A Theory of Interpretable Approximations

Marco Bressan
Università degli Studi di Milano, Italy
marco.bressan@unimi.it

Nicolò Cesa-Bianchi
Università degli Studi di Milano, Italy
Politecnico di Milano, Italy
nicolo.cesa-bianchi@unimi.it

Emmanuel Esposito
Università degli Studi di Milano, Italy
Istituto Italiano di Tecnologia, Italy
emmanuel@emmanuelesposito.it

Yishay Mansour
Tel Aviv University, Israel
Google Research
mansour.yishay@gmail.com

Shay Moran
Technion, Israel
Google Research
smoran@technion.ac.il

Maximilian Thiessen
TU Wien, Austria
maximilian.thiessen@tuwien.ac.at

Abstract

Can a deep neural network be approximated by a small decision tree based on simple features? This question and its variants are behind the growing demand for machine learning models that are *interpretable* by humans. In this work we study such questions by introducing *interpretable approximations*, a notion that captures the idea of approximating a target concept c by a small aggregation of concepts from some base class $\mathcal{H}$. In particular, we consider the approximation of a binary concept c by decision trees based on a simple class $\mathcal{H}$ (e.g., of bounded VC dimension), and use the tree depth as a measure of complexity. Our primary contribution is the following remarkable trichotomy. For any given pair of $\mathcal{H}$ and c, exactly one of these cases holds: (i) c cannot be approximated by $\mathcal{H}$ with arbitrary accuracy; (ii) c can be approximated by $\mathcal{H}$ with arbitrary accuracy, but there exists no universal rate that bounds the complexity of the approximations as a function of the accuracy; or (iii) there exists a constant κ that depends only on $\mathcal{H}$ and c such that, for *any* data distribution and *any* desired accuracy level, c can be approximated by $\mathcal{H}$ with a complexity not exceeding κ. This taxonomy stands in stark contrast to the landscape of supervised classification, which offers a complex array of distribution-free and universally learnable scenarios. We show that, in the case of interpretable approximations, even a slightly nontrivial a-priori guarantee on the complexity of approximations implies approximations with constant (distribution-free and accuracy-free) complexity. We extend our trichotomy to classes $\mathcal{H}$ of unbounded VC dimension and give characterizations of interpretability based on the algebra generated by $\mathcal{H}$.

1 Introduction

Many machine learning techniques, such as deep neural networks, produce large and complex models whose inner workings are difficult to grasp. In sectors such as healthcare and law enforcement, where the stakes of automated decisions are high, this is a serious problem: complex models make it hard to explain the rationale behind an outcome, or why two similar inputs produce different outcomes. In

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024. This work is published at the Conference on Learning Theory (COLT 2024).

those cases, *interpretable* models may become the preferred choice. Although there is an ongoing debate around the notion of interpretability [Erasmus, Brunet, and Fisher, 2021], decision trees are typically considered as the quintessential example of interpretable models [Molnar, 2022]: ones that favor a transparent decision-making process, and that allow users to understand how individual features influence predictions. A line of research in this area studies the extent to which small decision trees can approximate some specific learning models, such as neural networks [Craven and Shavlik, 1995] and k-means classifiers [Dasgupta, Frost, Moshkovitz, and Rashtchian, 2020]. Inspired by these results, we develop a general theory of interpretability viewed as approximability via simple decision trees. Our guiding principle can be summarized as follows.

Interpretable approximations = Small aggregations of simple hypotheses.

In analogy with PAC learning, we focus on binary classification tasks and view a classifier (e.g., a neural network) as a concept $c \subseteq X$, where X is the data domain. Now let $\mathcal{H} \subseteq 2^X$ be a family of simple hypotheses, for instance decision stumps or halfspaces. Our goal is to understand how well c can be approximated by aggregating a small set of elements in $\mathcal{H}$. To formalize this goal in the language of decision trees we introduce two notions. First, we say that c is *approximable* by $\mathcal{H}$ if, under any given data distribution, there exists a finite decision tree using splitting functions from $\mathcal{H}$ that approximates c arbitrarily well. Moreover, if the approximation can be always achieved using a shallow tree, we say that c is *interpretable* by $\mathcal{H}$. It is easy to see that, depending on c and $\mathcal{H}$, one may have interpretability, approximability but not interpretability, or even non-approximability. In Section 4, we give explicit examples of pairs $(c, \mathcal{H})$ for each one of the three above cases.

Note that in this initial study on the general structure of interpretable approximations we focus on the fundamental question of what conditions ensure the existence of accurate approximations and interpretations. Important topics, such as the informational or computational complexity of obtaining accurate interpretations, are not addressed in this work. Note also that we do not make any specific assumption on the data distribution P. Our approach is thus in line with standard notions and theories in machine learning—e.g., universal Bayes consistency [Devroye, Györfi, and Lugosi, 2013], PAC learnability [Shalev-Shwartz and Ben-David, 2014], and universal learnability [Bousquet, Hanneke, Moran, van Handel, and Yehudayoff, 2021]—as it encompasses both distribution-free and distribution-dependent guarantees.

While our primary focus is not algorithmic, our work reveals profound connections within the algorithmic framework of boosting. Indeed, there is a clear relationship between boosting, which involves the aggregation of weak hypotheses to learn a target concept, and interpretable approximation, which concerns the aggregation of simple hypotheses to approximate a target concept. However, our work uncovers and exploits deeper links at a technical level. In particular, our general construction that gives decision trees whose depth depends logarithmically on the accuracy is based on boosting decision trees, and its analysis uses potential functions from this line of work. Our improved bound for the VC classes, which provides approximating decision trees with constant (accuracy- and distribution-free) depth, is somewhat more subtle; it is also based on a boosting perspective, this time using majority-vote based algorithms and the minimax theorem. However, to eliminate the dependency on the accuracy, we utilize tools from VC theory, particularly uniform convergence.

1.1 Contributions

Degrees of interpretability (Section 4). We introduce our learning-theoretic notions of approximability and interpretability. Informally speaking, we use the depth of the shallowest approximating tree to measure the extent to which a certain concept c is interpretable by a given class $\mathcal{H}$ (e.g., hyperplanes or single features). Approximability is our weakest notion, as we do not constrain the rate at which the rate of the shallowest approximating tree grows as a function of the desired accuracy. Our strongest notion is instead interpretability with a tree depth that is constant with respect to both accuracy and data distribution. In between these two extremes, a wide variety of behaviors is possible, as the tree depth may grow at different rates that may be uniform, or depend on the data distribution (similarly to the distinction between PAC learning and universal learning).

Collapse of the degrees (Section 5). We prove that the range of possible behaviors collapses dramatically, and only three cases are actually possible: c is uniformly interpretable by $\mathcal{H}$, c is approximable but not interpretable by $\mathcal{H}$, c is not approximable by $\mathcal{H}$. If the class $\mathcal{H}$ of splits has

bounded VC dimension, which conforms to our request that $\mathcal{H}$ be simple, we show that whenever c is interpretable (possibly with a distribution-dependent rate) then it is uniformly interpretable by $\mathcal{H}$ *at constant depth*. This means that, for every data distribution P and every accuracy $\varepsilon > 0$, there exists an $\mathcal{H}$-based decision tree that approximates c with accuracy ε and whose depth is bounded *by a constant* depending only on $c, \mathcal{H}$ (but not on P, ε). Thus, whenever c is interpretable at some arbitrary rate, it is in fact interpretable at a constant rate. We show a similar collapse for classes $\mathcal{H}$ of unbounded VC dimension: in this case, we show that interpretability collapses to uniform interpretability at logarithmic depth $\mathcal{O}\big(\log \frac{1}{\varepsilon}\big)$.

Algebraic characterizations (Section 6). We prove that the trichotomy described above can be characterized in terms of algebras and closures over $\mathcal{H}$. For example, we show that if $\mathcal{H}$ has bounded VC dimension, then c is interpretable at constant depth if and only if c is in the algebra generated by the *closure* of $\mathcal{H}$, i.e., the family of all the concepts that can be approximated arbitrarily well by single hypotheses of $\mathcal{H}$. We also present a simpler characterization when the domain X is countable.

Extension to other complexity measures (Appendix C). Finally, we exploit the equivalence between $\mathcal{H}$-based decision trees and Boolean formulae over $\mathcal{H}$ to show that the trichotomy above holds for a large class of complexity measures, including not only tree-depth but also, for example, circuit size. In particular, we show that for any complexity measure in our class, interpretability collapses to uniform interpretability at constant complexity rate for VC classes and at polynomial complexity rate for non-VC classes.

2 Related Work

According to Molnar [2022], there are different approaches to interpretability in learning. One important distinction is between local explanation, where we explain the prediction of the model on a single data point, and global interpretation, where we explain the model itself. In this work we focus on the latter. A common approach to global interpretation is to use simpler "interpretable" models (e.g., decision trees) to approximate more complex ones [Craven and Shavlik, 1995]. This is known as *post-hoc interpretability* [Molnar, 2022]. For example, Zhang, Yang, Ma, and Wu [2019] used decision trees to interpret convolutional neural networks. Formally, interpretability can be modeled as a property of a classifier. For example, Dziugaite, Ben-David, and Roy [2020] define a variant of empirical risk minimization (ERM), where each classifier in a given class $\mathcal{H}$ is either interpretable or not, and the task is to learn an interpretable one even though the target concept is not necessarily interpretable. We generalize this setup by assigning a complexity measure to each classifier, e.g., the depth for decision trees. This allows to trade-off the desired accuracy ε and the maximum depth of a decision tree one is willing to call interpretable. Learning-theoretic perspectives on interpretability are rare and typically not covered in standard books and surveys. One important line of work initiated by Dasgupta et al. [2020] deals with the problem of approximating a given k-means or k-median clustering with decision trees. From this perspective, our setup can be seen as a generalization from clusterings to arbitrary concepts. However, that line of work focuses on efficient algorithms to compute decision trees with k leaves and approximation guarantees in terms of the k-means or k-medians cost function, not in terms of classification error under a distribution as we do here. Bastani, Kim, and Bastani [2017] discuss a related problem setup where a given classifier is approximated using a decision tree. Under strong assumptions, the authors state convergence results for the proposed decision tree. However, they do not state bounds on the required depth which is assumed to be given as a hyperparameter. Some algorithmic analyses exist for specific cases of hypothesis spaces and standard explainers. For example, Garreau and Luxburg [2020] analyse LIME [Ribeiro, Singh, and Guestrin, 2016], one of the most used explanation techniques. Li, Nagarajan, Plumb, and Talwalkar [2021] discuss generalization bounds for local explainers. Blanc, Lange, and Tan [2021] introduce a local variant of our setup with the goal of explaining the classification $f(x)$ of a single instance x using a conjunction with small size (i.e., a small decision list). Their results cannot be used for our goal of global interpretation as one would have to take the union of all the local conjunctions for all (potentially infinite) instances x. Closer to our setup, Moshkovitz, Yang, and Chaudhuri [2021] state bounds on the depth of a decision tree required to fit a linear classifier with margin. Similarly to us, they also strongly rely on boosting arguments. Vidal and Schiffer [2020] give upper bounds on the number of nodes of a single decision tree to approximate an ensemble of trees. While mainly focusing on local explainability, Blanc et al. [2021] also state bounds on the depth of a decision tree required to fit an arbitrary classifier $f\colon \{0,1\}^d \to \{0,1\}$ under the uniform

distribution on $\{0,1\}^d$. They do so by relying on classical bounds on the depth in terms of certificate complexity [Smyth, 2002, Tardos, 1989]. As we focus instead on general hypothesis classes and distributions, their results are not directly comparable to ours.

3 Preliminaries and Notation

Let X be any domain. We denote by P a distribution[1] on X and by $\mathcal{P}(X)$ the set of all distributions on X, by $\mathcal{H} \subseteq 2^X$ a hypothesis class on X, and by $\mathrm{VC}(X, \mathcal{H})$ its VC dimension. We denote by $\mathrm{Alg}(\mathcal{H})$ the algebra generated by $\mathcal{H}$, i.e., the smallest set system $\mathcal{A} \subseteq 2^X$ closed under complements and finite unions such that $\mathcal{H} \subseteq \mathcal{A}$ and $\emptyset, X \in \mathcal{A}$. The σ-algebra $\sigma(\mathcal{H})$ is the smallest algebra containing $\mathcal{H}$ that is closed under countable unions. We denote by $c \in 2^X$ an arbitrary concept (not necessarily in $\mathcal{H}$). As usual we also view c as a binary classification function $c\colon X \to \{0,1\}$. Our goal is to understand how well c can be approximated using aggregations of hypotheses in $\mathcal{H}$. We let $\mathbb{N}$ denote the naturals including 0, and $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$.

A decision tree over X is a full finite binary tree T with nodes $\mathcal{V}(T)$, where every leaf $z \in \mathcal{L}(T)$ holds a label $\ell_z \in \{0,1\}$ and every internal node $v \in \mathcal{V}(T) \setminus \mathcal{L}(T)$ holds a *decision stump* $f_v\colon X \to \{0,1\}$. The depth (or height) of T is denoted as $\mathrm{depth}(T)$. We say T is *$\mathcal{H}$-based* if $f_v \in \mathcal{H}$ for all $v \in \mathcal{V}(T)$, and we denote by $\mathcal{T}_{\mathcal{H}}$ the set of all $\mathcal{H}$-based decision trees. We also use T to denote the binary classifier $T\colon X \to \{0,1\}$ induced by T in the standard way. Note that $\mathcal{T}_{\mathcal{H}} \equiv \mathrm{Alg}(\mathcal{H})$, as any $\mathcal{H}$-based tree T can be rewritten as a Boolean formula and vice versa. For every $d \in \mathbb{N}_+$ we let $\mathrm{Alg}_d(\mathcal{H}) = \{T \in \mathrm{Alg}(\mathcal{H}) : \mathrm{depth}(T) \leq d\}$. Given $P \in \mathcal{P}(X)$ and a concept $c \in 2^X$, the *loss* of T with respect to c under P is $L_P(T, c) = \mathbb{P}_{x \sim P}(T(x) \neq c(x)) = P\big(T^{-1}(1) \triangle c\big)$, where $A \triangle B = (A \setminus B) \cup (B \setminus A)$ is the symmetric difference between A and B.

An ε-accurate $\mathcal{H}$-approximation of c under P is an $\mathcal{H}$-based decision tree T with $L_P(T, c) \leq \varepsilon$. The set of all such trees is denoted as $\mathcal{T}^c_{\mathcal{H}}(\varepsilon \mid P)$, which is also known as the ε-Rashomon set [Fisher, Rudin, and Dominici, 2019], and their minimal depth is

$$\mathrm{depth}^c_{\mathcal{H}}(\varepsilon \mid P) = \inf_{T \in \mathcal{T}^c_{\mathcal{H}}(\varepsilon|P)} \mathrm{depth}(T) \ . \tag{1}$$

4 Approximability and Interpretability

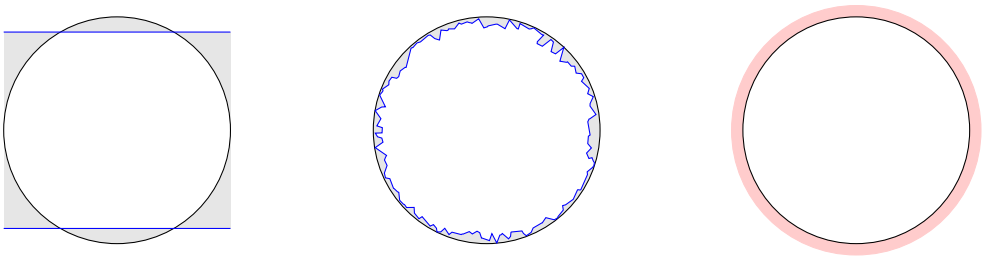
This section introduces the key definitions used in our results. We start with the definition of approximability.

Definition 1 (Approximability)**.** *A concept c is approximable by $\mathcal{H}$ if $\mathcal{T}^c_{\mathcal{H}}(\varepsilon \mid P) \neq \emptyset$ for every distribution $P \in \mathcal{P}(X)$ and every $\varepsilon > 0$.*

Approximability is our weakest notion, as it only requires that for any desired accuracy value a tree approximating c exists under any distribution, without any constraint on its depth. In fact, there may not even exist a function f such that $\mathrm{depth}^c_{\mathcal{H}}(\varepsilon \mid P)$ is bounded by $f(\varepsilon)$ for all distributions P.

For example, for $X = \mathbb{R}^d$ let c be the unit d-dimensional Euclidean ball centered at the origin and $\mathcal{H}$ be the family of affine halfspaces whose boundary is orthogonal to, say, the d-th dimension. Then, any finite aggregation T of such halfspaces is unable to discern points that are aligned along the i-th dimension for any $i \neq d$ and, thus, necessarily incurs a constant $L_P(T, c)$ for some distribution P. On the other hand, if we extend $\mathcal{H}$ to be the family of all halfspaces in $X = \mathbb{R}^d$, then it is possible to show that we can approximate the unit ball c up to any accuracy under any distribution. Indeed, it is known that a variant of the 1-nearest neighbour (1-NN) algorithm is universally strongly Bayes consistent in essentially separable metric spaces [Hanneke, Kontorovich, Sabato, and Weiss, 2021], and any 1-NN classifier corresponds to a finite Voronoi partition which can be represented as an $\mathcal{H}$-based decision tree. However, we expect the number of Voronoi cells, and thus the depth of the $\mathcal{H}$-based decision tree representing it, to grow larger as the distribution P concentrates around the decision boundary (that is, the surface of the unit ball c). Consider, for instance, the family of distributions P_α with

[1] By default we assume a fixed but otherwise arbitrary σ-algebra on X and that all functions/sets discussed in our theorems are measurable. We also borrow standard assumptions on the underlying σ-algebra which allow us to use the VC Theorem [Vapnik and Chervonenkis, 1971]. See, e.g., Blumer, Ehrenfeucht, Haussler, and Warmuth [1989].



(a) Non-approximable (b) Approximable but not interpretable (c) Uniformly interpretable

Figure 1: Approximating a disk with halfspaces: the approximation error is the grey-shaded area, while the pink area is the margin region. In (a), we show inapproximability with x-axis-aligned halfspaces. In (b), we show the disk is approximable (but not interpretable) with arbitrary halfspaces, via a Voronoi tessellation with one-sided error. In (c), we show the disk with margin is uniformly interpretable with halfspaces.

$\alpha \in (0,1)$, where each P_α has support corresponding to the spherical shell $B^d(1+\alpha) \setminus B^d(1-\alpha)$ with inner radius $1-\alpha$ and outer radius $1+\alpha$ (here we denote by $B^d(r) = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ the origin-centered Euclidean ball of radius $r > 0$ in $\mathbb{R}^d$). Then, we expect the number of Voronoi cells defining the decision boundary of the 1-NN classifiers that guarantee loss at most $0 < \varepsilon \leq 1$ to grow as $\alpha \to 0^+$. Figures 1a and 1b illustrate these examples in $\mathbb{R}^2$.

Next, we define interpretability. Recall that we view an interpretation as an approximation via a tree of small depth. We formalize "small" by requiring the existence of a function that bounds the depth of the tree in terms of its accuracy.

Definition 2 (Interpretability)**.** *A concept c is* interpretable *by $\mathcal{H}$ if there is a function $f\colon (0,1] \to \mathbb{N}$ such that, for every distribution $P \in \mathcal{P}(X)$, there exists $\varepsilon_P > 0$ for which*

$$\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq f(\varepsilon) \qquad \textit{for all} \quad 0 < \varepsilon \leq \varepsilon_P \ .$$

If this is the case, then we say that c is interpretable by $\mathcal{H}$ at depth rate f.
A concept c is uniformly interpretable *by $\mathcal{H}$ if there is a function $f'\colon (0,1] \to \mathbb{N}$ such that*

$$\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq f'(\varepsilon) \qquad \textit{for all} \quad P \in \mathcal{P}(X) \quad \textit{and} \quad 0 < \varepsilon \leq 1 \ .$$

If this is the case, then we say that c is uniformly interpretable by $\mathcal{H}$ at depth rate f'.

Note that interpretability requires the bound on the depth to hold only for values of ε that are smaller than a certain threshold ε_P which may depend on the distribution P. Uniform interpretability, instead, requires the depth bound to hold for all ε irrespective of the distribution.

Recalling the above example with the Euclidean space $X = \mathbb{R}^d$ as domain and the family of halfspaces as the hypothesis class $\mathcal{H}$, if the concept c corresponds to the unit Euclidean ball with margin $\mu > 0$ then c is uniformly interpretable. More formally, such a concept c can be modeled as a partial function $c\colon X \to \{0,1\}$ with natural domain $\widetilde{X} \subset X$, where points in the margin belong to $X \setminus \widetilde{X} = B^d(1+\mu) \setminus B^d(1)$ and $c^{-1}(1) = B^d(1)$. Then, without loss of generality, the same definitions and results apply as if the concept c was a total function by restricting the domain to $\widetilde{X}$ and, for every distribution $P \in \mathcal{P}(X)$, considering the distribution $\widetilde{P}(\cdot) = P(\cdot \mid \widetilde{X})$ instead. This follows from the fact that we incur no mistakes for any labeling of points that do not belong to the domain of the "partial" concept c, and the loss of any $\mathcal{H}$-approximation T of c is thus given by $L_{\widetilde{P}}(T, c)$. By reusing geometric results on the approximation of convex bodies, there exists a polytope Q such that $B^d(1) \subseteq Q \subseteq B^d(1+\mu)$, whose (finite) number of vertices is bounded from above by a function of d and μ [Naszódi, 2019]. The polytope Q thus separates the positively labeled points $B^d(1)$ from the negatively labeled ones—achieving loss 0 under any distribution with support $\widetilde{X}$—and it is equivalently representable as an $\mathcal{H}$-based decision tree with depth bounded by a function of d and μ only (i.e., the intersection of halfspaces associated to the facets of Q). See Figure 1c for an illustration of this example in $\mathbb{R}^2$.

At first glance, our notions of interpretability may appear a little narrow. Suppose that, for every distribution P, a concept c is interpretable by $\mathcal{H}$ at polynomial depth rate, but the degree can grow

unbounded with P. In other words, for every $d \in \mathbb{N}_+$ there exists P_d such that c is interpretable by $\mathcal{H}$ at polynomial depth rate with degree d, but not at polynomial depth rate with any smaller degree $d' < d$. Then c is not interpretable by $\mathcal{H}$ according to our definition, but we could still say that c is interpretable at *polynomial* depth rate. More formally, we could consider the family $\mathcal{F}$ of all polynomials, and require that for every P there is some $f \in \mathcal{F}$ that bounds $\mathrm{depth}^c_{\mathcal{H}}(\cdot \mid P)$. By varying $\mathcal{F}$, we obtain a vast range of interpretability rates: logarithmic, sublinear, linear, polynomial, exponential, and so on. Surprisingly, our results show that this hierarchy collapses: an approximable concept c is either not interpretable at all, or is uniformly interpretable at logarithmic rate.

5 A Trichotomy for Interpretability

This section states our main result: as soon as a concept is interpretable at *some* rate, then it is uniformly interpretable at a constant rate for VC classes, and at a logarithmic rate in general.

Theorem 3 (Interpretability trichotomy)**.** *Let X be any domain. For every concept c and every VC hypothesis class $\mathcal{H}$ over X exactly one of the following cases holds:*

(1) c is not approximable by $\mathcal{H}$.

(2) c is approximable but not interpretable by $\mathcal{H}$.

(3) c is uniformly interpretable by $\mathcal{H}$ at constant depth rate.

If $\mathrm{VC}(X, \mathcal{H}) = \infty$ then all claims above hold true, but with (3) replaced by:

(3′) c is uniformly interpretable by $\mathcal{H}$ at depth rate at most logarithmic.

Moreover, all cases are nonempty.

We emphasize again that Theorem 3 is in stark contrast with the behavior of excess risk in terms of training set size observed in statistical learning, where, in the non-uniform (or universal) setting, both exponential and linear rates are possible. It should also be noted that we do not know if case (3′) collapses into case (3)—that is, if a constant depth rate holds also for non-VC classes—or if a non-constant rate is in general unavoidable. This is one of the questions the present work leaves open.

We further observe that, while point (3) of Theorem 3 shows that c is uniformly interpretable by $\mathcal{H}$ at a constant depth rate, this does not necessarily imply the existence of a single $\mathcal{H}$-based decision tree providing such a guarantee for all values of $\varepsilon > 0$. For example, consider a domain $\mathcal{X} = \mathbb{N}$, a concept $c = \{0\}$, and a hypothesis class $\mathcal{H} = \{\{1, \ldots, n\} : n \in \mathbb{N}^+\}$. Now let P be the distribution such that $P(0) = 0.5$ and $P(x) = 2^{-(x+1)}$ for all $x \in \mathbb{N}^+$. For any $\varepsilon > 0$ the depth-1 decision tree with splitting criterion $h = \{1, \ldots, \lceil \log_2(1/\varepsilon) \rceil\}$ is an ε-accurate approximation of c under P, but no $\mathcal{H}$-based tree is an ε-approximation of c for all ε simultaneously.

Our proof of Theorem 3 combines a variety of techniques from different contexts. The first step involves identifying a criterion which can be thought of as a form of "weak interpretability" (items (a) and (b) in the proof). The rest of the proof demonstrates that if a concept c fails to satisfy this criterion, then it is not interpretable by $\mathcal{H}$, and if it does, then it is uniformly interpretable by $\mathcal{H}$. The former impossibility result entails establishing a lower bound on the interpretation rate for an arbitrarily small accuracy with respect to a *fixed* and carefully tailored distribution. This type of lower bounds are more intricate than distribution-free lower bounds (such as those outlined in the No-Free-Lunch Theorem in the PAC setting) and were studied, e.g., by Antos and Lugosi [1998], Bousquet, Hanneke, Moran, Shafer, and Tolstikhin [2023]. In the complementary case, when c satisfies the weak interpretability criterion with respect to $\mathcal{H}$, we prove that c is in fact uniformly interpretable by $\mathcal{H}$ with logarithmic depth, and if $\mathcal{H}$ has a finite VC dimension, then c is interpretable with constant depth. The logarithmic construction and its analysis builds on ideas and techniques originating from boosting algorithms for decision trees [Kearns and Mansour, 1999, Takimoto and Maruoka, 2003]. The derivation of constant depth approximation when $\mathcal{H}$ is a VC class relies on a uniform convergence argument [Vapnik and Chervonenkis, 1971] combined with the Minimax Theorem [von Neumann, 1928]. This derivation is also linked to boosting theory and resembles the boosting-based sample compression scheme by Moran and Yehudayoff [2016].

The proof of case (3′) of Theorem 3 uses the following result. Its proof can be found in Appendix D, and is an adaptation of the results by Kearns and Mansour [1999] and Takimoto and Maruoka [2003]

on boosting decision trees. The main difference is that, via an adequate modification of the `TopDown` algorithm [Kearns and Mansour, 1999], we bound the depth rather than the size of the boosted decision tree.

Theorem 4. *Let X be any domain. For any concept c and any hypothesis class $\mathcal{H}$ over X, if there exist $\gamma \in (0, \frac{1}{2})$ and $d \in \mathbb{N}$ such that $\mathrm{depth}^c_{\mathcal{H}}\left(\frac{1}{2} - \gamma \mid P\right) \leq d$ for all $P \in \mathcal{P}(X)$, then $\mathrm{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq \frac{d}{2\gamma^2} \log \frac{1}{2\varepsilon}$ for all $P \in \mathcal{P}(X)$ and all $\varepsilon > 0$.*

6 Algebraic Characterizations

In this section we show that the notions of approximability and interpretability admit set-theoretical and measure-theoretical characterizations based on properties of $\mathcal{H}$ and the algebras it generates.

To begin with, we need a notion of closure of $\mathcal{H}$. Loosely speaking, we want to include all concepts that, under every distribution, can be approximated arbitrarily well by single elements of $\mathcal{H}$. In other words, these are the concepts that are approximable by $\mathcal{H}$ using decision trees of depth 1.

Definition 5. *The* closure *of $\mathcal{H} \subseteq 2^X$ is*

$$\mathrm{clos}(\mathcal{H}) = \left\{ h \subseteq X \,\middle|\, \forall P \in \mathcal{P}(X),\, \exists h_1, h_2, \ldots \in \mathcal{H} \text{ s.t. } \lim_{n\to\infty} P(h \triangle h_n) = 0 \right\} . \tag{2}$$

Observe that $\mathrm{clos}(\mathcal{H}) \supseteq \mathcal{H}$ by definition. To illustrate the closure let us discuss the hypothesis class $\mathcal{H}$ of rational halfspaces in $\mathbb{R}^2$, i.e., sets of the form $\left\{\{(x, y) \mid ax + by + d \geq 0\} : a, b, d \in \mathbb{Q}\right\}$. Every concept $c\colon \mathbb{R}^2 \to \{0, 1\}$ is approximable by $\mathcal{H}$, as before, relying on the 1-NN algorithm. Halfspaces with real coefficients such as $\{(x, y) \mid x + y \geq \sqrt{2}\}$ are not in $\mathcal{H}$ but are interpretable by $\mathcal{H}$ with depth 1. In general, the closure is related to the concept of universally measurable sets.

We now state the algebraic characterization of the concepts that are approximable by a given hypothesis class $\mathcal{H}$ on some domain X.

Theorem 6 (Algebraic characterization of approximability)**.** *Let X be any domain and $\mathcal{H}$ any hypothesis class over X. A concept $c \subseteq X$ is approximable by $\mathcal{H}$ if and only if $c \in \mathrm{clos}(\sigma(\mathcal{H}))$.*

Furthermore, we manage to prove an algebraic characterization for the concepts that are uniformly interpretable, given a VC class $\mathcal{H}$ on some domain X.

Theorem 7 (Characterization of uniform interpretability for VC classes)**.** *Let X be any domain and let $\mathcal{H}$ be a VC hypothesis class over X. A concept c is uniformly interpretable (at a constant depth) if and only if $c \in \bigcup_{d=1}^{\infty} \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$.*

If the domain X is countable, then closure reduces to pointwise convergence and our algebraic characterization becomes simpler. This is formalized in the next theorem, whose proof relies on some technical lemmas and can be found below. Namely, we show that $\mathrm{clos}(\sigma(\mathcal{H})) = \sigma(\mathcal{H})$ and $\bigcup_{d=1}^{\infty} \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H})) = \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.

Theorem 8 (Characterization for VC classes and countable domains)**.** *Let X be any countable domain, let c be any concept, and let $\mathcal{H}$ be a VC hypothesis class over X. Then:*

1. *c is approximable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H})$.*
2. *c is approximable but not interpretable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H}) \setminus \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*
3. *c is uniformly interpretable by $\mathcal{H}$ if and only if $c \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*

More details for the algebraic characterization can be found in Appendix B and a generalization to representations beyond decision trees is in Appendix C.

Acknowledgment

MB, NCB, and EE acknowledge the financial support from the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme and the the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237, project ELIAS (European Lighthouse of AI for Sustainability). SM is supported by a Robert J. Shillman Fellowship, by ISF grant 1225/20, by BSF grant 2018385, by an Azrieli Faculty

Fellowship, by Israel PBC-VATAT, and by the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. YM has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation, the Yandex Initiative for Machine Learning at Tel Aviv University and a grant from the Tel Aviv University Center for AI and Data Science (TAD). MT acknowledges support from a DOC fellowship of the Austrian academy of sciences (ÖAW).

References

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Mach. Learn.*, 30(1):31–56, 1998. doi: 10.1023/A:1007454427662. URL https://doi.org/10.1023/A:1007454427662.

P. Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33(3):233–282, 1983. doi: 10.5802/aif.938. URL http://www.numdam.org/articles/10.5802/aif.938/.

O. Bastani, C. Kim, and H. Bastani. Interpretability via model extraction. *FAT/ML Workshop 2017 (arXiv preprint arXiv:1705.08504v6)*, 2017. URL https://arxiv.org/abs/1705.08504v6.

G. Blanc, J. Lange, and L.-Y. Tan. Provably efficient, succinct, and precise explanations. *Advances in Neural Information Processing Systems*, 34:6129–6141, 2021.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989. doi: 10.1145/76359.76371. URL https://doi.org/10.1145/76359.76371.

O. Bousquet, S. Hanneke, S. Moran, R. van Handel, and A. Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 532–541, 2021.

O. Bousquet, S. Hanneke, S. Moran, J. Shafer, and I. O. Tolstikhin. Fine-grained distribution-dependent learning curves. In G. Neu and L. Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 5890–5924. PMLR, 2023. URL https://proceedings.mlr.press/v195/bousquet23a.html.

M. Craven and J. Shavlik. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8, 1995.

S. Dasgupta, N. Frost, M. Moshkovitz, and C. Rashtchian. Explainable k-means and k-medians clustering. In *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria*, pages 12–18, 2020.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

R. M. Dudley. Central Limit Theorems for Empirical Measures. *The Annals of Probability*, 6 (6):899 – 929, 1978. doi: 10.1214/aop/1176995384. URL https://doi.org/10.1214/aop/1176995384.

G. K. Dziugaite, S. Ben-David, and D. M. Roy. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*, 2020.

A. Erasmus, T. D. Brunet, and E. Fisher. What is interpretability? *Philosophy & Technology*, 34(4): 833–862, 2021.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

D. Garreau and U. Luxburg. Explaining the explainer: A first theoretical analysis of LIME. In *International conference on artificial intelligence and statistics*, pages 1287–1296, 2020. URL https://arxiv.org/abs/2008.11092v3.

P. R. Halmos. *Measure theory*, volume 18. Springer, 2013. doi: 10.1007/978-1-4684-9440-2. URL https://doi.org/10.1007/978-1-4684-9440-2.

S. Hanneke, A. Kontorovich, S. Sabato, and R. Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129 – 2150, 2021. doi: 10.1214/20-AOS2029.

M. J. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. *J. Comput. Syst. Sci.*, 58(1):109–128, 1999. doi: 10.1006/JCSS.1997.1543. URL https://doi.org/10.1006/jcss.1997.1543.

J. Li, V. Nagarajan, G. Plumb, and A. Talwalkar. A learning theoretic perspective on local explainability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7aL-OtQrBWD.

C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

S. Moran and A. Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, 2016. doi: 10.1145/2890490. URL https://doi.org/10.1145/2890490.

M. Moshkovitz, Y.-Y. Yang, and K. Chaudhuri. Connecting interpretability and robustness in decision trees through separation. In *International Conference on Machine Learning*, pages 7839–7849. PMLR, 2021.

M. Naszódi. Approximating a convex body by a polytope using the epsilon-net theorem. *Discrete & Computational Geometry*, 61:686–693, 2019.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

C. Smyth. Reimer's inequality and Tardos' conjecture. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, pages 218–221, 2002.

E. Takimoto and A. Maruoka. Top-down decision tree learning as information based boosting. *Theoretical Computer Science*, 292(2):447–464, 2003. ISSN 0304-3975. doi: 10.1016/S0304-3975(02)00181-0. Theoretical Aspects of Discovery Science.

G. Tardos. Query complexity, or why is it difficult to separate $\mathrm{NP}^A \cap \mathrm{coNP}^A$ from P^A by random oracles A? *Combinatorica*, 9:385–392, 1989.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL https://doi.org/10.1137/1116025.

T. Vidal and M. Schiffer. Born-again tree ensembles. In *International Conference on Machine Learning*, pages 9743–9753, 2020.

J. von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. URL http://eudml.org/doc/159291.

Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu. Interpreting CNNs via decision trees. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6261–6270, 2019.

A Main Proof

Proof of Theorem 3. We start by proving the cases (1)-(3). Suppose (1) fails, so $\mathrm{depth}^c_{\mathcal{H}}(\varepsilon \mid P) < \infty$ for all $P \in \mathcal{P}(X)$ and all $\varepsilon > 0$. This implies that, for any fixed $\gamma \in (0, \frac{1}{2})$, exactly one of the following two cases holds:

(a) for every $d \in \mathbb{N}$ there exists a distribution P_d such that $\mathrm{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P_d\big) > d$;

(b) there exists $d \in \mathbb{N}$ such that $\mathrm{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \leq d$ for all distributions P.

Suppose (a) holds; we show this implies case (2) of the trichotomy. To this end, we prove that there is no function $r\colon (0,1] \to \mathbb{N}$ such that c is interpretable by $\mathcal{H}$ at depth rate r. Choose indeed any such r. For every $n \in \mathbb{N}^+$ let $d_n = r(2^{-n}(\frac{1}{2} - \gamma))$, and consider the following distribution over X:

$$P^* = \sum_{n \in \mathbb{N}^+} 2^{-n} \cdot P_{d_n} \ . \tag{3}$$

Since P_{d_n} appears in P^* with coefficient 2^{-n}, this implies that, for $\varepsilon_n = 2^{-n}(\frac{1}{2} - \gamma)$, any ε_n-accurate $\mathcal{H}$-interpretation of c under P^* is $(\frac{1}{2} - \gamma)$-accurate under P_{d_n} and so has depth larger than $d_n = r(\varepsilon_n)$. Hence,

$$\mathrm{depth}^c_{\mathcal{H}}(\varepsilon_n \mid P^*) \geq \mathrm{depth}^c_{\mathcal{H}}\left(\frac{1}{2} - \gamma \mid P_{d_n}\right) > d_n = r(\varepsilon_n) \tag{4}$$

holds for all $n \in \mathbb{N}^+$. We conclude that c is not interpretable by $\mathcal{H}$ at depth rate r, as desired.

Now suppose (b) holds; we show this implies case (3) of the trichotomy. Let $\mathcal{T}$ be the set of all binary classifiers that are represented by $\mathcal{H}$-based decision trees of depth at most d, where $d \in \mathbb{N}$ satisfies $\mathrm{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \leq d$ for all $P \in \mathcal{P}(X)$. It is known that $\mathrm{VC}(X, \mathcal{H}) < \infty$ implies $\mathrm{VC}(X, \mathcal{T}) < \infty$ [Dudley, 1978].

We will first prove the claim by taking as domain an arbitrary but finite subset $U \subseteq X$. Later on we will choose U appropriately as a function of the distribution $P \in \mathcal{P}(X)$, and this will prove the theorem's claim. Fix then any such U, and let $\mathcal{P}(U)$ be the family of all distributions over U. By definition of d,

$$\sup_{P \in \mathcal{P}(U)} \inf_{T \in \mathcal{T}} L_P(T, c) \leq \frac{1}{2} - \gamma \ . \tag{5}$$

By von Neumann's minimax theorem, recalling that the value of the game does not change if the column player uses a pure strategy, we have that

$$\sup_{P \in \mathcal{P}(U)} \inf_{T \in \mathcal{T}} L_P(T, c) = \inf_{D \in \mathcal{P}(\mathcal{T})} \sup_{x \in U} \mathbb{E}_{T \sim D} L_{\delta_x}(T, c) \ , \tag{6}$$

where $\mathcal{P}(\mathcal{T})$ is the set of all distributions over $\mathcal{T}$, δ_x is the Dirac delta at $x \in U$, and $\mathbb{E}_{T \sim D} L_{\delta_x}(T, c)$ is thus the expected loss on x of a tree T drawn from D. Hence, there exists $D^* \in \mathcal{P}(\mathcal{T})$ for which

$$\mathbb{E}_{T \sim D^*} L_{\delta_x}(T, c) \leq \frac{1}{2} - \gamma \qquad \forall x \in U \ , \tag{7}$$

and therefore, since $c(x), T(x) \in \{0, 1\}$ for all x and T,

$$\big|c(x) - \mathbb{P}_{T \sim D^*}(T(x) = 1)\big| = \mathbb{P}_{T \sim D^*}(T(x) \neq c(x)) \leq \frac{1}{2} - \gamma \qquad \forall x \in U \ . \tag{8}$$

Let $(\mathcal{T}, U)$ be the dual set system of $(U, \mathcal{T})$. Note that $\mathrm{VC}(\mathcal{T}, U) \leq \mathrm{VC}(\mathcal{T}, X) < 2^{\mathrm{VC}(X, \mathcal{T})+1} < \infty$, where the second inequality shows a known relation [Assouad, 1983] between the primal VC dimension $\mathrm{VC}(X, \mathcal{T})$ of $(X, \mathcal{T})$ and its dual VC dimension $\mathrm{VC}(\mathcal{T}, X)$. By the classic uniform convergence result of Vapnik and Chervonenkis [1971], there exists a multiset $R \subseteq \mathcal{T}$ with $|R| \leq r := r(\mathrm{VC}(X, \mathcal{T}), \gamma, d)$ such that, for every $x \in U$,

$$\left| \frac{|\{T \in R : T(x) = 1\}|}{|R|} - \mathbb{P}_{T \sim D^*}(T(x) = 1) \right| < \frac{\gamma}{2} \ . \tag{9}$$

Together with (8) and (9) this yields

$$\left|\frac{|\{T \in R : T(x) = 1\}|}{|R|} - c(x)\right| < \frac{1}{2} - \frac{\gamma}{2} \tag{10}$$

by the triangle inequality.[2] We now build a $\mathcal{H}$-based decision tree T^*_U that computes the majority vote over all $T \in R$. This tree can be constructed as follows. Let $T_1, \ldots, T_{|R|}$ be the trees in R. Replace each leaf of T_1 with a copy of T_2; in the resulting tree replace every leaf with a copy of T_3, and so on until obtaining T^*_U. For each leaf $z \in \mathcal{L}(T^*_U)$ of T^*_U, define its label ℓ_z as the majority vote given by leaves of (the copies of) $T_1, \ldots, T_{|R|}$ that are encountered on the path from the root of T^*_U to z. Note that T^*_U has depth bounded by rd and, by (10), computes $c(x)$ for all $x \in U$. Thus, $L_U(T^*_U, c) = 0$ where L_U is the expected loss over the uniform distribution over U.

We now choose the set U appropriately. Let $\mathcal{T}^*$ be the family of all $\mathcal{H}$-based decision trees whose depth is at most rd. Because, once again, $\mathrm{VC}(X, \mathcal{T}^*) < \infty$, by uniform convergence there is a finite multiset $U \subseteq X$ such that, for all $T \in \mathcal{T}^*$, $|L_P(T, c) - L_U(T, c)| \leq \varepsilon$. Since $T^*_U \in \mathcal{T}^*$ and $L_U(T^*_U, c) = 0$, it follows that $L_P(T^*_U, c) \leq \varepsilon$. This completes the proof of case (3). Case (3′) follows from Theorem 4 below, assuming (b) holds.

It remains to prove that all cases are nonempty. For (1) let $X = \{a, b\}$, $\mathcal{H} = \{X\}$, $c = \{a\}$, and note that under the uniform distribution no $\mathcal{H}$-interpretation of c is ε-accurate for $\varepsilon < \frac{1}{2}$. For (3) consider any $X, \mathcal{H}$ with $\mathcal{H} \neq \emptyset$ and choose any $c \in \mathcal{H}$; this holds for (3′) too if $\mathcal{H}$ is not a VC class. For (2) we show $c, \mathcal{H}$ that satisfy case (a) above. Let $X = \mathbb{N}$, $c = \mathbb{N}^+$, and $\mathcal{H} = \{\{i\} : i \in \mathbb{N}^+\}$. For every $n \in \mathbb{N}^+$ consider the distribution P_n with support $\{0, \ldots, n\}$ such that $P_n(0) = \frac{1}{2}$ and that $P_n(i) = \frac{1}{2n}$ for every $i \in \{1, \ldots, n\}$. To conclude note that $\mathrm{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P_n\big)$ is unbounded as a function of n for any constant $\gamma \in (0, \frac{1}{2})$. □

B Algebraic Characterizations

In this section we show that the notions of approximability and interpretability admit set-theoretical and measure-theoretical characterizations based on properties of $\mathcal{H}$ and the algebras it generates.

To begin with, we need a notion of closure of $\mathcal{H}$. Loosely speaking, we want to include all concepts that, under every distribution, can be approximated arbitrarily well by single elements of $\mathcal{H}$. In other words, these are the concepts that are approximable by $\mathcal{H}$ using decision trees of depth 1.

Definition 9. *The* closure *of* $\mathcal{H} \subseteq 2^X$ *is*

$$\mathrm{clos}(\mathcal{H}) = \left\{h \subseteq X \,\middle|\, \forall P \in \mathcal{P}(X),\, \exists h_1, h_2, \ldots \in \mathcal{H} \text{ s.t. } \lim_{n\to\infty} P(h \triangle h_n) = 0\right\} . \tag{11}$$

Observe that $\mathrm{clos}(\mathcal{H}) \supseteq \mathcal{H}$ by definition. To illustrate the closure let us discuss the hypothesis class $\mathcal{H}$ of rational halfspaces in $\mathbb{R}^2$, i.e., sets of the form $\big\{\{(x, y) \mid ax + by + d \geq 0\} : a, b, d \in \mathbb{Q}\big\}$. Every concept $c\colon \mathbb{R}^2 \to \{0, 1\}$ is approximable by $\mathcal{H}$, as before, relying on the 1-NN algorithm. Halfspaces with real coefficients such as $\{(x, y) \mid x + y \geq \sqrt{2}\}$ are not in $\mathcal{H}$ but are interpretable by $\mathcal{H}$ with depth 1. In general, the closure is related to the concept of universally measurable sets.

We start with the following lemma, which is derived from well-known results in measure theory.

Lemma 10. *Let* X *be any domain and* $\mathcal{H} \subseteq 2^X$*. Then,* $\mathrm{clos}(\sigma(\mathcal{H})) = \mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$.

Proof. We begin with the proof of the first identity in the statement. The inclusion $\mathrm{clos}(\mathrm{Alg}(\mathcal{H})) \subseteq \mathrm{clos}(\sigma(\mathcal{H}))$ immediately follows by definition of closure. We now show that the converse is also true. Let $T \in \mathrm{clos}(\sigma(\mathcal{H}))$. Fix a distribution $P \in \mathcal{P}(X)$ and $\varepsilon > 0$. By definition of closure, there exists a sequence $A_1, A_2, \ldots \in \sigma(\mathcal{H})$ such that $\lim_{i\to\infty} P(A_i \triangle T) = 0$. Consequently, for every $\varepsilon > 0$ there exists some $i \in \mathbb{N}^+$ such that $P(A_i \triangle T) \leq \varepsilon$. Thus, we can assume without loss of generality that the sequence $(A_i)_{i\in\mathbb{N}^+}$ satisfies $P(A_i \triangle T) \leq \varepsilon_i$ for the choice $\varepsilon_i = 2^{-i}$, for each $i \in \mathbb{N}^+$ (as we can select such a subsequence). Denote the restriction of P to $\sigma(\mathcal{H})$ as $P\big|_{\sigma(\mathcal{H})}$, that is $P\big|_{\sigma(\mathcal{H})}\colon \sigma(\mathcal{H}) \to \mathbb{R}_{\geq 0}$ and $P\big|_{\sigma(\mathcal{H})}(A) = P(A)$ for all $A \in \sigma(\mathcal{H})$. It is well

[2] Note that, if no D^* achieves the infimum of the r.h.s. of eq. (6), the same result holds with, say, $(1 - \gamma)/2$ as the r.h.s. of eq. (7) because it suffices to show that the l.h.s. of eq. (10) is strictly less than $1/2$ for our purposes.

known that, for each i, we can select an element $B_i \in \mathrm{Alg}(\mathcal{H})$ with $P\big|_{\sigma(\mathcal{H})}(B_i \triangle A_i) \leq \varepsilon_i$ (see, e.g., Halmos [2013, Theorem D, Section 13]); hence, $P(B_i \triangle A_i) \leq \varepsilon_i$. By the triangle inequality $P(T \triangle B_i) \leq 2\varepsilon_i = 2^{-i+1}$ for any i, which also implies that $\lim_{i\to\infty} P(T \triangle B_i) = 0$ for the sequence $(B_j)_{j\in\mathbb{N}^+}$ in $\mathrm{Alg}(\mathcal{H})$. Therefore, $T \in \mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$. □

We now state the algebraic characterization of the concepts that are approximable by a given hypothesis class $\mathcal{H}$ on some domain X.

Theorem 6 (Algebraic characterization of approximability)**.** *Let X be any domain and $\mathcal{H}$ any hypothesis class over X. A concept $c \subseteq X$ is approximable by $\mathcal{H}$ if and only if $c \in \mathrm{clos}(\sigma(\mathcal{H}))$.*

Proof. Suppose c is universally approximable by $\mathcal{H}$. Let $P \in \mathcal{P}(X)$ be any distribution. Then, for every $\varepsilon > 0$ there exists an ε-accurate $\mathcal{H}$-approximation $T \in \mathrm{Alg}(\mathcal{H})$ of c under P. Then $P(T \triangle c) = L_P(T, c) \leq \varepsilon$. Consider now the sequence $T_1, T_2, \ldots \in \mathrm{Alg}(\mathcal{H})$ such that, for each $n \in \mathbb{N}_+$, T_n is an ε_n-accurate $\mathcal{H}$-approximation of c under P with the choice $\varepsilon_n = 2^{-n}$. The sequence $(T_n)_{n\in\mathbb{N}_+}$ is such that $\lim_{n\to\infty} P(T_n \triangle c) \leq \lim_{n\to\infty} 2^{-n} = 0$, and thus $c \in \mathrm{clos}(\mathrm{Alg}(\mathcal{H})) = \mathrm{clos}(\sigma(\mathcal{H}))$, where the latter equality follows by Lemma 10.

Now suppose $c \in \mathrm{clos}(\sigma(\mathcal{H})) = \mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$. Fix a distribution $P \in \mathcal{P}(X)$ and $\varepsilon > 0$. By definition of closure, and because $\mathrm{Alg}(\mathcal{H}) \equiv \mathcal{T}_{\mathcal{H}}$, there exists a sequence $T_1, T_2, \ldots \in \mathrm{Alg}(\mathcal{H})$ of trees such that $\lim_{n\to\infty} P(T_n \triangle c) = 0$, and thus there exists some $i \in \mathbb{N}_+$ such that $P(T_i \triangle c) \leq \varepsilon$. This implies that T_i is an ε-accurate $\mathcal{H}$-approximation of c under P with finite depth. As this holds for every P and every $\varepsilon > 0$, it follows that c is universally approximable by $\mathcal{H}$. □

Furthermore, we manage to prove an algebraic characterization for the concepts that are uniformly interpretable, given a VC class $\mathcal{H}$ on some domain X.

Theorem 7 (Characterization of uniform interpretability for VC classes)**.** *Let X be any domain and let $\mathcal{H}$ be a VC hypothesis class over X. A concept c is uniformly interpretable (at a constant depth) if and only if $c \in \bigcup_{d=1}^{\infty} \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$.*

Proof. Since $\mathrm{VC}(X, \mathcal{H}) < \infty$, item (3) of Theorem 3 implies that there exists $d \in \mathbb{N}$ such that, for all $P \in \mathcal{P}(X)$ and all $\varepsilon > 0$, $\mathrm{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq d$. Using an argument similar to the one used in the proof of Theorem 6, we then conclude that $c \in \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$. Hence, the set of concepts that are uniformly interpretable by $\mathcal{H}$ is precisely $\bigcup_{d=1}^{\infty} \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$. □

If the domain X is countable, then closure reduces to pointwise convergence and our algebraic characterization becomes simpler. This is formalized in the next theorem, whose proof relies on some technical lemmas and can be found below. Namely, we show that $\mathrm{clos}(\sigma(\mathcal{H})) = \sigma(\mathcal{H})$ and $\bigcup_{d=1}^{\infty} \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H})) = \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.

Theorem 8 (Characterization for VC classes and countable domains)**.** *Let X be any countable domain, let c be any concept, and let $\mathcal{H}$ be a VC hypothesis class over X. Then:*

1. *c is approximable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H})$.*
2. *c is approximable but not interpretable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H}) \setminus \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*
3. *c is uniformly interpretable by $\mathcal{H}$ if and only if $c \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*

Proof. Item 1 follows from Lemma 11 Theorems 6 and 7, and items 2 and 3 from Lemma 15. □

Lemma 11. *Let X be any countable domain and $\mathcal{H}$ be any hypothesis class over X. Then,* $\mathrm{clos}(\mathrm{Alg}(\mathcal{H})) = \sigma(\mathcal{H})$.

Proof. Clearly $\sigma(\mathcal{H}) \subseteq \mathrm{clos}(\sigma(\mathcal{H})) = \mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$ by Lemma 10. Now we prove the converse. Let $A \in \mathrm{clos}(\sigma(\mathcal{H}))$ and let $P \in \mathcal{P}(X)$ such that $P(x) > 0$ for all $x \in X$; note that P exists as X is countable and it also means that $\mathrm{supp}(P) = X$. By definition of $\mathrm{clos}(\sigma(\mathcal{H}))$, there exists a sequence $(A_i)_{i\in\mathbb{N}^+}$ in $\sigma(\mathcal{H})$ such that

$$\lim_{i\to\infty} P(A \triangle A_i) = 0 \ . \tag{12}$$

By selecting an appropriate subsequence, we can assume $P(A \triangle A_i) \leq 2^{-i}$ for all $i \in \mathbb{N}^+$ without loss of generality. Define

$$B_i = \bigcap_{j \geq i} A_j \qquad \forall i \in \mathbb{N}^+ \tag{13}$$

and observe that $B_i \in \sigma(\mathcal{H})$ for each $i \in \mathbb{N}^+$. Note that

$$P(A \triangle B_i) = P\left(A \triangle \bigcap_{j \geq i} A_j\right) \leq \sum_{j \geq i} P(A \triangle A_j) \leq 2^{-i+1} \ . \tag{14}$$

Note also that $B_i \subseteq A$ for all $i \in \mathbb{N}^+$. Suppose indeed this was not the case, then $B_i \setminus A \neq \emptyset$. Hence, by definition of B_i, there exists some $x \in A_j \setminus A$ for all $j \geq i$. Since $P(x) > 0$ by the choice of P, we have the contradiction

$$\lim_{i \to \infty} P(A \triangle A_i) \geq P(x) > 0 \ . \tag{15}$$

Now consider the set

$$B = \bigcup_{i \in \mathbb{N}^+} B_i = \lim_{i \to \infty} B_i \ . \tag{16}$$

Note that by construction $B \in \sigma(\mathcal{H})$.[3] Moreover, since $B_i \subseteq B_{i+1}$ and $B_i \subseteq A$ for all $i \in \mathbb{N}^+$, we have that the sequence $(A \triangle B_i)_{i \in \mathbb{N}^+}$ is downward monotone and thus

$$P(A \triangle B) = P\left(\bigcap_{i \in \mathbb{N}^+} A \triangle B_i\right) = \lim_{i \to \infty} P(A \triangle B_i) = 0 \ . \tag{17}$$

Given that P has full support, this implies $A = B$. □

Definition 12. *Let X be any set. A sequence $(h_i)_{i \in \mathbb{N}}$ in 2^X is* pointwise convergent *to $h \in 2^X$ if*

$$\forall x \in X \quad \exists i_x \in \mathbb{N} : \ \forall i \geq i_x \qquad x \in h_i \iff x \in h \ . \tag{18}$$

Proposition 13. *If X is countable then every infinite sequence $(h_i)_{i \in \mathbb{N}}$ in 2^X contains an infinite subsequence that is pointwise convergent.*

Let $\mathcal{H} \subseteq 2^X$ and let $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H})$ be the family of all subsets of X that are the pointwise limit of some sequence in $\mathcal{H}$. Clearly $\mathcal{H} \subseteq \mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) \subseteq 2^X$.

Lemma 14. *If X is countable then* $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) = \mathrm{clos}(\mathcal{H})$.

Proof. To see that $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) \subseteq \mathrm{clos}(\mathcal{H})$, recall the definition of pointwise convergence, and note how it implies that if a sequence $(h_i)_{i \in \mathbb{N}}$ converges pointwise to h then $\lim_{i \to \infty} P(h_i \triangle h) = 0$ for every $P \in \mathcal{P}(X)$. To see that $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) \supseteq \mathrm{clos}(\mathcal{H})$, choose any sequence $(h_i)_{i \in \mathbb{N}}$ that converges to some $h \in \mathrm{clos}(\mathcal{H})$ under an appropriate distribution $P \in \mathcal{P}(X)$ such that $\mathrm{supp}(P) = X$ (which exists as X is countable); observe that this implies the pointwise convergence of $(h_i)_{i \in \mathbb{N}}$ to h. □

Lemma 15. *If X is countable and $\mathcal{H}$ is a VC class over X, then* $\mathrm{clos}(\mathrm{Alg}_d(\mathcal{H})) \subseteq \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$ *for every $d \in \mathbb{N}$.*

Proof. Let $d \in \mathbb{N}$ and $c \in \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$. By Lemma 14, $c \in \mathrm{clos}_{\mathrm{pw}}(\mathrm{Alg}_d(\mathcal{H}))$, so there exists an infinite sequence of trees $(T_i)_{i \in \mathbb{N}}$ in $\mathrm{Alg}_d(\mathcal{H})$ that converge pointwise to c. Without loss of generality, we may assume that every T_i is a complete tree of depth d.[4] Now consider the sequence $(h_i^1)_{i \in \mathbb{N}}$ of decision rules used by the first node (say, the root) of those trees. By Proposition 13 there is an infinite subsequence $(h_{i_j}^1)_{j \in \mathbb{N}}$ that is pointwise convergent to some $h^1 \in \mathcal{H}$. Now consider the infinite sequence of trees $(T_{i_j})_{j \in \mathbb{N}}$, and repeat the argument for the second node (say, a child of the root corresponding to a specific output of the decision stump at the root). By repeating the argument $2^d - 1$ times (one for every internal node of the trees) we obtain an infinite sequence $(T_i^*)_{i \in \mathbb{N}}$ of trees in $\mathrm{Alg}_d(\mathcal{H})$ that converge pointwise to c and such that at every node v the decision rules converge

[3] In particular, $B = \liminf_{i \to \infty} A_i$.

[4] One can always complete T_i using internal nodes that hold, e.g., the decision rule of the root.

pointwise to some h^v. Now let T^* be the decision tree obtained by using h^v as decision rule at v. We observe that $T^* = c$. Let $x \in X$. By Definition 12, for each node v there exists i_x^v such that $x \in h_i^v$ iff $x \in h^v$ for every $i \geq i_x^v$, where h_i^v is the stump used at v by T_i^*. By letting $i_x = \max_v i_x^v$ it follows that $x \in h_i^v$ iff $x \in h^v$ for every $i \geq i_x$ and all nodes v simultaneously. Therefore all trees T_i^* with $i \geq i_x$ send x to the same leaf, and moreover that leaf remains the same if we use h^v at v. Note also that, since $(T_i^*)_{i\in\mathbb{N}}$ is infinite, then we can assume that every leaf predicts the same label in all T_i^* (since there is certainly an infinite subsequence that satisfies such a constraint). It follows that $(T_i^*)_{i\in\mathbb{N}}$ converges pointwise the tree T^* that uses the limit stump h^v at v. But the labeling of $(T_i^*)_{i\in\mathbb{N}}$ converges pointwise to c, too. We conclude that $T = T^*$. Finally, note that by construction $h^v \in \mathrm{clos}_{\mathrm{pw}}(\mathcal{H})$, and thus by Lemma 14 $h^v \in \mathrm{clos}(\mathcal{H})$, for all v, hence $T^* \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$. It follows that $c \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$. □

C General Representations

Although shallow decision trees are the blueprint of interpretable models, our theory naturally extends to ways of measuring the complexity of elements in $\mathrm{Alg}(\mathcal{H})$ different from the tree depth. Next, we define a set of minimal conditions (satisfied, e.g., by tree depth) that a function must satisfy to be used as a complexity measure for $\mathrm{Alg}(\mathcal{H})$.

Definition 16. *Let X be any domain and $\mathcal{H}$ a hypothesis class over X. A function $\Gamma\colon\ \mathrm{Alg}(\mathcal{H}) \to \mathbb{N}$ is a* graded complexity measure *if:*

1. $\Gamma(f) = 0$ *for all* $f \in \mathcal{H}$,
2. $\Gamma(f_1 \cup f_2) \leq 1 + \Gamma(f_1) + \Gamma(f_2)$ *for all* $f_1, f_2 \in \mathrm{Alg}(\mathcal{H})$,
3. $\Gamma(f_1 \cap f_2) \leq 1 + \Gamma(f_1) + \Gamma(f_2)$ *for all* $f_1, f_2 \in \mathrm{Alg}(\mathcal{H})$, *and*
4. $\Gamma(X \setminus f) \leq 1 + \Gamma(f)$ *for all* $f \in \mathrm{Alg}(\mathcal{H})$.

The minimal complexity of an ε-accurate $\mathcal{H}$-interpretation of c under P is

$$\Gamma^c_{\mathcal{H}}(\varepsilon \mid P) = \inf_{T\in\mathrm{Alg}(\mathcal{H})\colon\, L_P(T,c)\leq\varepsilon} \Gamma(T) \ . \tag{19}$$

The definitions of approximability, interpretability, and uniform interpretability are readily generalized to an arbitrary graded complexity measure, by simply replacing $\mathrm{depth}(\cdot)$ with $\Gamma(\cdot)$. We can then prove the following extension of Theorem 3.

Theorem 17 (Interpretability trichotomy for general representations)**.** *Let X be any domain and let Γ be any graded complexity measure. Then, for every concept c and every VC hypothesis class $\mathcal{H}$ over X exactly one of the following cases holds:*

(1) c is not approximable by $\mathcal{H}$.

(2) c is approximable by $\mathcal{H}$ but not interpretable by $\mathcal{H}$.

(3) c is uniformly interpretable by $\mathcal{H}$ at constant Γ-complexity rate.

If $\mathrm{VC}(X, \mathcal{H}) = \infty$ then all claims above hold true, but with (3) replaced by:

(3′) c is uniformly interpretable by $\mathcal{H}$ at a Γ-complexity rate $\mathcal{O}\big(\frac{1}{\varepsilon^d}\big)$ for some $d \in \mathbb{N}$.

Unlike Theorem 3, cases (2) and (3) might collapse for certain choices of Γ even when $\mathcal{H}$ is not a VC class. Indeed, according to our definition, Γ is not forced to grow at any specific rate, and thus $\Gamma(f)$ might be bounded by some constant uniformly over $\mathrm{Alg}(\mathcal{H})$. In an extreme case one might in fact set $\Gamma \equiv 0$, although clearly this would not yield any interesting result.

Proof of Theorem 17. The proof is similar to the proof of Theorem 3. Suppose (1) fails, so $\Gamma^c_{\mathcal{H}}(\varepsilon \mid P) < \infty$ for all $\varepsilon > 0$ and all distributions P. This implies that, for any fixed $\gamma \in (0, \frac{1}{2})$, exactly one of the following two cases holds:

(a) for every $k \in \mathbb{N}$ there exists a distribution P_k such that $\Gamma^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P_k\big) > k$;

(b) there exists $k \in \mathbb{N}$ such that $\Gamma^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \leq k$ for all distributions P. 5

Suppose (a) holds; we show this implies case (2) of the trichotomy. Choose any function $r\colon (0,1] \to \mathbb{N}$. For every $n \in \mathbb{N}^+$ let $d_n = r(2^{-n}(\frac{1}{2} - \gamma))$, and consider the following distribution over X:

$$P^* = \sum_{n \in \mathbb{N}^+} 2^{-n} \cdot P_{d_n} \ . \tag{20}$$

Since P_{d_n} appears in P^* with coefficient 2^{-n}, this implies that, for $\varepsilon_n = 2^{-n}(\frac{1}{2} - \gamma)$, any ε_n-accurate interpretation of c under P^* is $(\frac{1}{2} - \gamma)$-accurate under P_{d_n}, and thus

$$\Gamma^c_{\mathcal{H}}(\varepsilon_n \mid P^*) \geq \Gamma^c_{\mathcal{H}}\left(\frac{1}{2} - \gamma \mid P_{d_n}\right) > d_n = r(\varepsilon_n) \ . \tag{21}$$

Hence, $\Gamma^c_{\mathcal{H}}(\varepsilon_n \mid P^*) > r(\varepsilon_n)$ for all $n \in \mathbb{N}^+$.

Suppose now (b) holds; we show this implies case (3) of the trichotomy. Define the family $\mathcal{A}_k = \big\{A \in \mathrm{Alg}(\mathcal{H}) \,:\, \Gamma(A) \leq k\big\}$. Fix any $P \in \mathcal{P}(X)$ and $\varepsilon > 0$. Following the same argument as in the proof of case (3) in Theorem 3, there exists an $\mathcal{A}_k$-based decision tree T such that $L_P(T, c) \leq \varepsilon$ and $\mathrm{depth}(T) \leq d$ for some $d \in \mathbb{N}$ independent of P and ε. Now we rewrite T as an element of $\mathrm{Alg}(\mathcal{H})$. Let $A_v \in \mathcal{A}_k$ be the decision stump T used at $v \in \mathcal{V}(T)$ and, denoting by $\mathcal{L}(T)$ the set of leaves of T, let $\ell_z \in \{0,1\}$ be the label of the leaf $z \in \mathcal{L}(T)$ in T. For every $v \in \mathcal{V}(T)$, define

$$A^T_v = \begin{cases} X & v \in \mathcal{L}(T),\ \ell_v = 1 \\ \emptyset & v \in \mathcal{L}(T),\ \ell_v = 0 \\ (A_v \cap A^T_u) \cup (\overline{A}_v \cap A^T_w) & v \notin \mathcal{L}(T) \end{cases} \tag{22}$$

where u and w are, respectively, the left and right child of v when $v \notin \mathcal{L}(T)$. Let $A = A^T_r$ where r is the root of T. Observe that A is equivalent to T, and that $A \in \mathrm{Alg}(\mathcal{H})$. Moreover, $\Gamma(A^T_v) \leq 4 + 2\Gamma(A_v) + \Gamma(A^T_u) + \Gamma(A^T_w)$ by the properties of Γ (see Definition 16). Therefore,

$$\Gamma(A) = \mathcal{O}\left(\sum_{v \in \mathcal{V}(T)} (\Gamma(A_v) + 1)\right) = (k+1) \times \mathcal{O}(|\mathcal{V}(T)|) = \mathcal{O}(|\mathcal{V}(T)|) \ , \tag{23}$$

where we used the fact that $\Gamma(A_v) \leq k$ because $A_v \in \mathcal{A}_k$. To conclude the proof, note that the above bound on $\mathrm{depth}(T)$ implies $\mathcal{O}(|\mathcal{V}(T)|) = \mathcal{O}(2^{\mathrm{depth}(T)}) = \mathcal{O}(2^d)$, where both d and the constants in the $\mathcal{O}(\cdot)$ notation depend neither on P nor on ε.

As for case (3′), assume again (b) holds. Then, Theorem 4 applied to the class $\mathcal{A}_k$ implies the existence of an $\mathcal{A}_k$-based decision tree T such that $L_P(T, c) \leq \varepsilon$ and $\mathrm{depth}(T) \leq d \log \frac{1}{2\varepsilon}$ for all P and $\varepsilon > 0$, where $d = \frac{1}{2\gamma^2}$. Constructing again $A \in \mathrm{Alg}(\mathcal{H})$ equivalent to T as above and using the bound on $\mathrm{depth}(T)$, we have $\Gamma(A) = \mathcal{O}(|\mathcal{V}(T)|) = \mathcal{O}(2^{\mathrm{depth}(T)}) = \mathcal{O}\big(\frac{1}{\varepsilon^d}\big)$ where both d and the constants in the $\mathcal{O}(\cdot)$ notation are independent of P and ε. □

D Boosting Decision Trees with Bounded Depth

Theorem 4. *Let X be any domain. For any concept c and any hypothesis class $\mathcal{H}$ over X, if there exist $\gamma \in (0, \frac{1}{2})$ and $d \in \mathbb{N}$ such that $\mathrm{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \leq d$ for all $P \in \mathcal{P}(X)$, then $\mathrm{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq \frac{d}{2\gamma^2} \log \frac{1}{2\varepsilon}$ for all $P \in \mathcal{P}(X)$ and all $\varepsilon > 0$.*

We use a surrogate loss $G(q) = \sqrt{q(1-q)}$, where $0 \leq q \leq 1$. Since $\min\{q, 1-q\} \leq G(q)$, the surrogate loss bounds from above the classification error of the majority vote. For a distribution $P \in \mathcal{P}(X)$, let $G_P(c) = G\big(P(c=1)\big)$. Let the conditional surrogate loss of $f\colon X \to \{0,1\}$ be

$$G_P(c \mid f) = P(f=0)G\big(P(c=1 \mid f=0)\big) + P(f=1)G\big(P(c=1 \mid f=1)\big) \ . \tag{24}$$

Finally, given a decision tree T with leaves $\mathcal{L}(T)$, define the conditional surrogate loss of T as

$$H_P(c \mid T) = \sum_{z \in \mathcal{L}(T)} P(z) G(p_{c|z}) \ , \tag{25}$$

where $P(z)$ is the probability that $x \sim P$ is mapped to leaf z in the tree T and $p_{c|z} = P(c=1 \mid z)$. Our goal is to construct an $\mathcal{H}$-based decision tree T such that $H_P(c \mid T) \leq \varepsilon$, implying that

$L_P(T, c) \leq \varepsilon$ because $G\big(p_{c|z}\big)$ bounds from above the probability that $T(x) \neq c(x)$ conditioned on x being mapped to z in T.

Our variant of `TopDown`, called `TopDownLBL` (TopDown Level-By-Level), starts from a single-leaf tree T with a majority-vote label and works in phases. In each phase, we replace each leaf $z \in \mathcal{L}(T)$ of the current tree T with a suitably chosen $\mathcal{H}$-based d-depth tree T_z using the same criterion as `TopDown`. The main difference is that the weak learners adopted by `TopDownLBL` consist of $\mathcal{H}$-based trees of depth bounded by d, which generalize from the individual decision stumps of $\mathcal{H}$ as in `TopDown` (corresponding to the case $d = 1$). Hence, at the end of each phase, the depth of T increases by at most d. The algorithm stops if and when $H_P(c \mid T) \leq \varepsilon$.

We use the two following lemmas.

Lemma 18 (Takimoto and Maruoka [2003, Lemma A.1])**.** *Let P be a balanced distribution, i.e., $P(c = 1) = P(c = 0) = \frac{1}{2}$. Let $f \colon X \to \{0, 1\}$ be such that $L_P(f, c) \leq \frac{1}{2} - \gamma$ for some $\gamma \in (0, \frac{1}{2})$. Then, $G_P(c \mid f) \leq (1 - 2\gamma^2)G_P(c)$.*

Lemma 19 (Takimoto and Maruoka [2003, Proposition 5])**.** *Let P be a distribution and P' its balanced version. If $G_{P'}(c \mid h) \leq (1 - \beta)G_{P'}(c)$ for some $\beta > 0$ then $G_P(c \mid h) \leq (1 - \beta)G_P(c)$.*

Proof of Theorem 4. Our algorithm `TopDownLBL` can be equivalently viewed as building a $\mathcal{H}'$-based tree T', where $\mathcal{H}'$ is the class of $\mathcal{H}$-based d-depth trees. Any $\mathcal{H}'$-based tree T' can be transformed into a $\mathcal{H}$-based tree T in a top-down fashion simply by listing the nodes at each level of T' starting from the root, and iteratively replacing every decision stump $h' \in \mathcal{H}'$ with the corresponding $\mathcal{H}$-based tree $T_{h'}$. Then, each leaf $z \in \mathcal{L}(T_{h'})$ of $T_{h'}$ is replaced by copies of the left or right subtree of the decision stump h' in T' based on the values (0 or 1) of the label ℓ_z of z. Clearly, the depth of T is at most d times the depth of T'.

We now bound the drop in $H_P(c \mid T')$ when a leaf z in the $\mathcal{H}'$-based tree T' is replaced by a decision stump in $\mathcal{H}'$. Let P the distribution over X conditioned on x being mapped to z and let P' its "balanced" version satisfying $P'(c = 1) = P'(c = 0) = \frac{1}{2}$. Because of our weak learning assumption, we know there exists $h'_z \in \mathcal{H}'$ with error at most $1/2 - \gamma$ on P'. By Lemma 18, $G_{P'}(c \mid h'_z) \leq (1 - 2\gamma^2)G_{P'}(p'_{c|z})$, where $p'_{c|z} = \frac{1}{2}$ because of the balanced property of P'. Hence, by Lemma 19,

$$G_P(c \mid h'_z) \leq (1 - 2\gamma^2)G_P(p_{c|z}) \ . \tag{26}$$

Let T'_z be the tree T' in which we replaced a leaf $z \in \mathcal{L}(T')$ with the decision stump $h'_z \in \mathcal{H}'$. Using Equation (26),

$$H_P(c \mid T') - H_P(c \mid T'_z) = \big(G_P(p_{c|z}) - G_P(c \mid h'_z)\big)P(z) \geq 2\gamma^2 G_P(p_{c|z})P(z) \ . \tag{27}$$

Now let T'_i be the tree after the algorithm has run for i phases. Using the above inequality for each $z \in \mathcal{L}(T'_i)$, we obtain

$$H_P(c \mid T'_i) - H_P(c \mid T'_{i+1}) \geq \sum_{z \in \mathcal{L}(T'_i)} 2\gamma^2 G_P(p_{c|z})P(z) = 2\gamma^2 H_P(c \mid T'_i) \ . \tag{28}$$

Hence, after m phases,

$$L_P(T'_m, c) \leq H_P(c \mid T'_m) \leq \big(1 - 2\gamma^2\big)^m H_P(c \mid T'_0) \leq \frac{1}{2}e^{-2m\gamma^2} \ , \tag{29}$$

where T'_0 is the initial tree consisting of a single leaf z and, in the last inequality, we used the fact that $H_P(c \mid T'_0) = G_P(p_{c|z}) \leq \frac{1}{2}$ and the inequality $1 - x \leq e^{-x}$. The proof is concluded by noting that $\frac{1}{2}e^{-2m\gamma^2} \leq \varepsilon$ for $m \geq \frac{1}{2\gamma^2} \log \frac{1}{2\varepsilon}$. □

We remark that we recover the standard setting of boosting decision trees when $d = 1$. In this special case, our result matches the depth lower bound mentioned by Kearns and Mansour [1999], while guaranteeing a $\mathcal{O}\big(2^{\mathrm{depth}^{\mathrm{c}}_{\mathcal{H}}(\varepsilon|P)}\big) = \mathcal{O}\big((1/\varepsilon)^{1/(2\gamma^2)}\big)$ tree-size upper bound that is analogous to the ones by Kearns and Mansour [1999] and Takimoto and Maruoka [2003].

E Remarks on the Graded Complexity Measure Results

In Section C we demonstrated more general guarantees for any graded complexity measure Γ, given any domain X and any hypothesis class $\mathcal{H}$ over X. Observe that, when $\mathcal{H}$ is a non-VC class, item (3′) of Theorem 17 states an upper bound on the Γ-complexity rate of order $\mathcal{O}\big(\frac{1}{\varepsilon^d}\big)$ for a constant $d \in \mathbb{N}$. This bound is indeed larger compared to the previous guarantee of $\mathcal{O}(\log(1/\varepsilon))$ on the depth of $\mathcal{H}$-based decision trees (Theorem 3) and it has to do with the generality of the definition of graded complexity measure.

Keeping this in mind, we remark that it is possible to recover the $\mathcal{O}(\log(1/\varepsilon))$ Γ-complexity rate bound under a stronger assumption on the graded complexity measure Γ. In particular, it is sufficient for Γ to satisfy

$$\Gamma(f_1 \cup f_2) \leq 1 + \max\{\Gamma(f_1), \Gamma(f_2)\} \qquad \forall f_1, f_2 \in \mathrm{Alg}(\mathcal{H}) \ . \tag{30}$$

Note that this condition is satisfied when Γ corresponds to the depth of $\mathcal{H}$-based decision trees. For example, consider a similar representation of trees as in Equation (22) using directly $\mathcal{H}$ for the decision rules of the internal nodes.

Thus, we can follow the same steps as in the proof of Theorem 17 with a particular focus on the construction of A from the decision tree T in Equation (22). It immediately follows that $\Gamma(A_v^T) \leq 3 + \Gamma(A_v) + \max\big\{\Gamma(A_u^T), \Gamma(A_w^T)\big\}$ for any internal node $v \notin \mathcal{L}(T)$, where u and w are, respectively, the left and right child of v. Now, let $\rho(z) \subseteq \mathcal{V}(T)$ be the nodes along the path from the root of T to the leaf $z \in \mathcal{L}(T)$. We can thus show that

$$\Gamma(A) = \mathcal{O}\left(\max_{z \in \mathcal{L}(T)} \sum_{v \in \rho(z)} (\Gamma(A_v) + 1)\right) = \mathcal{O}\big((k+1) \cdot \mathrm{depth}(T)\big) = \mathcal{O}\Big(\log \frac{1}{\varepsilon}\Big) \ , \tag{31}$$

where we used the fact that T has $\mathrm{depth}(T) \leq \frac{1}{2\gamma^2} \log \frac{1}{2\varepsilon}$ and that $\Gamma(A_v) \leq k$ for any internal node v of T.