaging with respect to an arbitrary measure: see Theorem 23 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 uniform bounds on a measurable family to bounds on the expectation class, proving Theorem 16.

4.2 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 each class of sets or functions in a measurable family and the dimensions of the expectation class of the family. 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 each class in a family $\mathcal{F}$ through mean width, the same bound will apply to $E\mathcal{F}$. This strategy stems from Ben Yaacov (2009, Theorem 4.1), where it was applied to Gaussian mean width.

We can now make a note about how width of the set of vectors induced by a measurable family behaves under expectation.

Lemma 21 *Let (Ω, Σ, μ) be a probability space, and let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes on X. Fix $\bar{x} \in X^n$ and $\vec{b} \in \mathcal{R}^n$. Then*

$$w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \vec{b}) \leq \mathbb{E}_\mu[w(\mathcal{H}_\omega(\bar{x}, \Theta), \vec{b})].$$

Here recall from Definition 20 that $\mathcal{H}_\omega(\bar{x}, \Theta)$ is the image of the set of vectors for hypothesis class $\mathcal{H}_\omega$ at $\bar{x}$, as y ranges over Θ.

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

$$\begin{aligned} w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \vec{b}) &= \sup_{p \in \Theta} \vec{b} \cdot (\mathbb{E}\mathcal{F}(\bar{x}, p)) \\ &= \sup_{p \in \Theta} \mathbb{E}_\mu[\vec{b} \cdot (\mathcal{H}_\omega(\bar{x}, p))] \\ &\leq \mathbb{E}_\mu \left[\sup_{p \in \Theta} \vec{b} \cdot \mathcal{H}_\omega(\bar{x}, p) \right] \\ &= \mathbb{E}_\mu[w(\mathcal{H}_\omega(\bar{x}, \Theta), \vec{b})]. \end{aligned}$$

which proves the lemma. [2] ■

And we can now extend that statement about width of sets to a statement about mean width.

Lemma 22 *Let (Ω, Σ, μ) be a probability space, and let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes on X.*

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

$$w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \beta) \leq \mathbb{E}_\mu[w(\mathcal{H}_\omega(\bar{x}, \Theta), \beta)].$$

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

$$w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \vec{b}) \leq \mathbb{E}_\mu[w(\mathcal{H}_\omega(\bar{x}, \Theta), \vec{b})].$$

Proof To prove this, we only need to unfold the definition of $w(A, \beta)$ and apply Lemma 21 and Fubini's Theorem.

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

■

We are now ready to bound the Rademacher mean width of an expectation using the Rademacher mean width of the underlying class:

Theorem 23 (Pushing a Mean Width Bound through an Expectation) *Let (Ω, Σ, μ) be a probability space, and let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes on X.*

Then

$$\mathcal{R}_{\mathbb{E}\mathcal{F}}(n) \leq \sup_\omega \mathcal{R}_{\mathcal{H}_\omega}(n).$$

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

$$\sup_{\bar{x} \in X^n} w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \beta) \leq \sup_\omega \sup_{\bar{x} \in X^n} w(\mathcal{H}_\omega(\bar{x}, \Theta), \beta),$$

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

$$w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \beta) \leq \sup_\omega w(\mathcal{H}_\omega(\bar{x}, \Theta), \beta),$$

which follows from Lemma 22 as

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

■

4.3 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 Theorem 4.10 (Wainwright, 2019) in terms of GC-dimension, using the fact that for any probability measure μ on X and any hypothesis class $\mathcal{H}$ on X parameterized by Θ, the *Rademacher complexity* $\frac{1}{n}\mathbb{E}_{\mu^n}[w_{\mathcal{R}}(\mathcal{H}(\bar{x}, \Theta))]$ is at most $\frac{1}{n}\mathcal{R}_{\mathcal{H}}(n)$.

Fact 11 *Let $\mathcal{H}$ be a hypothesis class on X parameterized by Θ. For any $\delta > 0$ and n, then*

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

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

Lemma 24 (From Rademacher Width to GC of Expectation Class) *Let (Ω, Σ, μ) be a probability space, and let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes on X.*

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

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

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

Then by Theorem 23, we see that n is also large enough that $\frac{\mathcal{R}_{\mathbb{E}\mathcal{F}}(n)}{n} \leq \frac{\epsilon}{4}$. Then setting $\gamma = \epsilon - \frac{2\mathcal{R}_{\mathbb{E}\mathcal{F}}(n)}{n}$, our assumption implies that $\gamma \geq \frac{\epsilon}{2}$. Plugging in γ for δ in Fact 11 we have $GC_{\mathbb{E}\mathcal{F}}\left(\epsilon, \exp\left(-\frac{n\gamma^2}{2}\right)\right) \leq n$. 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. ■

4.4 Proof of Theorem 16 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 $\mathcal{R}_{\mathcal{C}}(n)$ in terms of VC-dimension, defining $\mathcal{R}_{\mathcal{C}}(n)$

Fact 12 (Wainwright 2019, Lemma 4.14 and Equation 4.24) *Assume that $\mathcal{C}$ is a concept class with VC-dimension at most $d_{\mathcal{C}}$. Then for $n \geq 1$,*

$$\mathcal{R}_{\mathcal{C}}(n) \leq 2 \cdot \sqrt{d_{\mathcal{C}} \cdot n \cdot \log(n+1)}.$$

This fact will give us the bound on Rademacher mean width that we can plug in to the "pushing through expectation lemma", Lemma 24.

Recall that $\mathcal{F} = (\mathcal{C}_\omega : \omega \in \Omega)$ is a measurable family where each $\mathcal{C}_\omega$ is a $\{0,1\}$-valued class (that is, a concept class) with VC-dimension at most d. Putting Fact 12 together with Lemma 24, we see that to bound the GC-dimension of the expectation class $\mathbb{E}\mathcal{F}$, 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

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

and

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

By Fact 10, this completes the proof of Theorem 16 in the case of concept classes.

4.5 Extension to the real-valued case

We now extend to get sample complexity bounds for random objects, but where we start with a measurable family of hypothesis classes. Our aim will be:

Theorem 25 *For any $\epsilon, \delta > 0$, if $\mathcal{F}$ is a measurable family of hypothesis classes such that each class $\mathcal{H} \in \mathcal{F}$ has $\frac{\epsilon}{50}$ fat-shattering dimension at most d, the sample complexity of agnostic PAC learning on $\mathbb{E}\mathcal{F}$ is bounded by:*

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

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 24.

We will use covering numbers. The ℓ_q norm on $\mathcal{R}^n$ is $(\sum_{i=1}^n x_i^q)^{1/q}$, while ℓ_∞ is $\max_{i=1}^n x_i$.

Definition 26 *For $A \subseteq \mathcal{R}^n$, $\gamma > 0$, and $1 \leq q \leq \infty$, we let $\mathcal{N}_q(\gamma, A)$, the γ* covering number *of A, be the minimum number of γ-balls in the ℓ_q-metric that cover A.*

We also let $\mathcal{N}_q(\gamma, \mathcal{H}, n)$ denote $\sup_{\bar{x} \in X^n} \mathcal{N}_q(\gamma, \mathcal{H}(\bar{x}, \Theta))$, with $\mathcal{H}(\bar{x}, \Theta) \subseteq [0,1]^n$ defined as in Definition 20.

We have defined these for arbitrary q, but from now we will use only $q = 2$ and $q = \infty$. 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 $\mathcal{H}$ and n,

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

We can bound covering numbers using fat-shattering:

Fact 13 (From the proof of Alon et al. 1997, Lemma 3.5) *Let $\mathcal{H}$ be a hypothesis class on X parameterized by Θ. Let d be the $\frac{\gamma}{4}$ fat-shattering dimension of $\mathcal{H}$. Then*

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

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 27 *Let $\beta = (\beta_1, \ldots, \beta_n)$, where the σ_is are independent Gaussian variables with distribution $N(0,1)$. We define the* Gaussian mean width*, denoted $w_{\mathcal{G}}(A)$, 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 14 (Wainwright 2019, Exercise 5.5) *For any $A \subseteq [0,1]^n$,*

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

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

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

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

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

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

Corollary 28 *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 16, using the following bound on Glivenko-Cantelli dimension:

Theorem 29 *For any $\epsilon, \delta > 0$, if $\mathcal{F}$ is a measurable family of hypothesis classes such that each class $\mathcal{H} \in \mathcal{F}$ has $\frac{\epsilon}{50}$ fat-shattering dimension at most d, then*

$$GC_{\mathbb{E}\mathcal{F}}(\epsilon, \delta) = O\left(\frac{d}{\epsilon^2}\log^2\frac{d}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$$

.

Proof By Lemma 24, it suffices to show that there is a constant $C > 0$ such that if

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

then

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

As an intermediate bound, we can use Corollary 28 to bound the Rademacher complexity in terms of covering numbers, using $\gamma = \frac{\epsilon}{\sqrt{32\pi}}$:

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

Thus it suffices to show that for suitably large n,

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

or equivalently,

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

By Fact 13, we see that for all n,

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

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

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

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 $n \geq Cb^{-1}\log^2(ab^{-1})$, then, (setting $x = ab^{-1}$ for notation's sake,)

$$\frac{\log^2(an)}{bn} = \frac{\log^2(a \cdot Cb^{-1}\log^2(x))}{b \cdot Cb^{-1}\log^2(x)} = \frac{\log^2(Cx\log^2(x))}{C\log^2(x)},$$

so it suffices to observe that for large C,

$$\log(Cx\log^2(x)) \leq \log(C) + 3\log(x) \leq \sqrt{C}\log(x).$$

We then see that

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

which completes the proof.

We thank an anonymous reviewer for improving this bound relative to an earlier version of the paper. ■

Theorem 16 for general real-valued hypothesis classes follows from Theorem 29 using Fact 10.

4.6 Simpler arguments for agnostic PAC learnability of the distribution function class

A much simpler argument is available that provides bounds for the distribution class or dual distribution class, but which does not extend to provide a bound for general measurable families as in Theorem 16. 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 $\frac{x_1+\ldots+x_k}{k}$ 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 expectation class of a measurable family, 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 Θ such that $\mathcal{C}$ has finite VC dimension $d_{\mathcal{C}}$ and dual VC dimension $d^*_{\mathcal{C}}$.

We let $\chi^{\mathcal{C}}_x$ be the dual family of characteristic functions: the family of functions on Θ, indexed by elements of X, given by $\chi^{\mathcal{C}}_x(c) = 1$ 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 $\chi^{\mathcal{C}}_m$ be the composed class, indexed by $x_1 \dots x_m \in X$, with each function taking $c \in \Theta$ to

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

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

Fact 16 (Attias and Kontorovich 2024) *The γ fat-shattering dimension of the composed class $\chi^{\mathcal{C}}_m$ 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 $\chi^{\mathcal{C}}_m$ 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 $\mathcal{D}ualDistr_{\mathcal{C}}$ is the dual distribution function class for the concept class $\mathcal{C}$. Suppose the γ fat-shattering dimension of $\mathcal{D}ualDistr_{\mathcal{C}}$ 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 $\vec{c}$, the percentage of elements in the set satisfying $\mathcal{C}_{\vec{c}}$ is within $\frac{\gamma}{2}$ of the measure of the set. Then the $\frac{\gamma}{2}$ fat-shattering dimension of the class $\chi^{\mathcal{C}}_{n_k}$ would be above k'.

Thus $k' \leq J(n_k, d^*_{\mathcal{C}})$, or restated, the γ fat-shattering dimension of $\mathcal{D}ualDistr_{\mathcal{C}}$ 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 17 *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 $\mathcal{D}ualDistr_{\mathcal{C}}$ 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 $\mathcal{D}istr_{\mathcal{C}}$ is bounded by:

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

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 $\tilde{O}(\frac{1}{\epsilon^{\lambda+1}})$ 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 the expectation class of a measurable family. We do not compute sample complexity bounds explicitly for this alternative proof, but they are similar to those given above.

Theorem 30 *Let $\mathcal{H}$ be a class of functions over X, indexed by Θ. 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 Θ, 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 $\mathrm{Avg}_m(v_1\dots v_m)$ be the average of $v_1\dots v_m$. Let $\mathrm{Avg}_m(\mathcal{H}^*)$ be the composed class, indexed by $x_1\dots x_m$ in X^m taking c to $\mathrm{Avg}_m(h_{x_1}(c)\dots h_{x_m}(c))$.

This again fits the composition framework in Theorem 1 of Attias and 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 $m\cdot\mathrm{FatSHDim}_\gamma(\mathcal{H}^*)$.

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

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

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

Fix γ, and again let $\mathcal{D}ualDistr_{\mathcal{H}}$ be the dual distribution function class for $\mathcal{H}$. Suppose the γ fat-shattering dimension of $\mathcal{D}ualDistr_{\mathcal{H}}$ 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 $\mathcal{D}ualDistr_{\mathcal{H}}$ is bounded by $J(n_k, \text{FatSHDim}_\gamma(\mathcal{H}^*))$.

It suffices to take n_k large enough that, for some fixed $0 < \delta < 1$, $GC_{\mathcal{H}}\left(\frac{\gamma}{2}, \delta\right) \leq n_k$, 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

$$n_k = O\left(\frac{1}{\epsilon^2} \cdot \text{FatSHDim}_{\frac{\epsilon}{9}}(\mathcal{H}) \cdot \log^2\left(\frac{1}{\epsilon}\right)\right).$$

4.7 The Realizable Case for PAC Learning

The previous subsections showed preservation of agnostic PAC learnability for the distribution class construction. We now show by contrast that the distribution class constructions do not preserve PAC learnability in the realizable case.

Proposition 31 *There is a hypothesis class $\mathcal{H}$ that is realizable PAC learnable, but the distribution function class and dual distribution function 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 Example 1 (Attias et al., 2023). Let X be a set, partitioned into nonempty pieces $X_0, X_1, \ldots$, 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 $b = (b_0, b_1, \ldots)$, 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 32 *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 $\vec{b}_1 \ldots \vec{b}_n$, with value at most $\frac{1}{4}$, we will be able to exclude some partition elements X_i, but we will have infinitely many X_i possible. ■

We now show the same thing for the distribution class.

Lemma 33 *The distribution class of $\mathcal{H}_0$ is 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 $\frac{1}{8}$-graph dimension at least n. The definition of graph dimension states that this holds when we have $x_0, \ldots, x_{n-1} \in X$, $f_1, \ldots, f_{n-1} \in [0,1]$, 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 $\frac{1}{8}$-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 $1_n = (1, 1, \ldots, 1, 0, 0, \ldots) \in \{0,1\}^{\mathbb{N}}$ 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 zeros, that is, $\beta'_i = 0$ for $i \geq n$. For $b \in \{0,1\}^{\mathbb{N}}$, let

$$c_b = \frac{1}{8} \sum_{i=0}^{\infty} b_i \cdot 2^{-i},$$

so that for any $x_i \in X_i$,

$$h_b(x_i) = \frac{3}{4} b_i + c_b.$$

We note that for all b,

$$0 \leq c_b \leq \frac{1}{8} \sum_{i=0}^{\infty} 2^{-i} = \frac{1}{4},$$

and if there are only finitely many i such that $b_i = 1$, as is the case if $b = \beta'$ for some $\beta \in \{0,1\}^n$ or $b = 1_n$ for some n, then c_b is 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

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

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

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

and if $\beta_i = 1$, we have both $h_{\beta'}(x_i), h_{1_n}(x_i) \geq \frac{3}{4}$, so $h'_\beta(x_i) = \lambda \cdot h_{\beta'}(x_i) + (1-\lambda) \cdot h_1(x_i) \geq \frac{3}{4}$. Thus in particular $|h'_\beta(x_i) - f_i| > \frac{1}{8}$ as required. ■

We now show that we can "fix" these anomalies by dealing not with a single hypothesis classes, but a family with reasonable closure properties. If we consider classes closed under composition with continuous bijections $[0,1] \to [0,1]$, then there is no difference between the realizable and agnostic case, and we do have preservation under moving to statistical classes.

Proposition 34 *Let $\mathcal{F}$ be a family of hypothesis classes on range space X parametrized by Θ closed under applying continuous bijections: for any $\mathcal{H} \in \mathcal{F}$, the composition of $\mathcal{H}$ with a continuous bijection f is another class in $\mathcal{F}$. If $\mathcal{F}$ contains a hypothesis class with infinite γ fat-shattering dimension for some $\gamma > 0$, there is another hypothesis class in $\mathcal{F}$ which is not realizable PAC learnable. Thus in particular if every hypothesis class in $\mathcal{F}$ is realizable PAC learnable, then every class in $\mathcal{F}$ is agnostic PAC learnable, and thus every distribution class or dual distribution class arising from a class in $\mathcal{F}$ is agnostic PAC learnable and realizable PAC learnable.*

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 $r < s$, where $s - r$ is bounded below by a function of γ:

Fact 18 (Alon et al. 1997, Theorem 4.2) *For every $\gamma > 0$, there is some $\beta > 0$ such that the following holds:*

Consider a real-valued hypothesis class $\mathcal{H}$ consisting of functions h_p for p ranging over the parameter set. Suppose that $\mathcal{H}$ has infinite γ fat-shattering dimension. Then for every natural number d, there are also $0 \leq r < s \leq 1$ such that $s - r \geq \beta$ and there are $x_1, \ldots, x_d \in X$ and $(p_b : b \in \{0,1\}^d)$ such that for each b and i, if $b(i) = 0$, then $h(x_i; p_b) \leq r$, and if $b(i) = 1$, then $h(x_i; p_b) \geq s$.

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

We now begin the proof of Proposition 34:

Proof Fix a family $\mathcal{F}$ of hypothesis classes that is closed under applying continuous bijections. Assume there is a class $\mathcal{H} \in \mathcal{F}$ that is not agnostic PAC learnable, and we will prove that there is some $\mathcal{H}' \in \mathcal{F}$ that is not realizable PAC learnable.

In particular, we assume that $\mathcal{H}$ has infinite γ fat-shattering dimension for some $\gamma > 0$. Then by Fact 18, there is some $\beta > 0$ such that for every natural number d, there are also $0 \leq r < s \leq 1$ such that $s - r \geq \beta$ and there are $x_1, \ldots, x_d \in X$ and $(p_b : b \in \{0,1\}^d)$ such that for each b and i, if $b(i) = 0$, then $\mathcal{H}_{p_b}(x_i) \leq r$, and if $b(i) = 1$, then $\mathcal{H}_{p_b}(x_i) \geq s$.

Let $f : [0,1] \to [0,1]$ be a continuous bijection such that $f(r) = 0$ and $f(s) = 1$. Then consider the hypothesis class $\mathcal{H}' = \{h'_b : b \in \Theta\}$ defined by $h'_b(x) = f(h_b(x))$. Because $\mathcal{F}$ is closed under applying continuous bijections, $\mathcal{H}' \in \mathcal{F}$, and we will show that $\mathcal{H}$ is not realizable PAC learnable.

To do this, we show that the $\frac{1}{8}$-graph dimension of $\mathcal{H}'$ is infinite. (In fact, the $1-$graph dimension is.) For every d, there are there are $x_1, \ldots, x_d \in X$ and $(p_b : b \in \{0,1\}^d)$ such that for each b and i, if $b(i) = 0$, then $h'_{p_b}(x_i) = 0$, and if $b(i) = 1$, then $h_{p_b}(x_i) = 1$.

To show that the $\frac{1}{8}$-graph dimension is greater than d, by the characterization in the proof of Lemma 33, it now suffices to define $f_1, \ldots, f_{n-1} \in [0,1]$ to all be 0, which ensures that

- $h'_{p_b}(x_i) = f_i$ when $b_i = 0$
- $|h'_{p_b}(x_i) - f_i| > \frac{1}{8}$ when $b_i = 1$,

because $|h'_{p_b}(x_i) - f_i| = h'_{p_b}(x_i)$. ■

5. Online learnability of statistical classes

We will now be interested in getting an analog of Theorem 16 for online learning, deriving bounds on learnability for statistical classes based on dimensions of the base class. As we did in the agnostic PAC case, we will work via a broader result on measurable families. Recall from Fact 6 that agnostic PAc learnability is controlled by the sequential fat-shattering dimension of a hypothesis class. We will calculate the following regret bound on the expectation class of a measurable family in terms of a bound on sequential fat-shattering dimension for the members in the family:

Theorem 35 *For any $\epsilon, \delta > 0$, if $\mathcal{F}$ is a measurable family of hypothesis classes such that each class $\mathcal{H} \in \mathcal{F}$ has $\frac{\epsilon}{50}$ sequential fat-shattering dimension at most d, the minimax regret of online learning for $\mathbb{E}\mathcal{F}$ 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 the classes in $\mathcal{F}$ are $\{0,1\}$-valued and have Littlestone dimension at most d, the minimax regret of online learning for $\mathbb{E}\mathcal{F}$ over runs of length n is at most $O\left(\sqrt{d \cdot n}\right)$.

Note that we can again apply Proposition 12 to apply this to the distribution class of a fixed hypothesis class that has finite sequential fat-shattering dimensions. 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 36 *If function class $\mathcal{H}$ is agnostic online learnable, in the sense of having sublinear minimax regret, then so are the distribution class and the dual distribution class.*

For the corollary, we use the fact that given any class $\mathcal{H}$ of finite γ sequential fat-shattering dimension d, every element of the the parameter randomized family of $\mathcal{H}$ (see Definition 9) has γ sequential fat-shattering dimension at most d. So its expectation class, $\mathcal{H}(\text{CompatParamRV}(\mathcal{H}))$, as in Definition 10, is agnostic online learnable. The distribution class then embeds into $\mathcal{H}(\text{CompatParamRV}(\mathcal{H}))$. To deal with the dual distribution class, we also use the fact that agnostic online learnability is closed under dualization by Proposition 3.

5.1 Proof of the Main Theorems on Agnostic Online Learnability of Statistical Classes, with Quantitative Bounds

We now present the proof of Theorem 35.

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 expectation 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 37 (Sequential Rademacher mean width) *For any n, let $\{1,-1\}^{<n} = \bigcup_{t=0}^{n-1}\{1,-1\}^t$. For each sequence $s = (s_1,\ldots,s_n) \in \{1,-1\}^n$, let $v_s \in \mathcal{R}^{\{1,-1\}^{<n}}$ be the vector such that for $1 \leq t \leq n$, $(v_s)_{(s_1,\ldots,s_{t-1})} = s_{t+1}$, 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 20. Define the* sequential Rademacher mean width *of a set $A \subseteq \mathcal{R}^{\{1,-1\}^{<n}}$, $w^S_{\mathcal{R}}(A)$, to be $w(A,\beta)$ where β is the uniform distribution on the 2^n elements of T_n.*

For a hypothesis class $\mathcal{H}$ on X, we then define the sequential Rademacher mean width*, $\mathcal{R}^{Seq}{}_f(n)$, to be $\sup_{\bar{x}\in X^{\{1,-1\}^{<n}}} w_{\mathcal{R}^{Seq}}(\mathcal{H}(\bar{x},\Theta)$. 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 19 (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*

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

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

Theorem 38 *Let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes on X. Then*

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

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 $A \subseteq \mathcal{R}^{\{1,-1\}^{<n}}$,

$w^S_{\mathcal{R}}(A)$ is $w(A, \beta)$. For a function $f(x, y)$, the sequential Rademacher mean width was defined by

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

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

$$\sup_{\bar{x} \in X^{\{-1,1\}^{<n}}} w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \beta) \leq \sup_{\omega} \sup_{\bar{x} \in X^{\{-1,1\}^{<n}}} w(\mathcal{H}_\omega(\bar{x}, \Theta), \beta),$$

and the proof of this is essentially identical to the proof of Theorem 23, but with a new distribution β. ■

From this, we can prove a bound on regret for the expectation class in terms of the sequential fat-shattering dimension, proving the first part of Theorem 35.

Theorem 39 *Let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes on X.*

The minimax regret of online learning for $\mathbb{E}\mathcal{F}$ 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 38, any uniform (in ω) bound on $\mathcal{R}^{Seq}{}_{\mathcal{H}_\omega}(n)$ also applies to the expectation class, so regret for the expectation class is bounded by

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

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

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

Fact 20 (Alon et al. (2021, See Lemma 6.4 and Theorem 12.2)) *If $\mathcal{C}$ 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 expectation class of a family of concept classes with Littlestone dimension bounded by d is also at most $O(\sqrt{d \cdot n})$, which proves the second part of Theorem 35.

5.2 Preservation of Online Learnability in Moving to Statistical Classes, Via Stability

Above we showed that online learnability is preserved in moving to statistical classes. We now give a variant without the quantitative bounds, but one which derives from prior

results stated in the context of model theory. We can show that agnostic online learnability is equivalent to the notion of *stability*, a notion originating in logic. We can then use prior results (Ben Yaacov, 2009, 2013) on the preservation of stability in moving to an expectation class.

Definition 40 *A concept class $\mathcal{C}$ over X parameterized by Θ is* stable *if there do not exist arbitrarily large sequences $a_i, b_i : 1 \leq n$ with $a_i \in X, b_i \in \Theta$ such that $\forall i, j \leq n$ $a_i \in \mathcal{C}_{b_j}$ if and only if $i < j$.*

Roughly speaking stability says that $\mathcal{C}$ does not define arbitrarily large linear orders. This can be phrased as $\mathcal{C}$ having finite *threshold dimension*, as it is called in Ghazi et al. (2021), where this dimension is the largest size of a finite sequence of pairs (a_i, b_i) with the above property.

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 21 (Chase and Freitag 2019) *A concept class is stable if and only if it 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 the class having finite Littlestone dimension by Fact 6.

Thus far we are reviewing a connection between stability and online learnability for concept classes, which is already known. We now turn to real-valued classes. The notion of a stable hypothesis class generalizes to this setting, but now requires a real parameter. It can be defined in terms of either of the following dimensions:

Definition 41 *Let $\mathcal{H}$ be a hypothesis class on X.*

For any $\gamma > 0$ and any d, say that $\mathcal{H}$ has $\gamma-$threshold dimension at least *d when there are $a_1, \ldots, a_d \in X$, $h_1, \ldots, h_d \in \mathcal{H}$ such that for all $i < j$,*

$$|h_j(a_i) - h_i(a_j)| \geq \gamma.$$

For any $r < s$ and any d, say that $\mathcal{H}$ has $(r, s)-$threshold dimension at least *d when there are $a_1, \ldots, a_d \in X$, $h_1, \ldots, h_d \in \mathcal{H}$ such that for all $i < j$, $h_j(a_i) \leq r$ and $h_i(a_j) \geq s$.*

Call a hypothesis class $\mathcal{H}$ γ-stable *when it has finite γ-threshold dimension, and call $\mathcal{H}$* (r, s)-stable *when it has finite (r, s)-threshold dimension.*

These dimensions give two equivalent definitions of overall stability for a hypothesis class, as shown in the model-theoretic context in Ben Yaacov and Usvyatsov (2010, Section 7):

Fact 22 (Ben Yaacov and Usvyatsov 2010, Section 7) *If $\mathcal{H}$ is a hypothesis class, the following are equivalent:*

- *For every $\gamma > 0$, $\mathcal{H}$ is γ-stable.*

- *For every $r < s$, $\mathcal{H}$ is $(r,s)-$stable.*

We call such a class *stable*. Roughly speaking stability says that we cannot use gaps in function values, discretized up to some γ, to define arbitrarily large linear orders.

We will give a slightly more explicit version of this equivalence, using largely the same proof as in Ben Yaacov and Usvyatsov (2010), but with only finitary Ramsey theory, in order to define stability uniformly over a family of hypothesis classes.

Definition 42 *If $\mathcal{F}$ is a family of hypothesis classes on X parameterized by Θ, and $\gamma > 0$, say that $\mathcal{F}$ is* uniformly γ-stable *when there is some d such that every $\mathcal{H} \in \mathcal{F}$ has $\gamma-$threshold dimension at most d.*

For $r < s$, say that $\mathcal{F}$ is uniformly (r,s)-stable *when there is some d such that every $\mathcal{H} \in \mathcal{F}$ has $(r,s)-$threshold dimension at most d.*

These two notions are closely related:

Lemma 43 *For any $\gamma > 0$ and d, if $\mathcal{H}$ has γ-threshold dimension less than d, then for all r, s with $r + \gamma \leq s$, $\mathcal{H}$ has (r,s)-threshold dimension less than d.*

For any $0 < \delta < \gamma$ and d, there is d' such that if $\mathcal{H}$ has (r,s)-threshold dimension less than d whenever $r + \delta \leq s$, then $\mathcal{H}$ has γ-threshold dimension less than d'.

Proof It is straightforward to see that if $\mathcal{H}$ has (r,s)-threshold dimension at least d, then the same $a_1, \ldots, a_d \in X$, $h_1, \ldots, h_d \in \mathcal{H}$ that recognize this show that $\mathcal{H}$ has $(s-r)$-threshold dimension at least d.

Now fix $0 < \delta < \gamma$ and d. Let $n = \lceil \frac{1}{\gamma - \delta} \rceil$. Then by Ramsey's theorem, there is some d' such that any coloring of the complete graph on d' vertices with $2n$ colors admits a monochromatic set of size d.

We now show that if $\mathcal{H}$ has γ-threshold dimension at least d', then there are some $r < s$ with $r + \delta \leq s$ such that $\mathcal{H}$ has (r,s)-threshold dimension at least d.

By our assumption, there are $a_1, \ldots, a_{d'} \in X$, $h_1, \ldots, h_{d'} \in \mathcal{H}$ such that for all $i < j$, $|h_j(a_i) - h_i(a_j)| \geq \gamma$.

Thus for each $i < j$, because $\frac{1}{n} \leq \gamma - \delta$, there is some $0 \leq k < n$ such that either $h_j(a_i) \leq \frac{k}{n}, -h_i(a_j) \geq \frac{k}{n} + \delta$, or $h_i(a_j) \leq \frac{k}{n}, -h_j(a_i) \geq \frac{k}{n} + \delta$. This gives $2n$ possible cases, so we may color the complete graph on d' with $2n$ colors such that every edge (i,j) with $i < j$ given a particular color falls into the same case. We may thus find a monochromatic subset $I \subseteq \{1, \ldots, d'\}$ of size d. Let k be the k appearing in the case corresponding to the color of that monochromatic subset, and then let $r = \frac{k}{n}, s = \frac{k}{n} + \delta$. Then either for all $i < j$ in I, $h_j(a_i) \leq r, -h_i(a_j) \geq s$, in which case we are done, or for all $i < j$ in I, $h_i(a_j) \leq r, -h_j(a_i) \geq s$, in which case we only need to reverse the order of our witnesses. ■

Thus if we universally quantify over all the parameters, these two notions are the same:

Corollary 44 *If $\mathcal{F}$ is a family of hypothesis classes on X parameterized by Θ, then the following are equivalent:*

- *For every $\gamma > 0$, $\mathcal{F}$ is uniformly γ-stable.*

- *For every $r < s$, $\mathcal{F}$ is uniformly (r, s)-stable.*

This lets us define two equivalent properties of a class $\mathcal{F}$ - if either holds, we call $\mathcal{F}$ *uniformly stable*.

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

Theorem 45 *A hypothesis class $\mathcal{H}$ is stable if and only if for every $\gamma > 0$, the sequential fat-shattering dimension $\mathrm{FatSHDim}^{Seq}_{\gamma}(\mathcal{H})$ is finite, and a family $\mathcal{F}$ of hypothesis classes is uniformly stable if and only if for every $\gamma > 0$, there is some d such that for each $\mathcal{H} \in \mathcal{F}$, the sequential fat-shattering dimension $\mathrm{FatSHDim}^{Seq}_{\gamma}(\mathcal{H})$ is at most d.*

Specifically, the following implications hold:

- *If $\mathcal{H}$ has γ sequential fat-shattering dimension less than d, and $r + \gamma \leq s$, then $\mathcal{H}$ has (r, s)-threshold dimension less than $2^{d+1} - 1$.*
- *If $0 < \delta < \frac{\gamma}{2}$, then for every d there is some d' such that if $\mathcal{H}$ has δ-threshold dimension less than d, then $\mathcal{H}$ has γ sequential fat-shattering dimension less than d'.*

The interest in the above theorem is that stability has already been shown to be preserved in moving to from a measurable family to its expectation class:

Fact 23 (Ben Yaacov and Keisler 2009; Ben Yaacov 2013) *If a measurable family $\mathcal{F}$ of hypothesis classes on X parameterized by Θ is uniformly stable, then the expectation class $\mathbb{E}\mathcal{F}$ is also stable.*

The result above is phrased very differently in the cited source, in terms of continuous logic. In Appendix D we explain how to get from the statement in the sources to this version.

Thus we can combine Theorem 45, Fact 23, 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 46 *For any measurable family $\mathcal{F}$ of hypothesis classes on X parameterized by Θ, the expectation class $\mathbb{E}\mathcal{F}$ is agnostic online learnable if and only if for every $\gamma > 0$, there is some d such that for each $\mathcal{H}$ in the range of $\mathcal{F}$, $\mathcal{H}$ has γ sequential fat-shattering dimension at most d.*

In particular, if a hypothesis class $\mathcal{H}$ is agnostic online learnable, so are its distribution class and dual distribution class.

We now turn to proving Theorem 45:

Proof Assume that $\mathcal{H}$ has (r, s)-threshold dimension at least $2^{d+1} - 1$, with $\gamma = s - r$. We will show that $\mathcal{H}$ 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 t_1 of length k with $k < d$, and any string t_2 strictly extending t_1, if $(t_2)_k = -1$ then $t_1 < t_2$, and if $(t_2)_k = 1$ then $t_2 > t_1$. By the assumption that $\mathcal{H}$ has (r, s)-threshold dimension at least $2^{d+1} - 1$, as the set $\{-1, 1\}^{\leq d}$ has size $2^{d+1} - 1$, there are $((a_t, h_t) : t \in \{-1, 1\}^{\leq d})$ such that for all $i < j$ in this order, $h_j(a_i) \leq r$ and $h_i(a_j) \geq s$.

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 $E_{<t} = (E(0), \ldots, E(t-1))$. We consider h_E, and see that for all $0 \leq t < d$, if $E(t) = -1$, then $E < E_{<t}$, so $h_E(a_{E_{<t}}) \leq r$, while if $E(t) = 1$, we have $E > E_{<t}$ and $h_E(a_{E_{<t}}) \geq s$. As $s - r \geq \gamma$, this tree is γ-shattered. Thus we have finished the proof of the first dimension implication in Theorem 45.

To prove the other direction, we fix $0 < \delta < \frac{\gamma}{2}$. We will show by induction on d that if $k > (\gamma - 2\delta)^{-1}$ and $\mathcal{H}$ γ-shatters a binary tree T in X of depth $\frac{k^{d+1}-1}{k-1}$, then there are $x_1, \ldots, x_d \in X$ and $h_1, \ldots, h_d \in \mathcal{H}$ 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 47 (Subtree) *Let $T : \{-1,1\}^{<d} \to X$ 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 $(E_0, \ldots, E_t, -1)$. If instead E' extends $(E_0, \ldots, E_t, 1)$, we call y a* right descendant *of x.*

Then a subtree *of T of depth $d' \leq d$ is a map $T' : \{-1,1\}^{<d'} \to X$ 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 24 (Alon et al. 2019, Lemma 16) *If p, q are positive integers, $T : \{-1,1\}^{<p+q-1} \to X$ 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 48 (Tree Ramsey) *If $d_1, \ldots, d_k$ are positive integers, $T : \{-1,1\}^{<d_1+\cdots+d_k-k+1} \to X$ is a binary tree in X of depth $d_1 + \cdots + d_k - k + 1$, and the elements of X are partitioned into k sets $X_1, \ldots, X_k$, 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 $x_1, \ldots, x_d \in X$ and $h_1, \ldots, h_d \in \mathcal{H}$ 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 $I_1, \ldots, I_k$ each of width at most $\gamma - 2\delta$, and then partition X into sets $X_1, \ldots, X_k$ where if $x \in X_i$, then $h(x) \in I_i$. Then by Corollary 48, 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 $x_1, \ldots, x_d$ of x' in T' and $h_1, \ldots, h_d \in \mathcal{H}$ 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 45. ■

.

5.3 Realizable Online Learning for Statistical Classes

We now turn to preservation of realizable online learning for statistical classes. As with realizable PAC learning, our results will be negative.

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 49 *[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 $T : \{-1, 1\}^{<d} \to X$, a real-valued binary tree $\tau : \{-1, 1\}^{<d} \to [0, 1]$, and a $\mathcal{H}$-labelling of each branch in such a tree, $\Lambda : \{-1, 1\}^d \to \mathcal{H}$, 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 $t_0, \ldots, t_{d-1}$, we have $\sum_{i=0}^{d-1} \tau(t_i) > D$.*

Finiteness of online dimensions characterizes realizable online learnability.

Fact 25 (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 50 *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 $\tau : \{-1,1\}^{<d} \to [0,1]$ 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

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

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 $d(\gamma - \epsilon)$. ■

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 51 *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, we see that:

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 52 *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 53 *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 Θ, 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 $f : [0,1] \to [0,1]$ such that $f \circ \mathcal{H} = \mathcal{H}'$, as functions $X \times \Theta \to [0,1]$.*

In terms of learning, this means that both classes are agnostic online learnable, but only $\mathcal{H}'$ is realizable online learnable. In particular, this show 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 $\Gamma = \gamma_0, \gamma_1, \ldots$ of positive reals to be determined, with $\lim_d \gamma_d = 0$. We define $\mathcal{H}^\Gamma$, 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 $\mathcal{H}^\Gamma$ has finite γ sequential fat-shattering dimension. Specifically, if $\mathcal{H}^\Gamma$ 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 50, $\mathcal{H}^\Gamma$ 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 $\mathcal{H}^\Gamma$ is at most $\sum_{i=0}^{\infty} 2^i \gamma_i$, and in particular, $\mathcal{H}^\Gamma$ 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 $t_0, \ldots, t_{d-1}$. We can bound the sum of the weights of that branch by grouping the indices i by the value of $T(t_i)$:

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

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 $\sum_{x \in X} \left(\sum_{0 \leq i < d-1 \colon T(t_i) = x} \tau(t_i) \right)$. If x has length ℓ and i is such that $T(t_i) = x$, then $\tau(t_i) \leq \gamma_\ell$, so

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

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

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

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

We now claim that if $\mathcal{H}^\Gamma$ is constructed from the sequence $\Gamma = \gamma_1, \gamma_2, \ldots$ while $\mathcal{H}^{\Gamma'}$ is constructed in the same way from the sequence $\Gamma' = \gamma'_1, \gamma'_2, \ldots$, then there is an increasing homeomorphism $f : [0, 1] \to [0, 1]$ 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 $\lim_{x \to 0} f(x) = 0$, so defining $f(0) = 0$ will maintain continuity.

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

Using the same family of examples, we now prove Proposition 52:

Proof Let $\Gamma = \gamma_i : i > 0$ be a sequence such that the $i \cdot \gamma_i$ is unbounded. Let $\mathcal{H}^\Gamma$ be the class from Theorem 53. Recall that range points are prefixes x (finite sequences). Hypotheses are parameterized by infinite sequences b and the value of a hypothesis h_b on a prefix x is either zero or $\frac{1}{\gamma_n}$ for n the length of x. Then, as proven in Theorem 53 H_γ is not online learnable in the realizable case.

Let D_γ be the dual class. So points are now ω-sequences b, hypotheses are prefixes x, and the value of a hypothesis h_x at a sequence b is 0 if b does not extend x and is $\frac{1}{\gamma_n}$ if b does extend x, where again n is the length of prefix x. 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 b. And at a move for learner with previous adversary range points $b_1 \dots b_k$, learner knows the values $v_1 \dots v_{k-1}$ of a D_γ-consistent hypothesis for $b_1 \dots b_{k-1}$. 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 b_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. ■

5.4 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.

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 $\sum_{i \le T} |y_i' - y_i|$. Here we let the loss be $\sum_{i \le T} \ell(|y_i' - y_i|)$, for certain $\ell : [0,1] \to [0,\infty)$ 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 54 (Online learnability for a general loss function) *Given a loss function $\ell : [0,1] \to [0,\infty)$, 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 55 *Let $C > 0$, and suppose $\ell_1, \ell_2 : [0,1] \to [0,1]$ are loss functions with $C\ell_1(x) \le \ell_2(x)$ 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. ■

We now define the loss functions we will focus on:

Definition 56 (ϵ-truncated loss functions) *Given $\epsilon > 0$, define $\ell_\epsilon, L_\epsilon : [0,1] \to [0,\infty)$ by*

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

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 57 *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 $x \in [0,1]$,

$$\epsilon L_\epsilon(x) \leq \ell_{\text{id}}(x)$$
$$\ell_\epsilon(x) \leq L_\epsilon(x),$$

so the result follows by Lemma 55. ■

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 58 *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 59 *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 for both the distribution class and the dual distribution class.*

We now work towards the proof of Theorem 58.

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 60 *[Online dimension for a general loss function] In the context of a loss function $\ell : [0,1] \to [0,1]$, 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 $T : \{-1,1\}^{<d} \to X$, a real-valued binary tree $\tau : \{-1,1\}^{<d} \to [0,1]$, and a $\mathcal{H}$-labelling of each branch in such a tree, $\Lambda : \{-1,1\}^d \to \mathcal{H}$, 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*

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

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

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

Fact 26 (Attias et al. 2023, Theorem 4) *Let $\mathcal{H}$ be a hypothesis class on a set X. Then when $\ell : [0,1] \to [0,1]$ is a loss function, $\mathcal{H}$ has bounded regret for realizable online learning if and only if $\mathrm{Onl}_\ell(\mathcal{H}) < \infty$.*

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

Lemma 61 *For any non-decreasing loss function $\ell : [0,1] \to [0,1]$, if a hypothesis class $\mathcal{H}$ on X γ sequentially fat-shatters a binary tree in X of depth d, then $\mathrm{Onl}_\ell(\mathcal{H}) \geq d \cdot \ell(\gamma)$, 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 $\mathrm{Onl}_\ell(\mathcal{H}) \geq d \cdot \ell(\gamma)$. This is only a slight modification of Lemma 50.

As in that proof, suppose T is the γ-fat-shattered tree. Let the binary tree $s : \{-1,1\}^{<d} \to \mathcal{R}$ and the branch labelling Λ which labels each branch $b \in \{-1,1\}$ 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 $\tau : \{-1,1\}^{<d} \to [0,1]$ 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 $d(\ell(\gamma) - \epsilon)$. ■

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

Lemma 62 *For any hypothesis class $\mathcal{H}$ and any $\epsilon > 0$, if $\mathrm{Onl}_{L_\epsilon}(\mathcal{H})$ 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 Section 5.2.

Specifically we will show that for $0 < \delta < \frac{\gamma}{2}$, $\mathcal{H}$ is not δ-stable, which suffices by Theorem 45.

Because L_ϵ is $\{0,1\}$-valued, if D is an integer and $\mathrm{Onl}_{L_\epsilon} \geq D$, then there is a tree $T : \{-1,1\}^{<d} \to X$, a $\{0,1\}$-valued binary tree $\tau : \{-1,1\}^{<d} \to \{0,1\}$, and an $\mathcal{H}$-labelling of each branch in the tree, $\Lambda : \{-1,1\}^d \to \mathcal{H}$, 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 $t_0, \ldots, t_{d-1}$, 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 47.

We claim that for any integer D, if every branch of a finite-depth $\{0,1\}$-valued binary tree $t : \{-1,1\}^{<d} \to \{0,1\}$ 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 $\iota : \{-1,1\}^{<D} \to \{-1,1\}^{<d}$, 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 $t : \{-1,1\}^{<d} \to \{0,1\}$ 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 $t' : \{-1,1\}^{<d'} \to \{0,1\}$, 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 $v = (v_0, \ldots, v_\ell)$, and let $b = (b_0, \ldots, b_{d'-1})$ - we then define this new branch to be $b' = (v_0, \ldots, v_\ell, -1, b_0, \ldots, b_{d'-1})$. 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 $\iota_{-1} : \{-1,1\}^{<D} \to \{-1,1\}^{<d'}$ 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 $\iota : \{-1,1\}^{<D} \to \{-1,1\}^{<d}$ 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 45. Fix $0 < \delta < \frac{\epsilon}{2}$ and $k > (\epsilon - 2\delta)^{-1}$. We wish to show that for all d, there are $x_1, \ldots, x_d \in X$ and $h_1, \ldots, h_d \in \mathcal{H}$ 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 $x_1, \ldots, x_d \in X$ and $h_1, \ldots, h_d \in \mathcal{H}'$ 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.

Consider the base case, $d = 1$. 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 spreads-shatters a binary tree of depth $\frac{k^{d+1}-1}{k-1}$, then there are $x_1, \ldots, x_d \in X$ and $h_1, \ldots, h_d \in \mathcal{H}'$ 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 spread-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 $I_1, \ldots, I_k$ each of width at most $\epsilon - 2\delta$, and then partition X into sets $X_1, \ldots, X_k$ where if $x \in X_i$, then $h(x) \in I_i$. Then by our Ramsey result for trees, Corollary 48, 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 $\{h_R(x') : h_R \in \mathcal{H}_R\}$ 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 $x_1, \ldots, x_d$ of x' in T' and $h_1, \ldots, h_d \in \mathcal{H}_L$ 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 $|h_i(x') - h(x_i)| \geq \delta$. ■

We are now ready to prove Theorem 58:

Proof Many of these implications follow from Lemma 57. 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 61, 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 62, 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 57 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 61. ■

6. Related Work

As noted earlier, what we refer to as the "dual distribution class" is from Hu et al. (2022), where it is defined only for concept classes. The basic result of Hu et al. (2022) is that agnostic PAC learnability of a base concept class implies agnostic PAC learnability of the dual distribution class, with accompanying sample complexity bounds that are asymptotically looser than ours. The extension of dual distribution classes to base classes consisting of real-valued values, as well as the notion of the distribution class, is new to our work.

We have presented our arguments without relying on logical notions. But our work is inspired by work on learnability of hypothesis classes coming from logical formulas, so we say more about the connection here. Given a structure $\mathfrak{M}$ and a logical formula $\phi(x_1 \ldots x_k)$,

one can look at the k-tuples from the domain of $\mathfrak{M}$ that satisfy ϕ: this is a *definable set of tuples* from $\mathfrak{M}$. Given a first-order formula $\phi(x_1 \dots x_j; y_1 \dots y_k)$ in which the free variables are partitioned, we get a family of sets of j-tuples, by varying $\vec{y}$ over all k-tuples in $\mathfrak{M}$. Thus each $\phi(\vec{x}; \vec{y})$ gives a concept class with range space the j-tuples from the domain of $\mathfrak{M}$, and parameter space the k-tuples from the structure. For example the family of rectangles in the reals and the family of intervals in the reals are definable families. A partitioned formula is said to be *NIP* if the corresponding concept class has finite VC dimension.

Model theorists have identified a number of structures where the concept class arising from each partitioned formula has finite VC dimension (or equivalently, is agnostic PAC learnable): these are called *NIP structures*: see, for example, Simon (2015). As discussed in our paper, agnostic online learnability corresponds to a hypothesis class being stable: see Definition 40. A partitioned formula $\phi(\vec{x}; \vec{y})$ in a structure $\mathfrak{M}$ is called stable exactly when the corresponding concept class is stable in the sense of Definition 40. The study of stable formulas has been developed over many decades by model theorists, and many structures have been identified in which every partitioned formula is stable: see, for example Shelah (1990). There is an analogous notion of "continuous logic", in which formulas are real-valued, and in continuous logic partitioned formulas give real-valued hypothesis classes. The results in works such as Ben Yaacov (2009); Ben Yaacov and Keisler (2009) are phrased in terms of NIP and stability of continuous logic structures.

The notion of measurable family and its expectation class was introduced in Ben Yaacov (2009). It relates to a transformation that is similar to the one we perform here: moving from a logical structure to its "randomization", another structure where the elements are random variables. The notion of randomization originates in Keisler (1999), and was developed further in Ben Yaacov (2009); Keisler (1999); Ben Yaacov and Keisler (2009). Many of our results on both agnostic PAC learning and agnostic online learning of statistical classes can be seen as refinements of results in these works. Our refinements include: translating these results outside of their original context of hypothesis classes coming from logic to general hypothesis classes, and presenting sample complexity bounds. In the agnostic online learning setting we generalize results proven in prior work only for concept classes to the setting of real-valued hypothesis classes.

Model-theoretic characterizations of which partitioned formulas are learnable in other learning models (e.g. Private PAC learning) are provided in Alon et al. (2019).

7. 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 a theory of randomized families of classes, stemming from work in model theory: there we move from a random hypothesis class to its expectation class. We have proven that these transformations preserve agnostic PAC learnability and agnostic online learnability, refining results from both learning of database queries (Hu et al., 2022) and model theory (Ben Yaacov and 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.

Our motivation concerns distribution classes, but we obtain our positive results by embedding into a strictly more general setting, the expectation class of a measurable family. This setting 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 52.

Acknowledgments

The first author was partially supported by the UCLA Dissertation Year Fellowship, and NSF grants DMS-1651321, DMS-2246598, and DMS-2054379.

Appendix A. Agnostic Online Learnabilty and Duality: Proof of Proposition 3

Recall Proposition 3:

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. We can use the characterization in terms of stability of a hypothesis class in Theorem 45: The theorem shows that a hypothesis class $\mathcal{H}$ on X parameterized by Θ is online learnable if and only if:

For all $\gamma > 0$, there is some d such that there are no $a_1, \ldots, a_d \in X$, $b_1, \ldots, b_d \in \Theta$, such that for all $i < j$,

$$|h_{b_j}(a_i) - h_{b_i}(a_j)| \geq \gamma.$$

Clearly, if the roles of range and parameter space are swapped, the same thing holds.

Appendix B. Justifying and Generalizing Example 2

Recall that in the body of the paper we presented Example 2, an example of a hypothesis class of functions. We claimed that this was agnostic PAC learnable. This can be argued directly via computation of fat-shattering dimensions. Here we give a more general argument, deriving from logic.

Consider the real numbers with operations addition, multiplication, and the inequality: this is a real-closed order field. We recall what it means for a set of tuples in the reals to be *first-order definable* over this vocabulary: The *terms* of first-order logic are built up from real constants and variables by applying the functions addition and multiplication: that is, they are polynomials with real coefficients. The *atomic formulas* are inequalities between terms. The *formulas* of first-order logic are built up from atomic formulas via the Boolean operations $\wedge, \vee, \neg$ and the quantifiers $\exists x$, $\forall x$.

A *partitioned formula* $\phi(x_1 \ldots x_j; w_1 \ldots w_\ell)$ is a formula where the free variables are partitioned into two subsets, $\vec{x}$ and $\vec{y}$. Such a formula is associated with a family of subsets of $\mathcal{R}^j$, indexed by $\mathcal{R}^\ell$, in the obvious way. We can similarly talk about a formula with a partition of the free variables into three parts.

A 3-partitioned formula of the form $\phi(x_1 \ldots x_j; y_1 \ldots y_k; w_1 \ldots w_\ell)$ is a *definable family of functions* if for every $\vec{w}$ and $\vec{x}$, there is exactly one $\vec{y}$ that makes the formula true. Such a formula is associated with a family of functions from $\mathcal{R}^j$ to $\mathcal{R}^k$, indexed by $\mathcal{R}^\ell$.

Note that the class of Example 2 is of this form.

The following proposition follows from the fact that the real ordered field is an "NIP structure":

Proposition 63 *Every definable family of functions over the real ordered field has finite γ fat-shattering dimension for each $\gamma > 0$, and is thus agnostic PAC learnable.*

Although this is also probably well known, we sketch a proof:

Proof To use the terminology within model theory, the real-ordered field is o-minimal, and for every o-minimal structure over the reals, it is shown in van den Dries (1998) that every definable family family of subsets in an o-minimal structure has finite VC dimension. Thus

if we have $\phi(\vec{x};\vec{w})$, with ϕ a formula over the real field, the hypothesis class consisting of functions $h_{\vec{w}}$, the characteristic function $\phi(\vec{x};\vec{w})$ has finite γ fat-shattering dimension for each γ. Thus the same holds for finite sums (with fixed number of summands) of such characteristic functions. A definable family of functions, like the family in Example 2, can be approximated within any tolerance in the sup norm, uniformly in the parameters $\vec{w}$ by finite sums of characteristic functions. Thus the proposition holds. ■

Appendix C. Measurability Issues

In the body of this paper, there are two important steps where our arguments ignored some issues related to measurability assumptions, and two points at which we cite other work without clarifying the measurability assumptions. We spell out these steps and the assumptions here.

Let us first discuss the agnostic PAC case. We have given two arguments for preservation of agnostic PAC learnability to statistical classes in the body of the paper. Both of them proceed by proving bounds on certain combinatorial dimensions, like γ fat-shattering. Our story was that this implies finite GC-dimension, which in turn implies agnostic PAC learnability, a standard chain of deductions. However, the argument from combinatorial bounds to GC-dimension bounds in prior work requires some measurability assumptions, which are often not spelled out in detail in the literature: see, e.g. Lothar Sebastian Krapp and Laura Wirth (2025) for an overview of these issues.

We will not provide exact measurability conditions. But we will show that *countability of the parameter space is sufficient for our sample complexity arguments about the distribution class, and similarly countability of the range space for the arguments about the sample complexity of the dual distribution class.*

Recall that a key point of the argument is to bound the GC dimensions of the expectation class of a measurable family. We will argue that:

- for this argument it suffices that the parameter space is *separable*: informally, this means we can approximate the behaviour by looking at a countable subset of the parameter space.
- when we revisit the embedding of the distribution class as a measurable family from Section 3.2, it produces a separable measurable family as long as the parameter space of the base class is countable.

Similarly, in agnostic online case, our proof of Theorem 35 requires additional assumptions, including those required by results cited from Rakhlin et al. (2015b). We will see that these arguments also work in the context of a separable parameter space.

C.1 The Necessary Measurability Assumptions

Recall Lemma 21:

Let (Ω, Σ, μ) be a probability space, and let $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ be a measurable family of hypothesis classes on X.

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

$$w(\mathbb{E}\mathcal{F}(\bar{x}, \Theta), \beta) \leq \mathbb{E}_\mu[w(\mathcal{H}_\omega(\bar{x}, \Theta), \beta)].$$

The first issue is that for $\mathbb{E}_\mu[w(\mathcal{H}_\omega(\bar{x}, \Theta), \vec{b})]$ in the statement of Lemma 21 to be well-defined, we need the measurable family $\mathcal{F}$ to satisfy a stronger measurability assumption:

Definition 64 *Assume that (Ω, Σ, μ) is a probability space. Say that a measurable family $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ of hypothesis classes on X parameterized by Θ is* strongly measurable *when for each $\bar{x} \in X^n$ and $\vec{b} \in \mathcal{R}^n$, the function $\omega \mapsto w(\mathcal{H}_\omega(\bar{x}, \Theta), \vec{b})$ is measurable.*

Another place measurability is required is defining Glivenko-Cantelli dimensions. The Glivenko-Cantelli dimensions of a hypothesis class $\mathcal{H}$ are defined in terms of the measures of the sets

$$\left\{(x_1, ..., x_m) \mid \exists h \in \mathcal{H}, \left|\frac{1}{m} \cdot (\Sigma_{i=1}^m h(x_i)) - \int h(u) dD(u)\right| > \epsilon\right\}$$

for $\varepsilon > 0, m$, and a distribution D, which we could rewrite as the union

$$\bigcup_{h \in \mathcal{H}} \left\{(x_1, ..., x_m) \mid \left|\frac{1}{m} \cdot (\Sigma_{i=1}^m h(x_i)) - \int h(u) dD(u)\right| > \epsilon\right\},$$

which in general need not be measurable, although they are if for each m, the following function is measurable:

$$(x_1, \ldots, x_m) \mapsto \sup_{h \in \mathcal{H}} \left|\frac{1}{m} \cdot (\Sigma_{i=1}^m h(x_i)) - \int h(u) dD(u)\right|.$$

In summary, when we refer to the Glivenko-Cantelli dimensions of $\mathcal{H}$, we implicitly assume that $\mathcal{H}$ has the following property:

Definition 65 (well-defined GC dimensions) *Say a hypothesis class $\mathcal{H}$ on X* has well-defined Glivenko-Cantelli dimensions *if for all $\varepsilon > 0$, all m, and all distributions D on measure space on X induced by $\mathcal{H}$, the set*

$$\left\{(x_1, ..., x_m) \mid \exists h \in \mathcal{H}, \left|\frac{1}{m} \cdot (\Sigma_{i=1}^m h(x_i)) - \int h(u) dD(u)\right| > \epsilon\right\}$$

is measurable with respect to the measure space on X induced by $\mathcal{H}$.

If $\mathcal{H}$ has well-defined Glivenko-Cantelli dimensions, then the bounds in Fact 10 on the learnability of $\mathcal{H}$ hold.

The two results we cited without measurability assumptions are Facts 11 and 19:

Fact 11 is used in the proof of Theorem 16 to deduce bounds on Glivenko-Cantelli dimensions from Rademacher mean width. The result this rephrases, Theorem 4.10 (Wainwright, 2019), assumes measurability in the form of well-defined Glivenko-Cantelli dimensions. It is observed at the beginning of Section 4.4 (Wainwright, 2019) that this measurability assumption holds in the case of a parameter space which is either countable or, in an appropriate metric, separable.

Fact 19 is used in the proof of Theorem 35 to deduce online learnability from sequential Rademacher mean width. It is cited from Rakhlin et al. (2015b), where X and Θ are assumed to be separable metric spaces, with $\mathcal{H} : X \times \Theta \to [0,1]$ continuous (or at least, lower-semicontinuous). In the case where X and Θ are countable, this can be satisfied trivially by giving each the discrete metric ($d(a,b) = 1$ when $a \neq b$). We will show that this assumption is also satisfied when X is countable and Θ is a space of random variables on a countable set.

It is worth noting that Theorem 30 and Corollary 46 provide alternate proofs that agnostic PAC and online learning respectively are preserved under passing to the distribution class. The proofs of both of these theorems simply show that finite (sequential) fat-shattering dimension is preserved, from which one can deduce that learnability is as well. However, this latter deduction is once again subject to measurability constraints. In the PAC case, well-defined Glivenko-Cantelli dimensionssuffice, and in the online case, separability suffices.

C.2 Countable Parameter Spaces

We now check that if Θ is countable, then these measurability assumptions are satisfied.

Lemma 66 *If $\mathcal{F}$ is a measurable family of hypothesis classes on X parameterized by Θ, and Θ is countable, then $\mathcal{F}$ is strongly measurable.*

Proof Assume Θ is countable, and fix $\bar{x} = (x_1, \ldots, x_n) \in X^n$, $\vec{b} = (b_1, \ldots, b_n) \in \mathcal{R}^n$. For any $\omega \in \Omega$, we can view $\mathcal{H}_\omega$ as a function $\mathcal{H}_\omega : X \times \Theta \to [0,1]$. Then expanding definitionally,

$$w(\mathcal{H}_\omega(\bar{x}, \Theta), \vec{b}) = \sup_{p \in \Theta} \sum_{i=1}^{n} \mathcal{H}_\omega(x_i, p) b_i.$$

The assumption that $\mathcal{F}$ is measurable means that for each $x \in X, p \in \Theta$, the function $\omega \mapsto \mathcal{H}_\omega(x,p)$ is measurable, from which we see that for each $p \in \Theta$, the function $\omega \mapsto \sum_{i=1}^{n} \mathcal{H}_\omega(x_i, p) b_i$ is measurable. A countable supremum of measurable functions is measurable, so $\omega \mapsto w(\mathcal{H}_\omega(\bar{x}, \Theta), \vec{b})$ is measurable, making $\mathcal{F}$ a strongly measurable class. ■

It is also clear that if $\mathcal{H}$ is a class parameterized by a countable set Θ, the Glivenko-Cantelli dimensions are well-defined, because the sets in question are countable unions.

We can also conclude that if $\mathcal{F}$ is a measurable family of hypothesis classes parameterized by countable Θ, then Theorem 16 holds. Because a countable set is also separable in any metric, the measurability requirements for Fact 19 are satisfied, so Theorem 35 holds.

C.3 From Countable Parameter Spaces to the Parameters of Random Variable Classes

When we analyze a class $\mathcal{H}$ under the assumption that its parameter set Θ is countable, we eventually have to consider the class $\mathcal{H}(\text{CompatParamRV}(\mathcal{H}))$, parameterized by

CompatParamRV($\mathcal{H}$), which need not be countable. However, we will show that it is separable in a suitable metric, and show that this suffices for the rest of our measurability requirements.

Lemma 67 *If Θ is countable, then the set* CompatParamRV($\mathcal{H}$) *is separable with the metric where*

$$d(\Theta_1, \Theta_2) = \sum_{p \in \Theta} |\mathbb{P}[\Theta_1 = p] - \mathbb{P}[\Theta_2 = p]|.$$

Proof Let A be the set of all $P \in \text{CompatParamRV}(\mathcal{H})$ such that for all $p \in \Theta$, $\mathbb{P}[P = p] \in \mathbb{Q}$, for all but finitely many $p \in \Theta$, $\mathbb{P}[P = p] = 0$. This set is countable, and to show that it is dense, we will show that for all $P \in \text{CompatParamRV}(\mathcal{H})$, and $\varepsilon > 0$, there is some $P_\varepsilon \in A$ such that $d(P, P_\varepsilon) \leq \varepsilon$.

As $\sum_{p \in \Theta} \mathbb{P}[P = p] = 1$, there is some finite set $S \subseteq \Theta$ with $\sum_{p \in S} \mathbb{P}[P = p] \geq 1 - \frac{\varepsilon}{3}$. Then by density of the rational numbers, we can find nonnegative rational numbers $(r_p : p \in S)$ such that $\sum_{p \in S} r_p = 1$, and $\sum_{p \in S} |\mathbb{P}[P = p] - r_p| \leq \frac{2\varepsilon}{3}$. We then define P_ε so that if $p \in S$, then $\mathbb{P}[P_\varepsilon = p] = r_p$, and otherwise, $\mathbb{P}[P_\varepsilon = p] = 0$.

Then $d(P, P_\varepsilon) = \sum_{p \in S} \sum_{p \in S} |\mathbb{P}[P = p] - r_p| + \sum_{p \in \Theta \setminus S} |\mathbb{P}[P = p] - 0| \leq \frac{2\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$. ■

We now check that separability in this particular metric suffices to imply well-defined Glivenko-Cantelli dimension, justifying the application of Fact 11 in the proof of Theorem 16.

Theorem 68 *If Θ is countable, then for any $\delta, \varepsilon > 0$, $GC_{\varepsilon,\delta}(\mathcal{H}(\text{CompatParamRV}(\mathcal{H})))$ is well-defined.*

Proof It suffices to show that for all m, the function

$$(x_1, \ldots, x_m) \mapsto \sup_{\Theta' \in \text{CompatParamRV}(\mathcal{H})} \left| \frac{1}{m} \cdot (\Sigma_{i=1}^m \mathbb{E}_\mu[h_{\Theta'}(x_i)]) - \int \mathbb{E}_\mu[h_{\Theta'}(u)] dD(u) \right|.$$

is measurable. To do this, let A be a countable dense set of CompatParamRV($\mathcal{H}$) in the metric from Lemma 67, and we will show that the functions f_P defined by

$$f_P(x_1, \ldots, x_m) = \left| \frac{1}{m} \cdot (\Sigma_{i=1}^m \mathbb{E}_\mu[h_{\Theta'}(x_i)]) - \int \mathbb{E}_\mu[h_{\Theta'}(u)] dD(u) \right|$$

are themselves measurable, and that $\{f_{\Theta'} : \Theta' \in A\}$ is dense in $\{f_{\Theta'} : \Theta' \in \text{CompatParamRV}(\mathcal{H})\}$, in the sup-metric, so any $f_{\Theta'}$ is a uniform limit of elements of $\{f_{\Theta'} : \Theta' \in A\}$, from which we conclude that

$$\sup_{\Theta' \in \text{CompatParamRV}(\mathcal{H})} f_{\Theta'} = \sup_{\Theta' \in A} f_{\Theta'},$$

and the latter countable supremum is measurable.

First, let $P \in \text{CompatParamRV}(\mathcal{H})$. The function f_P is measurable, because it is the absolute value of a sum of measurable functions, as $x \mapsto \mathbb{E}_\mu[h_{\Theta'}(x)]$ is the uniform limit of linear combinations of measurable functions $h_p(x)$ for $p \in \Theta'$.

Then we show that $\{f_{\Theta'} : \Theta' \in A\}$ is dense in $\{f_{\Theta'} : \Theta' \in \text{CompatParamRV}(\mathcal{H})\}$, in the sup-metric, by showing that if $\Theta_1, \Theta_2 \in \text{CompatParamRV}(\mathcal{H})$, then $\sup_{\bar{x}} |f_{\Theta_1}(\bar{x}) - f_{\Theta_2}(\bar{x})| \leq 2d(\Theta_1, \Theta_2)$.

We see that for every $x \in X$,

$$\begin{aligned}\mathbb{E}_\mu[|h_{\Theta_1}(x) - h_{\Theta_2}(x)|] &\leq \sum_{p \in \Theta} h_p(x) \, |\mathbb{P}[\Theta_1 = p] - \mathbb{P}[\Theta_1 = p]| \\ &\leq \sum_{p \in \Theta} |\mathbb{P}[\Theta_1 = p] - \mathbb{P}[\Theta_1 = p]| \\ &= d(\Theta_1, \Theta_2),\end{aligned}$$

so for every $\bar{x} \in X^m$,

$$\begin{aligned}&f_{\Theta_1}(\bar{x}) - f_{\Theta_2}(\bar{x}) \\ &= \left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\mu[h_{\Theta_1}(x_i)] - \int \mathbb{E}_\mu[h_{\Theta_1}(u)] dD(u) \right| - \left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\mu[h_{\Theta_2}(x_i)] - \int \mathbb{E}_\mu[h_{\Theta_2}(u)] dD(u) \right| \\ &\leq \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\mu \left[|h_{\Theta_1}(x_i) - h_{\Theta_2}(x_i)| \right] + \int \mathbb{E}_\mu \left[|h_{\Theta_1}(u) - h_{\Theta_2}(u)| \right] dD(u) \\ &\leq 2d(\Theta_1, \Theta_2).\end{aligned}$$

■

Furthermore, to show that the assumptions of Fact 19 are satisfied, we check the following:

Lemma 69 *Suppose $\mathcal{H}$ is a hypothesis class with range space X and parameter space Θ, which are both countable. Then $\mathcal{H}(\text{CompatParamRV}(\mathcal{H})) : X \times \text{CompatParamRV}(\mathcal{H}) \to [0, 1]$ is continuous, where X has the discrete metric and $\text{CompatParamRV}(\mathcal{H})$ has the above separable metric.*

Proof Because X has the discrete metric, it suffices to show that for each $x \in X$, the function

$$f_x : P \mapsto (\mathcal{H}(\text{CompatParamRV}(\mathcal{H})))_P(x)$$

is continuous.

In fact, we will see that it is 1-Lipschitz. Suppose $P_1, P_2 \in \text{CompatParamRV}(\mathcal{H})$. Then

$$\begin{aligned}
|f_x(P_1) - f_x(P_2)| &= \left| \sum_{p \in \Theta} \mathbb{P}[P_1 = p] h_p(x) - \sum_{p \in \Theta} \mathbb{P}[P_1 = p] h_p(x) \right| \\
&= \left| \sum_{p \in \Theta} \left(\mathbb{P}[P_1 = p] - \mathbb{P}[P_1 = p] \right) h_p(x) \right| \\
&\leq \sum_{p \in \Theta} \left| \mathbb{P}[P_1 = p] - \mathbb{P}[P_1 = p] \right| h_p(x) \\
&\leq \sum_{p \in \Theta} \left| \mathbb{P}[P_1 = p] - \mathbb{P}[P_1 = p] \right| \\
&= d(P_1, P_2).
\end{aligned}$$

■

C.4 Without Countability Assumptions

If we place no countability or separability assumptions on Θ, then we can still prove results which transfer from a base class to a statistical class, but which only deal with combinatorial dimensions, rather than learnability. Here would be one transfer result for the agnostic PAC case:

Theorem 70 *If $\mathcal{F}$ is a measurable family of hypothesis classes with uniformly bounded γ (sequential) fat-shattering dimension for all $\gamma > 0$, then $\mathbb{E}\mathcal{F}$ has finite γ (sequential) fat-shattering dimension for all $\gamma > 0$.*

Proof We prove this first for fat-shattering dimension, and then consider the sequential version.

If the class $\mathbb{E}\mathcal{F}$ has infinite γ fat-shattering dimension for some $\gamma > 0$, this is witnessed by a countable subset $\Theta' \subseteq \Theta$. If we restrict the parameters of every class in $\mathcal{F}$ to Θ', the resulting measurable family $\mathcal{F}'$ is strongly measurable. Thus $\mathbb{E}\mathcal{F}'$ having infinite γ fat-shattering dimension implies, by Theorems 23 and and the connection between finite fat-shattering dimensions and Rademacher mean width, that there must be some $\delta > 0$ for which the classes $\mathcal{H} \in \mathcal{F}'$ have unbounded fat-shattering dimension, and the same applies to $\mathcal{F}$.

For the sequential version, the argument is the same, but citing Theorem 38. ■

However, Theorems 16 and 35 are still subject to the measurability assumptions of Facts 11 and 19 respectively.

Appendix D. More Detail on Fact 23

Recall that in the body we stated Fact 23, about passing from a uniformly stable family to stability of its expectation class. We cited Ben Yaacov and Keisler (2009); Ben Yaacov

(2013) for this, and mention that the language used in these works is quite different. We now explain how to translate from the setting of these works to Fact 23.

Our statement is that if a measurable family $\mathcal{F}$ of hypothesis classes on X parameterized by Θ is uniformly stable, then the expectation class $\mathbb{E}\mathcal{F}$ is also stable. The statements in Ben Yaacov and Keisler (2009); Ben Yaacov (2013) show that stability, as a property of a formula in a metric structure of continuous logic, is preserved under a construction known as "randomization". For the definitions of continuous logic and metric structure, see Ben Yaacov (2013); Ben Yaacov and Keisler (2009) and references therein.

Given a measurable family $\mathcal{F}$ as above, we will construct such a metric structure, and check that stability of $\mathbb{E}\mathcal{F}$ follows from stability of a corresponding formula in the randomization structure.

In order to state this properly, we need to explain what it means for a formula to be stable in a metric structure or theory of continuous logic. We define this in terms of hypothesis classes to match the rest of this paper, but the resulting definition matches Definition 7.1 (Ben Yaacov and Usvyatsov, 2010).

Definition 71 *In any metric structure $\mathcal{M}$, given a formula $\phi(x;p)$ of continuous logic, let $\mathcal{H}_{\mathcal{M},\phi}$ be the class on the x-sort of $\mathcal{M}$ parameterized by the p-sort of $\mathcal{M}$ and given by $(\mathcal{H}_{\mathcal{M},\phi})_p(x) = \phi(x,p)$.*

We say that $\phi(x;p)$ is stable *in $\mathcal{M}$ when $\mathcal{H}_{\mathcal{M},\phi}$ is. We say that $\phi(x;p)$ is* stable *in a theory T when it is stable in every model $\mathcal{M} \models T$.*

We begin by recalling some definitions and results related to randomization from Ben Yaacov (2013). The special case of concept classes is worked out in Ben Yaacov and Keisler (2009).

Given a theory T of continuous logic in the language $\mathcal{L}$, there is a language $\mathcal{L}_R$, known as the *randomization* of $\mathcal{L}$, and an $\mathcal{L}_R$-theory T_R of continuous logic, known as the *randomization* of T, such that for every $\mathcal{L}$-formula $\phi(\bar{x})$ in the language of T, there is a corresponding $\mathcal{L}_R$-formula $\mathbb{E}[\phi(\bar{x})]$ in the language of T_R.

We defer to Ben Yaacov (2013) for the exact construction of $\mathcal{L}_R$ and T_R. Below we summarize their most relevant properties:

Definition 72 (See Ben Yaacov 2013, Definitions 3.4, 3.23) *Let (Ω, Σ, μ) be a probability space, and let $(\mathcal{M}_\omega : \omega \in \Omega)$ be a family of metric structures in a relational language of continuous logic. We call $(\mathcal{M}_\omega : \omega \in \Omega)$ a* random family of structures *when for all relation symbols $R(\bar{x})$ in this language, and tuples $\bar{a}$ of random variables $a : \omega \to \bigcup_\omega \mathcal{M}_\omega$ where $a(\omega) \in \mathcal{M}_\omega$, the function $\omega \mapsto R(\bar{a}(\omega))$ is measurable.*

Fact 27 (Ben Yaacov 2013, Corollary 3.24) *Let T be a theory of continuous logic, and let $(\mathcal{M}_\omega : \omega \in \Omega)$ be a random family of models of T. Then there is a metric $\mathcal{L}_R$-structure $\mathcal{R} \models T_R$, consisting of the random variables $a : \omega \to \bigcup_\omega \mathcal{M}_\omega$ where $a(\omega) \in \mathcal{M}_\omega$, where for every formula $\phi(\bar{x})$,*

$$(\mathbb{E}[\phi(\bar{a})])^{\mathcal{R}} = \mathbb{E}_\mu \left[\phi(\bar{a}(\omega)))^{\mathcal{M}_\omega} \right].$$

Fact 28 (Ben Yaacov 2013, Theorem 4.9) *If $\phi(\bar{x})$ is stable in T, then $\mathbb{E}[\phi(\bar{x})]$ is stable in T_R.*

Recall that $\mathcal{F} = (\mathcal{H}_\omega : \omega \in \Omega)$ consists of hypothesis classes $\mathcal{H}_\omega$ on X parameterized by Θ, where (Ω, Σ, μ) is an atomless probability space with respect to which each function $\omega \mapsto \mathcal{H}_\omega(x, p)$ with $x \in X, p \in \Theta$ is measurable.

We now construct a measurable family of structures based on $\mathcal{H}_\omega$. Given $\omega \in \Omega$, let $\mathcal{M}_\omega$ be a metric structure with two sorts: X, Θ. To each sort we assign the discrete metric ($d(x, y) = 1$ when $x \neq y$), so all functions on these sorts will be uniformly continuous. Then we make this a structure in a language with one relation symbol, $\phi(x, p)$, where x is a variable of sort X and p is a variable of sort Θ. We interpret this symbol in $\mathcal{M}_\omega$ by $\phi(x, p) = (\mathcal{H}_\omega)_p(x)$.

Now let T be the shared theory of all the structures $\mathcal{M}_\omega$.

Lemma 73 *The partitioned formula $\phi(x; p)$ is stable in the theory T.*

Proof This means that for every $\mathcal{M} \vDash T$, the class $\mathcal{H}_{\mathcal{M},\phi}$ on the X-sort of $\mathcal{M}$ parameterized by the Θ-sort of $\mathcal{M}$ and given by $(\mathcal{H}_{\mathcal{M},\phi})_p(x) = \phi(x, p)$ is stable.

Note that the class $\mathcal{H}_{\mathcal{M}_\omega,\phi}$ is just $\mathcal{H}_\omega$.

Because T is the common theory of the structures $\mathcal{M}_\omega$, and for each $r < s$, the (r, s)-threshold dimensions of the classes $\mathcal{H}_{\mathcal{M}_\omega,\phi}$ are uniformly bounded, the same bound is implied by some condition proven by the theory T, so this same bound also holds for any other $\mathcal{M} \vDash T$, so $\phi(x; p)$ is stable in T. ■

This implies also that $\mathbb{E}[\phi(x, p)]$ is stable in T_R. This allows us to construct stable hypothesis classes from models of T_R. We see that there is a model $\mathcal{R} \vDash T_R$ consisting of random variables where for each pair a, b of random variables in the X-sort and Θ-sort respectively,

$$(\mathbb{E}[\phi(a, b)])^{\mathcal{R}} = \mathbb{E}_\mu \left[\phi(a(\omega), b(\omega)))^{\mathcal{M}_\omega}\right].$$

We now claim that the hypothesis class $\mathbb{E}\mathcal{F}$ embeds into the class defined by $\mathbb{E}[\phi(x, p)]$ in this structure, and is thus stable. To see this, note that for every $x \in X$, $p \in \Theta$, the constant random variables a, b taking values x, p respectively are elements of the appropriate sorts of $\mathcal{R}$, by the construction in Ben Yaacov (2013). Thus for these elements of $\mathcal{R}$,

$$(\mathbb{E}[\phi(a, b)])^{\mathcal{R}} = \mathbb{E}_\mu \left[\phi(x, p)^{\mathcal{M}_\omega}\right] = \mathbb{E}_\mu \left[(\mathcal{H}_\omega)_p(x)\right] = \mathbb{E}\mathcal{F}_p(x).$$

References

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

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 Velegkas. 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.

Itaï Ben Yaacov. Continuous and random Vapnik-Chervonenkis classes. *Israel Journal of Mathematics*, 173(1):309–333, 2009.

Itaï Ben Yaacov. On theories of random variables. *Israel Journal of Mathematics*, 194: 957–1012, 2013.

Itaï Ben Yaacov and H. Jerome Keisler. Randomizations of models as metric structures. *Confluentes Mathematici*, 1(2):197–223, 2009.

Itaï Ben Yaacov and Alexander Usvyatsov. Continuous first order logic and local stability. *Transactions of the American Mathematical Society*, 362(10):5213–5259, 2010.

Hunter Chase and James Freitag. Model theory and machine learning. *The Bulletin of Symbolic Logic*, 25(3):319–332, 2019.

David Heaver Fremlin. *Measure Theory*. Torres Fremlin, 2002. URL `http://www.essex.ac.uk/maths/staff/fremlin/int.htm`.

Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Near-tight closure bounds for the Littlestone and threshold dimensions. In *Algorithmic learning theory*, pages 686–696. PMLR, 2021.

Xiao Hu, Yuxi Liu, Haibo Xiu, Pankaj K. Agarwal, Debmalya Panigrahi, Sudeepa Roy, and Jun Yang. Selectivity functions of range queries are learnable. In *SIGMOD*, 2022.

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 Kleer 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.

Lothar Sebastian Krapp and Laura Wirth. Measurability in the Fundamental Theorem of Statistical Learning, 2025. https://arxiv.org/abs/2410.10243.

Daniel Raban. The Glivenko-Cantelli Theorem and Introduction to VC Dimension, 2023. URL `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.

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.