Proceedings of Machine Learning Research vol 195:1–67, 2023 36th Annual Conference on Learning Theory

Universal Rates for Multiclass Learning

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Shay Moran SMORAN@TECHNION.AC.IL
Technion

Qian Zhang ZHAN3761@PURDUE.EDU
Purdue University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study universal rates for multiclass classification, establishing the optimal rates (up to log factors) for all hypothesis classes. This generalizes previous results on binary classification (Bousquet, Hanneke, Moran, van Handel, and Yehudayoff, 2021), and resolves an open question studied by Kalavasis, Velegkas, and Karbasi (2022) who handled the multiclass setting with a bounded number of class labels. In contrast, our result applies for any countable label space. Even for finite label space, our proofs provide a more precise bounds on the learning curves, as they do not depend on the number of labels. Specifically, we show that any class admits exponential rates if and only if it has no infinite Littlestone tree, and admits (near-)linear rates if and only if it has no infinite Daniely-Shalev-Shwartz-Littleston (DSL) tree, and otherwise requires arbitrarily slow rates. DSL trees are a new structure we define in this work, in which each node of the tree is given by a pseudo-cube of possible classifications of a given set of points. Pseudo-cubes are a structure, rooted in the work of Daniely and Shalev-Shwartz (2014) and recently shown by Brukhim, Carmon, Dinur, Moran, and Yehudayoff (2022) to characterize PAC learnability (i.e., uniform rates) for multiclass classification. We also resolve an open question of Kalavasis, Velegkas, and Karbasi (2022) regarding the equivalence of classes having infinite Graph-Littlestone (GL) trees versus infinite Natarajan-Littlestone (NL) trees, showing that they are indeed equivalent.

Keywords: Multiclass learning, Universal rates, Learning curve, Statistical learning, Online learning

1. Introduction

Multiclass classification, i.e., classifying data into multiple classes in some label (class) space $\mathcal{Y}$ is a fundamental task in machine learning with direct application in a wide range of scenarios including image recognition (Rawat and Wang, 2017), natural language processing (Young et al., 2018), protein structure classification (Dietmann and Holm, 2001), etc. In practice, the number of classes ($|\mathcal{Y}|$) could be huge or infinite; e.g., in statistical language models (Song and Croft, 1999), $|\mathcal{Y}|$ is the vocabulary size; for count data prediction (Hellerstein and Mendelsohn, 1993), $\mathcal{Y}$ is the set of natural numbers. Thus, the study of multiclass learnability and error rates has been a crucial problem in learning theory. However, even under the renowned PAC (Probably Approximately Correct) learning framework (Valiant, 1984), until recently solved by Brukhim et al. (2022), the characterization of multiclass learnability for infinite number of classes ($|\mathcal{Y}| = \infty$) remained to be a challenging problem for decades after the characterization of PAC learnability of binary classification ($|\mathcal{Y}| = 2$) with the finiteness of the Vapnik-Chervonenkis (VC) dimension (Vapnik and

Chervonenkis, 1971; Blumer et al., 1989). Natarajan and Tadepalli (1988); Natarajan (1989) defined two extentions of the VC dimension in multiclass learning, the Natarajan dimension ($\dim_N$) and the Graph dimension ($\dim_G$) which both characterize the multiclass PAC learnability for finite number of classes ($|\mathcal{Y}| < \infty$). Though the Graph dimension was shown to be unable to characterize the multiclass PAC learnability when $|\mathcal{Y}| = \infty$, it was conjectured if the Natarajan dimension would do (Natarajan, 1989). Daniely and Shalev-Shwartz (2014) defined a new dimension named the *Daniely-Shalev-Shwartz (DS) dimension* ($\dim$) by Brukhim et al. (2022) and showed that finite DS dimension is a necessary condition for PAC learnability. Recently, Brukhim et al. (2022) proved that the DS dimension fully characterizes PAC learnability in the multiclass setting by proposing an algorithm achieving $O(\frac{\dim(\mathcal{H})^{3/2}\log^2(n)}{n})$ (see Section 1.1 for details) error rate for any hypothesis class $\mathcal{H}$ under the PAC framework. They also refuted the conjecture that the Natarajan dimension characterizes multiclass PAC learnability by providing a hypothesis class with the Natarajan dimension 1 and an infinite DS dimension.

In terms of the *learning curve*, i.e., the error rate (measured on test data) as a function of the number of training examples, due to its distribution-free nature, the PAC framework, however, fails to capture the fine-grained and potentially faster *distribution-dependent* learning curves of hypothesis classes. In the realizable setting, PAC learning considers the best worst-case (uniform) performance of any algorithm on a hypothesis class against any realizable distribution. While in real-world problems, the distribution for data generation is often fixed in one task and the study of learning curves under fixed distributions is concerned. These thoughts motivate the proposition of *universal learning* in the work of Bousquet et al. (2021), where they consider the distribution-dependent error rate of a learning algorithm on a hypothesis class, holding universally for all realizable distributions. They showed that for binary classification, the following *trichotomy* exists for any hypothesis class $\mathcal{H}$ with $|\mathcal{H}| > 3$: $\mathcal{H}$ is either universally learnable with optimal rate e^{-n} (exponential rate), universally learnable with optimal rate $1/n$ (linear rate), or requires arbitrarily slow rates (see Section 1.1 for details), which is fully determined by the combinatorial properties of $\mathcal{H}$ (the nonexistence of certain infinite trees). Compared to the dichotomy in PAC learning: $\mathcal{H}$ is either PAC-learnable with a linear uniform rate ($1/n$) or is not PAC-learnable at all, universal learning provides more insights of the learning curve in binary classification.

A natural direction is to extend the framework of universal learning to multiclass classification that would bring fine-grained disribution-dependent analysis of learning curves in multiclass problems. Recently, Kalavasis et al. (2022) proved the same trichotomy for multiclass universal learining assuming *finite* label space ($|\mathcal{Y}| < \infty$): a hypothesis class with finite label space is either universally learnable with optimal rate e^{-n}, universally learnable with optimal rate $1/n$, or requires arbitrarily slow rates, depending on the nonexistence of an infinite Littlestone tree and an infinite Natarajan-Littlestone (NL) tree they defined (see Section 1.2 for details). However, their analysis for the linear universal rate based on NL trees cannot be extended to the setting of countable label spaces. As is pointed out in Kalavasis et al. (2022), it is an important next step to characterize multiclass universal learning with *infinite* label space ($|\mathcal{Y}| = \infty$).

However, for general uncountable label spaces, the existence of a universally measurable learning algorithm that is universally consistent (see Section 1.1 for details), i.e., with an error rate converging to zero for any realizable distributions, remains unsolved to our knowledge, which is an important problem in itself. Thus, we focus on countable label spaces in this paper and summarize our contributions below.

Contributions. In this paper, we study multiclass universal learning for general *countable* label spaces ($|\mathcal{Y}|$ can be infinite). We prove in Theorem 8 that a hypothesis class with a countable label space is either universally learnable with optimal rate e^{-n}, universally learnable with optimal rate in $\widetilde{\Theta}(1/n)$ (near-linear rate), or requires arbitrarily slow rates, which is fully characterized by the nonexistence of an infinite *Littlestone tree* and an infinite *Daniely-Shalev-Shwartz-Littleston (DSL) tree* proposed by us (see Section 1.2 for details). In particular, we propose different universally measurable learning algorithms that achieve the exponential and near-linear rates in those corresponding settings. We also show that the NL tree does *not* characterize the near-linear rate by proving the existence of a hypothesis class that has an infinite DSL tree but has no NL tree of depth 2 for countable label space in Theorem 9. Finally, we solve the first question in Kalavasis et al. (2022, Open question 1) by proving in Theorem 10 that a hypothesis class with finite label space ($|\mathcal{Y}| < \infty$) has an infinite NL tree if and only if it has an infinite Graph-Littlestone (GL) tree defined in Kalavasis et al. (2022, Definition 8), which implies that the GL tree is equivalent to the NL tree in determining the universal rate of multiclass learning with finite label space.

Outline. In Section 1.1, we formally define the multiclass learning problem considered in this paper and the universal error rate which is compared to the uniform error rate in PAC learning. In Section 1.2, we introduce the definitions of the different tree structures of a hypothesis class and state the main theoretical results. In Section 1.3, we discuss some future research directions in multiclass learning. In Section 2, we provide three examples of the multiclass learning problem, each corresponding to a different universal rate in the trichotomy. In Section 3, we summarize the key technical details and the proof sketches of the main results. The complete proofs are included in the appendix.

1.1. The multiclass learning problem and the universal rates

In this section, we introduce the multiclass learning problem considered in this paper and the concept of universal learning. We refer readers to Appendix A.1 for the notation we used throughout the paper. Let $\mathcal{X}$ denote the domain (feature space), $\mathcal{Y}$ denote the codomain (label space), and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ denote the hypothesis class. To avoid measurability issues, we assume that $\mathcal{X}$ is a Polish space and $\mathcal{Y}$ is countable with $|\mathcal{Y}| \geqslant 2$ throughout the paper.

A classifier in multiclass learning is a universally measurable function $h : \mathcal{X} \to \mathcal{Y}$. For any probability distribution P on $\mathcal{X} \times \mathcal{Y}$, we define the error rate of h under P as

$$\mathrm{er}(h) = \mathrm{er}_P(h) := P(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : h(x) \neq y\}).$$

In this paper, we focus on realizable distributions: a distribution P is called ($\mathcal{H}$-)realizable if $\inf_{h \in \mathcal{H}} \mathrm{er}_P(h) = 0$. We use $\mathrm{RE}(\mathcal{H})$ to denote the set of all $\mathcal{H}$-realizable distributions. A multiclass learning algorithm is a sequence of universally measurable functions[1]

$$H_n : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathcal{Y}, \quad n \in \mathbb{N}_0.$$

For a sequence of independent P-distributed samples $((X_1, Y_1))_{i \in \mathbb{N}}$, the learning algorithm outputs a data-dependent function for each $n \in \mathbb{N}_0$

$$\widehat{h}_n : \mathcal{X} \to \mathcal{Y},\ x \mapsto H_n((X_1, Y_1), \ldots, (X_n, Y_n), x).$$

1. For notational convenience, we only defines deterministic algorithms here. However, our results still hold when randomized algorithms are allowed, as all algorithms we construct to show the upper bounds are deterministic and all proofs of lower bounds apply to randomized algorithms.

The objective of multiclass learning is to design a learning algorithm such that the expected error rate of the output classifier $\mathbf{E}[\mathrm{er}(\widehat{h}_n)]$ decreases as fast as possible with the size of the input sequence n. Since $\mathcal{X}$ is Polish and $\mathcal{Y}$ is countable, for $\mathcal{H}$ defined as the set of all measurable functions in $\mathcal{Y}^{\mathcal{X}}$, there exists a universally consistent learning algorithm, i.e., a learning algorithm such that $\mathbf{E}[\widehat{h}_n] \to 0$ for all realizable distributions P (Hanneke et al., 2021) [2]. Then, it is natural to ask about the rate of the convergence.

Under PAC learning, the uniform error rate over all realizable distributions is concerned. For multiclass learning, the following upper and lower bounds of the uniform rate is proved:

$$\Omega\left(\frac{\dim(\mathcal{H})}{n}\right) \leqslant \inf_{\widehat{h}_n} \sup_{P \in \mathrm{RE}(\mathcal{H})} \mathbf{E}[\mathrm{er}_P(\widehat{h}_n)] \leqslant O\left(\frac{\dim(\mathcal{H})^{3/2} \log^2(n)}{n}\right), \tag{1}$$

where the upper bound can be derived from the proof of Brukhim et al. (2022, Theorem 1) (see Corollary 64) and the lower bound can be found in Daniely and Shalev-Shwartz (2014). However, the worst-case analysis of PAC learning is too pessimistic to reflect many practical machine learning scenarios where the sample distribution keeps unchanged with the increase of the sample size, resulting in much faster decay in the error rate. Thus, Bousquet et al. (2021) proposed the concept of universal learning to characterize the distribution-dependent universal error rate of a hypothesis class. We state the definition of universal rates below.

Definition 1 (Universal rate, Bousquet et al. 2021, Definition 1.4) *Let $\mathcal{H}$ be a hypothesis class. Let $R : \mathbb{N} \to [0, 1]$ with $R(n) \to 0$ be a rate function.*

- *$\mathcal{H}$ is* learnable at rate R *if there is a learning algorithm $\widehat{h}_n$ such that for every realizable distribution P, there exist $C, c > 0$ for which $\mathbf{E}[\mathrm{er}(\widehat{h}_n)] \leqslant CR(cn)$ for all n.*
- *$\mathcal{H}$ is* not learnable at rate faster than R *if for every learning algorithm, there exists a realizable distribution P and $C, c > 0$ for which $\mathbf{E}[\mathrm{er}(\widehat{h}_n)] \geqslant CR(cn)$ for infinitely many n.*
- *$\mathcal{H}$ is* learnable with optimal rate R *if $\mathcal{H}$ is learnable at rate R and $\mathcal{H}$ is not learnable at rate faster than R.*
- *$\mathcal{H}$* is learnable but requires arbitrarily slow rates *if there is a learning algorithm $\widehat{h}_n$ such that $\mathbf{E}[\mathrm{er}(\widehat{h}_n)] \to 0$ for every realizable distribution P, and for every $R(n) \to 0$, $\mathcal{H}$ is not learnable faster than R.*

Note that in Definition 1, we define "$\mathcal{H}$ is learnable but requires arbitrarily rates" instead of defining "$\mathcal{H}$ requires arbitrarily rates" (Bousquet et al., 2021, Definition 1.4) to emphasize the existence of a universally consistent learning algorithm for $\mathcal{X}$ being Polish and $\mathcal{Y}$ being countable (Hanneke et al., 2021). Thus, the case that $\mathcal{H}$ is not universally learnable does not exist. As is formalized in the definition, the term "universal" refers to the requirement that the rate function R is universal for all realizable distributions. The major difference between universal rates and uniform rates is that the constants c and C can depend on the distribution P for universal rates, while the constants must be distribution-independent (i.e., uniform) for uniform rates. As is depicted in Bousquet et al. (2021, Figure 1), the distinction may results in the collapsing of exponential universal rates to linear uniform rates; e.g., in Example 1, we provide an example in multiclass learning where an exponential universal rate is achieved by the proposed algorithm, which is much faster than the

2. Actually, Hanneke et al. (2021) establishes the existence of a universally consistent learning algorithm assuming $\mathcal{X}$ is essentially separable and $\mathcal{Y}$ is countable. Any Polish space, being separably metrizable, is essentially separable.

linear uniform rate for finite label spaces. Bousquet et al. (2021) successfully characterized the fined-grained trichotomy in the optimal universal rates of binary classification problems, which motivates us to study the characterization of universal rates in multiclass learning with potentially infinite label spaces.

1.2. Main results

In this section, we state the main results together with some key definitions. First, we rule out some trivial hypothesis classes by considering $\mathcal{H}$ that is "nondegenerate" specified in the following definition.

Definition 2 (Nondegenerate hypothesis class) *A hypothesis class $\mathcal{H} \in \mathcal{Y}^{\mathcal{X}}$ is called* nondegenerate *if there exist $h_1, h_2 \in \mathcal{H}$ and $x, x' \in \mathcal{X}$ such that $h_1(x) = h_2(x)$ and $h_1(x') \neq h_2(x')$. $\mathcal{H}$ is called* degenerate *if it is not nondegenerate.*

Indeed, for $\mathcal{H}$ that is degenerate, if $h_1, h_2 \in \mathcal{H}$ satisfy $h_1 \neq h_2$, then, we have $h_1(x) \neq h_2(x)$ for any $x \in \mathcal{X}$. Thus, one sample suffices to reach zero error rate under any realizable distributions.

For the measurability of the learning algorithms we design in this paper, we need the following definition regarding the measurability of the hypothesis class $\mathcal{H}$.

Definition 3 (Measurable hypothesis class, Bousquet et al. 2021, Definition 3.3) *A hypothesis class $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathcal{Y}$ on Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ is said to be* measurable *if there is a Polish space Θ and a Borel-measurable map $\mathsf{h} : \Theta \times \mathcal{X} \to \mathcal{Y}$ so that $\mathcal{H} = \{\mathsf{h}(\theta, \cdot) : \theta \in \Theta\}$.*

As is discussed in Bousquet et al. (2021), the above definition is standard in the literature and almost any $\mathcal{H}$ considered in practice is measurable. Bousquet et al. (2021) and Kalavasis et al. (2022) also assume measurable hypothesis classes in their results.

The following theorem depicts the trichotomy in the universal rates of multiclass learning for general countable label spaces.

Theorem 4 *For any nondegenerate measurable hypothesis class $\mathcal{H}$, exactly one of the following holds:*

- *$\mathcal{H}$ is learnable with optimal rate e^{-n}.*
- *$\mathcal{H}$ is learnable with optimal rate in $\widetilde{\Theta}(1/n)$.*
- *$\mathcal{H}$ is learnable but requires arbitrarily slow rates.*

Then, we characterize the complexity measures of $\mathcal{H}$ that determine the universal rates of it: the nonexistence of certain tree structures of $\mathcal{H}$. We start with the Littlestone tree defined below.

Definition 5 (Littlestone tree) *A* Littlestone tree *for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a complete binary tree of depth $d \leqslant \infty$ whose internal nodes are labelled by $\mathcal{X}$, and whose two edges connecting a node to its two children are labelled by two different labels from $\mathcal{Y}$, such that every finite path emanating from the root is consistent with a concept $h \in \mathcal{H}$.*

Equivalently, a Littlestone tree of depth $d \leqslant \infty$ for $\mathcal{H}$ can also be represented as a collection

$$\left\{(x_{\mathbf{u}}, y_{\mathbf{u}}^0, y_{\mathbf{u}}^1) \in \widetilde{\mathcal{X}} : \mathbf{u} \in \{0,1\}^k, 0 \leqslant k < d\right\} \subseteq \widetilde{\mathcal{X}} := \{(x, y, y') \in \mathcal{X} \times \mathcal{Y}^2 : y \neq y'\} \tag{2}$$

such that for any $\boldsymbol{\eta} \in \{0,1\}^d$ and $0 \leqslant n < d$, there exists a concept $h \in \mathcal{H}$ such that $h(x_{\boldsymbol{\eta}_{\leqslant k}}) = y_{\boldsymbol{\eta}_{\leqslant k}}^{\eta_{k+1}}$ for each $0 \leqslant k \leqslant n$, where $\boldsymbol{\eta}_{\leqslant k} := (\eta_1, \ldots, \eta_k)$. We say that $\mathcal{H}$ has an *infinite Littlestone tree* if there is a Littlestone tree for $\mathcal{H}$ of depth $d = \infty$.

The definition of the Littlestone tree was first proposed by Daniely et al. (2015) to generalize the Littleston dimension to multiclass hypothesis classes, where they assume that $\mathcal{X}$ and $\mathcal{Y}$ are countable. Bousquet et al. (2021) restricted the definition to binary hypothesis classes and emphasized the difference between having of an infinite Littlestone tree and having an infinite Littlestone dimension (i.e., having Littlestone trees of arbitrarily large depth), where they prove that the nonexistence of the former distinguishes the exponential rate and the linear rate. Kalavasis et al. (2022) restricted the definition to multiclass hypothesis classes with finite label spaces ($|\mathcal{Y}| < \infty$) and proved that the nonexistence of an infinite Littlestone tree distinguishes the exponential rate and the linear rate for finite $\mathcal{Y}$.

Next, we introduce a new tree structure, the Daniely-Shalev-Shwartz-Littleston (DSL) tree which builds on the concept of pseudo-cubes in the definition of the DS dimension (Brukhim et al., 2022). For completeness, we state the definition of pseudo-cubes below.

Definition 6 (Pseudo-cube, Brukhim et al. 2022, Definition 5) *For any $d \in \mathbb{N}$, a class $C \subseteq \mathcal{Y}^d$ is called a* pseudo-cube *of dimension d if it is non-empty, finite, and for every $h \in C$ and $i \in [d]$, there is an i-neighbor of $g \in C$ of h (i.e., $g(i) \neq h(i)$ and $g(j) = h(j)$ for all $j \in [d]\backslash\{i\}$).*

For any $d \in \mathbb{N}$ and hypothesis class $H \subseteq \mathcal{Y}^d$, let $\mathsf{PC}(H)$ denote the collection of all d-dimensional pseudo-cubes contained in H. Then, we provide the definition of DSL trees below.

Definition 7 (DSL tree) *A* DSL tree *for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of depth $d \leqslant \infty$ is a tree of depth d satisfying the following properties.*

- *For each integer k such that $0 \leqslant k < d$ and each node v in level k of the tree (assume that the level of the root node is 0), node v is labelled with some $\mathbf{x} \in \mathcal{X}^{k+1}$. Moreover, there exists some pseudo-cube $C \in \mathsf{PC}(\mathcal{H}|_{\mathbf{x}})$ such that node v has exactly $|C|$ children and each edge connecting node v to its children is labelled with a unique element in C.*
- *For each integer k such that $0 \leqslant k < d$ and each node v in level k, denote the label of v with $\mathbf{x}_k \in \mathcal{X}^{k+1}$. Denote the labels of the nodes and the labels of the edges along the path emanating from the root node to node v with $\mathbf{x}_0 \in \mathcal{X}^1, \ldots, \mathbf{x}_{k-1} \in \mathcal{X}^k$ and $\mathbf{y}_0 \in \mathcal{Y}^1, \ldots, \mathbf{y}_{k-1} \in \mathcal{Y}^k$ correspondingly. Denote the number of the children of node v with n and the labels of the edges connecting node v to its children with $\mathbf{y}_{k,1}, \ldots, \mathbf{y}_{k,n} \in \mathcal{Y}^{k+1}$. Then, for each $i \in [n]$, there exists some $h \in \mathcal{H}$ such that $h|_{\mathbf{x}_t} = \mathbf{y}_t$ for all $0 \leqslant t \leqslant k-1$ and $h|_{\mathbf{x}_k} = \mathbf{y}_{k,i}$.*

Similarly, we say that $\mathcal{H}$ has an *infinite DSL tree* if there is a DSL tree for $\mathcal{H}$ of depth $d = \infty$. The definition of the DSL tree resembles those of the VCL tree (Bousquet et al., 2021, Definition 1.8), the NL tree (Kalavasis et al., 2022, Definition 6), and the GL tree (Kalavasis et al., 2022, Definition 8). Each node in level k is labelled with a sequence of $k+1$ points in $\mathcal{X}$ for $k \in \mathbb{N}_0$. However, for VCL trees and NL trees, the edges connecting a node to its children correspond to a copy of the Boolean-cube while they correspond to a pseudo-cube for DSL trees. Thus, the structure of a DSL tree is much more complicated since the sizes of pseudo-cubes of fixed dimension are not fixed, and it is hard to directly formulate a DSL tree like VCL trees or NL trees. For completeness and future reference, we state the definitions of the NL tree and the GL tree in Appendix A.2.

Now, we are ready to present the characterization of the multiclass universal rates in terms of those definitions.

Theorem 8 *For any nondegenerate measurable hypothesis class $\mathcal{H}$, the followings hold:*

- *If $\mathcal{H}$ does not have an infinite Littlestone tree, then $\mathcal{H}$ is learnable with optimal rate e^{-n}.*
- *If $\mathcal{H}$ has an infinite Littlestone tree but does not have an infinite DSL tree, then $\mathcal{H}$ is learnable at rate $\frac{\log^2 n}{n}$ and is not learnable at rate faster than $\frac{1}{n}$.*
- *If $\mathcal{H}$ has an infinite DSL tree, then $\mathcal{H}$ is learnable but requires arbitrarily slow rates.*

Since Theorem 4 follows immediately from Theorem 8, we directly prove Theorem 8 in this paper. A major difference between Theorem 8 and Kalavasis et al. (2022, Theorem 2) lies in the complexity measure that distinguishes the (near-)linear rate and arbitrarily slow rates: Kalavasis et al. (2022, Theorem 2) uses the nonexistence of an infinite NL tree. Then, a natural question is whether having an infinite DSL tree is equivalent to having an infinite NL tree for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $|\mathcal{Y}| = \infty$. Generalizing Brukhim et al. (2022, Theorem 2), we are able to show that they are not equivalent even for countably infinite $\mathcal{X}$ and $\mathcal{Y}$ in the following theorem.

Theorem 9 *There exist some countable sets $\mathcal{X}$ and $\mathcal{Y}$, and a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $\mathcal{H}$ has an infinite DSL tree but does not have any NL tree of depth 2.*

Thus, the nonexistence of an infinite NL tree does not distinguish the near-linear rate and arbitrarily slow rates for infinite label space ($|\mathcal{Y}| = \infty$).

We briefly comment on the $\log^2 n$ gap between the upper and lower bounds of the optimal universal rate in the second case (i.e., $\mathcal{H}$ has an infinite Littlestone tree but does not have an infinite DSL tree) of Theorem 8. It is worth pointing out that the $\frac{\log^2 n}{n}$ universal rate follows from the $\frac{\log^2 n}{n}$ uniform rate in (1). In fact, we prove in Theorem 66 that roughly speaking, a learning algorithm achieving some uniform rate for hypothesis classes with finite DS dimensions implies a learning algorithm achieving the same universal rate for any hypothesis class that does not have an infinite DSL tree. The $\frac{\log^2 n}{n}$ rate proved in Brukhim et al. (2022) is currently the sharpest uniform rate to our knowledge, and a sharper uniform rate will narrow the gap between the upper and lower bounds of the optimal universal rate. Nevertheless, the gap may also be narrowed by improving the lower bound. We list this problem as a future direction in Section 1.3.

Furthermore, we solve the first question in Kalavasis et al. (2022, Open question 1) which asks whether the existence of an infinite NL tree is equivalent to the existence of an infinite GL tree for finite label spaces ($|\mathcal{Y}| < \infty$). We prove that it is equivalent in the following theorem.

Theorem 10 *Let $K \in \mathbb{N} \backslash \{1\}$, and let $\mathcal{H} \subseteq [K]^{\mathcal{X}}$. Then, $\mathcal{H}$ has an infinite NL tree if and only if it has an infinite GL tree.*

Since it is not hard to see from definitions that a NL tree for $\mathcal{H}$ can be converted into a DSL tree for $\mathcal{H}$ of the same depth, and a DSL tree for $\mathcal{H}$ can be converted into a GL tree for $\mathcal{H}$ of the same depth, we immediately obtain the following corollary for $|\mathcal{Y}| < \infty$.

Corollary 11 *If $|\mathcal{Y}| < \infty$, then for any $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the followings are equivalent:*

- *$\mathcal{H}$ has an infinite NL tree.*
- *$\mathcal{H}$ has an infinite DSL tree.*
- *$\mathcal{H}$ has an infinite GL tree.*

Thus, the term "infinite Natarajan-Littlestone tree" in Kalavasis et al. (2022, Theorem 2) can be replaced with "infinite DSL tree" or "infinite GL tree".

1.3. Future direction

There are three immediate future directions following our current results in this paper. The first direction is to bridge the gap between the near-linear upper bound and linear lower bound of the optimal universal rate for hypothesis classes that have an infinite Littlestone tree but do not have an infinite DSL tree. As is already pointed out, tighter analysis of the uniform rate for hypothesis classes with finite DS dimensions would directly help in solving this problem. The second direction is to analyze the universal rates for uncountable label spaces. We believe that the major difficulty lies in proving the universal measurability of the learning algorithm constructed, and establishing the existence of a universally measurable learning algorithm that is universally consistent for general uncountable label spaces would shed light on this problem. Finally, it is an important next step to extend the results to the agnostic setting.

2. Examples

In this section, we present three examples in multiclass learning with different universal rates.

Example 1 (Multiclass linear classifier on $\mathbb{N}^d$) *For $d \in \mathbb{N}$, $K \in \mathbb{N}\backslash\{1\}$, $\mathcal{X} = \mathbb{N}^d$, and $\mathcal{Y} = [K]$, consider the following hypothesis class*

$$\mathcal{H} := \Big\{\mathcal{X} \to \mathcal{Y}, \mathbf{x} \mapsto \max(\operatorname*{argmax}_{k\in[K]} \mathbf{w}_k \cdot \mathbf{x} - b_k) : \\ \mathbf{w}_1 = \mathbf{0}, (\mathbf{w}_k)_j \leqslant (\mathbf{w}_{k+1})_j,\ \forall\, k \in [K], j \in [d],\ (b_1, \ldots, b_K) \in (0,\infty)^K\Big\}. \tag{3}$$

Consider any sequence $((\mathbf{x}_i, y_i))_{i\in\mathbb{N}} \in (\mathcal{X}\times\mathcal{Y})^\infty$ that is consistent with $\mathcal{H}$; i.e., for any $n \in \mathbb{N}$ and $S_n := ((\mathbf{x}_i, y_i))_{i\in[n]}$, there exists some $h_n \in \mathcal{H}$ with $h_n(\mathbf{x}_i) = y_i$ for all $i \in [n]$. For any $n \in \mathbb{N}$ and $\mathbf{x} \in \mathcal{X}$, we define the set

$$\mathsf{Y}_{S_n,\mathbf{x}} := \Big\{k \in [K] : \exists \mathbf{z}' \in [0,\infty)^d \text{ such that } \mathbf{x} - \mathbf{z}' \in \mathrm{Conv}(\{\mathbf{x}_i : (\mathbf{x}_i, k) \in S_n, i \in [n]\})\Big\}$$

where $\mathrm{Conv}(\varnothing) := \varnothing$ *and for any $t \in \mathbb{N}$ and set $\{\mathbf{z}_1, \ldots, \mathbf{z}_t\} \subseteq \mathcal{X}$,*

$$\mathrm{Conv}(\{\mathbf{z}_1, \ldots, \mathbf{z}_t\}) := \left\{\sum_{i=1}^t \alpha_i \mathbf{z}_t : (\alpha_1, \ldots, \alpha_t) \in [0,1]^t, \sum_{i=1}^t \alpha_i = 1\right\}$$

denotes the convex hull of the set $\{z_1, \ldots, z_t\}$. Then, we define the data-dependent classifier $\widehat{h}_n : \mathcal{X} \to \mathcal{Y}$ by

$$\widehat{h}_n(\mathbf{x}) := \begin{cases} \min \mathsf{Y}_{S_n,\mathbf{x}}, & \text{if } \mathsf{Y}_{S_n,\mathbf{x}} \neq \varnothing, \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

We prove the following proposition in Appendix G.

Proposition 12 *$(\widehat{h}_n)_{n\in\mathbb{N}}$ defined in (4) only makes finitely many mistakes for any consistent sequence $((\mathbf{x}_n, y_n))_{n\in\mathbb{N}}$. Moreover, if $\widehat{h}_n(\mathbf{x}_{n+1}) = y_{n+1}$, then we have $\widehat{h}_{n+1} = \widehat{h}_n$.*

Thus, by the construction and proofs given in Bousquet et al. (2021, Section 4.1), such an adversarial algorithm implies an online learning algorithm with exponential rate. By Theorem 8, $\mathcal{H}$ is learnable with optimal rate e^{-n} and $\mathcal{H}$ does not have an infinite Littlestone tree.

Example 2 (Multiclass linear classifier on $\mathbb{R}^d$) *For $d \in \mathbb{N}$, $K \in \mathbb{N}\backslash\{1\}$, $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{Y} = [K]$, consider the hypothesis class $\mathcal{H}$ defined by (3). Notice that the class of threshold functions constructed in Bousquet et al. (2021, Example 2.2) can be obtained from $\mathcal{H}$ by restricting $(\mathbf{w}_k)_1 = 1$, $(\mathbf{w}_k)_j = 0$, and $b_k = b_K$ for all $j \in [d]\backslash\{1\}$ and $k \in [K]\backslash\{1\}$. Thus, $\mathcal{H}$ has an infinite Littlestone tree. By Daniely and Shalev-Shwartz (2014, Theorem 7), we have that $\dim_N(\mathcal{H}) < \infty$. By Bendavid et al. (1995); Daniely and Shalev-Shwartz (2014), we have*

$$\dim(\mathcal{H}) \leqslant \dim_G(\mathcal{H}) \leqslant 5\log_2(K)\dim_N(\mathcal{H}) \tag{5}$$

which actually holds for any hypothesis class. It follows that $\dim(\mathcal{H}) < \infty$ and $\mathcal{H}$ does not have an infinite DSL tree. Then, by Theorem 8, $\mathcal{H}$ is learnable with optimal rate in $\widetilde{\Theta}(\frac{1}{n})$.

Example 3 (A class with an infinite DSL tree but no NL tree of depth 2) *Theorem 9 guarantees the existence of a hypothesis class $\mathcal{H}$ that has an infinite DSL tree but does not have any NL tree of depth 2 (see the proof of Theorem 9 in Appendix E for the construction of $\mathcal{H}$). Then, by Theorem 8, $\mathcal{H}$ is learnable but requires arbitrarily slow rates.*

3. Technical Overview

In this section, we briefly discuss some key technical points in the proofs of our main results.

3.1. Exponential rates

We sketch the proof of the following theorem in this subsection.

Theorem 13 *For any nondegenerate measurable hypothesis class $\mathcal{H}$, if $\mathcal{H}$ does not have an infinite Littlestone tree, then $\mathcal{H}$ is learnable with optimal rate e^{-n}.*

The complete proof is provided in Appendix B. Since $\mathcal{H}$ is nondegenerate, according to Bousquet et al. (2021, Lemma 4.2) and its proof, we can show that $\mathcal{H}$ is not learnable at rate faster than the exponential rate e^{-n}. The main point of the proof is to construct a learning algorithm that achieves the exponential universal rate if $\mathcal{H}$ does not have an infinite Littlestone tree. We follow the framework in Bousquet et al. (2021) for the construction. First, we consider an adversarial online learning game $\bar{\mathcal{B}}$ played in rounds between an adversary $\bar{P}_a$ and a learner $\bar{P}_l$ defined in Appendix B.1. If we prove that for $\mathcal{H}$ that does not have an infinite Littlestone tree, there exists a universally measurable strategy for the learner $\bar{P}_l$ in $\bar{\mathcal{B}}$ that only makes finitely many mistakes against any adversary $\bar{P}_a$ and only changes its prediction function when a mistake happens, then by the analysis in Bousquet et al. (2021, Section 4.1), there is a learning algorithm that achieves the exponential universal rate.

From (2), we can naturally relate Littlestone trees to the following adversarial game $\mathcal{B}$ between two players P_A and P_L. In each round $\tau \in \mathbb{N}$:

- Player P_A chooses a three-tuple $\xi_\tau = (x_\tau, y_\tau^0, y_\tau^1) \in \widetilde{\mathcal{X}}$ and shows it to Player P_L.
- Player P_L chooses a point $\eta_\tau \in \{0, 1\}$.

Player P_L wins the game in round $\tau \in \mathbb{N}$ if $\mathcal{H}_{\xi_1(1),\xi_1(\eta_1+2),\ldots,\xi_\tau(1),\xi_\tau(\eta_\tau+2)} = \varnothing$ (see Appendix A.1 for explanations of notation). Player P_A wins the game if the game continues indefinitely. We prove in Lemma 19 that a winning strategy of P_A is equivalent to an infinite Littlestone tree of $\mathcal{H}$. According to Bousquet et al. (2021, Theorem B.1), P_L has a universally measurable winning strategy if $\mathcal{H}$ has no infinite Littlestone tree. However, this winning strategy cannot be directly applied for the construction of a strategy for $\bar{P}_l$ in $\bar{\mathcal{B}}$ as in Bousquet et al. (2021, Section 3.2) because P_A chooses two labels y_τ^0 and y_τ^1 in each round τ while $\bar{P}_a$ does not provide this information (for the binary case, $\{y_\tau^0, y_\tau^1\}$ is trivially $\{0, 1\}$).

We tackle this problem by first defining the value function on the positions of $\bar{\mathcal{B}}$, which extends the value function defined on the positions of $\mathcal{B}$ (see Section B.1 for the terminologies and definitions). Then, by Bousquet et al. (2021, Proposition B.8), for each round in $\bar{\mathcal{B}}$, whatever point $\bar{P}_a$ picks, there is at most one point in $\mathcal{Y}$ such that the value function does not decrease. Then, we can define the function (7) which informally speaking, maps the current position and a point $x \in \mathcal{X}$ to the point in $\mathcal{Y}$ that does not decrease the value function. For Polish $\mathcal{X}$, countable $\mathcal{Y}$, and measurable $\mathcal{H}$, we prove that this function is universally measurable. Moreover, when $\mathcal{H}$ has no infinite Littlestone tree, we can prove that there is no infinite value-decreasing sequence of positions by the well-ordering of the ordinals (Karel and Thomas, 2017). Then, by playing the strategy induced from that defined function, $\bar{P}_l$ will only make finitely many mistakes because otherwise there will be an infinite value-decreasing sequence.

3.2. Near-linear rates

In this subsection, we sketch the proof of the following theorem.

Theorem 14 *For any nondegenerate measurable hypothesis class $\mathcal{H}$, if $\mathcal{H}$ has an infinite Littlestone tree but does not have an infinite DSL tree, then $\mathcal{H}$ is learnable at rate $\frac{\log^2 n}{n}$ and is not learnable at rate faster than $\frac{1}{n}$.*

The complete proof is provided in Appendix C. The fact that $\mathcal{H}$ is not learnable at rate faster than $\frac{1}{n}$ if it has an infinite Littlestone tree can be proved by generalizing the techniques used in the proof of Bousquet et al. (2021, Theorem 4.6). The key difficulty is to construct a learning algorithm that achieves $\frac{\log^2 n}{n}$ universal rate when $\mathcal{H}$ does not have an infinite DSL tree. As is discussed in Section 1.2, we show in Theorem 66 that a learning algorithm achieving some uniform rate for any hypothesis class with a finite DS dimension implies a learning algorithm achieving the same universal rate for any hypothesis class that does not have an infinite DSL tree. Since a learning algorithm that achieves $O(\frac{\dim(H)^{3/2}\log^2(n)}{n})$ uniform error rate for any hypothesis class $H \subseteq \mathcal{Y}^{\mathcal{X}}$ has been constructed (Brukhim et al., 2022), it suffices to prove Theorem 66. We follow the framework in Bousquet et al. (2021, Section 5). Similar to the case of exponential rates, we relate that the DSL tree to the following game $\mathfrak{B}$ between player P_A and P_L. At each round $\tau \in \mathbb{N}$:

- Player P_A chooses a sequence $\mathbf{x}_\tau = (x_\tau^0, \ldots, x_\tau^{\tau-1}) \in \mathcal{X}^\tau$ and a set $C_\tau \in \mathsf{PC}(\mathcal{Y}^\tau)$.
- Player P_L chooses a sequence $\mathbf{y}_\tau = (y_\tau^0, \ldots, y_\tau^{\tau-1}) \in \mathcal{Y}^\tau$.

Player P_L wins the game in round τ if

- either $C_\tau \notin \mathsf{PC}(\mathcal{H}|_{\mathbf{x}_\tau})$
- or $\mathbf{y}_s \in C_s$ for all $1 \leqslant s \leqslant \tau$ and $\mathcal{H}_{\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_\tau,\mathbf{y}_\tau} = \varnothing$, where

$$\mathcal{H}_{\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_\tau,\mathbf{y}_\tau} := \left\{h \in \mathcal{H} : h\left(x_s^i\right) = y_s^i \text{ for } 0 \leqslant i < s, 1 \leqslant s \leqslant \tau\right\}.$$

Player P_A wins the game if the game continues indefinitely. We emphasize the subtlety in the winning rule of P_L. In this way, we can ensure that $\mathfrak{B}$ is a Gale-Stewart game and an infinite DSL tree is equivalent to a winning strategy for P_A (Lemma 49). Similar to the analysis of exponential rates, there exists a mismatch between a winning strategy of P_L and a "pattern avoidance function" required in the template for constructing learning algorithms in the probabilistic setting in Bousquet et al. (2021, Section 5.2): in the adversarial learning problem, the adversary does not provide a pseudo-cube as P_A does. Thus, it is tricky to construct pattern avoidance functions which successfully rule out label patterns from their mappings of the feature patterns for any $\mathcal{H}$-consistent sequence in a finite number of steps, and keeps unchanged after the success. We provide our definition of pattern avoidance functions in (8). Informally, for a consistent sequence, given the current position in $\mathfrak{B}$ as well as the current feature pattern and label pattern from the sequence, we traverse all pseudo-cubes contained in the projection of $\mathcal{H}$ on the feature pattern, where by a feature (label) pattern we refer to a consecutive subsequence of the feature (label) sequence ending at the current point. If the value function defined on positions in $\mathfrak{B}$ deceases after adding the feature pattern, the current pseudo-cube, and label pattern into the position, we accept this new position, proceed one round in $\mathfrak{B}$, and stop the traverse. If the value function never decreases after the traverse, we still use the original position and does not change the round in $\mathfrak{B}$. Then, the feature pattern and label pattern are updated accordingly. Now, we define the current pattern avoidance function as the mapping from the current position and feature pattern to the set of all label patterns for which the position will be updated after traversing all the pseudo-cubes in the projection of $\mathcal{H}$ on the feature pattern. Then, with the similar idea of showing contradiction with nonexistence of infinite value-decreasing sequences, we can prove the desired pattern avoidance property of the set functions we defined. The next step is to show the universal measurability. Unlike the pattern avoidance function in Bousquet et al. (2021); Kalavasis et al. (2022), our pattern avoidance function maps to a set of patterns. This increases the difficulty in proving the universal measurability of the pattern avoidance functions we define since then we need to pay attention to the topology on power sets. One key point to notice is that since pseudo-cubes are finite by definition, $\mathsf{PC}(\mathcal{Y}^\tau)$ is countable as the set of finite subsets of a countable set is also countable. We can use this point to show that certain sets served as the building blocks in the pull-back set of the pattern avoidance functions are analytic. We also note that the universal measurability of the winning strategy for P_L or some value-decreasing function defined in $\mathfrak{B}$ does not obviously imply the universal measurability of the pattern avoidance functions since there are repetitions when feeding the data sequence as inputs to the game $\mathfrak{B}$; both Bousquet et al. (2021) and Kalavasis et al. (2022) does not provide a proof for this step (Bousquet et al., 2021, Remark 5.4). Thus, we provide an explicit and complete derivation of the universal measurability of the pattern avoidance functions that covers this step in our even more complicated setting, which also turns out to be very tricky.

There are still several big technical obstacles in plugging the pattern avoidance functions and a learning algorithm A with a uniform rate guarantee for hypothesis classes of finite DS dimensions into the template algorithm in Bousquet et al. (2021, Section 5.2). We first upper bound the DS dimension of the hypothesis class (11) constructed through a pattern avoidance function with its length (i.e., the length of the pattern the function seeks to avoid) in Lemma 58, where informally, (11) consists of projections of hypotheses in $\mathcal{H}$ on a given feature sequence such that any ordered subsequence of the projection is avoided by the pattern avoidance function. Then, we prove in Lemma 60 that informally, the uniform distribution over an independent and identically distributed (i.i.d.) data sequence that defines (11) is realizable with the class (11) almost surely if the pattern avoidance function that defines (11) avoids an i.i.d. data sequence with probability 1. Then, we would like to apply the the uniform learning algorithm A to (11) with that uniform distribution as the realizable distribution. However, for the usage of A in the template specified in Theorem 66, informally, given a sequence $(X_1, Y_1, \ldots, X_n, Y_n)$ and a feature X_{n+1}, the training data for A are drawn from the uniform distribution only over $\{(1, Y_1), \ldots, (n, Y_n)\}$ as we do not know Y_{n+1}, but the test data is always fixed, i.e., $n + 1$. In Lemma 65, we upper bound the error rate in this setting with twice the error rate in the standard setting (i.e., both the training data and the test data are drawn i.i.d. from the uniform distribution only over $\{(1, Y_1), \ldots, (n + 1, Y_{n+1})\}$ with Y_{n+1} being the label of X_{n+1}). Similar to Theorem 66, Lemma 65 is interesting in itself for dealing with partial training distributions.

3.3. Arbitrarily slow rates

In this subsection, we sketch the proof of the following theorem.

Theorem 15 *If $\mathcal{H}$ has an infinite DSL tree, then $\mathcal{H}$ is learnable but requires arbitrary slow rates.*

The complete proof is provided in Appendix D. The proof follows the framework for the construction of distributions in Bousquet et al. (2021, Theorem 5.11). Since for DSL trees, the numbers of the children of the nodes are not fixed in each level, to even formulate a uniform distribution over the paths in the infinite DSL tree is non-trivial. The key for the proof is to show (18), which holds trivially for both VCL trees (Bousquet et al., 2021) and NL trees (Kalavasis et al., 2022) since the labels of the edges connecting a node to its children consist a copy of the Boolean cube. However, such result for pseudo-cubes is novel; it actually implies an elegant proof for the $\Omega(\frac{\dim(\mathcal{H})}{n})$ lower bound of the uniform rate in (1). There are two key steps to show (18). We first prove that for any pseudo-cube, any position, and any label, the proportion of hypotheses in the pseudo-cube that maps that position to that label is at most half. Then, we prove that when restricting some arbitrary positions to some arbitrary pattern, a pseudo-cube, projected to the unrestricted positions, is still a pseudo-cube. Both steps follow from careful examination of the definition of pseudo-cubes.

Now, Theorem 8 directly follows from Theorem 13, Theorem 14, and Theorem 15.

3.4. Proof sketch of Theorem 9

The complete proof of Theorem 9 is provided in Appendix E. We use the disjoint pseudo-cubes of all dimensions on disjoint finite label spaces constructed in the proof of Brukhim et al. (2022, Theorem 2) as our starting point. We first build an infinite complete tree using these pseudo-cubes as blocks and take the disjoint unions to construct a countable label space, a countable feature space, and a hypothesis class. Then, we add to the label space a unique new element $\star$ used for extending the domain of a hypothesis to the whole feature space. Specifically, in a top-down manner of the tree constructed, we extend the definition of a hypothesis which corresponds to an edge in the tree to be consistent with the the hypotheses in the path eliminating from the root to its edge. Then, we define its value to be $\star$ on any other features. The fact that this class has an infinite DSL tree directly follow from the tree we constructed and the way we extend the definitions of hypotheses. Then, we prove that the class has a NL dimension 1 by considering the projection of the class on two arbitrary features, which requires more sophisticated discussion compared with the proof of Brukhim et al. (2022, Theorem 2).

3.5. Proof sketch of Theorem 10

The complete proof of Theorem 10 is provided in Appendix F. The fact that a GL tree can be obtained from a NL tree is obvious. The key is to construct an infinite NL tree from an infinite GL tree. For each node in the infinite GL tree except the root, we can associate it with a hypothesis in $\mathcal{H}$ that witnesses the requirement of the GL tree. Then, the rough idea is to construct for "each" node a distinct new sequence of labels for which each edge between this node and its children corresponds to a unique concept in the Boolean cube formed by this new sequence and the sequence provided by the GL tree, and the associated hypothesis of each descendant of this node along the path starting with this edge is consistent with the concept of the edge on this node. Here, by "each" we do not mean to construct for each element in the infinite GL tree, but we actually mean to select a node in the infinite GL tree for each position in the infinite NL tree to build.

We first deal with the consistency. In an infinite GL tree, for a node and an edge between the node and one of its children, the associated hypotheses of its descendants along the path starting with the chosen edge can predict differently on the chosen node, and the prediction can be used to color each descendant of the chosen node starting with the chosen edge. Then, we obtain an infinite colored subtree. Since $|\mathcal{Y}| = K < \infty$, the total number of colorings is finite. Thus, by the Milliken's tree theorem (Milliken, 1979), there is a strongly embedded subtree whose edges have the same color. But we still need to prune this subtree so that it has the same structure as the original subtree, after which we replace the original subtree with the monochromatic subtree. Now, the prediction made by the associated hypothesis of each descendant along the path starting with the chosen edge is the same on the chosen node. This step is formally presented in Lemma 70.

However, we still face the fact that the predictions specified for each edge of a given node in the previous step do not necessary make a copy of a Boolean cube. For this problem, we observe that all the predictions make a hypothesis class with its Graph dimension greater than d for d denoting the length of the feature sequence of the given node. By (5), this class has a Natarajan dimension greater than $d/(5\log_2(K))$, which implies a Boolean cube of dimension greater than $d/(5\log_2(K))$. Thus, by skipping $\lceil 5\log_2(K) \rceil$ levels in choosing nodes from the infinite GL tree in a top-down manner, we are able to ensure the existence of a copy of a Boolean cube of required dimension for each level of the NL tree constructed by some proper pruning. The proof is formalized by induction.

Acknowledgments

Shay Moran is a Robert J. Shillman Fellow; he acknowledges support by ISF grant 1225/20, by BSF grant 2018385, by an Azrieli Faculty Fellowship, by Israel PBC-VATAT, by the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by the 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.

References

S. Bendavid, N. Cesabianchi, D. Haussler, and P.M. Long. Characterizations of learnability for classes of {0, ..., n}-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86,

1995. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1995.1008. URL https://www.sciencedirect.com/science/article/pii/S0022000085710082.

Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, oct 1989. ISSN 0004-5411. doi: 10.1145/76359.76371. URL https://doi.org/10.1145/76359.76371.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 532–541, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451087. URL https://doi.org/10.1145/3406325.3451087.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *Journal of Machine Learning Research*, 16:2377–2404, 2015.

Sabine Dietmann and Liisa Holm. Identification of homology in protein structure classification. *nature structural biology*, 8(11):953–957, 2001.

David Gale and F. M. Stewart. Infinite games with perfect information. In *Contributions to the theory of games, vol. 2*, Annals of Mathematics Studies, no. 28, pages 245–266. Princeton University Press, Princeton, N.J., 1953.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129 – 2150, 2021. doi: 10.1214/20-AOS2029. URL https://doi.org/10.1214/20-AOS2029.

Daniel Hellerstein and Robert Mendelsohn. A theoretical foundation for count data models. *American journal of agricultural economics*, 75(3):604–611, 1993.

Alkis Kalavasis, Grigoris Velegkas, and Amin Karbasi. Multiclass learnability beyond the pac framework: Universal rates and partial concept classes. *arXiv preprint arXiv:2210.02297*, 2022.

Hrbacek Karel and Jech Thomas. *Introduction to Set Theory: Revised and Expanded*. Crc Press, 2017.

Keith R. Milliken. A ramsey theorem for trees. *Journal of Combinatorial Theory, Series A*, 26(3):215–237, 1979. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(79)90101-8. URL https://www.sciencedirect.com/science/article/pii/0097316579901018.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

Balas K Natarajan and Prasad Tadepalli. Two new frameworks for learning. In *Machine Learning Proceedings 1988*, pages 402–415. Elsevier, 1988.

Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017. doi: 10.1162/neco_a_00990.

Fei Song and W Bruce Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321, 1999.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, nov 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018. doi: 10.1109/MCI.2018.2840738.

Appendix A. Preliminaries

In this section, we describe the notation used in this paper and present the definitions of NL trees and GL trees in the general multiclass setting.

A.1. Notation

We use the following notation throughout the paper. $\mathbb{N}$ denotes the set of positive integers. $\mathbb{N}_0$ denotes the set of non-negative integers. For any $n \in \mathbb{N}$, we define $[n] := \{1, \ldots, n\}$. For any $a, b \in \mathbb{R}$, we define $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For a set A, $|A|$ denotes its cardinality and 2^A denotes its power set. For any sets X, Y and hypothesis class $F \subseteq Y^X$, let $\dim(F)$ denote the Daniely-Shalev-Shwartz (DS) dimension of F, $\dim_N(F)$ denote the Natarajan dimension of F, and $\dim_G$ denote the Graph dimension of F. For any $n \in \mathbb{N}$, any sequence $S = (x_1, \ldots, x_n) \in X^n$, and any function $f : X \to Y$, we define the projection of f to S as $f|_S := (f(x_1), \ldots, f(x_n)) \in Y^n$ and use $S(i)$ to denote the i-th element in S (i.e., $S(i) = x_i$) for any $i \in [n]$. By convention, $f|_\varnothing = \varnothing$. Then, we define the projection of $F \subseteq Y^X$ to S as $F|_S := \{f|_S : f \in F\} \subseteq Y^n$. For any $x_1, \ldots, x_n \in X$ and $y_1, \ldots, y_n \in Y$, we define

$$F_{x_1,y_1,\ldots,x_n,y_n} := \{f \in F : f(x_1) = y_1, \ldots, f(x_n) = y_n\}.$$

A.2. NL trees and GL trees

In this section, we define NL trees and GL trees for the general multiclass setting.

Definition 16 (NL tree) *A Natarajan-Littlestone (NL) tree for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of depth $d \leqslant \infty$ is the following collection*

$$\cup_{0\leqslant n<d}\left\{(\mathbf{x_u}, \mathbf{s}_\mathbf{u}^{(0)}, \mathbf{s}_\mathbf{u}^{(1)}) \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} \times \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\right\}$$

such that for any $0 \leqslant n < d$ and $\mathbf{u} = (u_1^0, (u_2^0, u_2^1), \ldots, (u_n^0, \ldots, u_n^{n-1})) \in \prod_{l=1}^n\{0,1\}^l$, the followings hold:

- $s_\mathbf{u}^{(0)i} \neq s_\mathbf{u}^{(1)i}$ *for all $0 \leqslant i \leqslant n$.*
- *If $n \geqslant 1$, then there exists some $h_\mathbf{u} \in \mathcal{H}$ such that $h_\mathbf{u}(x^i_{\mathbf{u}_{\leqslant l}}) = s^{(0)i}_{\mathbf{u}_{\leqslant l}}$ if $u^i_{l+1} = 0$ and $h_\mathbf{u}(x^i_{\mathbf{u}_{\leqslant l}}) = s^{(1)i}_{\mathbf{u}_{\leqslant l}}$ otherwise for all $0 \leqslant i \leqslant l$ and $0 \leqslant l < n$, where*

$$\mathbf{u}_{\leqslant l} := (u_1^0, (u_2^0, u_2^1), \ldots, (u_l^0, \ldots, u_l^{l-1})),\ x_{\mathbf{u}_{\leqslant l}} := (x^0_{\mathbf{u}_{\leqslant 1}}, \ldots, x^l_{\mathbf{u}_{\leqslant 1}}).$$

We call

$$\cup_{1\leqslant n<d}\left\{h_\mathbf{u} \in \mathcal{H} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\right\}$$

the associated hypothesis set of the NL tree. We say that $\mathcal{H}$ has an infinite NL tree if it has a NL tree of depth $d = \infty$.

Definition 17 (GL tree) *A Graph-Littlestone (GL) tree for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of depth $d \leqslant \infty$ is the following collection*

$$\cup_{0\leqslant n<d}\left\{(\mathbf{x_u}, \mathbf{s_u}) \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\right\}$$

such that for any $0 \leqslant n < d$ and $\mathbf{u} = (u_1^0, (u_2^0, u_2^1), \ldots, (u_n^0, \ldots, u_n^{n-1})) \in \prod_{l=1}^n\{0,1\}^l$, the following holds:

- *If $n \geqslant 1$, then there exists some $h_\mathbf{u} \in \mathcal{H}$ such that $h_\mathbf{u}(x^i_{\mathbf{u}_{\leqslant l}}) = s^i_{\mathbf{u}_{\leqslant l}}$ if $u^i_{l+1} = 0$ and $h_\mathbf{u}(x^i_{\mathbf{u}_{\leqslant l}}) \neq s^i_{\mathbf{u}_{\leqslant l}}$ otherwise for all $0 \leqslant i \leqslant l$ and $0 \leqslant l < n$, where*

$$\mathbf{u}_{\leqslant l} := (u_1^0, (u_2^0, u_2^1), \ldots, (u_l^0, \ldots, u_l^{l-1})),\ x_{\mathbf{u}_{\leqslant l}} := (x^0_{\mathbf{u}_{\leqslant 1}}, \ldots, x^l_{\mathbf{u}_{\leqslant 1}}).$$

We call

$$\cup_{1\leqslant n<d}\left\{h_\mathbf{u} \in \mathcal{H} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\right\}$$

the associated hypothesis set of the GL tree. We say that $\mathcal{H}$ has an infinite GL tree if it has a GL tree of depth $d = \infty$.

Appendix B. Exponential Rates

In this section, we prove Theorem 13.

B.1. Adversarial learning algorithm

We propose and analyze an adversarial learning algorithm in this section. Define $\widetilde{\mathcal{Y}^2} := \{(y, y') \in \mathcal{Y}^2 : y \neq y'\}$ and $\widetilde{\mathcal{X}} := \mathcal{X} \times \widetilde{\mathcal{Y}^2}$. For any $\xi \in \widetilde{\mathcal{X}}$, there exist $x \in \mathcal{X}$ and $y^0,\ y^1 \in \mathcal{Y}$ such that $\xi = (x, y^0, y^1)$. Then, we let $\xi(1)$ denote x and $\xi(i+2)$ denote y^i for $i \in \{0, 1\}$. Then, a Littlestone tree can be equivalently represented as the following collection

$$\left\{\xi_{\mathbf{u}} = (x_{\mathbf{u}}, y^0_{\mathbf{u}}, y^1_{\mathbf{u}}) \in \widetilde{\mathcal{X}} : \mathbf{u} \in \{0,1\}^k, 0 \leqslant k < d\right\} \subseteq \widetilde{\mathcal{X}}$$

such that for any $\boldsymbol{\eta} \in \{0,1\}^d$ and $0 \leqslant n < d$, there exists a concept $h \in \mathcal{H}$ such that $h(\xi_{\boldsymbol{\eta}_{\leqslant k}}(1)) = \xi_{\boldsymbol{\eta}_{\leqslant k}}(\eta_{k+1} + 2)$.

For the multiclass online learning problem, we can define the following online learning game $\bar{\mathcal{B}}$ played in rounds between an adversary $\bar{P}_a$ and the learner $\bar{P}_l$. In each round $t \geqslant 1$:

- The adversary $\bar{P}_a$ chooses a point $x_t \in \mathcal{X}$.
- The learner $\bar{P}_l$ makes a prediction $\widehat{y}_t \in \mathcal{Y}$.
- The adversary $\bar{P}_a$ reveals the true label $y_t = h(x_t)$ for some concept $h \in \mathcal{H}$ such that h is consistent with the previous points: $y_1 = h(x_1),\ \ldots,\ y_{t-1} = h(x_{t-1})$.

We would like to prove the following theorem.

Theorem 18 *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. For any hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, we have the following dichotomy.*

- *If $\mathcal{H}$ has an infinite Littlestone tree, then there is a strategy for the adversary $\bar{P}_a$ in $\bar{\mathcal{B}}$ such that $\widehat{y}_t \neq y_t$ in each round $t \geqslant 1$ against any learner $\bar{P}_l$.*
- *If $\mathcal{H}$ does not have an infinite Littlestone tree, then there is a strategy for the learner $\bar{P}_l$ in $\bar{\mathcal{B}}$ that only makes finitely many mistakes against any adversary $\bar{P}_a$.*

Consider the following game $\mathcal{B}$ between two players P_A and P_L. In each round $\tau \in \mathbb{N}$:

- Player P_A chooses a three-tuple $\xi_\tau = (x_\tau, y^0_\tau, y^1_\tau) \in \widetilde{\mathcal{X}}$ and shows it to Player P_L.
- Player P_L chooses a point $\eta_\tau \in \{0, 1\}$.

We say that player P_L wins the game in round $\tau \in \mathbb{N}$ if $\mathcal{H}_{\xi_1(1),\xi_1(\eta_1+2),\ldots,\xi_\tau(1),\xi_\tau(\eta_\tau+2)} = \varnothing$, where $\mathcal{H}_{x_1,y_1,\ldots,x_t,y_t} := \{h \in \mathcal{H} : h(x_1) = y_1, \ldots, h(x_t) = y_t\}$ for any $x_1, \ldots, x_t \in \mathcal{X}$ and $y_1, \ldots, y_t \in \mathcal{Y}$. We say that player P_A wins the game if the game continues indefinitely. We say a strategy for P_A is winning if playing that strategy, P_A wins the game no matter what strategy P_L plays. We define a winning strategy for P_L analogously. According to the rule of $\mathcal{B}$, the set of winning sequence of P_L is

$$\mathsf{W} := \left\{(\boldsymbol{\xi}, \boldsymbol{\eta}) \in \left(\widetilde{\mathcal{X}} \times \{0,1\}\right)^{\infty} : \mathcal{H}_{\xi_1(1),\xi_1(\eta_1+2),\ldots,\xi_\tau(1),\xi_\tau(\eta_\tau+2)} = \varnothing \text{ for some } \tau \in \mathbb{N}\right\}.$$

Since W is finitely decidable (i.e., for any $(x_1, y_1, x_2, y_2, \ldots) \in \mathsf{W}$, there exists $n \in \mathbb{N}$ such that $(x_1, y_1, \ldots, x_n, y_n, x'_{n+1}, y'_{n+1}, \ldots) \in \mathsf{W}$ for all $(x'_{n+1}, y'_{n+1}, x'_{n+2}, y'_{n+2}, \ldots) \in (\mathcal{X} \times \mathcal{Y})^\infty$), $\mathcal{B}$ is a Gale-Stewart game; then, either P_A or P_L has a winning strategy (Gale and Stewart, 1953). We refer readers to Bousquet et al. (2021, Appendix A.1) for detailed descriptions of the notion we use above and Gale-Stewart games.

Then, we prove the following lemma that relates a winning strategy of P_A to an infinite Littlestone tree of $\mathcal{H}$.

Lemma 19 *Player P_A has a winning strategy in the game $\mathcal{B}$ if and only if $\mathcal{H}$ has an infinite Littlestone tree.*

Proof Suppose that $\mathcal{H}$ has an infinite Littlestone tree represented by

$$\left\{(x_{\mathbf{u}}, y^0_{\mathbf{u}}, y^1_{\mathbf{u}}) : 0 \leqslant k < \infty, \mathbf{u} \in \{0,1\}^k\right\}.$$

Define a strategy for P_A by $\xi_\tau(\eta_1, \ldots, \eta_{\tau-1}) := (x_{\eta_1,\ldots,\eta_{\tau-1}}, y^0_{\eta_1,\ldots,\eta_{\tau-1}}, y^1_{\eta_1,\ldots,\eta_{\tau-1}})$ for any $\tau \in \mathbb{N}$. By the definition of Littlestone tree, we have that $\mathcal{H}_{\xi_1(1),\xi_1(\eta_1+2),\ldots,\xi_\tau(1),\xi_\tau(\eta_\tau+2)} \neq \varnothing$ for any $\tau \in \mathbb{N}$. Thus, P_A has a winning strategy.

Suppose that P_A has a winning strategy $\xi_\tau(\eta_1, \ldots, \eta_{\tau-1})$ for any $\eta_1, \ldots, \eta_{\tau-1} \in \{0,1\}$ and $1 \leqslant \tau < \infty$. Define an infinite binary tree represented by $\left\{(x_{\mathbf{u}}, y^0_{\mathbf{u}}, y^1_{\mathbf{u}}) : \mathbf{u} \in \{0,1\}^k, 0 \leqslant k < \infty\right\}$ with

$$(x_{u_1,\ldots,u_k}, y^0_{u_1,\ldots,u_k}, y^1_{u_1,\ldots,u_k}) := \xi_{k+1}(u_1, \ldots, u_k).$$

By the definition of winning strategy of P_A in $\mathcal{B}$, the tree defined above is an infinite Littlestone tree of $\mathcal{H}$. ■

For any $n \in \mathbb{N}_0$, define $\mathsf{P}_n := \left(\tilde{\mathcal{X}} \times \{0,1\}\right)^n$ to be the set of positions of length n in the game $\mathcal{B}$, where a position of a game is a finite sequence of plays made by the two players alternatively from the start to some round and $\mathsf{P}_0 = \varnothing$ by convention. A position is called active if P_L has not won yet after this position. Then, the set of active positions of length n in the game $\mathcal{B}$ can be written as

$$\mathsf{A}_n := \cup_{\mathbf{w} \in (\tilde{\mathcal{X}} \times \{0,1\})^\infty} \{\mathbf{v} \in \mathsf{P}_n : (\mathbf{v}, \mathbf{w}) \in \mathsf{W}^c\}.$$

Then, we define $\mathsf{P} := \cup_{0 \leqslant n < \infty} \mathsf{P}_n$ to be the set of all positions and $\mathsf{A} := \cup_{0 \leqslant n < \infty} \mathsf{A}_n$ to be the set of all active positions in the game $\mathcal{B}$.

Analogously, for any $n \in \mathbb{N}_0$, we define $\bar{\mathsf{P}}_n := (\mathcal{X} \times \mathcal{Y})^n$ to be the set of all positions of length n in the game $\bar{\mathcal{B}}$. For notational convenience, we also define $\mathsf{P}_\infty := \left(\tilde{\mathcal{X}} \times \{0,1\}\right)^\infty$ and $\bar{\mathsf{P}}_\infty := (\mathcal{X} \times \mathcal{Y})^\infty$.

As in Bousquet et al. (2021), we need to describe for how many rounds the game can be kept active starting from an arbitrary position. The following definitions of decision trees and active decision trees are the direct restriction of Bousquet et al. (2021, Definition B.4) for P in our setting.

Definition 20 (Bousquet et al. 2021, Definition B.4) *Given a position $\mathbf{v} \in \mathsf{P}_k$ of length $k \in \mathbb{N}_0$:*

- *A decision tree of depth n with starting position $\mathbf{v}$ is a collection of points*

$$\mathbf{t} = \left\{\xi_{\boldsymbol{\eta}} \in \tilde{\mathcal{X}} : \boldsymbol{\eta} \in \{0,1\}^t, 0 \leqslant t < n\right\}.$$

By convention, we call $\mathbf{t} = \varnothing$ a decision tree of depth 0.

- $\mathbf{t}$ *is called active if* $(\mathbf{v}, \xi_{\varnothing}, \eta_{k+1}, \xi_{\eta_{k+1}}, \eta_{k+2}, \ldots, \xi_{\eta_{k+1},\ldots,\eta_{k+n-1}}, \eta_{k+n}) \in \mathsf{A}_{k+n}$ *for all choices of* $(\eta_{k+1}, \ldots, \eta_{k+n}) \in \{0,1\}^n$.
- *We denote by* $\mathsf{T}_{\mathbf{v}}$ *the set of all decision trees with starting position* $\mathbf{v}$ *(and any depth* $n \in \mathbb{N}_0$*), and by* $\mathsf{T}_{\mathbf{v}}^{\mathsf{A}} \subseteq \mathsf{T}_{\mathbf{v}}$ *the set of all active trees.*

Note that $\mathsf{T}_{\mathbf{v}} = \mathsf{T}_{\mathbf{v}'}$ for any $\mathbf{v}, \mathbf{v}' \in \mathsf{P}$ by the above definition.

We use ORD to denote the set of of all ordinals. We use -1 to denote an element that is smaller than every ordinal and Ω to denote an element that is are larger than every ordinal. Define $\mathrm{ORD}^* := \mathrm{ORD} \cup \{\Omega, -1\}$. We refer readers to Bousquet et al. (2021, Appendix A) for brief introductions about the concepts of ordinals, well-founded relations, ranks, Polish spaces, universally measurability, analytic sets, etc.

For any $\mathbf{v} \in \mathsf{A}$, we define a relation $\prec_{\mathbf{v}}$ on $\mathsf{T}_{\mathbf{v}}^{\mathsf{A}}$. For $\mathbf{t}, \mathbf{t}' \in \mathsf{T}_{\mathbf{v}}^{\mathsf{A}}$, we say that $\mathbf{t}' \prec_{\mathbf{v}} \mathbf{t}$ if and only if the tree $\mathbf{t}$ is obtained from $\mathbf{t}'$ by removing its leaves. Let $\rho_{\prec_{\mathbf{v}}} : \mathsf{T}_{\mathbf{v}}^{\mathsf{A}} \to \mathrm{ORD}$ denote the rank function of the relation $\prec_{\mathbf{v}}$. Then, we define the following game value of on P as in Bousquet et al. (2021).

Definition 21 (Bousquet et al. 2021, Definition B.5) *The* game value $\mathrm{val} : \mathsf{P} \to \mathrm{ORD}^*$ *is defined as follows.*

- $\mathrm{val}(\mathbf{v}) = -1$ *if* $\mathbf{v} \notin \mathsf{A}$.
- $\mathrm{val}(\mathbf{v}) = \Omega$ *if* $\mathbf{v} \in \mathsf{A}$ *and* $\prec_{\mathbf{v}}$ *is not well-founded.*
- $\mathrm{val}(\mathbf{v}) = \rho_{\prec_{\mathbf{v}}}(\varnothing)$ *if* $\mathbf{v} \in \mathsf{A}$ *and* $\prec_{\mathbf{v}}$ *is well-founded.*

According to Lemma 19 and Definition 21, we have the following Lemma about $\mathrm{val}(\varnothing)$.

Lemma 22 *We have* $\mathrm{val}(\varnothing) > -1$. *If* $\mathcal{H}$ *does not have an infinite Littlestone tree, then* $\mathrm{val}(\varnothing) < \Omega$.

Proof Obviously, $\varnothing \in \mathsf{A}$. If $\mathcal{H}$ does not have an infinite Littlestone tree, by Lemma 19, P_A does not have a winning strategy. Thus, $\prec_{\varnothing}$ is well-defined. By Definition 21, we have $\mathrm{val}(\varnothing) < \Omega$. ■

In order to define game values on $\bar{\mathsf{P}}$, we prove the following lemma.

Lemma 23 *For any* $k_1, k_2 \in \mathbb{N}_0$, $\mathbf{v}_a \in \mathsf{P}_{k_1}$, $\mathbf{v}_b \in \mathsf{P}_{k_1}$, $x \in \mathcal{X}$, *and* y, y', $y'' \in \mathcal{Y}$ *such that* $y' \neq y$ *and* $y'' \neq y$, *we have* $\mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y,y'),0,\mathbf{v}_b)} = \mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y'',y),1,\mathbf{v}_b)} = \mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y,y''),0,\mathbf{v}_b)}$. *In particular,* $\mathrm{val}(\mathbf{v}_a, (x,y,y'), 0, \mathbf{v}_b) = \mathrm{val}(\mathbf{v}_a, (x,y'',y), 1, \mathbf{v}_b) = \mathrm{val}(\mathbf{v}_a, (x,y,y''), 0, \mathbf{v}_b)$.

Proof It suffices to show that $\mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y,y'),0,\mathbf{v}_b)} = \mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y'',y),1),\mathbf{v}_b}$ and $\mathrm{val}(\mathbf{v}_a, (x,y,y'), 0, \mathbf{v}_b) = \mathrm{val}(\mathbf{v}_a, (x,y'',y), 1, \mathbf{v}_b)$. Indeed, since $y \neq y''$, the above results immediately imply that

$$\mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y,y''),0,\mathbf{v}_b)} = \mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y'',y),1,\mathbf{v}_b)} = \mathsf{T}^{\mathsf{A}}_{(\mathbf{v}_a,(x,y,y'),0,\mathbf{v}_b)}$$

and

$$\mathrm{val}(\mathbf{v}_a, (x,y,y''), 0, \mathbf{v}_b) = \mathrm{val}(\mathbf{v}_a, (x,y'',y), 1, \mathbf{v}_b) = \mathrm{val}(\mathbf{v}_a, (x,y,y'), 0, \mathbf{v}_b).$$

Let $k = k_1 + k_2$. Since $\mathbf{v}_a \in \mathsf{P}_{k_1}$ and $\mathbf{v}_b \in \mathsf{P}_{k_2}$, we have $\mathbf{v}_a = (\xi_1, \eta_1, \ldots, \xi_{k_1}, \eta_{k_1})$ and $\mathbf{v}_b = (\xi_{k_1+2}, \eta_1, \ldots, \xi_{k+1}, \eta_{k+1})$ for some $(\xi_1, \ldots, \xi_{k_1}) \in \widetilde{\mathcal{X}}^{k_1}$, $(\xi_{k_1+2}, \ldots, \xi_{k+1}) \in \widetilde{\mathcal{X}}^{k_2}$, $(\eta_1, \ldots, \eta_{k-1}) \in \{0,1\}^{k_1}$, and $(\eta_{k_1+2}, \ldots, \eta_{k+1}) \in \mathcal{Y}^{k_2}$. Define $\xi^0_{k_1+1} := (x, y, y')$, $\eta^0_{k_1+1} := 0$,

$\xi^1_{k_1+1} := (x, y'', y)$, $\eta^1_{k_1+1} := 1$, $\mathbf{v}_0 := (\mathbf{v}_a, (x, y, y'), 0, \mathbf{v}_b)$, and $\mathbf{v}_1 := (\mathbf{v}_a, (x, y'', y), 1, \mathbf{v}_b)$. For any decision tree

$$\mathbf{t} = \left\{\xi_{\boldsymbol{\eta}} \in \widetilde{\mathcal{X}} : \boldsymbol{\eta} \in \{0,1\}^t, 0 \leqslant t < n\right\} \in \mathsf{T}_{\mathbf{v}_0}$$

of depth n ($0 \leqslant n < \infty$), we have $\mathbf{t} \in \mathsf{T}_{\mathbf{v}_1}$.

If $\mathbf{t} \in \mathsf{T}^{\mathsf{A}}_{\mathbf{v}_0}$, for any $\boldsymbol{\eta} = (\eta_{k+2}, \dots, \eta_{k+n+1}) \in \{0,1\}^n$, we have

$$\mathbf{v}_{0,\mathbf{t},\boldsymbol{\eta}} := (\mathbf{v}_0, \xi_{\varnothing}, \eta_{k+2}, \xi_{\eta_{k+2}}, \dots, \xi_{\eta_{k+2},\dots,\eta_{k+n}}, \eta_{k+n+1}) \in \mathsf{A}_{k+n+1}.$$

By the definition of A_{k+n+1}, there exists

$$\mathbf{w} = (\xi_{k+n+2}, \eta_{k+n+2}, \xi_{k+n+3}, \eta_{k+n+3}, \dots) \in \left(\widetilde{\mathcal{X}} \times \{0,1\}\right)^{\infty}$$

such that $(\mathbf{v}_{0,\mathbf{t},\boldsymbol{\eta}}, \mathbf{w}) \in \mathsf{W}^c$.

For each $t \in [k_1] \cup \{k_1+2, \dots, k+1\} \cup \{k+n+2, k+n+3, \dots\}$, define $x_t := \xi_t(1)$ and $y_t := \xi_t(\eta_t + 2)$. Define $x_{k_1+1} := \xi^0_{k_1+1}(1) = x$ and $y_{k_1+1} := \xi^0_{k_1+1}(\eta^0_{k_1+1} + 2) = y$. Define $x_{k+2} := \xi_{\varnothing}(1)$ and $y_{k+2} := \xi_{\varnothing}(\eta_{k+2} + 2)$. For each $t \in \{k+3, \dots, k+n+1\}$, define $x_t := \xi_{\eta_{k+2},\dots,\eta_{t-1}}(1)$ and $y_t := \xi_{\eta_{k+2},\dots,\eta_{t-1}}(\eta_t + 2)$.

Since $(\mathbf{v}_{0,\mathbf{t},\boldsymbol{\eta}}, \mathbf{w}) \in \mathsf{W}^c$, by the definition of W, for any $0 \leqslant \tau < \infty$, there exists $h \in \mathcal{H}$ such that $h(x_t) = y_t$ for any $1 \leqslant t \leqslant \tau$. Since $\xi^1_{k_1+1}(1) = x = \xi^0_{k_1+1}(1) = x_{k_1+1}$ and $\xi^1_{k_1+1}(\eta^1_{k_1+1} + 2) = y = \xi^0_{k_1+1}(\eta^0_{k_1+1} + 2) = y_{k_1+1}$, we have $(\mathbf{v}_{1,\mathbf{t},\boldsymbol{\eta}}, \mathbf{w}) \in \mathsf{W}^c$ where

$$\mathbf{v}_{1,\mathbf{t},\boldsymbol{\eta}} := (\mathbf{v}_1, \xi_{\varnothing}, \eta_{k+2}, \xi_{\eta_{k+2}}, \dots, \xi_{\eta_{k+2},\dots,\eta_{k+n}}, \eta_{k+n+1}).$$

Thus, $\mathbf{v}_{1,\mathbf{t},\boldsymbol{\eta}} \in \mathsf{A}_{k+n+1}$ for any $\boldsymbol{\eta} \in \{0,1\}^n$. By the definition of $\mathsf{T}^{\mathsf{A}}_{\mathbf{v}_1}$, we have $\mathbf{t} \in \mathsf{T}^{\mathsf{A}}_{\mathbf{v}_1}$. Since it holds for any $\mathbf{t} \in \mathsf{T}^{\mathsf{A}}_{\mathbf{v}_0}$, we have $\mathsf{T}^{\mathsf{A}}_{\mathbf{v}_0} \subseteq \mathsf{T}^{\mathsf{A}}_{\mathbf{v}_1}$.

By symmetry, we can also show that $\mathsf{T}^{\mathsf{A}}_{\mathbf{v}_1} \subseteq \mathsf{T}^{\mathsf{A}}_{\mathbf{v}_0}$, which implies that $\mathsf{T}^{\mathsf{A}}_{\mathbf{v}_0} = \mathsf{T}^{\mathsf{A}}_{\mathbf{v}_1}$. Since $\mathsf{T}_{\mathbf{v}_0} = \mathsf{T}_{\mathbf{v}_1}$, we also have $\mathrm{val}(\mathbf{v}_0) = \mathrm{val}(\mathbf{v}_1)$. ■

Now, we can define game values on $\bar{\mathsf{P}}$ using game values on P.

Definition 24 *The* game value $\mathrm{val} : \bar{\mathsf{P}} \to \mathrm{ORD}^*$ *is defined as follows. For* $\varnothing$, $\mathrm{val}(\varnothing)$ *is defined by Definition 21. For any* $n \in \mathbb{N}$ *and* $\mathbf{z} = (x_1, y_1, \dots, x_n, y_n) \in \bar{\mathsf{P}}_n$, *pick a sequence* $y'_1, \dots, y'_n$ *such that* $y'_1 \neq y, \dots,$ *and* $y'_n \neq y_n$. *Define* $\mathbf{v} := (\xi_1, \eta_1, \dots, \xi_n, \eta_n) \in \mathsf{P}_n$ *with* $\xi_i := (x_i, y_i, y'_i)$ *and* $\eta_i := 0$ *for any* $i \in [n]$. *Define* $\mathrm{val}(\mathbf{z}) := \mathrm{val}(\mathbf{v})$.

By Lemma 23, val is well-defined on $\bar{\mathsf{P}}$ and the following corollary holds.

Corollary 25 *For any* $0 \leqslant n < \infty$ *and* $(\xi_1, \eta_1, \dots, \xi_n, \eta_n) \in \left(\widetilde{\mathcal{X}} \times \{0,1\}\right)^n$, *we have*

$$\mathrm{val}(\xi_1(1), \xi_1(\eta_1 + 2), \dots, \xi_n(1), \xi_n(\eta_n + 2)) = \mathrm{val}(\xi_1, \eta_1, \dots, \xi_n, \eta_n). \tag{6}$$

Proof For $n = 0$, we have $\text{val}(\varnothing) = \text{val}(\varnothing)$. For any $n \geqslant 1$, by Definition 24 and Lemma 23, we have

$$\begin{aligned}
&\text{val}(\xi_1(1), \xi_1(\eta_1+2), \dots, \xi_n(1), \xi_n(\eta_n+2)) \\
=&\text{val}((\xi_1(1), \xi_1(\eta_1+2), \xi_1(3-\eta_1), 0, \dots, (\xi_n(1), \xi_n(\eta_n+2), \xi_n(3-\eta_n), 0) \\
=&\text{val}((\xi_1(1), \xi_1(\eta_1+2), \xi_1(3-\eta_1), 0, \dots, (\xi_n(1), \xi_n(2), \xi_n(3)), \eta_n) \\
=&\text{val}((\xi_1(1), \xi_1(\eta_1+2), \xi_1(3-\eta_1), 0, \dots, \xi_n, \eta_n) \\
&\vdots \\
=&\text{val}(\xi_1, \eta_1, \dots, \xi_n, \eta_n),
\end{aligned}$$

which gives (6). ∎

According to Bousquet et al. (2021, Lemma B.7), we have the following Lemma.

Lemma 26 *If* W *is coanalytic, then for any* $\mathbf{v} \in \mathsf{P}$*, either* $\text{val}(\mathbf{v}) = \Omega$ *or* $\text{val}(\mathbf{v}) < \omega_1$*. In particular, it follows that either* $\text{val}(\varnothing) = \Omega$ *or* $\text{val}(\varnothing) < \omega_1$.

According to Bousquet et al. (2021, Proposition B.8), we have the following Proposition.

Proposition 27 *Fix* $0 \leqslant n < \infty$ *and* $\mathbf{v} \in \mathsf{P}_n$ *such that* $0 \leqslant \text{val}(\mathbf{v}) < \Omega$*. For any* $\xi = (x, y^0, y^1) \in \tilde{\mathcal{X}}$*, there exists* $\eta \in \{0, 1\}$ *such that* $\text{val}(\mathbf{v}, \xi, \eta) < \text{val}(\mathbf{v})$.

For any $0 \leqslant n < \infty$, define

$$\mathsf{D}_{n+1} := \{(\mathbf{v}, \xi, \eta) \in \mathsf{P}_{n+1} : \text{val}(\mathbf{v}, \xi, \eta) < \min\{\text{val}(\mathbf{v}), \text{val}(\varnothing)\}\}$$

and

$$\bar{\mathsf{D}}_{n+1} := \left\{(\mathbf{z}, x, y) \in \bar{\mathsf{P}}_{n+1} : \text{val}(\mathbf{z}, x, y) < \min\{\text{val}(\mathbf{z}), \text{val}(\varnothing)\}\right\}.$$

The following lemma relates $\bar{\mathsf{D}}_{n+1}$ to D_{n+1}.

Lemma 28 *For any* $0 \leqslant n < \infty$*, we have*

$$\bar{\mathsf{D}}_{n+1} = \{(\xi_1(1), \xi_1(\eta_1+2), \dots, \xi_{n+1}(1), \xi_{n+1}(\eta_{n+1}+2)) : \ (\xi_1, \eta_1, \dots, \xi_{n+1}, \eta_{n+1}) \in \mathsf{D}_{n+1}\}.$$

Proof For any $(\xi_1, \eta_1, \dots, \xi_{n+1}, \eta_{n+1}) \in \mathsf{D}_{n+1}$, we have

$$\text{val}(\xi_1, \eta_1, \dots, \xi_{n+1}, \eta_{n+1}) < \min\{\text{val}(\xi_1, \eta_1, \dots, \xi_n, \eta_n), \text{val}(\varnothing)\}.$$

By Corollary 25, we have

$$\begin{aligned}
&\text{val}(\xi_1(1), \xi_1(\eta_1+2), \dots, \xi_{n+1}(1), \xi_{n+1}(\eta_{n+1}+2)) \\
=&\text{val}(\xi_1, \eta_1, \dots, \xi_{n+1}, \eta_{n+1}) \\
<&\min\{\text{val}(\xi_1, \eta_1, \dots, \xi_n, \eta_n), \text{val}(\varnothing)\}. \\
=&\min\{\text{val}(\xi_1(1), \xi_1(\eta_1+2), \dots, \xi_n(1), \xi_n(\eta_n+2), \text{val}(\varnothing)\}
\end{aligned}$$

which implies that $(\xi_1(1), \xi_1(\eta_1+2), \ldots, \xi_{n+1}(1), \xi_{n+1}(\eta_{n+1}+2)) \in \bar{\mathsf{D}}_{n+1}$.

On the other hand, for any $(x_1, y_1, \ldots, x_{n+1}, y_{n+1}) \in \bar{\mathsf{D}}_{n+1}$, define $\xi_i := (x_i, y_i, y_i')$ for arbitrary $y_i' \in \mathcal{Y}$ satisfying $y_i' \neq y_i$ and $\eta_i := 0$ for each $i \in [n]$. Then, by Corollary 25, we have

$$\begin{aligned}
&\operatorname{val}(\xi_1, \eta_1, \ldots, \xi_{n+1}, \eta_{n+1}) \\
=&\operatorname{val}(\xi_1(1), \xi_1(\eta_1+2), \ldots, \xi_{n+1}(1), \xi_{n+1}(\eta_{n+1}+2)) \\
=&\operatorname{val}(x_1, y_1, \ldots, x_{n+1}, y_{n+1}) \\
<&\min\{\operatorname{val}(x_1, y_1, \ldots, x_n, y_n), \min(\varnothing)\} \\
=&\min\{\operatorname{val}(\xi_1, \eta_1, \ldots, \xi_n, \eta_n), \min(\varnothing)\}
\end{aligned}$$

Thus, $(\xi_1, \eta_1, \ldots, \xi_n, \eta_n) \in \mathsf{D}_{n+1}$ and $\xi_i(1) = x_i$, $\xi_i(\eta_i+2) = y_i$ for all $i \in [n]$. Therefore,

$$\begin{aligned}
&(x_1, y_1, \ldots, x_{n+1}, y_{n+1}) \\
\in&\left\{(\xi_1(1), \xi_1(\eta_1+2), \ldots, \xi_{n+1}(1), \xi_{n+1}(\eta_{n+1}+2)) : \ (\xi_1, \eta_1, \ldots, \xi_{n+1}, \eta_{n+1}) \in \mathsf{D}_{n+1}\right\}.
\end{aligned}$$

In conclusion, Lemma 28 is proved. ■

According to Bousquet et al. (2021, Lemma B.10), we have the following lemma.

Lemma 29 *For any $0 \leqslant n < \infty$, $\mathbf{v} \in \mathsf{P}_n$, and $\kappa \in \mathrm{ORD}$, we have $\operatorname{val}(\mathbf{v}) > \kappa$ if and only if there exists $\xi \in \widetilde{\mathcal{X}}$ such that $\operatorname{val}(\mathbf{v}, \xi, \eta) \geqslant \kappa$ for all $\eta \in \{0, 1\}$.*

Then, by Corollary 25 and Lemma 29, the following corollary holds.

Corollary 30 *For any $0 \leqslant n < \infty$, $\mathbf{z} \in \bar{\mathsf{P}}_n$, and $\kappa \in \mathrm{ORD}$, we have $\operatorname{val}(\mathbf{z}) > \kappa$ if and only if there exist $x \in \widetilde{\mathcal{X}}$ and $(y, y') \in \widetilde{\mathcal{Y}^2}$ such that $\operatorname{val}(\mathbf{z}, x, y) \geqslant \kappa$ and $\operatorname{val}(\mathbf{z}, x, y) \geqslant \kappa$ and $\operatorname{val}(\mathbf{z}, x, y') \geqslant \kappa$.*

Define

$$\bar{\mathsf{W}} := \{(x_1, y_1, \ldots) \in (\mathcal{X} \times \mathcal{Y})^\infty : \mathcal{H}_{x_1, y_1, \ldots, x_\tau, y_\tau} = \varnothing \text{ for some } 0 \leqslant \tau < \infty\}.$$

Then, we can show that $\bar{\mathsf{W}}$ is coanalytic under the assumption that $\mathcal{H}$ is measurable.

Lemma 31 *If $\mathcal{X}$ and $\mathcal{Y}$ are Polish and $\mathcal{H}$ is measurable, then $\bar{\mathsf{W}}$ is coanalytic.*

Proof According to Definition 3, we have

$$\begin{aligned}
\bar{\mathsf{W}}^c &= \{(x_1, y_1, \ldots) \in (\mathcal{X} \times \mathcal{Y})^\infty : \mathcal{H}_{x_1, y_1, \ldots, x_\tau, y_\tau} \neq \varnothing \text{ for all } \tau < \infty\} \\
&= \cap_{\tau=1}^{\infty} \cup_{\theta \in \Theta} \cap_{t=1}^{\tau} \left\{(x_1, y_1, \ldots) \in (\mathcal{X} \times \mathcal{Y})^\infty : \mathsf{h}(\theta, x_t) = y_t\right\}.
\end{aligned}$$

For any $h \in \mathcal{H}$ and $1 \leqslant t < \infty$, define $\widetilde{h}_t : \Theta \times (\mathcal{X} \times \mathcal{Y})^\infty \to \mathcal{Y}$, $(\theta, x_1, y_1, \ldots) \mapsto h(\theta, x_t)$ and $l_t : \Theta \times (\mathcal{X} \times \mathcal{Y})^\infty \to \mathbb{R}$, $(\theta, x_1, y_1, \ldots) \mapsto \mathbb{1}\{y_t \neq \widetilde{h}_t(\theta, x_1, y_1, \ldots)\}$.

Since $\mathcal{H}$ is measurable, $h \in \mathcal{H}$ is Borel-measurable. Thus, $\widetilde{h}_t$ is also Borel-measurable. Since the mapping $\Theta \times (\mathcal{X} \times \mathcal{Y})^\infty \to \mathcal{Y}$, $(\theta, x_1, y_1, \ldots) \mapsto y_t$ is Borel-measurable, the mapping $\Theta \times (\mathcal{X} \times \mathcal{Y})^\infty \to \mathcal{Y}^2$, $(\theta, x_1, y_1, \ldots) \mapsto (y_t, \widetilde{h}_t(\theta, x_1, y_1, \ldots))$ is also Borel-measurable, which, together with the fact that the mapping $\mathcal{Y}^2 \to \{0, 1\}$, $(y, y') \mapsto \mathbb{1}\{y = y'\}$ is Borel-measurable, implies that l_t is Borel-measurable. Since

$$\{(\theta, x_1, y_1, \ldots) \in \Theta \times (\mathcal{X} \times \mathcal{Y})^\infty : h(\theta, x_t) = y_t\} = l_t^{-1}(\{1\}),$$

we have that $\{(\theta, x_1, y_1, \dots) \in \Theta \times (\mathcal{X} \times \mathcal{Y})^\infty : h(\theta, x_t) = y_t\}$ is Borel for any $1 \leqslant t < \infty$ and $h \in \mathcal{H}$. Thus, for any $1 \leqslant \tau < \infty$, $\cap_{t=1}^{\tau} \{(\theta, x_1, y_1, \dots) \in \Theta \times (\mathcal{X} \times \mathcal{Y})^\infty : h(\theta, x_t) = y_t\}$ is Borel. Since the union over $\theta \in \Theta$ corresponds to a projection and the intersection over τ is countable, the set $\bar{\mathsf{W}}^c$ is analytic. ■

Define

$$\bar{\mathsf{A}}_n := \cup_{\mathbf{w} \in (\mathcal{X} \times \mathcal{Y})^\infty} \left\{ \mathbf{z} \in \bar{\mathsf{P}}_n : (\mathbf{z}, \mathbf{w}) \in \bar{\mathsf{W}}^c \right\}.$$

Since $\mathcal{X}$ and $\mathcal{Y}$ are Polish, we have that $\bar{\mathsf{P}}_n$ is Polish for any $0 \leqslant n \leqslant \infty$. If $\bar{\mathsf{W}}$ is coanalytic, then $\bar{\mathsf{A}}_n$ is an analytic subset of $\bar{\mathsf{P}}_n$ for any $0 \leqslant n < \infty$.

Define $\bar{\mathsf{A}} := \cup_{0 \leqslant n < \infty} \bar{\mathsf{A}}_n$. We have the following lemma.

Lemma 32 $\mathrm{val}(\mathbf{z}) > -1$ *for any* $\mathbf{z} \in \bar{\mathsf{A}}$.

Proof Since $\mathbf{z} \in \bar{\mathsf{A}}$, there exists $0 \leqslant n < \infty$ such that $\mathbf{z} \in \bar{\mathsf{A}}_n$. There exist $(x_1, y_1, \dots, x_n, y_n) \in \bar{\mathsf{P}}_n$ and $(x_{n+1}, y_{n+1}, x_{n+2}, y_{n+2}, \dots) \in (\mathcal{X} \times \mathcal{Y})^\infty$ such that $\mathbf{z} = (x_1, y_1, \dots, x_n, y_n)$ and

$$\mathcal{H}_{x_1, y_1, \dots, x_\tau, y_\tau} \neq \varnothing$$

for all $1 \leqslant \tau < \infty$.

For any $1 \leqslant i < \infty$, define $\eta_i = 0$ and $\xi_i = (x_i, y_i, y_i')$ for arbitrary $y_i' \in \mathcal{Y}$ such that $y_i' \neq y_i$. It follows that $\mathcal{H}_{\xi_1(1), \xi_1(\eta_1+2), \dots, \xi_\tau(1), \xi_\tau(\eta_\tau+2)} \neq \varnothing$ for all $1 \leqslant \tau < \infty$. Thus,

$$\mathbf{v} := (\xi_1, \eta_1, \dots, \xi_n, \eta_n) \in \mathsf{A}_n$$

and by Definition 21 and Definition 24, $\mathrm{val}(\mathbf{z}) = \mathrm{val}(\mathbf{v}) > -1$. ■

Now, the corollary below holds.

Corollary 33 *If* $\bar{\mathsf{W}}$ *is coanalytic, then the set*

$$\bar{\mathsf{A}}_n^\kappa := \{\mathbf{z} \in \bar{\mathsf{A}}_n : \mathrm{val}(\mathbf{z}) > \kappa\}$$

is analytic for every $0 \leqslant n < \infty$ *and* $-1 \leqslant \kappa < \omega_1$.

Proof Since $\bar{\mathsf{W}}$ is coanalytic, we have that $\bar{\mathsf{A}}_n$ is analytic. For $\kappa = -1$, since $\mathrm{val}(\mathbf{z}) > -1$ for any $\mathbf{z} \in \bar{\mathsf{A}}_n$ by Lemma 32, we have that $\bar{\mathsf{A}}_n^{-1} = \bar{\mathsf{A}}_n$ is analytic for any $0 \leqslant n < \infty$. For $\kappa > -1$, suppose that for all $-1 \leqslant \lambda < \kappa$, $\bar{\mathsf{A}}_n^\lambda$ is analytic for every $0 \leqslant n < \infty$. According to Corollary 30, for any $0 \leqslant n < \infty$, we have

$$\begin{aligned}
\bar{\mathsf{A}}_n^\kappa &= \cup_{(x,y,y') \in \mathcal{X} \times \widetilde{\mathcal{Y}^2}} \left\{ \mathbf{z} \in \bar{\mathsf{A}}_n : \mathrm{val}(\mathbf{z}, x, y) \geqslant \kappa \text{ and } \mathrm{val}(\mathbf{z}, x, y') \geqslant \kappa \right\} \\
&= \cup_{(x,y,y') \in \mathcal{X} \times \mathcal{Y}^2} \left(\left\{ \mathbf{z} \in \bar{\mathsf{A}}_n : \mathrm{val}(\mathbf{z}, x, y) \geqslant \kappa \text{ and } \mathrm{val}(\mathbf{z}, x, y') \geqslant \kappa \text{ and } y \neq y' \right\} \right)
\end{aligned}$$

Consider the function $f : \bar{\mathsf{P}}_n \times \mathcal{X} \times \mathcal{Y}^2 \to \bar{\mathsf{P}}_n \times \mathcal{X} \times \mathcal{Y}^2$, $(\mathbf{z}, x, y^0, y^1) \mapsto (\mathbf{z}, x, y^1, y^0)$. f is a continuous function. Since by the induction hypothesis, $\bar{\mathsf{A}}_{n+1}^\lambda$ is analytic for any $-1 \leqslant \lambda < \kappa$, we have that $f(\bar{\mathsf{A}}_{n+1}^\lambda)$ is also analytic. Thus,

$$\begin{aligned}
&\left\{ (\mathbf{z}, x, y, y') \in \bar{\mathsf{A}}_n \times \mathcal{X} \times \mathcal{Y}^2 : \mathrm{val}(\mathbf{z}, x, y) \geqslant \kappa \text{ and } \mathrm{val}(\mathbf{z}, x, y') \geqslant \kappa \text{ and } y \neq y' \right\} \\
&= \cap_{-1 \leqslant \lambda < \kappa} \left(\bar{\mathsf{A}}_{n+1}^\lambda \cap f(\bar{\mathsf{A}}_{n+1}^\lambda) \right) \cap \bar{\mathsf{A}}_n \times \mathcal{X} \times \widetilde{\mathcal{Y}^2}
\end{aligned}$$

is also analytic. Since $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces, we have that $\bar{\mathsf{A}}_n^\kappa$ is analytic. By induction, $\bar{\mathsf{A}}_n^\kappa$ for any $0 \leqslant n < \infty$ and $-1 \leqslant \kappa < \omega_1$. ■

For any $0 \leqslant n < \infty$, define

$$\bar{\mathsf{D}}_{n+1} := \left\{(\mathbf{z}, x, y) \in \bar{\mathsf{P}}_{n+1} : \mathrm{val}(\mathbf{z}, x, y) < \min\{\mathrm{val}(\mathbf{z}), \mathrm{val}(\varnothing)\}\right\}.$$

Then, we can show the following corollary.

Corollary 34 *If* $\mathrm{val}(\varnothing) < \omega_1$ *and* $\bar{\mathsf{W}}$ *is coanalytic, then* $\bar{\mathsf{D}}_{n+1}$ *is universally measurable for any* $0 \leqslant n < \infty$.

Proof By the definition of $\bar{\mathsf{D}}_{n+1}$, we have

$$\begin{aligned}\bar{\mathsf{D}}_{n+1} &= \cup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)} \left\{(\mathbf{z}, x, y) \in \bar{\mathsf{P}}_{n+1} : \mathrm{val}(\mathbf{z}, x, y) \leqslant \kappa \text{ and } \mathrm{val}(\mathbf{z}) > \kappa\right\} \\ &= \cup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)} \left\{(\mathbf{z}, x, y) \in \bar{\mathsf{P}}_{n+1} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c \text{ and } \mathbf{z} \in \bar{\mathsf{A}}_n^\kappa\right\} \\ &= \cup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)} \left((\bar{\mathsf{A}}_{n+1}^\kappa)^c \cap \bar{\mathsf{A}}_n^\kappa \times \mathcal{X} \times \mathcal{Y}\right)\end{aligned}$$

with $\bar{\mathsf{A}}_n^\kappa$ defined in Corollary 33. According to Corollary 33, $(\bar{\mathsf{A}}_{n+1}^\kappa)^c \cap \bar{\mathsf{A}}_n^\kappa \times \mathcal{X} \times \mathcal{Y}$ is universally measurable for any $-1 \leqslant \kappa < \omega_1$. Since $\mathrm{val}(\varnothing) < \omega_1$, the union over $-1 \leqslant \kappa < \mathrm{val}(\varnothing)$ is countable. Thus, $\bar{\mathsf{D}}_{n+1}$ is universally measurable. ■

However, for the universal measurability of the learning strategy we defined, the above corollary does not directly apply. We need more refined analysis of the projection set of $\bar{\mathsf{D}}_n$. For any $0 \leqslant n < \infty$ and $y \in \mathcal{Y}$, define

$$\begin{aligned}\bar{\mathsf{D}}_{n+1}^y &:= \left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : \mathrm{val}(\mathbf{z}, x, y) < \min\{\mathrm{val}(\mathbf{z}), \mathrm{val}(\varnothing)\}\right\} \\ &= \left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : (\mathbf{z}, x, y) \in \bar{\mathsf{D}}_{n+1}\right\}.\end{aligned}$$

Then, we can proceed to show the following corollary.

Corollary 35 *If* $\mathrm{val}(\varnothing) < \omega_1$ *and* $\bar{\mathsf{W}}$ *is coanalytic, then* $\bar{\mathsf{D}}_{n+1}^y$ *is universally measurable for any* $0 \leqslant n < \infty$ *and* $y \in \mathcal{Y}$.

Proof By the definition of $\bar{\mathsf{D}}_{n+1}^y$, we have

$$\begin{aligned}\bar{\mathsf{D}}_{n+1}^y &= \cup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)} \left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : \mathrm{val}(\mathbf{z}, x, y) \leqslant \kappa \text{ and } \mathrm{val}(\mathbf{z}) > \kappa\right\} \\ &= \cup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)} \left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c \text{ and } \mathbf{z} \in \bar{\mathsf{A}}_n^\kappa\right\} \\ &= \cup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)} \left(\left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c\right\} \cap \bar{\mathsf{A}}_n^\kappa \times \mathcal{X}\right)\end{aligned}$$

with $\bar{\mathsf{A}}_n^\kappa$ defined in Corollary 33. Note that

$$\begin{aligned}&\left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c\right\} \\ &= \left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : (\mathbf{z}, x, y) \in \bar{\mathsf{A}}_{n+1}^\kappa\right\}^c \\ &= \left(\cup_{y'\in\mathcal{Y}} \left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : (\mathbf{z}, x, y') \in \bar{\mathsf{A}}_{n+1}^\kappa \cap \bar{\mathsf{P}}_n \times \mathcal{X} \times \{y\}\right\}\right)^c\end{aligned}$$

According to Corollary 33, $\bar{\mathsf{A}}_n^\kappa \times \mathcal{X}$ is an analytic subset of $\bar{\mathsf{P}}_n \times \mathcal{X}$ and $\bar{\mathsf{A}}_{n+1}^\kappa \cap \bar{\mathsf{P}}_n \times \mathcal{X} \times \{y\}$ is an analytic subset of $\bar{\mathsf{P}}_{n+1}$ for any $-1 \leqslant \kappa < \omega_1$. Thus, $\left\{(\mathbf{z}, x) \in \bar{\mathsf{P}}_n \times \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c\right\}$ is coanalytic for any $-1 \leqslant \kappa < \omega_1$. Since $\mathrm{val}(\varnothing) < \omega_1$, the union over $-1 \leqslant \kappa < \mathrm{val}(\varnothing)$ is countable. Thus, $\bar{\mathsf{D}}_{n+1}^y$ is universally measurable. ■

For any $0 \leqslant n < \infty$, $\mathbf{z} \in \bar{\mathsf{P}}_n$, and $y \in \mathcal{Y}$, define

$$\begin{aligned}\bar{\mathsf{D}}_{n+1}^{\mathbf{z},y} &:= \{x \in \mathcal{X} : \mathrm{val}(\mathbf{z}, x, y) < \min\{\mathrm{val}(\mathbf{z}), \mathrm{val}(\varnothing)\}\} \\ &= \left\{x \in \mathcal{X} : (\mathbf{z}, x, y) \in \bar{\mathsf{D}}_{n+1}\right\}.\end{aligned}$$

Then, we can show the following corollary.

Corollary 36 *If* $\mathrm{val}(\varnothing) < \omega_1$ *and* $\bar{\mathsf{W}}$ *is coanalytic, then* $\bar{\mathsf{D}}_{n+1}^{\mathbf{z},y}$ *is universally measurable for any* $0 \leqslant n < \infty$*,* $\mathbf{z} \in \bar{\mathsf{P}}_n$*, and* $y \in \mathcal{Y}$*.*

Proof By the definition of $\bar{\mathsf{D}}_{n+1}^{\mathbf{z},y}$, we have

$$\begin{aligned}\bar{\mathsf{D}}_{n+1}^{\mathbf{z},y} &= \cup_{-1 \leqslant \kappa < \mathrm{val}(\varnothing)} \{x \in \mathcal{X} : \mathrm{val}(\mathbf{z}, x, y) \leqslant \kappa \text{ and } \mathrm{val}(\mathbf{z}) > \kappa\} \\ &= \cup_{-1 \leqslant \kappa < \mathrm{val}(\varnothing)} \left\{x \in \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c \text{ and } \mathbf{z} \in \bar{\mathsf{A}}_n^\kappa\right\} \\ &= \cup_{\kappa: -1 \leqslant \kappa < \mathrm{val}(\varnothing),\ \mathbf{z} \in \bar{\mathsf{A}}_n^\kappa} \left\{x \in \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c\right\}\end{aligned}$$

with $\bar{\mathsf{A}}_n^\kappa$ defined in Corollary 33. Note that

$$\begin{aligned}&\left\{x \in \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c\right\} \\ =& \left\{x \in \mathcal{X} : (\mathbf{z}, x, y) \in \bar{\mathsf{A}}_{n+1}^\kappa\right\}^c \\ =& \left(\cup_{(\mathbf{w}, y') \in \bar{\mathsf{P}}_n \times \mathcal{Y}} \left\{x \in \mathcal{X} : (\mathbf{w}, x, y') \in \bar{\mathsf{A}}_{n+1}^\kappa \cap \{\mathbf{z}\} \times \mathcal{X} \times \{y\}\right\}\right)^c\end{aligned}$$

According to Corollary 33, $\bar{\mathsf{A}}_{n+1}^\kappa \cap \{\mathbf{z}\} \times \mathcal{X} \times \{y\}$ is an analytic subset of $\bar{\mathsf{P}}_{n+1}$ for any $-1 \leqslant \kappa < \omega_1$. Thus, $\left\{x \in \mathcal{X} : (\mathbf{z}, x, y) \in (\bar{\mathsf{A}}_{n+1}^\kappa)^c\right\}$ is coanalytic for any $-1 \leqslant \kappa < \omega_1$. Since $\mathrm{val}(\varnothing) < \omega_1$, the union over κ is countable. Thus, $\bar{\mathsf{D}}_{n+1}^{\mathbf{z},y}$ is universally measurable. ■

Now, we are ready to define a value-decreasing function. For any $1 \leqslant t < \infty$, $\mathbf{z} \in \bar{\mathsf{P}}_{t-1}$, and $x \in \mathcal{X}$, define the set $G_{t,\mathbf{z},x} := \left\{y \in \mathcal{Y} : (\mathbf{z}, x, y) \notin \bar{\mathsf{D}}_t\right\}$. When $\mathcal{Y}$ is uncountable, define the mapping $g_t : \bar{\mathsf{P}}_{t-1} \times \mathcal{X} \to \mathcal{Y}$ by

$$g_t(\mathbf{z}, x) := \begin{cases} \text{arbitrary } y \in G_{t,\mathbf{z},x}, & \text{if } G_{t,\mathbf{z},x} \neq \varnothing, \\ \text{arbitrary } y \in \mathcal{Y}, & \text{if } G_{t,\mathbf{z},x} = \varnothing. \end{cases} \tag{7}$$

When $\mathcal{Y}$ is countable, we can enumerate $\mathcal{Y}$ as $\{y^1, y^2, y^3, \dots, \}$. Then, the mapping $g_t : \bar{\mathsf{P}}_{t-1} \times \mathcal{X} \to \mathcal{Y}$ is defined as

$$g_t(\mathbf{z}, x) := \begin{cases} y^i, & \text{if } G_{t,\mathbf{z},x} \neq \varnothing \text{ and } y^j \notin G_{t,\mathbf{z},x} \text{ for all } 1 \leqslant j \leqslant i-1,\ y^i \in G_{t,\mathbf{z},x}, \\ y^1, & \text{if } G_{t,\mathbf{z},x} = \varnothing. \end{cases}$$
$$= \begin{cases} y^i, & \text{if } (\mathbf{z}, x) \in \bar{\mathsf{D}}_t^{y^j} \text{ for all } 1 \leqslant j \leqslant i-1 \text{ and } (\mathbf{z}, x) \notin \bar{\mathsf{D}}_t^{y^i}, \\ y^1, & \text{if } G_{t,\mathbf{z},x} = \varnothing. \end{cases}$$
$$= \begin{cases} y^i, & \text{if } (\mathbf{z}, x) \in \left(\cap_{j=1}^{i-1} \bar{\mathsf{D}}_t^{y^j} \right) \cap \left(\bar{\mathsf{D}}_t^{y^i} \right)^c, \\ y^1, & \text{if } (\mathbf{z}, x) \in \cap_{j=1}^{\infty} \bar{\mathsf{D}}_t^{y^j}. \end{cases}$$

Corollary 37 *If $\mathcal{Y}$ is countable,* $\mathrm{val}(\varnothing) < \omega_1$*, and $\bar{\mathsf{W}}$ is coanalytic, then g_t is universally measurable for any $1 \leqslant t < \infty$.*

Proof For any $2 \leqslant i < \infty$, we have

$$g_t^{-1}(y^i) = \left(\cap_{j=1}^{i-1} \bar{\mathsf{D}}_t^{y^j} \right) \cap \left(\bar{\mathsf{D}}_t^{y^i} \right)^c$$

which is universally measurable by Corollary 35. For $i = 1$, we have

$$g_t^{-1}(y^1) = \left(\cap_{j=1}^{\infty} \bar{\mathsf{D}}_t^{y^j} \right) \cup \left(\bar{\mathsf{D}}_t^{y^1} \right)^c$$

which is also universally measurable by Corollary 35. ■

For any $1 \leqslant t < \infty$, $\mathbf{z} \in \bar{\mathsf{P}}_{t-1}$, and $x \in \mathcal{X}$, define the mapping $g_{t,\mathbf{z}} : \mathcal{X} \to \mathcal{Y}$, $x \mapsto g_t(\mathbf{z}, x)$. Then, we have the following corollary.

Corollary 38 *If $\mathcal{Y}$ is countable,* $\mathrm{val}(\varnothing) < \omega_1$*, and $\bar{\mathsf{W}}$ is coanalytic, then $g_{t,\mathbf{z}}$ is universally measurable for any $1 \leqslant t < \infty$ and $\mathbf{z} \in \bar{\mathsf{P}}_{t-1}$.*

Proof By the definition of $g_{t,\mathbf{z}}$, we have

$$g_{t,\mathbf{z}}(x) = \begin{cases} y^i, & \text{if } x \in \left(\cap_{j=1}^{i-1} \bar{\mathsf{D}}_t^{\mathbf{z},y^j} \right) \cap \left(\bar{\mathsf{D}}_t^{\mathbf{z},y^i} \right)^c, \\ y^1, & \text{if } x \in \cap_{j=1}^{\infty} \bar{\mathsf{D}}_t^{\mathbf{z},y^j}. \end{cases}$$

Thus, for $2 \leqslant i < \infty$, we have

$$g_{t,\mathbf{z}}^{-1}(y^i) = \left(\cap_{j=1}^{i-1} \bar{\mathsf{D}}_t^{\mathbf{z},y^j} \right) \cap \left(\bar{\mathsf{D}}_t^{\mathbf{z},y^i} \right)^c$$

which is universally measurable by Corollary 36. For $i = 1$, we have

$$g_{t,\mathbf{z}}^{-1}(y^1) = \left(\cap_{j=1}^{\infty} \bar{\mathsf{D}}_t^{\mathbf{z},y^j} \right) \cup \left(\bar{\mathsf{D}}_t^{\mathbf{z},y^1} \right)^c$$

which is also universally measurable by Corollary 36. ■

For any $1 \leqslant t < \infty$, define the mapping $\bar{g}_t : \mathcal{X}^t \to \mathcal{Y}$,

$$(x_1, x_2, \ldots, x_t) \mapsto g_t(x_1, g_1(x_1), x_2, g_2(x_1, g_1(x_1), x_2), \ldots, x_{t-1}, g_{t-1}(x_1, g_1(x_1), \ldots, x_{t-1}), x_t)$$

We can show the following lemma.

Lemma 39 *For any $1 \leqslant t < \infty$, if g_i is universally measurable for all $i \in [t]$, then $\bar{g}_t$ is also universally measurable.*

Proof For each $i \in [t-1]$, define the mapping $\widetilde{g}_i : \bar{\mathsf{P}}_{i-1} \times \mathcal{X}^{t-i+1} \to \bar{\mathsf{P}}_i \times \mathcal{X}^{t-i}$,

$$\begin{aligned}&(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i, x_{i+1}, \ldots, x_t)\\ \mapsto&(x_1, y_1, \ldots, x_i, g_i(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i), x_{i+1}, x_{i+2}, \ldots, x_t).\end{aligned}$$

Then, we have $\bar{g}_t = g_t \circ \widetilde{g}_{t-1} \circ \cdots \circ \widetilde{g}_1$. Since g_t is universally measurable, it suffices to show that $\widetilde{g}_i$ is universally measurable for each $i \in [t-1]$. For any Polish space $\mathcal{E}_1$ and $\mathcal{E}_2$, let $\mathcal{F}(\mathcal{E}_1)$ denote the Borel σ-field of $\mathcal{E}_1$ and $\mathcal{F}(\mathcal{E}_1) \times \mathcal{F}(\mathcal{E}_2)$ denote the product σ-field of $\mathcal{F}(\mathcal{E}_1)$ and $\mathcal{F}(\mathcal{E}_2)$ on $\mathcal{E}_1 \times \mathcal{E}_2$. Since $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces, we have $\mathcal{F}(\bar{\mathsf{P}}_j \times \mathcal{X}^k) = (\mathcal{F}(\mathcal{X}) \times \mathcal{F}(\mathcal{Y}))^j \times \mathcal{F}(\mathcal{X})^k$ for any $0 \leqslant j, k < \infty$. Thus, it suffices to show that $\widetilde{g}_i^{-1}((\prod_{j=1}^i A_j \times B_j) \times (\prod_{k=i+1}^t A_k))$ is universally measurable in $\bar{\mathsf{P}}_{i-1} \times \mathcal{X}^{t-i+1}$ for any $A_j \in \mathcal{F}(\mathcal{X})$ with $j \in [t]$, any $B_j \in \mathcal{F}(\mathcal{Y})$ with $j \in [i]$, and any $i \in [t-1]$. By the definition of $\widetilde{g}_i$, we have

$$\begin{aligned}&\widetilde{g}_i^{-1}\left(\left(\prod_{j=1}^{i} A_j \times B_j\right) \times \left(\prod_{k=i+1}^{t} A_k\right)\right)\\ =&\Bigg\{(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i, x_{i+1}, \ldots, x_t) \in \left(\prod_{j=1}^{i-1} A_j \times B_j\right) \times \left(\prod_{k=i}^{t} A_k\right) :\\ &\quad g_i(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i) \in B_i\Bigg\}\\ =&\left(\prod_{j=1}^{i-1} A_j \times B_j\right) \times \left(\prod_{k=i}^{t} A_k\right) \cap g_i^{-1}(B_i) \times \left(\prod_{k=i+1}^{t} A_k\right)\end{aligned}$$

Since g_i is universally measurable, we have that $g_i^{-1}(B_i)$ is a universally measurable subset of $\bar{\mathsf{P}}_{i-1} \times \mathcal{X}$. It follows that $g_i^{-1}(B_i) \times \left(\prod_{k=i+1}^t A_k\right)$ is universally measurable in $\bar{\mathsf{P}}_{i-1} \times \mathcal{X}^{t-i+1}$. Thus, $\widetilde{g}_i^{-1}\left(\left(\prod_{j=1}^i A_j \times B_j\right) \times \left(\prod_{k=i+1}^t A_k\right)\right)$ is universally measurable in $\bar{\mathsf{P}}_{i-1} \times \mathcal{X}^{t-i+1}$. ■

The following corollary immediately follows from Corollary 37 and Lemma 39.

Corollary 40 *If $\mathcal{Y}$ is countable,* $\mathrm{val}(\varnothing) < \omega_1$, *and* $\bar{\mathsf{W}}$ *is coanalytic, then $\bar{g}_t$ is universally measurable for any $1 \leqslant t < \infty$.*

We have the following lemma.

Lemma 41 *For any $1 \leqslant t < \infty$, $\mathbf{z} = (x_1, y_1, \ldots, x_t, y_t) \in \bar{\mathsf{P}}_t$, we have*

$$\mathrm{val}(\mathbf{z}) = -1 \iff \mathcal{H}_{\mathbf{z}} = \varnothing.$$

Proof Define $\xi_i := (x_i, y_i, y_i')$ for arbitrary $y_i' \in \mathcal{Y}\backslash\{y_i'\}$ and $\eta_i := 0$ for each $i \in [t]$. Define $\mathbf{v} := (\xi_1, \eta_1, \ldots, \xi_t, \eta_t)$.

Assume that $\mathcal{H}_{\mathbf{z}} = \varnothing$. Then, for any $\mathbf{w} \in (\widetilde{\mathcal{X}} \times \{0,1\})^\infty$, we have

$$\mathcal{H}_{\xi_1(1),\xi_1(\eta_1+2),\ldots,\xi_t(1),\xi_t(\eta+2)} = \mathcal{H}_{\mathbf{z}} = \varnothing$$

which implies that $(\mathbf{v}, \mathbf{w}) \in \mathsf{W}$. Thus, we have $\mathbf{v} \notin \mathsf{A}$. By Definition 21, we have $\operatorname{val}(\mathbf{v}) = -1$. By Corollary 25, we have $\operatorname{val}(\mathbf{z}) = \operatorname{val}(\mathbf{v}) = -1$.

For the other direction, assume that $\operatorname{val}(\mathbf{z}) = -1$. By Corollary 25, we have $\operatorname{val}(\mathbf{v}) = \operatorname{val}(\mathbf{z}) = -1$. By Definition 21, we have $(\mathbf{v}, \mathbf{w}) \in \mathsf{W}$ for any $\mathbf{w} \in (\widetilde{\mathcal{X}} \times \{0,1\})^\infty$. Suppose that $\mathcal{H}_{\mathbf{z}} \neq \varnothing$. Choose arbitrary $h \in \mathcal{H}_{\mathbf{z}}$. Since $h \in \mathcal{Y}^{\mathcal{X}}$, there exists $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ such that $h(x) = y$. Choose arbitrary $y' \in \mathcal{Y}\backslash\{y\}$. Define $\xi_{t+i} := (x, y, y')$ and $\eta_{t+i} := 0$ for any $1 \leqslant i < \infty$, and $\mathbf{w} := (\xi_{t+1}, \eta_{t+1}, \xi_{t+2}, \eta_{t+2}, \ldots) \in (\widetilde{\mathcal{X}} \times \{0,1\})^\infty$. Then, for any $0 \leqslant \tau < \infty$, we have $h \in \mathcal{H}_{\xi_1(1),\xi_1(\eta_1+1),\ldots,\xi_\tau(1),\xi_\tau(\eta_\tau+1)}$. Then, we have $(\mathbf{v}, \mathbf{w}) \notin \mathsf{W}$. A contradiction. Thus, $\mathcal{H}_{\mathbf{z}} = \varnothing$. ■

Then, we can prove the following guarantee for g_t.

Proposition 42 *For any $(x_1, x_2, \ldots) \in \mathcal{X}^\infty$, any $y_1 \in \mathcal{Y}\backslash\{g_1(x_1)\}$, and any $y_t \in \mathcal{Y}$ such that $y_t \neq g_t(x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t)$ with $2 \leqslant t < \infty$, if $\operatorname{val}(\varnothing) < \Omega$, then there exists some positive integer τ $(1 \leqslant \tau < \infty)$ such that $\mathcal{H}_{x_1,y_1,\ldots,x_\tau,y_\tau} = \varnothing$.*

Proof By Lemma 22, we have $\operatorname{val}(\varnothing) \geqslant 0$. Define $\xi_t := (x_t, y_t, \bar{g}_t(x_1, \ldots, x_t))$, $\mathbf{v}_t = (\xi_1, 0, \ldots, \xi_t, 0)$, and $\mathbf{z}_t := (x_1, y_t, \ldots, x_t, y_t)$ for any $0 \leqslant t < \infty$ (when $t = 0$, we have $\mathbf{v}_0 = \varnothing$ and $\mathbf{z}_0 = \varnothing$).

We claim that for any $1 \leqslant t < \infty$, if $0 \leqslant \operatorname{val}(\mathbf{z}_{t-1}) \leqslant \operatorname{val}(\varnothing)$, we have $\operatorname{val}(\mathbf{z}_t) < \operatorname{val}(\mathbf{z}_{t-1})$. Indeed, by the definition of g_t, we have either $\operatorname{val}(\mathbf{z}_{t-1}, x_t, g_t(\mathbf{z}_{t-1}, x_t)) \geqslant \min\{\operatorname{val}(\mathbf{z}_{t-1}), \operatorname{val}(\varnothing)\}$ or $\operatorname{val}(\mathbf{z}_{t-1}, x_t, y) < \min\{\operatorname{val}(\mathbf{z}_{t-1}), \operatorname{val}(\varnothing)\}$ for all $y \in \mathcal{Y}$.

If $\operatorname{val}(\mathbf{z}_{t-1}, x_t, y) < \min\{\operatorname{val}(\mathbf{z}_{t-1}), \operatorname{val}(\varnothing)\}$ for all $y \in \mathcal{Y}$, it obviously follows that $\operatorname{val}(\mathbf{z}_t) < \min\{\operatorname{val}(\mathbf{z}_{t-1}), \operatorname{val}(\varnothing)\}$. If $\operatorname{val}(\mathbf{z}_{t-1}, x_t, g(\mathbf{z}_{t-1}, x_t)) \geqslant \min\{\operatorname{val}(\mathbf{z}_{t-1}), \operatorname{val}(\varnothing)\}$, since $\operatorname{val}(\mathbf{z}_{t-1}) \leqslant \operatorname{val}(\varnothing)$ by our assumption, we have $\operatorname{val}(\mathbf{z}_{t-1}, x_t, g(\mathbf{z}_{t-1}, x_t)) \geqslant \operatorname{val}(\mathbf{z}_{t-1})$. By Corollary 25, we have

$$\operatorname{val}(\mathbf{v}_{t-1}, \xi_t, 1) \geqslant \operatorname{val}(\mathbf{v}_{t-1}).$$

Then, by Proposition 27 and Corollary 25, we must have

$$\operatorname{val}(\mathbf{z}_t) = \operatorname{val}(\mathbf{v}_t) = \operatorname{val}(\mathbf{v}_{t-1}, \xi_t, 0) < \operatorname{val}(\mathbf{v}_{t-1}) = \operatorname{val}(\mathbf{z}_{t-1}).$$

Thus, the above claim holds.

Now we claim that $\operatorname{val}(\mathbf{z}_t) \leqslant \operatorname{val}(\varnothing)$ for $t = 0$ and $\operatorname{val}(\mathbf{z}_t) < \operatorname{val}(\varnothing)$ for any $1 \leqslant t < \infty$. Indeed, when $t = 0$, we have $\operatorname{val}(\mathbf{z}_0) = \operatorname{val}(\varnothing)$. Suppose $\operatorname{val}(\mathbf{z}_{t-1}) \leqslant \operatorname{val}(\varnothing)$ for some $1 \leqslant t < \infty$. If $\operatorname{val}(\mathbf{v}_{t-1}) = -1$, by Lemma 41, we have $\operatorname{val}(\mathbf{z}_t) = \operatorname{val}(\mathbf{z}_{t-1}) = -1 < \operatorname{val}(\varnothing)$. If $\operatorname{val}(\mathbf{z}_{t-1}) \geqslant 0$, we have $\operatorname{val}(\mathbf{z}_t) < \operatorname{val}(\mathbf{z}_{t-1}) \leqslant \operatorname{val}(\varnothing)$ by the first claim. Thus, by induction, the claim holds.

By the two claims, we can conclude that $\operatorname{val}(\varnothing) > \operatorname{val}(\mathbf{z}_1) > \operatorname{val}(\mathbf{z}_2) > \cdots > \operatorname{val}(\mathbf{z}_t)$ as long as $\operatorname{val}(\mathbf{z}_t) > -1$. If $\operatorname{val}(\varnothing) < \Omega$, by the well-ordering of ORD, there exists some finite positive integer τ such that $\operatorname{val}(\mathbf{z}_\tau) = -1$. Thus, by Lemma 41, we have $\mathcal{H}_{\mathbf{z}_\tau} = \varnothing$. ■

Now, we can present the proof of Theorem 18.

Proof of Theorem 18 Assume that $\mathcal{H}$ has an infinite Littlestone tree $\{\xi_{\mathbf{u}} : 0 \leqslant k < \infty, \mathbf{u} \in \{0,1\}^k\}$. Define the following strategy for the adversary $\bar{P}_a$: in each round $t \geqslant 1$, $\bar{P}_a$ chooses $x_t := \xi_{(\eta_1,\ldots,\eta_{t-1})}(1) \in \mathcal{X}$ with $\eta_i \in \{0,1\}$ $(1 \leqslant i \leqslant t-1)$ defined later (when $t = 1$, we have $\xi_{(\eta_1,\ldots,\eta_{t-1})} := \xi_{\varnothing}$). After the learner $\bar{P}_l$ makes the prediction $\widehat{y}_t$, define

$$\eta_t := \begin{cases} 0, & \text{if } \xi_{(\eta_1,\ldots,\eta_{t-1})}(2) \neq \widehat{y}_t, \\ 1, & \text{otherwise.} \end{cases}$$

Then, $\bar{P}_a$ reveals the true label $y_t := \xi_{(\eta_1,\ldots,\eta_{t-1})}(\eta_t + 2)$.

Since $\xi_{(\widehat{y}_1,\ldots,\widehat{y}_{t-1})}(2) \neq \xi_{(\widehat{y}_1,\ldots,\widehat{y}_{t-1})}(3)$, we have $y_t \neq \widehat{y}_t$ for each $t \geqslant 1$. Besides, by the definition of Littlestone tree (Definition 5), $(x_1, y_1, \ldots, x_t, y_t)$ is consistent with $\mathcal{H}$ for any $(\widehat{y}_1, \ldots, \widehat{y}_t) \in \mathcal{Y}^t$ and $t \geqslant 1$.

Assume that $\mathcal{H}$ does not have an infinite Littlestone tree. Consider the following strategy for the learner $\bar{P}_l$.

- Initialize $\tau \leftarrow 1$ and $f(x) \leftarrow g_1(x)$.
- For $t \leftarrow 1, 2, 3, \cdots$:
  - Predict $\widehat{y}_t = f(x_t)$.
  - If $\widehat{y}_t \neq y_t$:
    set $\widetilde{x}_\tau = x_t$, $\widetilde{y}_\tau = y_t$, $f(x) \leftarrow g_{\tau+1}(\widetilde{x}_1, \widetilde{y}_1, \ldots, \widetilde{x}_\tau, \widetilde{y}_\tau, x)$, and $\tau \leftarrow \tau + 1$.

Suppose that there exists some adversary $\bar{P}_a$ such that $\bar{P}_l$ makes infinitely many mistakes at $t_1, t_2, \ldots$ adopting the above strategy. Then according to Proposition 42, there exists some $1 \leqslant k < \infty$ such that $\mathcal{H}_{x_{t_1}, y_{t_1}, \ldots, x_{t_k}, y_{t_k}} = \varnothing$. However, this contradicts the rule of the online learning game $\bar{\mathcal{B}}$ because $\mathcal{H}_{x_1, y_1, \ldots, x_{t_k}, y_{t_k}} = \varnothing$. ■

Also, the universal measurability of the learning strategy can be proved.

Corollary 43 *If $\mathcal{X}$ is a Polish space, $\mathcal{Y}$ is countable, $\mathcal{H}$ is measurable as defined in Definition 3, and $\mathcal{H}$ does not have an infinite Littlestone tree, then the learning strategy of $\bar{P}_l$ specified in Theorem 18 is universally measurable.*

Proof Since $\mathcal{H}$ does not have an infinite Littlestone tree, according to Lemma 22, we have $\operatorname{val}(\varnothing) < \Omega$. Then, by Lemma 26, we have $\operatorname{val}(\varnothing) < \omega_1$. Since $\mathcal{H}$ is measurable, by Lemma 31, $\bar{\mathsf{W}}$ is coanalytic. Then, according to Corollary 37, g_t is universally measurable for any $1 \leqslant t < \infty$. According to Corollary 38, $f(x)$ is also universally measurable for any $1 \leqslant t < \infty$. Thus, the learning strategy for $\bar{P}_l$ specified in Theorem 18 is universally measurable. ■

B.2. Concluding proof

Proof of Theorem 13 First, according to Bousquet et al. (2021, Lemma 4.2), $\mathcal{H}$ is not learnable at rate faster than the exponential rate e^{-n}. Thus, the proof is completed once we construct a learning algorithm which, for $\mathcal{H}$ without an infinite Littlestone tree, achieves exponential rate for any realizable distribution P. We use the learning algorithm constructed in Bousquet et al. (2021, Section

4.1). According to Bousquet et al. (2021, Lemma 4.3, Lemma 4.4., and Corollary 4.5) and their proofs, for the learning algorithm to achieve exponential rate, it suffices to have an adversarial online learning algorithm with the properties that it only makes finitely many mistakes against any adversary and it only changes when a mistake is made. According to Theorem 18 and its proof, for $\mathcal{H}$ without an infinite Littlestone tree, the winning strategy constructed in the proof only makes finitely many mistakes against any adversary and changes only when a mistake happens. Then, the same proofs of Bousquet et al. (2021, Lemma 4.3, Lemma 4.4., and Corollary 4.5) can be applied to show that the constructed online learning algorithm achieves exponential rate. In conclusion, if $\mathcal{H}$ does not have an infinite Littlestone tree, then $\mathcal{H}$ is learnable with optimal rate e^{-n}. ■

Appendix C. Near-Linear Rates

In this section, we prove Theorem 14.

C.1. Slower than exponential is not faster than linear

In this subsection, we prove the following theorem.

Theorem 44 *If $\mathcal{H}$ has an infinite Littlestone tree, then for any learning algorithm A, there exists a $\mathcal{H}$-realizable distribution P such that for infinitely many n, $\mathbf{E}[\mathrm{er}(\widehat{h}_n)] \geqslant \frac{1}{33n}$ where $\widehat{h}_n = A(\mathcal{H}, S_n)$ with $S_n \sim P^n$. Thus, $\mathcal{H}$ is not learnable at rate faster than $\frac{1}{n}$.*

Proof Suppose that $\mathcal{H}$ has an infinite Littlestone tree

$$\left\{\xi_{\mathbf{u}} = (x_{\mathbf{u}}, y_{\mathbf{u}}^0, y_{\mathbf{u}}^1) : 0 \leqslant k < d, \mathbf{u} \in \{0,1\}^k\right\}.$$

Fix an arbitrary learning algorithm A. Let $\mathbf{u} = \{u_1, u_2, \dots\}$ be a sequence of i.i.d. Bernoulli($\frac{1}{2}$) random variables. Conditional on $\mathbf{u}$, define the distribution $P_{\mathbf{u}}$ on $\mathcal{X} \times \mathcal{Y}$ by

$$P_{\mathbf{u}}(\{x_{\mathbf{u}_{\leqslant k}}, y_{\mathbf{u}_{\leqslant k}}^{u_{k+1}}\}) = 2^{-k-1}, \ \forall\, k \geqslant 0.$$

Note that the mapping $\mathbf{u} \mapsto P_{\mathbf{u}}$ is measurable.

By the definition of Littlestone tree, for any $n \geqslant 0$, there exists a hypothesis $h_n \in \mathcal{H}$ such that $h_n(x_{\mathbf{u}_{\leqslant k}}) = y_{\mathbf{u}_{\leqslant k}}^{u_{k+1}}$ for any $0 \leqslant k \leqslant n$. Thus, we have

$$\mathrm{er}_{\mathbf{u}}(h_n) := P_{\mathbf{u}}(\{x, y) \in \mathcal{X} \times \mathcal{Y} : h_n(x) \neq y\}) \leqslant \sum_{k=n+1}^{\infty} 2^{-k-1} = 2^{-n-1}.$$

Then, $\inf_{h \in \mathcal{H}} \mathrm{er}_{\mathbf{u}}(h) = 0$ and $P_{\mathbf{u}}$ is $\mathcal{H}$-realizable.

Let $T, T_1, T_2, \dots$ be i.i.d. random variables with distribution Geometric($\frac{1}{2}$) (starting from 0). Define $X := x_{\mathbf{u}_{\leqslant T}}$, $Y := y_{\mathbf{u}_{\leqslant T}}^{u_{T+1}}$, $X_i := x_{\mathbf{u}_{\leqslant T_i}}$, and $Y_i := y_{\mathbf{u}_{\leqslant T_i}}^{u_{T_i+1}}$ for any $i \geqslant 1$. Then, conditional on $\mathbf{u}$, by the definition of $P_{\mathbf{u}}$, we know that $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ is a sequence of i.i.d. random variables with distribution $P_{\mathbf{u}}$. Now, define $\widehat{h}_n = A(\mathcal{H}, ((X_1, Y_1), \dots, (X_n, Y_n)))$. For any

$k \geqslant 1$, since $u_1, u_2, \dots$ are i.i.d. Bernoulli($\frac{1}{2}$) random variables, we have

$$\begin{aligned}
&\mathbf{P}(\widehat{h}_n(X) \neq Y, T = k, \max\{T_1, \dots, T_n\} < k) \\
=&\mathbf{P}(\widehat{h}_n(X) \neq y_{\mathbf{u}_{\leqslant k}}^{u_{k+1}}, T = k, \max\{T_1, \dots, T_n\} < k) \\
=&\mathbf{E}[\mathbf{P}(\widehat{h}_n(X) \neq y_{\mathbf{u}_{\leqslant k}}^{u_{k+1}} | X, T, T_1, \dots, T_n)\mathbb{1}\{T = k, \max\{T_1, \dots, T_n\} < k\}] \\
=&\mathbf{E}\left[\frac{1}{2}\mathbb{1}\{T = k, \max\{T_1, \dots, T_n\} < k\}\right] \\
=&\frac{1}{2}\mathbf{P}(T = k, \max\{T_1, \dots, T_n\} < k\}) \\
=&2^{-k-2}(1 - 2^{-k})^n.
\end{aligned}$$

Define $k_n := \lceil 1+\log_2(n) \rceil$ for $n \geqslant 1$. Then, we have $2^{-k_n-2} > \frac{1}{16n}$ and $(1-2^{-k_n})^n \geqslant (1-\frac{1}{2n})^n \geqslant \frac{1}{2}$, which, together with the above result, implies that

$$\begin{aligned}
\mathbf{P}(\widehat{h}_n(X) \neq Y, T = k_n) \geqslant&\mathbf{P}(\widehat{h}_n(X) \neq Y, T = k_n, \max\{T_1, \dots, T_n\} < k_n) \\
\geqslant&2^{-k_n-2}(1 - 2^{-k_n})^n \\
>&\frac{1}{32n}.
\end{aligned}$$

Since

$$n\mathbf{P}(\widehat{h}_n(X) \neq Y, T = k_n|\mathbf{u}) \leqslant n\mathbf{P}(T = k_n|\mathbf{u}) = n\mathbf{P}(T = k_n) = n2^{-k_n-1} \leqslant \frac{1}{4} \text{ a.s.,}$$

by Fatou's lemma, we have

$$\mathbf{E}[\limsup_{n\to\infty} n\mathbf{P}(\widehat{h}_n(X) \neq Y, T = k_n|\mathbf{u})] \geqslant \limsup_{n\to\infty} n\mathbf{P}(\widehat{h}_n(X) \neq Y, T = k_n) \geqslant \frac{1}{32}.$$

Since

$$\mathbf{P}(\widehat{h}_n(X) \neq Y, T = k_n|\mathbf{u}) \leqslant \mathbf{P}(\widehat{h}_n(X) \neq Y|\mathbf{u}) = \mathbf{E}[\mathrm{er}_{\mathbf{u}}(\widehat{h}_n)|\mathbf{u}] \text{ a.s.,}$$

we have $\mathbf{E}[\limsup_{n\to\infty} n\mathbf{E}[\mathrm{er}_{\mathbf{u}}(\widehat{h}_n)|\mathbf{u}]] \geqslant \frac{1}{32} > \frac{1}{33}$. Thus, there exists $\mathbf{u}' \in \{0,1\}^\infty$ such that $\mathbf{E}[\mathrm{er}_{\mathbf{u}'}(\widehat{h}_n)] \geqslant \frac{1}{33n}$ infinitely often. The proof is completed by setting $P = P_{\mathbf{u}'}$. ■

C.2. Pattern avoidance functions

In this subsection, we design pattern avoidance functions in the adversarial setting and analyze their properties. For any $n \in \mathbb{N}$ and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^n$, denote the collection of all n-dimensional pseudo-cubes of $\mathcal{H}$ with $\mathsf{PC}(\mathcal{H})$. For any $m \in \mathbb{N}$, denote the collection of all n-dimensional pseudo-cubes of $\mathcal{H}$ of size m with $\mathsf{PC}_m(\mathcal{H})$. Then, we have $\mathsf{PC}(\mathcal{H}) = \cup_{m=1}^\infty \mathsf{PC}_m(\mathcal{H})$. For any hypothesis class $F \subseteq \mathcal{Y}^n$, let $Q(F)$ denote the union of all the pseudo-cubes of dimension n that are subsets of F.

Consider the following game $\mathfrak{B}$ between player P_A and P_L. At each round $\tau \geqslant 1$:

- Player P_A chooses a sequence $\mathbf{x}_\tau = (x_\tau^0, \ldots, x_\tau^{\tau-1}) \in \mathcal{X}^\tau$ and a set $C_\tau \in \mathsf{PC}(\mathcal{Y}^\tau)$.
- Player P_L chooses a sequence $\mathbf{y}_\tau = (y_\tau^0, \ldots, y_\tau^{\tau-1}) \in \mathcal{Y}^\tau$.
- Player P_L wins the game in round τ if
  - either $C_\tau \notin \mathsf{PC}(\mathcal{H}|_{\mathbf{x}_\tau})$,
  - or $\mathbf{y}_s \in C_s$ for all $1 \leqslant s \leqslant \tau$ and $\mathcal{H}_{\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_\tau,\mathbf{y}_\tau} = \varnothing$, where $\mathcal{H}_{\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_\tau,\mathbf{y}_\tau} := \left\{h \in \mathcal{H} : h\left(x_s^i\right) = y_s^i \text{ for } 0 \leqslant i < s, 1 \leqslant s \leqslant \tau\right\}$.

The set of winning sequences of P_L in $\mathfrak{B}$ is

$$\mathsf{W}_{\mathfrak{B}} := \Bigg\{(\mathbf{x}_1, C_1, \mathbf{y}_1, \ldots) \in \prod_{t=1}^{\infty}(\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t) : \exists \tau \in \mathbb{N} \text{ such that}$$
$$\text{either } C_\tau \notin \mathsf{PC}(\mathcal{H}|_{\mathbf{x}_\tau}), \text{ or } \mathbf{y}_t \in C_t \text{ for all } t \in [\tau] \text{ and } \mathcal{H}_{\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_\tau,\mathbf{y}_\tau} = \varnothing\Bigg\}$$

Obviously, $\mathsf{W}_{\mathfrak{B}}$ is finitely decidable, which implies that $\mathfrak{B}$ is a Gale-Stewart game and according to Gale and Stewart (1953), either P_A or P_L has a winning strategy.

With regard to the universal measurability of the winning strategy, we assume that $\mathcal{X}$ is a Polish space, $\mathcal{Y}$ is countable, and $\mathcal{H}$ is measurable in the sense of Definition 3. We first prove the following lemma.

Lemma 45 *For any $t \in \mathbb{N}$ and $\mathbf{x}_t \in \mathcal{X}^t$, $\mathsf{PC}(\mathcal{Y}^t)$ and $\mathsf{PC}(\mathcal{H}|_{\mathbf{x}_t})$ are countable sets.*

Proof Since $\mathcal{Y}$ is countable, $\mathcal{Y}^t$ and $\mathcal{H}|_{\mathbf{x}_t}$ are also countable. By the definition of pseudo-cube, any pseudo-cube is a finite subset of the hypothesis class. Since the set of all finite subsets of a countable set is countable, $\mathsf{PC}(\mathcal{H}|_{\mathbf{x}_t})$ and $\mathsf{PC}(\mathcal{Y}^t)$ are countable sets. ■

For any $t \in \mathbb{N}$, define the set

$$\mathsf{XPC}_t := \cup_{\mathbf{x}_t \in \mathcal{X}^t}\{\mathbf{x}_t\} \times \mathsf{PC}(\mathcal{H}|_{\mathbf{x}_t}) \subseteq \mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t)$$

Then, we can prove the following property of XPC_t.

Lemma 46 *For any $t \in \mathbb{N}$, XPC_t is an analytic subset of the Polish space $\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t)$.*

Proof According to Lemma 45, $\mathsf{PC}(\mathcal{Y}^t)$ is countable. Thus, $\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t)$ is a Polish space. For any $t \in \mathbb{N}$, we have

$$\mathsf{XPC}_t = \cup_{n=1}^{\infty}\Big(\Big(\mathcal{X}^t \times \mathsf{PC}_n(\mathcal{Y}^t)\Big) \cap \cup_{(\theta_1,\ldots,\theta_n)\in\Theta^n}\left\{(\mathbf{x}, \mathbf{y}^1, \ldots, \mathbf{y}^n) \in \mathcal{X}^t \times \mathcal{Y}^{tn} : \mathsf{h}(\theta_i, \mathbf{x}) = \mathbf{y}^i \text{ for all } i \in [n]\right\}\Big)$$

where by $\mathsf{h}(\theta, (x_1, \ldots, x_t)) = (y_1, \ldots, y_t)$, we mean that $\mathsf{h}(\theta, x_\tau) = y_\tau$ for all $\tau \in [t]$. Indeed, for any $(\mathbf{x}, C) \in \mathsf{XPC}_t$, we have $\mathbf{x} \in \mathcal{X}^t$ and $C \in \mathsf{PC}(\mathcal{H}|_{\mathbf{x}})$. Then, by the definition of pseudo-cubes, there exists a finite $n \in \mathbb{N}$ such that $C \in \mathsf{PC}_n(\mathcal{H}|_{\mathbf{x}})$. Since $\mathcal{H}|_{\mathbf{x}} \subseteq \mathcal{Y}^t$, we have $C \in \mathsf{PC}_n(\mathcal{Y}^t)$ and $(\mathbf{x}, C) \in \mathcal{X}^t \times \mathsf{PC}_n(\mathcal{Y}^t)$. Moreover, since $C \subseteq \mathcal{H}|_{\mathbf{x}}$ with $|C| = n$, we can write $C = \{\mathbf{y}^1, \ldots, \mathbf{y}^n\}$

such that there exist $(h_1, \ldots, h_n) \in \mathcal{H}^n$ satisfying $h_i(\mathbf{x}) = \mathbf{y}^i$ for any $i \in [n]$. By Definition 3, there exist $(\theta_1, \ldots, \theta_n) \in \Theta^n$ such that $\mathsf{h}(\theta_i, \mathbf{x}) = \mathbf{y}^i$ for all $i \in [n]$.

On the other hand, if $(\mathbf{x}, \{\mathbf{y}^1, \ldots, \mathbf{y}^n\}) \in \mathcal{X}^t \times \mathsf{PC}_n(\mathcal{Y}^t)$ is such that there exist $(\theta_1, \ldots, \theta_n) \in \Theta^n$ satisfying $\mathsf{h}(\theta_i, \mathbf{x}) = \mathbf{y}^i$ for all $i \in [n]$, we have $C := \{\mathbf{y}^1, \ldots, \mathbf{y}^n\} \subseteq \mathcal{H}|_{\mathbf{x}}$ and C is a pseudo-cube of dimension t. Thus, $C \in \mathsf{PC}(\mathcal{H}|_{\mathbf{x}})$ and $(\mathbf{x}, \{\mathbf{y}^1, \ldots, \mathbf{y}^n\}) \in \mathsf{XPC}_t$.

We claim that the set

$$S_{t,n} := \left\{(\theta_1, \ldots, \theta_n, \mathbf{x}, \mathbf{y}^1, \ldots, \mathbf{y}^n) \in \Theta^n \times \mathcal{X}^t \times \mathcal{Y}^{tn} : \mathsf{h}(\theta_i, \mathbf{x}) = \mathbf{y}^i \text{ for all } i \in [n]\right\}$$

is a Borel set. The reason is as follows. For any $i \in [n]$, define the function

$$l : \Theta^n \times \mathcal{X}^t \times \mathcal{Y}^{tn} \to \{0, 1, \ldots, nt\}, \ \ (\theta_1, \ldots, \theta_n, \mathbf{x}, \mathbf{y}^1, \ldots, \mathbf{y}^n) \mapsto \sum_{i=1}^{n} \sum_{\tau=1}^{t} \mathbb{1}\{\mathsf{h}(\theta_i, x_\tau) \neq y_\tau^i\}.$$

Since h is Borel-measurable, we can conclude that l is also Borel-measurable with the argument analogous to that in the proof of Lemma 31. Thus, $S_{t,n} = l^{-1}(\{0\})$ is a Borel set. Then, the set

$$\begin{aligned} &\cup_{(\theta_1,\ldots,\theta_n)\in\Theta^n} \left\{(\mathbf{x}, \mathbf{y}^1, \ldots, \mathbf{y}^n) \in \mathcal{X}^t \times \mathcal{Y}^{tn} : \mathsf{h}(\theta_i, \mathbf{x}) = \mathbf{y}^i \text{ for all } i \in [n]\right\} \\ = &\cup_{(\theta_1,\ldots,\theta_n)\in\Theta^n} \left\{(\mathbf{x}, \mathbf{y}^1, \ldots, \mathbf{y}^n) \in \mathcal{X}^t \times \mathcal{Y}^{tn} : (\theta_1, \ldots, \theta_n, \mathbf{x}, \mathbf{y}^1, \ldots, \mathbf{y}^n) \in S_{t,n}\right\} \end{aligned}$$

is an analytic set for any $t, n \in \mathbb{N}$. Since $\mathsf{PC}(\mathcal{Y}^t)$ is countable, we know that $\mathsf{PC}_n(\mathcal{Y}^t)$ is countable. Since $\mathcal{X}$ is a Polish space, we have that $\mathcal{X}^t \times \mathsf{PC}_n(\mathcal{Y}^t)$ is an analytic set. In conclusion, XPC_t is an analytic set for any $t \in \mathbb{N}$. ■

Now, for any $0 \leqslant n \leqslant \infty$, define $\mathsf{P}_n := \prod_{t=1}^n (\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t)$ which is the set of positions of length n of the game $\mathfrak{B}$ and $\widetilde{\mathsf{P}}_n := \prod_{t=1}^n (\mathsf{XPC}_t \times \mathcal{Y}^t) \subseteq \mathsf{P}_n$. Define $\mathsf{P} := \cup_{n=0}^\infty \mathsf{P}_n$ (with $\mathsf{P}_0 := \varnothing$) to be the set of all positions of the Gale-Stewart game $\mathfrak{B}$. We can show the following results according to Lemma 46

Corollary 47 *For any $0 \leqslant n \leqslant \infty$, $\widetilde{\mathsf{P}}_n$ is an analytic subset of the Polish space P_n.*

Proof Since $\mathcal{Y}^t$ and $\mathsf{PC}(\mathcal{Y}^t)$ are countable and $\mathcal{X}$ is a Polish space, $\mathsf{P}_n = \prod_{t=1}^n (\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t)$ is also a Polish space for any $0 \leqslant n \leqslant \infty$. By Lemma 46, we know that $\mathsf{XPC}_t \times \mathcal{Y}^t$ is an analytic subset of $\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t$ for any $0 \leqslant t < \infty$. Then, we have that $\widetilde{\mathsf{P}}_n$ is an analytic subset of P_n for any $0 \leqslant n < \infty$. For $n = \infty$, we have that

$$\widetilde{\mathsf{P}}_\infty = \cap_{n=1}^\infty \left(\widetilde{\mathsf{P}}_n \times \prod_{t=n+1}^\infty (\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t)\right).$$

Since $\widetilde{\mathsf{P}}_n \times \prod_{t=n+1}^\infty (\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t)$ is an analytic subset of $\mathsf{P}_\infty = \prod_{t=1}^\infty (\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t)$ for any $1 \leqslant n < \infty$, we have that $\widetilde{\mathsf{P}}_\infty$ is also an analytic subset of P_∞. ■

Then, we can proceed to show that

Lemma 48 $\mathsf{P}_\infty \backslash \mathsf{W}_{\mathfrak{B}}$ *is an analytic set.*

Proof We have

$$\mathsf{P}_\infty\backslash\mathsf{W}_{\mathfrak{B}} = \Big\{(\mathbf{x}_1, C_1, \mathbf{y}_1, \dots) \in \prod_{t=1}^{\infty}(\mathsf{XPC}_t \times \mathcal{Y}^t) : \text{ for all } \tau \in \mathbb{N}, \text{ either } \mathcal{H}_{\mathbf{x}_1,\mathbf{y}_1,\dots,\mathbf{x}_\tau,\mathbf{y}_\tau} \neq \varnothing$$
$$\text{or } \exists\, s \in [\tau] \text{ s.t. } \mathbf{y}_s \notin C_s\Big\}$$
$$= \cap_{\tau=1}^{\infty}\bigg(\Big(\cup_{\theta\in\Theta} \cap_{s=1}^{\tau}\Big\{(\mathbf{x}, C_1, \mathbf{y}_1, \dots,) \in \prod_{t=1}^{\infty}(\mathsf{XPC}_t \times \mathcal{Y}^t) : \mathsf{h}(\theta, \mathbf{x}_s) = \mathbf{y}_s\Big\}\Big)$$
$$\cup \Big(\cup_{s=1}^{\tau}\Big\{(\mathbf{x}, C_1, \mathbf{y}_1, \dots,) \in \prod_{t=1}^{\infty}(\mathsf{XPC}_t \times \mathcal{Y}^t) : \mathbf{y}_s \notin C_s\Big\}\Big)\bigg).$$

By Lemma 45, Lemma 46, and Definition 3, for any $s \in \mathbb{N}$,

$$\Big\{(\theta, \mathbf{x}, C_1, \mathbf{y}_1, \dots,) \in \Theta \times \prod_{t=1}^{\infty}(\mathsf{XPC}_t \times \mathcal{Y}^t) : \mathsf{h}(\theta, \mathbf{x}_s) = \mathbf{y}_s\Big\}$$
$$= \Big(\Theta \times \prod_{t=1}^{\infty}(\mathsf{XPC}_t \times \mathcal{Y}^t)\Big) \cap \Big\{(\theta, \mathbf{x}, C_1, \mathbf{y}_1, \dots,) \in \Theta \times \prod_{t=1}^{\infty}(\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t) : \mathsf{h}(\theta, \mathbf{x}_s) = \mathbf{y}_s\Big\}$$

is an analytic set and

$$\Big\{(\mathbf{x}, C_1, \mathbf{y}_1, \dots,) \in \prod_{t=1}^{\infty}(\mathsf{XPC}_t \times \mathcal{Y}^t) : \mathbf{y}_s \notin C_s\Big\}$$
$$= \Big(\prod_{t=1}^{\infty}(\mathsf{XPC}_t \times \mathcal{Y}^t)\Big) \cap \Big\{(\mathbf{x}, C_1, \mathbf{y}_1, \dots,) \in \prod_{t=1}^{\infty}(\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \times \mathcal{Y}^t) : \mathbf{y}_s \notin C_s\Big\}$$

is also an analytic set. Thus, we have $\mathsf{P}_\infty\backslash\mathsf{W}_{\mathfrak{B}}$ is an analytic set. ■

We have the following lemma relating a winning strategy of P_A in $\mathfrak{B}$ to an infinite DSL tree of $\mathcal{H}$.

Lemma 49 *P_A has a winning strategy in $\mathfrak{B}$ if and only if $\mathcal{H}$ has an infinite DSL tree.*

Proof Suppose that P_A has a winning strategy $\xi_\tau : \prod_{t=1}^{\tau-1}\mathcal{Y}^t \to \mathcal{X}^\tau \times \mathsf{PC}(\mathcal{Y}^\tau)$ for all $\tau \in \mathbb{N}$ in $\mathfrak{B}$. Specifically, for any $(\mathbf{y}_1, \cdots, \mathbf{y}_{\tau-1}) \in \prod_{t=1}^{\tau-1}\mathcal{Y}^t$, we have $\xi_\tau(\mathbf{y}_1, \cdots, \mathbf{y}_{\tau-1}) = (\mathbf{x}_\tau, C)$ for some $\mathbf{x}_\tau \in \mathcal{X}^\tau$ and $C \in \mathsf{PC}(\mathcal{H}|_{\mathbf{x}_\tau})$. For notational convenience, let $\xi_\tau(\mathbf{y}_1, \cdots, \mathbf{y}_{\tau-1})(1)$ denote $\mathbf{x}_\tau$ and let $\xi_\tau(\mathbf{y}_1, \cdots, \mathbf{y}_{\tau-1})(2)$ denote C. Now, define the following infinite tree by induction.

- Let the root node of the tree be labelled with $\xi_1(\varnothing)(1) \in \mathcal{X}$ and have $|\xi_1(\varnothing)(2)|$ children such that each edge between the root node and its children is labelled with a unique element in $\xi_1(\varnothing)(2)$.
- Suppose that for some $\tau \in \mathbb{N}$, all the nodes in level $0, 1, \dots, \tau$ have been defined, all the nodes in level $0, 1, \dots, \tau - 1$ have been labelled, and the edges between each node in level k and its children have been labelled for all $k \in \{0, 1, \cdots, \tau - 1\}$.

Then, for each node v in level τ, denote the labels of the edges along the path eliminating from the root node to node v with $\mathbf{y}_1 \in \mathcal{Y}^1$, $\mathbf{y}_2 \in \mathcal{Y}^2$, …, and $\mathbf{y}_\tau \in \mathcal{Y}^\tau$. Now, let node v be labelled with $\xi_{\tau+1}(\mathbf{y}_1, \cdots, \mathbf{y}_\tau)(1)$ and have $|\xi_{\tau+1}(\mathbf{y}_1, \cdots, \mathbf{y}_\tau)(1)|$ children such that each edge between node v to one of its children is labelled with a unique element in $\xi_{\tau+1}(\mathbf{y}_1, \cdots, \mathbf{y}_\tau)(2)$.

By the definition of the winning strategy of P_A in $\mathfrak{B}$, the infinite tree defined above is an infinite DSL tree for $\mathcal{H}$.

For the other direction, suppose that $\mathcal{H}$ has an infinite DSL tree. For any $k \in \mathbb{N}_0$, denote the set of nodes in the level k of the infinite DSL tree with V_k. Note that if there exists some $\tau \in \mathbb{N}$ such that $C_t \subseteq \mathcal{H}|_{\mathbf{x}_t}$ for all $t \in [\tau]$, $\mathcal{H}|_{\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_{\tau-1},\mathbf{y}_{\tau-1}} \neq \varnothing$, and $\mathbf{y}_\tau \notin C_\tau$, then P_A wins in round τ of $\mathfrak{B}$. Define the following strategy $\xi_\tau : \prod_{t=1}^{\tau-1} \mathcal{Y}^t \to \mathcal{X}^\tau$ for P_A in $\mathfrak{B}$ and a corresponding node mapping $v_\tau : \prod_{t=1}^{\tau-1} \mathcal{Y}^t \to V_{\tau-1}$ by induction for all $\tau \in \mathbb{N}$.

- For $\tau = 1$, let $v_1(\varnothing)$ denote the root node, $\xi_1(\varnothing)(1)$ denote the label of the root node of the DSL tree, and $\xi_1(\varnothing)(2)$ denote the pseudo-cube in $\mathcal{H}|_{\xi_1(\varnothing)(1)}$ consisting of the labels of all the edges between the root node and its children.
- Suppose that for some $\tau \in \{2, 3, \ldots\}$, η_t and v_t has been defined for all $t \in [\tau - 1]$. For any $\mathbf{y}_1 \in \mathcal{Y}^1, \ldots, \mathbf{y}_{\tau-1} \in \mathcal{Y}^{\tau-1}$, there are two cases.
  - If P_A has not won in round $\tau - 1$, define $v_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})$ to be the node in $V_{\tau-1}$ which is the ending node of the path in the DSL tree eliminating from the root along the edges labelled with $\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1}$. Define $\xi_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})(1)$ to be the label of $v_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})$. Define $\xi_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})(2)$ to be the pseudo-cube in $\mathcal{H}|_{\xi_\tau(\mathbf{y}_1,\ldots,\mathbf{y}_{\tau-1})(1)}$ consisting of the labels of all the edges between $v_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})$ and its children.
  - If P_A has already won, define $v_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})$ to be the first child node of $v_{\tau-1}(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-2})$. Define $\xi_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})(1)$ to be the label of $v_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})$. Define $\xi_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})(2)$ to be the pseudo-cube in $\mathcal{H}|_{\xi_\tau(\mathbf{y}_1,\ldots,\mathbf{y}_{\tau-1})(1)}$ consisting of the labels of all the edges between $v_\tau(\mathbf{y}_1, \ldots, \mathbf{y}_{\tau-1})$ and its children.

According to the definition of DSL trees and the rules of $\mathfrak{B}$, $\{\xi_\tau\}_{\tau\in\mathbb{N}}$ is a winning strategy of P_A in $\mathfrak{B}$. ■

Moreover, we can ensure that there is a universally measurable winning strategy for P_L in $\mathfrak{B}$ when $\mathcal{H}$ does not have an infinite DSL tree.

Proposition 50 *If $\mathcal{H}$ does not have an infinite DSL tree, then there is a universally measurable winning strategy for P_L in $\mathfrak{B}$.*

Proof Since $\mathfrak{B}$ is a Gale-Stewart game, according to Lemma 49 and Bousquet et al. (2021, Theorem A.1), we have that if $\mathcal{H}$ does not have an infinite DSL tree, then there is a winning strategy for P_L in $\mathfrak{B}$.

According to Lemma 45, Corollary 47, and Lemma 48, we know that $\mathfrak{B}$ is a Gale-Stewart game such that the action sets of P_A ($\mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t)$, $t \in \mathbb{N}$) are Polish spaces, the action sets of P_L ($\mathcal{Y}^t$, $t \in \mathbb{N}$) are countable, and the set of winning sequences $\mathsf{W}_{\mathfrak{B}}$ for P_L is coanalytic. Then, according to Bousquet et al. (2021, Theorem B.1), P_L has a universally measurable winning strategy ($g_t : \mathsf{P}_{t-1} \times \mathcal{X}^t \times \mathsf{PC}(\mathcal{Y}^t) \to \mathcal{Y}^t$, $t \in \mathbb{N}$) for P_L in $\mathfrak{B}$ if $\mathcal{H}$ does not have an infinite DSL tree. For completeness, we provide the explicit definition of g_t below for $t \in \mathbb{N}$ according to the proof of Bousquet et al. (2021, Theorem B.1). For any $\mathbf{v} \in \mathsf{P}_{t-1}$, $\mathbf{x} \in \mathcal{X}^t$, and $C \in \mathsf{PC}(\mathcal{Y}^t)$, enumerate $\mathcal{Y}^t$ as $\{\mathbf{y}^{(t,i)}\}_{i\in\mathbb{N}}$ and define

$$g_t(\mathbf{v},\mathbf{x},C) := \begin{cases} \mathbf{y}^{(t,i)} & \text{if } \mathrm{val}(\mathbf{v},\mathbf{x},C,\mathbf{y}^{(t,j)}) \geqslant \min\{\mathrm{val}(\mathbf{v}),\mathrm{val}(\varnothing)\} \text{ for all } 1 \leqslant j < i \\ & \quad \text{and } \mathrm{val}(\mathbf{v},\mathbf{x},C,\mathbf{y}^{(t,i)}) < \min\{\mathrm{val}(\mathbf{v}),\mathrm{val}(\varnothing)\}, \\ \mathbf{y}^{(t,1)} & \text{if } \mathrm{val}(\mathbf{v},\mathbf{x},C,\mathbf{y}^{(t,j)}) \geqslant \min\{\mathrm{val}(\mathbf{v}),\mathrm{val}(\varnothing)\} \text{ for all } j \in \mathbb{N}. \end{cases}$$

■

From now on we assume that $\mathcal{H}$ does not have an infinite DSL tree. Analogous to Definition 21, we define the game value $\mathrm{val} : \mathsf{P} \to \mathrm{ORD}^*$ according to Bousquet et al. (2021, Definition B.5). For any $\tau \in \mathbb{N}$, define the mapping $\eta_\tau : \prod_{t=1}^{\tau} \left(\mathsf{XPC}_t \times \mathcal{Y}^t\right) \to \{0,1\}$ by

$$\eta_\tau(\mathbf{v},\mathbf{x},C,\mathbf{y}) := \begin{cases} 1 & \text{if } \mathrm{val}(\mathbf{v},\mathbf{x},C,\mathbf{y}) < \min\{\mathrm{val}(\mathbf{v}),\mathrm{val}(\varnothing)\}, \\ 0 & \text{otherwise,} \end{cases}$$

for any $\mathbf{v} \in \prod_{t=1}^{\tau-1} \left(\mathsf{XPC}_t \times \mathcal{Y}^t\right)$, $(\mathbf{x},C) \in \mathsf{XPC}_\tau$, and $\mathbf{y} \in \mathcal{Y}^\tau$. Define the following online algorithm which given a sequence of feature-label pairs $(x_1,y_1,x_2,y_2,\dots) \in (\mathcal{X}\times\mathcal{Y})^\infty$ chooses a sequence of elements in $\cup_{\tau=0}^{\infty} \prod_{t=1}^{\tau} (\mathsf{XPC}_t \times \mathcal{Y}^t)$ ("patterns"):

- Initialize $\tau_0 \leftarrow 1$.
- At every time step $t \in \mathbb{N}$:
  - Let $\tau_t \leftarrow \tau_{t-1}$.
  - For each $C \in \mathsf{PC}(\mathcal{H}|_{(x_{t-\tau_{t-1}+1},\dots,x_t)})$:
    * If

$$\eta_{\tau_{t-1}}\Big(\bar{\mathbf{x}}_1,\bar{C}_1,\bar{\mathbf{y}}_1,\dots,\bar{\mathbf{x}}_{\tau_{t-1}-1},\bar{C}_{\tau_{t-1}-1},\bar{\mathbf{y}}_{\tau_{t-1}-1}, (x_{t-\tau_{t-1}+1},\dots,x_t),C,(y_{t-\tau_{t-1}+1},\dots,y_t)\Big) = 1 :$$

      · Let $\bar{\mathbf{x}}_{\tau_{t-1}} \leftarrow (x_{t-\tau_{t-1}+1},\dots,x_t)$, $\bar{C}_{\tau_{t-1}} \leftarrow C$, $\bar{\mathbf{y}}_{\tau_{t-1}} \leftarrow (y_{t-\tau_{t-1}+1},\dots,y_t)$, and $\tau_t \leftarrow \tau_{t-1}+1$.
      · Break.

We use $\hat{\mathbf{y}}_t$ to denote the " pattern avoidance mapping" defined after time step t of the above algorithm; specifically, we define

$$\hat{\mathbf{y}}_t(x'_1,\dots,x'_{\tau_t}) := \cup_{C\in\mathsf{PC}(\mathcal{H}|_{(x'_1,\dots,x'_{\tau_t})})} \left\{\mathbf{y}' \in C : \eta_{\tau_t}(\bar{\mathbf{x}}_1,\bar{C}_1,\bar{\mathbf{y}}_1,\dots,\bar{\mathbf{x}}_{\tau_t-1},\bar{C}_{\tau_t-1},\bar{\mathbf{y}}_{\tau_t-1},(x'_1,\dots,x'_{\tau_t}),C,\mathbf{y}') = 1\right\}$$

for any $t \geqslant 0$ and $(x'_1,\dots,x'_{\tau_t}) \in \mathcal{X}^{\tau_t}$. From the above algorithm, we can also define the following functions for any $t \geqslant 0$,

$$T_t : (\mathcal{X}\times\mathcal{Y})^t \to \{1,\dots,t+1\}, \;\; (x_1,y_1,\dots,x_t,y_t) \mapsto \tau_t,$$

and

$$\widehat{\mathbf{Y}}_t : (\mathcal{X} \times \mathcal{Y})^t \times \cup_{s=1}^{t+1} \mathcal{X}^s \to \cup_{s=1}^{t+1} 2^{\mathcal{Y}^s}, \ \ (x_1, y_1, \ldots, x_t, y_t, x'_1, \ldots, x'_{\tau_t}) \mapsto \widehat{\mathbf{y}}_t(x'_1, \ldots, x'_{\tau_t}). \quad (8)$$

We have the following proposition.

Proposition 51 *For any sequence $x_1, y_1, x_2, y_2, \ldots$ that is consistence with $\mathcal{H}$, we have*

$$(y_{t-\tau_{t-1}+1}, \ldots, y_t) \notin \widehat{\mathbf{y}}_{t-1}(x_{t-\tau_{t-1}+1}, \ldots, x_t), \ \tau_{t-1} = \tau_t < \infty, \text{ and } \widehat{\mathbf{y}}_{t-1} = \widehat{\mathbf{y}}_t$$

for all sufficiently large t.

Proof Suppose that there is an infinite sequence of times $1 \leqslant t_1 < t_2 < \cdots$ such that

$$(y_{t_i-\tau_{t_i-1}+1}, \ldots, y_{t_i}) \in \widehat{\mathbf{y}}_{t_i-1}(x_{t_i-\tau_{t_i-1}+1}, \ldots, x_{t_i})$$

for any $i \in \mathbb{N}$. Define $\mathbf{x}_i := (x_{t_i-\tau_{t_i-1}+1}, \ldots, x_{t_i})$, $\mathbf{y}_i := (y_{t_i-\tau_{t_i-1}+1}, \ldots, y_{t_i})$, $C_i := \bar{C}_{\tau_{t_i-1}}$, and $\mathbf{v}_i := (\mathbf{x}_1, C_1, \mathbf{y}_1, \ldots, \mathbf{x}_i, C_i, \mathbf{y}_i)$ for $i \in \mathbb{N}$. Since $\mathcal{H}$ does not have an infinite DSL tree and $\mathsf{W}_{\mathfrak{B}}$ is coanalytic, we have $\mathrm{val}(\varnothing) < \omega_1$. Thus, there is no infinite value-decreasing sequence, which, together with the definition of $\widehat{\mathbf{y}}_t$, implies that $\mathrm{val}(\mathbf{v}_k) = -1$ for some $k \in \mathbb{N}$; i.e., P_L wins at round k of $\mathfrak{B}$ under the sequence of positions $\mathbf{v}_k$. Since we have ensured that $C_i \subseteq \mathsf{PC}(\mathcal{H}|_{\mathbf{x}_i})$ for all $i \in \mathbb{N}$, we must have $\mathcal{H}_{\mathbf{x}_1,\mathbf{y}_1,\ldots,\mathbf{x}_k,\mathbf{y}_k} = \varnothing$ by the winning rule of P_L in $\mathfrak{B}$. However, this contracts the assumption that the sequence $(x_1, y_1, x_2, y_2, \ldots)$ is consistent with $\mathcal{H}$. Thus, there exists some $t_0 \in \mathbb{N}$ such that

$$(y_{t-\tau_{t-1}+1}, \ldots, y_t) \notin \widehat{\mathbf{y}}_{t-1}(x_{t-\tau_{t-1}+1}, \ldots, x_t)$$

for all $t \geqslant t_0$. Then, according to the definition of τ_t and $\widehat{\mathbf{y}}_t$, we have $\tau_{t-1} = \tau_t \leqslant t_0 < \infty$ and $\widehat{\mathbf{y}}_{t-1} = \widehat{\mathbf{y}}_t$ for all $t \geqslant t_0$. ∎

C.3. Universal measurability

In this section, we prove the following proposition about the universal measurability of the functions T_t and $\widehat{\mathbf{Y}}_t$ defined in the previous section.

Proposition 52 *For any $t \geqslant 0$, T_t and $\widehat{\mathbf{Y}}_t$ are universally measurable.*

We start with some definitions of the building blocks for analyzing the universal measurability. For any $t \in \mathbb{N}$ and any $s \in [t]$, fix an arbitrary sequence $1 \leqslant j_1 \leqslant j_2 \leqslant \cdots \leqslant j_s \leqslant t - s + 1$. Define $j_0 := 0$ and $J_0 := 0$. For any $i \in [s]$, define

$$\begin{aligned} J_i :=& 1 + \sum_{k=2}^{i} \left[k - ((j_{k-1} + k - 2 - j_k + 1) \vee 0)\right] \\ =& \frac{i(i+1)}{2} - \sum_{k=2}^{i} ((j_{k-1} + k - 1 - j_k) \vee 0) \end{aligned}$$

and

$$I_i := J_{i-1} + 1 - ((j_{i-1} + i - 2 - j_i + 1) \vee 0) = J_{i-1} + ((j_i - j_{i-1} - i + 2) \wedge 1).$$

In this section, for any $k \in \mathbb{N}$, $i, j \in [k]$, and k-tuple $\mathbf{z} = (z^1, z^2, \ldots, z^k)$, let $z^{i:j}$ denote the subtuple $(z^i, z^{i+1}, \ldots, z^j)$ if $i \leqslant j$ and denote $\varnothing$ if $i > j$. We assume the convention that $\varnothing = \varnothing$.

For any $0 \leqslant i \leqslant s$, define

$$\begin{aligned}\mathsf{F}_{j_1,\ldots,j_i} := \Big\{(\mathbf{x}_1, C_1, \mathbf{y}_1 \ldots, \mathbf{x}_i, C_i, \mathbf{y}_i) \in \prod_{k=1}^{i}(\mathcal{X}^k \times \mathsf{PC}(\mathcal{Y}^k) \times \mathcal{Y}^k) :\\ x_k^{1:(j_{k-1}+k-1-j_k)} = x_{k-1}^{(j_k-j_{k-1}+1):(k-1)} \text{ and}\\ y_k^{1:(j_{k-1}+k-1-j_k)} = y_{k-1}^{(j_k-j_{k-1}+1):(k-1)} \text{ for all } 2 \leqslant k \leqslant i\Big\}.\end{aligned}$$

Then, $\mathsf{F}_{j_1,\ldots,j_i}$ is a closed subset of $\prod_{k=1}^{i}(\mathcal{X}^k \times \mathsf{PC}(\mathcal{Y}^k) \times \mathcal{Y}^k)$ and is also analytic. Define

$$\widetilde{\mathsf{F}}_{j_1,\ldots,j_i} := \mathsf{F}_{j_1\ldots,j_i} \times \prod_{k=i+1}^{\infty}(\mathcal{X}^k \times \mathsf{PC}(\mathcal{Y}^k) \times \mathcal{Y}^k).$$

Then, $\widetilde{\mathsf{F}}_{j_1,\ldots,j_i}$ is an analytic subset of $\prod_{k=1}^{\infty}(\mathcal{X}^k \times \mathsf{PC}(\mathcal{Y}^k) \times \mathcal{Y}^k)$.

For any $C_1 \in \mathsf{PC}(\mathcal{Y}^1), \ldots, C_i \in \mathsf{PC}(\mathcal{Y}^i)$, define Define

$$\mathsf{Z}_{C_1,\ldots,C_i} := \left(\prod_{k=1}^{i}(\mathcal{X}^k \times \{C_k\} \times \mathcal{Y}^k)\right)$$

and

$$\widetilde{\mathsf{Z}}_{C_1,\ldots,C_i} := \left(\prod_{k=1}^{i}(\mathcal{X}^k \times \{C_k\} \times \mathcal{Y}^k)\right) \times \left(\prod_{k=i+1}^{\infty}(\mathcal{X}^k \times \mathsf{PC}(\mathcal{Y}^k) \times \mathcal{Y}^k)\right).$$

Since $\mathsf{PC}(\mathcal{Y}^t)$ is countable by Lemma 45, we have that $\mathsf{Z}_{C_1,\ldots,C_i}$ is an analytic subset of $\prod_{k=1}^{i}(\mathcal{X}^k \times \mathsf{PC}(\mathcal{Y}^k) \times \mathcal{Y}^k)$. Thus, $\widetilde{\mathsf{Z}}_{C_1,\ldots,C_i}$ is an analytic subset of $\prod_{k=1}^{\infty}(\mathcal{X}^k \times \mathsf{PC}(\mathcal{Y}^k) \times \mathcal{Y}^k)$.

Define

$$\begin{aligned}\mathsf{XY}_{j_1,\ldots,j_i,C_1,\ldots,C_i} := \Big\{&(x_1, y_1, \ldots, x_{J_i}, y_{J_i}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} :\\ &(x_1, C_1, y_1, x_{I_2:(I_2+1)}, C_2, y_{I_2:(I_2+1)}, \ldots, x_{I_i:(I_i+i-1)}, C_i, y_{I_i:(I_i+i-1)}) \in \widetilde{\mathsf{P}}_i\Big\}.\end{aligned}$$

Then, we have

$$\begin{aligned}&\mathsf{XY}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\\ =& \bigcup_{(C_1',\ldots,C_i') \in \prod_{k=1}^{i}\mathsf{PC}(\mathcal{Y}^k)} \; \bigcup_{\left(x_2^{1:((j_1+1-j_2)\vee 0)}, y_2^{1:((j_1+1-j_2)\vee 0)}, \ldots, x_i^{1:((j_{i-1}+i-1-j_i)\vee 0)}, y_i^{1:((j_{i-1}+i-1-j_i)\vee 0)}\right) \in \prod_{k=1}^{i}(\mathcal{X}\times\mathcal{Y})^{1:((j_{k-1}+k-1-j_k)\vee 0)}}\\ &\Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} :\\ &\Big(x_1, C_1', y_1, (x_2^{1:((j_1+1-j_2)\vee 0)}, x_{(J_1+1):J_2}), C_2', (y_2^{1:((j_1+1-j_2)\vee 0)}, y_{(J_1+1):J_2}), \cdots,\\ &\quad (x_i^{1:((j_{i-1}+i-1-j_i)\vee 0)}, x_{(J_{i-1}+1):J_i}), C_i', (y_i^{1:((j_{i-1}+i-1-j_i)\vee 0)}, y_{(J_{i-1}+1):J_i})\Big)\\ &\in \widetilde{\mathsf{P}}_i \cap \mathsf{F}_{j_1,\ldots,j_i} \cap \mathsf{Z}_{C_1,\ldots,C_i}\Big\}.\end{aligned}$$

Since P_i, $\mathsf{F}_{j_1,\ldots,j_i}$, and $\mathsf{Z}_{C_1,\ldots,C_i}$ are analytic sets, we can conclude that $\mathsf{XY}_{j_1,\ldots,j_i,C_1,\ldots,C_i}$ is an analytic subset of $(\mathcal{X}\times\mathcal{Y})^{J_i}$.

Define the set

$$\begin{aligned}
&\mathsf{A}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\\
:=&\bigcup_{\mathbf{w}\in\prod_{t=i+1}^{\infty}(\mathsf{XPC}_t\times\mathcal{Y}^t)}\big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i})\in(\mathcal{X}\times\mathcal{Y})^{J_i}:\\
&(x_1,C_1,y_1,x_{I_2:(I_2+1)},C_2,y_{I_2:(I_2+1)},\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)},\mathbf{w})\in\mathsf{P}_\infty\backslash\mathsf{W}_{\mathfrak{B}}\big\}\\
=&\bigcup_{\mathbf{w}\in\prod_{t=i+1}^{\infty}(\mathsf{XPC}_t\times\mathcal{Y}^i)}\ \bigcup_{(C_1',\ldots,C_i')\in\prod_{k=1}^{i}\mathsf{PC}(\mathcal{Y}^k)}\\
&\bigcup_{\left(x_2^{1:((j_1+1-j_2)\vee 0)},y_2^{1:((j_1+1-j_2)\vee 0)},\ldots,x_i^{1:((j_{i-1}+i-1-j_i)\vee 0)},y_i^{1:((j_{i-1}+i-1-j_i)\vee 0)}\right)\in\prod_{k=1}^{i}(\mathcal{X}\times\mathcal{Y})^{1:((j_{k-1}+k-1-j_k)\vee 0)}}\\
&\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i})\in(\mathcal{X}\times\mathcal{Y})^{J_i}:\\
&\Big(x_1,C_1',y_1,\big(x_2^{1:((j_1+1-j_2)\vee 0)},x_{(J_1+1):J_2}\big),C_2',\big(y_2^{1:((j_1+1-j_2)\vee 0)},y_{(J_1+1):J_2}\big),\ \cdots,\\
&\big(x_i^{1:((j_{i-1}+i-1-j_i)\vee 0)},x_{(J_{i-1}+1):J_i}\big),C_i',\big(y_i^{1:((j_{i-1}+i-1-j_i)\vee 0)},y_{(J_{i-1}+1):J_i}\big),\mathbf{w}\Big)\\
&\in(\mathsf{P}_\infty\backslash\mathsf{W}_{\mathfrak{B}})\cap\widetilde{\mathsf{F}}_{j_1,\ldots,j_i}\cap\widetilde{\mathsf{Z}}_{C_1,\ldots,C_i}\Big\}.
\end{aligned}$$

By Lemma 48 and the analysis above, $\mathsf{P}_\infty\backslash\mathsf{W}_{\mathfrak{B}}$, $\widetilde{\mathsf{F}}_{j_1,\ldots,j_i}$, and $\widetilde{\mathsf{Z}}_{C_1,\ldots,C_i}$ are analytic subsets of P_∞. Moreover, $\prod_{k=1}^{i}(\mathcal{X}\times\mathcal{Y})^{1:((j_{k-1}+k-1-j_k)\vee 0)}$ is a Polish space. Thus, $\mathsf{A}_{j_1,\ldots,j_i,C_1,\ldots,C_i}$ is an analytic subsets of $(\mathcal{X}\times\mathcal{Y})^{J_i}$.

For any $\kappa\in\mathrm{ORD}\cup\{-1\}$, any $i\in\{2,3,\ldots,s\}$, and any $C_1\in\mathsf{PC}(\mathcal{Y}^1),\ldots,C_i\in\mathsf{PC}(\mathcal{Y}^i)$, define the sets

$$\mathsf{A}_i:=\bigcup_{\mathbf{w}\in\prod_{t=i+1}^{\infty}(\mathsf{XPC}_t\times\mathcal{Y}^t)}\{\mathbf{v}\in\mathsf{P}_i:(\mathbf{v},\mathbf{w})\in\mathsf{P}_\infty\backslash\mathsf{W}_{\mathfrak{B}}\},$$

$$\mathsf{A}_i^\kappa:=\{\mathbf{v}\in\mathsf{A}_i:\mathrm{val}(\mathbf{v})>\kappa\},$$

and

$$\begin{aligned}
&\mathsf{A}^\kappa_{j_1,\ldots,j_i,C_1,\ldots,C_i}\\
:=&\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i})\in\mathsf{A}_{j_1,\ldots,j_i,C_1,\ldots,C_i}:\\
&\quad\mathrm{val}\big(x_1,C_1,y_1,x_{I_2:(I_2+1)},C_2,y_{I_2:(I_2+1)},\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)}\big)>\kappa\Big\}.
\end{aligned}$$

When $\kappa = -1$, we have $\mathsf{A}^{-1}_{j_1,\ldots,j_i,C_1,\ldots,C_i} = \mathsf{A}_{j_1,\ldots,j_i,C_1,\ldots,C_i}$ and $\mathsf{A}_i^{-1} = \mathsf{A}_i$. According to Bousquet et al. (2021, Corollary B.11), A_i^{κ} is an analytic subset of P_i for any $-1 \leqslant \kappa < \omega_1$. Then, since

$$\begin{aligned}
&\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\\
=&\bigcup_{(C_1',\ldots,C_i')\in\prod_{k=1}^{i}\mathsf{PC}(\mathcal{Y}^k)}\\
&\bigcup_{\left(x_2^{1:((j_1+1-j_2)\vee 0)},y_2^{1:((j_1+1-j_2)\vee 0)},\ldots,x_i^{1:((j_{i-1}+i-1-j_i)\vee 0)},y_i^{1:((j_{i-1}+i-1-j_i)\vee 0)}\right)\in\prod_{k=1}^{i}(\mathcal{X}\times\mathcal{Y})^{1:((j_{k-1}+k-1-j_k)\vee 0)}}\\
&\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i})\in(\mathcal{X}\times\mathcal{Y})^{J_i}:\\
&\Big(x_1,C_1',y_1,\big(x_2^{1:((j_1+1-j_2)\vee 0)},x_{(J_1+1):J_2}\big),C_2',\big(y_2^{1:((j_1+1-j_2)\vee 0)},y_{(J_1+1):J_2}\big),\ \cdots,\\
&\big(x_i^{1:((j_{i-1}+i-1-j_i)\vee 0)},x_{(J_{i-1}+1):J_i}\big),C_i',\big(y_i^{1:((j_{i-1}+i-1-j_i)\vee 0)},y_{(J_{i-1}+1):J_i}\big)\Big)\\
&\in\mathsf{A}_i^{\kappa}\cap\mathsf{F}_{j_1,\ldots,j_i}\cap\mathsf{Z}_{C_1,\ldots,C_i}\Big\},
\end{aligned}$$

we can conclude that $\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}$ is analytic subset of $(\mathcal{X}\times\mathcal{Y})^{J_i}$ for any $-1 \leqslant \kappa < \omega_1$.

According to Lemma 49 and the definition of the game value (Bousquet et al., 2021, Definition B.5), we have $\mathrm{val}(\varnothing) < \Omega$ under the assumption that $\mathcal{H}$ does not have an infinite DSL tree. Then, by Bousquet et al. (2021, Lemma B.7), we immediately have the following lemma.

Lemma 53 $\mathrm{val}(\varnothing) < \omega_1$ *when $\mathcal{H}$ does not have an infinite DSL tree.*

Now, for any $m \in \mathbb{N}$ with $j_i \leqslant m \leqslant j_{i+1}$, any $(C_1,\ldots,C_{i+1}) \in \prod_{k=1}^{i+1}\mathsf{PC}(\mathcal{Y}^k)$, and any $\mathbf{y}' = (y_1',\ldots,y'_{(m-j_i+1)\wedge(i+1)}) \in \mathcal{Y}^{(m-j_i+1)\wedge(i+1)}$, define the set

$$\begin{aligned}
&\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1},\mathbf{y}'}\\
:=&\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},x'_{1:((m-j_i+1)\wedge(i+1))})\in(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}:\\
&(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x_k',y_k')_{k=1}^{((m-j_i+1)\wedge(i+1))})\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}},\\
&\mathrm{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)},\\
&\big(x_{(J_i+1-((j_i+i-m)\vee 0)):J_i},x'_{1:((m-j_i+1)\wedge(i+1))}\big),C_{i+1},\big(y_{(J_i+1-((j_i+i-m)\vee 0)):J_i},y'_{1:((m-j_i+1)\wedge(i+1))}\big)\big)\\
&<\min\left\{\mathrm{val}(\varnothing),\mathrm{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)}\big)\right\}\Big\}.
\end{aligned}$$

We prove the following result about the above set.

Lemma 54 *For any $m \in \mathbb{N}$ with $j_i \leqslant m \leqslant j_{i+1}$, any $(C_1,\ldots,C_{i+1}) \in \prod_{k=1}^{i+1}\mathsf{PC}(\mathcal{Y}^k)$, and any $\mathbf{y}' = (y_1',\ldots,y'_{(m-j_i+1)\wedge(i+1)}) \in \mathcal{Y}^{(m-j_i+1)\wedge(i+1)}$, $\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1},\mathbf{y}'}$ is universally measurable.*

Proof We can write

$$
\begin{aligned}
&\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1},\mathbf{y}'}\\
=&\bigcup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)}\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},x'_{1:((m-j_i+1)\wedge(i+1))})\in(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}:\\
&(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)})\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}},\\
&\mathrm{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)},\big(x_{(J_i+1-((j_i+i-m)\vee 0)):J_i},x'_{1:((m-j_i+1)\wedge(i+1))}\big),\\
&C_{i+1},\big(y_{(J_i+1-((j_i+i-m)\vee 0)):J_i},y'_{1:((m-j_i+1)\wedge(i+1))}\big)\big)\leqslant\kappa,\\
&\mathrm{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)}\big)>\kappa\Big\}\\
=&\bigcup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)}\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},x'_{1:((m-j_i+1)\wedge(i+1))})\in(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}:\\
&\big(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big)\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\backslash\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}},\\
&\big(x_1,y_1,\ldots,x_{J_i},y_{J_i}\big)\in\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\Big\}\\
=&\bigcup_{-1\leqslant\kappa<\mathrm{val}(\varnothing)}\Big(\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},x'_{1:((m-j_i+1)\wedge(i+1))})\in(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}:\\
&\big(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big)\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\backslash\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}\\
&\bigcap\Big(\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}\Big)\Big).
\end{aligned}
$$

By Lemma 53 and the previous results, for any $-1\leqslant\kappa<\mathrm{val}(\varnothing)<\omega_1$, we have that $\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}$ is an analytic subset of $(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}$. Moreover, for any $-1\leqslant\kappa<\mathrm{val}(\varnothing)<\omega_1$, we have

$$
\begin{aligned}
&\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},x'_{1:((m-j_i+1)\wedge(i+1))})\in(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}:\\
&\big(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big)\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\backslash\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}\\
=&\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},x'_{1:((m-j_i+1)\wedge(i+1))})\in(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}:\\
&\big(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big)\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}\\
&\Big\backslash\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},x'_{1:((m-j_i+1)\wedge(i+1))})\in(\mathcal{X}\times\mathcal{Y})^{J_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}:\\
&\quad\big(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big)\in\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}
\end{aligned}
$$

with

$$\Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, x'_{1:((m-j_i+1)\wedge(i+1))}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)} :$$
$$\big(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, (x'_k, y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big) \in \mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}$$
$$= \bigcup_{\mathbf{y}'' \in \mathcal{Y}^{(m-j_i+1)\wedge(i+1)}} \Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, x'_{1:((m-j_i+1)\wedge(i+1))}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)} :$$
$$\big(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, (x'_k, y''_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big) \in \mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$$
$$\bigcap \Big((\mathcal{X} \times \mathcal{Y})^{J_i} \times \prod_{k=1}^{(m-j_i+1)\wedge i} (\mathcal{X} \times \{y'_k\})\Big)\Big\}$$

and

$$\Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, x'_{1:((m-j_i+1)\wedge(i+1))}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)} :$$
$$\big(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, (x'_k, y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big) \in \mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}$$
$$= \bigcup_{\mathbf{y}'' \in \mathcal{Y}^{(m-j_i+1)\wedge(i+1)}} \Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, x'_{1:((m-j_i+1)\wedge(i+1))}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)} :$$
$$\big(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, (x'_k, y''_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big) \in \mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$$
$$\bigcap \Big((\mathcal{X} \times \mathcal{Y})^{J_i} \times \prod_{k=1}^{(m-j_i+1)\wedge i} (\mathcal{X} \times \{y'_k\})\Big)\Big\}.$$

Since we have proved that $\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$ and $\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$ are analytic subsets of $(\mathcal{X} \times \mathcal{Y})^{J_i+((m-j_i+1)\wedge(i+1))}$, we can conclude that the set

$$\Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, x'_{1:((m-j_i+1)\wedge(i+1))}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)} :$$
$$\big(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, (x'_k, y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big) \in \mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}$$

and the set

$$\Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, x'_{1:((m-j_i+1)\wedge(i+1))}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)} :$$
$$\big(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, (x'_k, y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big) \in \mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}$$

are both analytic subsets of $(\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)}$. Therefore, the set

$$\Big\{(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, x'_{1:((m-j_i+1)\wedge(i+1))}) \in (\mathcal{X} \times \mathcal{Y})^{J_i} \times \mathcal{X}^{(m-j_i+1)\wedge(i+1)} :$$
$$\big(x_1, y_1, \ldots, x_{J_i}, y_{J_i}, (x'_k, y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big) \in \mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}} \setminus \mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\}$$

is universally measurable. It follows from the fact that $\mathrm{val}(\varnothing) < \omega_1$ that $\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1},\mathbf{y}'}$ is universally measurable. ■

Next, we define the set

$$\begin{aligned}
&\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\\
:=&\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{((m-j_i+1)\wedge(i+1))})\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}:\\
&\operatorname{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)},\\
&\big(x_{((J_i+1-((j_i+i-m)\vee 0)):J_i)},x'_{1:((m-j_i+1)\wedge(i+1))}\big),C_{i+1},\big(y_{((J_i+1-((j_i+i-m)\vee 0)):J_i)},y'_{1:((m-j_i+1)\wedge(i+1))}\big)\big)\\
&<\min\big\{\operatorname{val}(\varnothing),\operatorname{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)}\big)\big\}\Big\}
\end{aligned}$$

and prove the following lemma.

Lemma 55 *For any $m\in\mathbb{N}$ with $j_i\leqslant m\leqslant j_{i+1}$ and any $(C_1,\ldots,C_{i+1})\in\prod_{k=1}^{i+1}\mathsf{PC}(\mathcal{Y}^k)$, $\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$ is universally measurable.*

Proof We have

$$\begin{aligned}
&\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\\
=&\bigcup_{-1\leqslant\kappa<\operatorname{val}(\varnothing)}\Big\{(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{((m-j_i+1)\wedge(i+1))})\in\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}:\\
&\operatorname{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)},\big(x_{((J_i+1-((j_i+i-m)\vee 0)):J_i)},x'_{1:((m-j_i+1)\wedge(i+1))}\big),\\
&C_{i+1},\big(y_{((J_i+1-((j_i+i-m)\vee 0)):J_i)},y'_{1:((m-j_i+1)\wedge(i+1))}\big)\big)\leqslant\kappa,\\
&\operatorname{val}\big(x_1,C_1,y_1,\ldots,x_{I_i:(I_i+i-1)},C_i,y_{I_i:(I_i+i-1)}\big)>\kappa\Big\}\\
=&\bigcup_{-1\leqslant\kappa<\operatorname{val}(\varnothing)}\Big\{\big(x_1,y_1,\ldots,x_{J_i},y_{J_i},(x'_k,y'_k)_{k=1}^{(m-j_i+1)\wedge(i+1)}\big)\in\\
&\qquad\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\backslash\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}:\big(x_1,y_1,\ldots,x_{J_i},y_{J_i}\big)\in\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\Big\}\\
=&\bigcup_{-1\leqslant\kappa<\operatorname{val}(\varnothing)}\Big(\big(\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\backslash\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\big)\\
&\qquad\bigcap\big(\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}\times\mathcal{X}^{(m-j_i+1)\wedge(i+1)}\big)\Big).
\end{aligned}$$

Since we have proved that $\mathsf{XY}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$, $\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$, and $\mathsf{A}^{\kappa}_{j_1,\ldots,j_i,C_1,\ldots,C_i}$ are analytic and $\operatorname{val}(\varnothing)<\omega_1$ (Lemma 53), we can conclude that $\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$ is universally measurable. ■

Then, the following set

$$\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i}:=\bigcup_{C_{i+1}\in\mathsf{PC}(\mathcal{Y}^{i+1})}\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$$

is also universally measurable because $\mathsf{PC}(\mathcal{Y}^{i+1})$ is countable.

Before proceeding to the next step, we will need the following lemmas regarding universal measurability. For any measure spaces $(A,\mathscr{A})$ and $(B,\mathscr{B})$, let $\mathscr{A}^*$ and $\mathscr{B}^*$ denote the universal completion of $\mathscr{A}$ and $\mathscr{B}$, respetively. Let $\mathscr{A}\times\mathscr{B}$ denote the product σ-algebra of $\mathscr{A}$ and $\mathscr{B}$ on

$(A \times B)$. Note that when A and B are Polish spaces, $\mathscr{A} \times \mathscr{B}$ is also the Borel σ-algebra of $A \times B$. A function $f : A \to B$ is called $\mathscr{A}/\mathscr{B}$-measurable if $f^{-1}(E) \in \mathscr{A}$ for any $E \in \mathscr{B}$. We prove the following lemmas.

Lemma 56 *For any two measurable space $(A, \mathscr{A})$ and $(B, \mathscr{B})$, any function $f : A \to B$ is $\mathscr{A}^*/\mathscr{B}^*$-measurable if and only if f is $\mathscr{A}^*/\mathscr{B}$-measurable.*

Proof Suppose that f is $\mathscr{A}^*/\mathscr{B}$-measurable. For any probability measure μ_A on $(A, \mathscr{A})$, let $(A, \mathscr{A}^*_{\mu_A}, \mu^*_A)$ denote the completion of $(A, \mathscr{A}, \mu_A)$. Then, $(A, \mathscr{A}^*, \mu^*_A)$ is also a probability space because $\mathscr{A}^* \subseteq \mathscr{A}^*_{\mu_A}$. Since f is $\mathscr{A}^*/\mathscr{B}$-measurable, we can define $\mu_B : \mathscr{B} \to [0, 1]$, $E \mapsto \mu^*_A(f^{-1}(E))$ which is the pushforward of μ^*_A by f.

For any $S \in \mathscr{B}^*$, there exist $U, V \in \mathscr{B}$ such that $U \subseteq S \subseteq V$ and $\mu_B(V \backslash U) = 0$. Then, we have $f^{-1}(U) \subseteq f^{-1}(S) \subseteq f^{-1}(V)$ and $f^{-1}(U), f^{-1}(V) \in \mathscr{A}^*$ which implies that there exist $U_l, U_u, V_l, V_u \in \mathscr{A}$ such that $U_l \subseteq f^{-1}(U) \subseteq U_u$, $V_l \subseteq f^{-1}(V) \subseteq V_u$ and $\mu_A(U_u \backslash U_l) = \mu_A(V_u \backslash V_l) = 0$. Moreover, it follows from the definition of μ_B that

$$\mu^*_A(f^{-1}(V) \backslash f^{-1}(U)) = \mu^*_A(f^{-1}(V \backslash U)) = \mu_B(V \backslash U) = 0.$$

Since μ^*_A is the completion of μ_A, there exists $K \in \mathscr{A}$ with $K \supseteq f^{-1}(V) \backslash f^{-1}(U)$ and $\mu_A(K) = 0$. Therefore, we have $U_l \subseteq f^{-1}(U) \subseteq f^{-1}(S) \subseteq f^{-1}(V) \subseteq V_u$ and

$$V_u \backslash U_l \subseteq (V_u \backslash V_l) \cup (U_u \backslash U_l) \cup K.$$

Since $\mu_A(V_u \backslash V_l) = \mu_A(U_u \backslash U_l) = \mu_A(K) = 0$, we have $\mu_A(V_u \backslash U_l) = 0$. Thus, by the arbitrariness of μ, we can conclude that $f^{-1}(S) \in \mathscr{A}^*$, which implies that f is $\mathscr{A}^*/\mathscr{B}^*$.

The other direction is trivial. ■

Lemma 57 *Consider any $n \in \mathbb{N}$ Polish space $A_1, \ldots, A_n$ with their Borel σ-algebras denoted as $\mathscr{A}_1, \ldots, \mathscr{A}_n$, respectively. For any $m \in [n]$, any sequence $1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n$, and any set $E \in (\prod_{k=1}^m \mathscr{A}_{i_k})^*$, then we have*

$$\widetilde{E} := \left\{ (x_1, \ldots, x_n) \in \prod_{j=1}^n A_j : (x_{i_1}, \ldots, x_{i_m}) \in E \right\} \in \left(\prod_{j=1}^n \mathscr{A}_j \right)^* .$$

Proof Consider the following collections of subsets of $\prod_{j=1}^n A_j$

$$\mathfrak{G} := \left\{ \left\{ (x_1, \ldots, x_n) \in \prod_{j=1}^n A_j : x_{i_1} \in B_1, \ldots, x_{i_m} \in B_m \right\} : B_1 \in \mathscr{A}_{i_1}, \ldots, B_m \in \mathscr{A}_{i_m} \right\}.$$

It is easy to see that $\mathfrak{G}$ is a π-system on $\prod_{j=1}^n A_j$. Define $\mathcal{G} := \sigma(\mathfrak{G})$ to be the σ-algebra generated by $\mathfrak{G}$ and define following the collection of subsets of $\prod_{j=1}^n A_j$

$$\mathcal{C} := \left\{ \left\{ (x_1, \ldots, x_n) \in \prod_{j=1}^n A_j : (x_{i_1}, \ldots, x_{i_m}) \in S \right\} : S \in \prod_{j=1}^m \mathscr{A}_{i_k} \right\}.$$

It is obvious that $\mathcal{C}$ is a σ-algebra on $\prod_{j=1}^{n} A_j$. Since $\mathfrak{G} \subseteq \mathcal{C}$, by Dynkin's π-λ theorem, we have $\mathcal{G} = \sigma(\mathfrak{G}) \subseteq \mathcal{C}$.

Next, define the following collection of subsets of $\prod_{k=1}^{m} A_{i_k}$

$$\mathfrak{F} := \left\{ G|_{\prod_{k=1}^{m} A_{i_k}} : G \in \mathcal{G} \right\}$$

where for any $G \subseteq \prod_{j=1}^{n} A_j$, we define

$$G|_{\prod_{k=1}^{m} A_{i_k}} := \left\{ (x_{i_1}, \ldots, x_{i_m}) : \exists (x'_1, \ldots, x'_n) \in G \text{ s.t. } x'_{i_k} = x_{i_k},\ \forall\, j \in [n] \right\}. \tag{9}$$

We now show that $\mathfrak{F}$ is a σ-algebra on $\prod_{k=1}^{m} A_{i_k}$.

- Since $\prod_{j=1}^{n} A_j \in \mathfrak{G}$, we have $\prod_{k=1}^{m} A_{i_k} \in \mathfrak{F}$.
- For any $G'_1, G'_2 \in \mathfrak{F}$ with $G'_1 \subseteq G'_2$, there exists $G_l \in \mathcal{G}$ such that $G'_l = G_l|_{\prod_{k=1}^{m} A_{i_k}}$ for $l = 1, 2$. By (9) and the facts that $G'_1 \subseteq G'_2$ and $\mathcal{G} \subseteq \mathcal{C}$, we must have $G_1 \subseteq G_2$. Since $\mathcal{G}$ is a σ-algebra, we have $G_2 \backslash G_1 \in \mathcal{G}$. By $\mathcal{G} \subseteq \mathcal{C}$ again, we have $(G_2 \backslash G_1)|_{\prod_{k=1}^{m} A_{i_k}} = G'_2 \backslash G'_1$, which implies that $G_2 \backslash G_1 \in \mathfrak{F}$.
- For any $G'_1, G'_2, \cdots \in \mathfrak{F}$, there exists $G_1, G_2, \cdots \in \mathcal{G}$ such that $G'_l = G_l|_{\prod_{k=1}^{m} A_{i_k}}$ for all $l \in \mathbb{N}$. Then, we have $\cup_{l=1}^{\infty} G_l \in \mathcal{G}$. Since $\mathcal{G} \subseteq \mathcal{C}$, we have $\cup_{l=1}^{\infty} G'_l = (\cup_{l=1}^{\infty} G_l)|_{\prod_{k=1}^{m} A_{i_k}} \in \mathfrak{F}$.

Thus, $\mathfrak{F}$ is a σ-algebra on $\prod_{k=1}^{m} A_{i_k}$.

By the definition of product σ-algebras, we have $\prod_{k=1}^{m} \mathscr{A}_{i_k} = \sigma(\mathfrak{C})$ where

$$\mathfrak{C} := \left\{ \prod_{k=1}^{m} B_k : B_1 \in \mathscr{A}_{i_1}, \ldots, B_m \in \mathscr{A}_{i_m} \right\}$$

is a π-system on $\prod_{k=1}^{m} A_{i_k}$. Obviously, $\mathfrak{C} \subseteq \mathfrak{F}$. Then, by Dynkin's π-λ theorem, we have $\prod_{k=1}^{m} \mathscr{A}_{i_k} = \sigma(\mathfrak{C}) \subseteq \mathfrak{F}$. By the definition of $\mathfrak{F}$ and $\mathcal{C}$ as well as the fact that $\mathcal{G} \subseteq \mathcal{C}$, we have $\mathcal{C} \subseteq \mathcal{G}$. It follows that $\mathcal{C} = \mathcal{G}$. Since $\mathfrak{G}$ is a subset of the collection of all rectangles on $\prod_{j=1}^{n} A_j$, we have $\mathcal{G} \subseteq \prod_{j=1}^{n} \mathscr{A}_j$. Thus, $\mathcal{C} \subseteq \prod_{j=1}^{n} \mathscr{A}_j$.

For any probability measure μ on $\prod_{j=1}^{n} \mathscr{A}_j$, consider its projection $\mu_{i_1,\ldots,i_m} := \mu|_{\prod_{k=1}^{m} A_{i_k}}$ on $\prod_{k=1}^{m} A_{i_k}$ defined by

$$\mu_{i_1,\ldots,i_m}(S) := \mu\left(\left\{ (x_1, \ldots, x_n) \in \prod_{j=1}^{n} A_j : (x_{i_1}, \ldots, x_{i_m}) \in S \right\} \right) \quad \forall S \in \prod_{k=1}^{m} \mathscr{A}_{i_k}. \tag{10}$$

Since we have proved $\mathcal{C} \subseteq \prod_{j=1}^{n} \mathscr{A}_j$, the above definition makes sense and

$$\left(\prod_{k=1}^{m} A_{i_k}, \prod_{k=1}^{m} \mathscr{A}_{i_k}, \mu_{i_1,\ldots,i_m} \right)$$

is indeed a probability space.

Since $E \in (\prod_{k=1}^m \mathscr{A}_{i_k})^*$ and $\mu_{i_1,\ldots,i_m}$ is a probability measure on $(\prod_{k=1}^m A_{i_k}, \prod_{k=1}^m \mathscr{A}_{i_k})$, there exist sets $U, V \in \prod_{k=1}^m \mathscr{A}_{i_k}$ such that $U \subseteq E \subseteq V$ and $\mu_{i_1,\ldots,i_m}(V \backslash U) = 0$. We define

$$\widetilde{U} := \left\{(x_1, \ldots, x_n) \in \prod_{j=1}^n A_j : (x_{i_1}, \ldots, x_{i_m}) \in U\right\}$$

and

$$\widetilde{V} := \left\{(x_1, \ldots, x_n) \in \prod_{j=1}^n A_j : (x_{i_1}, \ldots, x_{i_m}) \in V\right\}.$$

By definition, we have $\widetilde{U}, \widetilde{E} \in \mathcal{C} \subseteq \prod_{j=1}^n \mathscr{A}_j$ with $\widetilde{U} \subseteq \widetilde{E} \subseteq \widetilde{V}$. Moreover, by (10), we have

$$\mu(\widetilde{V} \backslash \widetilde{U}) = \mu_{i_1,\ldots,i_m}(V \backslash U) = 0.$$

Thus, we can conclude that $\widetilde{E} \in \left(\prod_{j=1}^n \mathscr{A}_j\right)^*$. ■

Now, we define the sets

$$\begin{aligned}
&\mathsf{D}^+_{t,(j_1,\ldots,j_i),m,(C_1,\ldots,C_i),C_{i+1}} \\
:=&\Big\{(x_1, y_1, \ldots, x_t, y_t) \in (\mathcal{X} \times \mathcal{Y})^t : \big(x_{j_1}, y_{j_1}, x_{((j_1+1)\vee j_2):(j_2+1)}, y_{((j_1+1)\vee j_2):(j_2+1)}, \ldots, \\
&x_{((j_{i-1}+i-1)\vee j_i):(j_i+i-1)}, y_{((j_{i-1}+i-1)\vee j_i):(j_i+i-1)}, x_{((j_i+i)\vee m):(m+i)}, y_{((j_i+i)\vee m):(m+i)}\big) \in \\
&\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}\Big\},
\end{aligned}$$

$$\begin{aligned}
&\mathsf{D}^-_{t,(j_1,\ldots,j_i),m,(C_1,\ldots,C_i)} \\
:=&\Big\{(x_1, y_1, \ldots, x_t, y_t) \in (\mathcal{X} \times \mathcal{Y})^t : \big(x_{j_1}, y_{j_1}, x_{((j_1+1)\vee j_2):(j_2+1)}, y_{((j_1+1)\vee j_2):(j_2+1)}, \ldots, \\
&x_{((j_{i-1}+i-1)\vee j_i):(j_i+i-1)}, y_{((j_{i-1}+i-1)\vee j_i):(j_i+i-1)}, x_{((j_i+i)\vee m):(m+i)}, y_{((j_i+i)\vee m):(m+i)}\big) \in \\
&\Big(\mathsf{XY}_{j_1,\ldots,j_i,C_1,\ldots,C_i} \times (\mathcal{X} \times \mathcal{Y})^{(m-j_i+1)\wedge(i+1)}\Big) \backslash \mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i}\Big\},
\end{aligned}$$

and

$$\begin{aligned}
&\mathsf{D}^\vee_{t,j_1,\ldots,j_i,C_1,\ldots,C_i,\mathbf{y}'} \\
:=&\bigcup_{C_{i+1}\in\mathsf{PC}(\mathcal{Y}^{i+1})} \Big\{(x_1, y_1, \ldots, x_t, y_t, \mathbf{x}') \in (\mathcal{X} \times \mathcal{Y})^t \times \mathcal{X}^{i+1} : \\
&\big(x_{j_1}, y_{j_1}, x_{((j_1+1)\vee j_2):(j_2+1)}, y_{((j_1+1)\vee j_2):(j_2+1)}, \ldots, \\
&x_{((j_{i-1}+i-1)\vee j_i):(j_i+i-1)}, y_{((j_{i-1}+i-1)\vee j_i):(j_i+i-1)}, \mathbf{x}'\big) \in \mathsf{D}_{j_1,\ldots,j_i,t+1,C_1,\ldots,C_i,C_{i+1},\mathbf{y}'}\Big\}.
\end{aligned}$$

Since we have proved that the sets $\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i,C_{i+1}}$, $\mathsf{XY}_{j_1,\ldots,j_i,C_1,\ldots,C_i}$, $\mathsf{D}_{j_1,\ldots,j_i,m,C_1,\ldots,C_i}$, and $\mathsf{D}_{j_1,\ldots,j_i,t+1,C_1,\ldots,C_i,C_{i+1},\mathbf{y}'}$ are universally measurable, by Lemma 57, we know that

$$\mathsf{D}^+_{t,(j_1,\ldots,j_i),m,(C_1,\ldots,C_i),C_{i+1}},\ \mathsf{D}^-_{t,(j_1,\ldots,j_i),m,(C_1,\ldots,C_i)},\ \text{and } \mathsf{D}^\vee_{t,j_1,\ldots,j_i,C_1,\ldots,C_i,\mathbf{y}'}$$

are also universally measurable.

For any $t \geqslant 1$, $0 \leqslant i < t$, $1 \leqslant j_1 \leqslant \cdots \leqslant j_i \leqslant t - i$, $(C_1, \ldots, C_i) \in \prod_{k=1}^{i} \mathsf{PC}(\mathcal{Y}^i)$, and $\mathbf{y}' \in \mathcal{Y}^{i+1}$, define the set

Define the following sets

$$\mathsf{V}_{t,j_1,\ldots,j_s,C_1,\ldots,C_s} := \big((\mathsf{D}^{-}_{1,\varnothing,1,\varnothing})^{j_1-1} \times \mathsf{D}^{+}_{1,\varnothing,1,\varnothing,C_1} \times (\mathcal{X} \times \mathcal{Y})^{t-j_1}\big) \bigcap \Big[\cap_{i=1}^{s-1} \big(\big(\cap_{m=j_i}^{j_{i+1}-1} \mathsf{D}^{-}_{t,(j_1,\ldots,j_i),m,(C_1,\ldots,C_i)}\big) \cap \mathsf{D}^{+}_{t,(j_1,\ldots,j_i),j_{i+1},(C_1,\ldots,C_i),C_{i+1}}\big)\Big] \bigcap \big(\cap_{m=j_s}^{t-s} \mathsf{D}^{-}_{t,(j_1,\ldots,j_s),m,(C_1,\ldots,C_s)}\big)$$

and

$$\mathsf{V}_{t,j_1,\ldots,j_s} := \bigcup_{(C_1,\ldots,C_s) \in \prod_{k=1}^{s} \mathsf{PC}(\mathcal{Y}^k)} \mathsf{V}_{t,j_1,\ldots,j_s,C_1,\ldots,C_s}.$$

By the results above, we can conclude that $\mathsf{V}_{t,j_1,\ldots,j_s,C_1,\ldots,C_s}$ is universally measurable. Thus, $\mathsf{V}_{t,j_1,\ldots,j_s}$ is also universally measurable.

Finally, we can complete the proof.

Proof of Proposition 52 First note that the function $T_0 : \varnothing \to \{1\}$ is obviously universally measurable.

For any $t \geqslant 1$ and any $1 \leqslant s \leqslant t+1$, we have

$$T_t^{-1}(s) = \cup_{1 \leqslant j_1 \leqslant \cdots \leqslant j_{s-1} \leqslant t-s+2} \mathsf{V}_{t,j_1,\ldots,j_{s-1}}$$

which is a universally measurable set according to the results proved above. Thus, T_t is universally measurable.

For any $t \geqslant 1$, any $1 \leqslant s \leqslant t$, any $\mathbf{y}^* = (y_1^*, \ldots, y_s^*) \in \mathcal{Y}^s$, define $\mathcal{S}_{\mathbf{y}^*} = \{S \subseteq \mathcal{Y}^s : \mathbf{y}^* \in S\}$. Then, we have

$$\widehat{\mathbf{Y}}_{t-1}^{-1}(\mathcal{S}_{\mathbf{y}^*}) = \bigcup_{1 \leqslant j_1 \leqslant \cdots \leqslant j_{s-1} \leqslant t-s+1} \; \bigcup_{(C_1,\ldots,C_{s-1}) \in \prod_{k=1}^{s-1} \mathsf{PC}(\mathcal{Y}^k)} \Big[\big(\mathsf{V}_{t-1,j_1,\ldots,j_{s-1},C_1,\ldots,C_{s-1}} \times \mathcal{X}^s\big) \cap \mathsf{D}^{\vee}_{t-1,j_1,\ldots,j_{s-1},C_1,\ldots,C_{s-1},\mathbf{y}^*}\Big]$$

which is a universally measurable set according to the results proved above. Then, by Lemma 56, $\widehat{\mathbf{Y}}_{t-1}$ is universally measurable. ■

C.4. Uniform rate implies universal rate

Now, we apply the pattern avoidance functions defined in the previous section into a template for building learning algorithms in the probabilistic setting. Any learning algorithm with some guaranteed uniform rate for finite DS dimensional hypothesis classes can be plugged into this template to construct a learning algorithm that achieves the same universal rate for classes without an infinite DSL tree.

For any $k \geqslant 1$, any $n \geqslant k$, any function $g : \mathcal{X}^k \to 2^{\mathcal{Y}^k}$, and any sequence $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$, define the concept set

$$\mathcal{H}(S, g) := \{h \in \mathcal{H}|_S : (h(i_1), \ldots, h(i_k)) \notin g(x_{i_1}, \ldots, x_{i_k}) \text{ for all distinct } 1 \leqslant i_1, \ldots, i_k \leqslant n\}.$$

For any $t \geqslant 0$, $n \geqslant \tau_t$, and any sequence $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$, define the concept set

$$\mathcal{H}(S, \widehat{\mathbf{y}}_t) := \{h \in \mathcal{H}|_S : (h(i_1), \ldots, h(i_{\tau_t})) \notin \widehat{\mathbf{y}}_t(x_{i_1}, \ldots, x_{i_{\tau_t}}) \text{ for all distinct } 1 \leqslant i_1, \ldots, i_{\tau_t} \leqslant n\}. \tag{11}$$

We have the following lemma.

Lemma 58 *For any $t \geqslant 0$ and any sequence $(x_1, y_1, \ldots, x_t, y_t) \in (\mathcal{X} \times \mathcal{Y})^t$ (where we say $(x_1, y_1, \ldots, x_t, y_t) = \varnothing$ if $t = 0$) that is consistent with $\mathcal{H}$, any $n \geqslant \tau_t$, and any $S := (x'_1, \ldots, x'_n) \in \mathcal{X}^n$, we have $\dim(\mathcal{H}(S, \widehat{\mathbf{y}}_t)) < \tau_t$, where $\dim(\mathcal{H}(S, \widehat{\mathbf{y}}_t))$ denotes the DS dimension of $\mathcal{H}(S, \widehat{\mathbf{y}}_t)$.*

Proof Define $k := \tau_t$. If $\dim(\mathcal{H}(S, \widehat{\mathbf{y}}_t)) \geqslant k$, then there exists a sequence $(i_1, \ldots, i_k)$ and pseudo-cube $C \in \mathsf{PC}(\mathcal{H}(S, \widehat{\mathbf{y}}_t)|_{(i_1, \ldots, i_k)})$. Define $\bar{\mathbf{x}}_k = (x_{i_1}, \ldots, x_{i_k})$. Then, by the definition of $\widehat{\mathbf{y}}_t$, for any $\mathbf{y}' \in C$, we have that

$$\begin{aligned} &\mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{k-1}, \bar{C}_{k-1}, \bar{\mathbf{y}}_{k-1}, \bar{\mathbf{x}}_k, C, \mathbf{y}') \\ &\geqslant \min\{\mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{k-1}, \bar{C}_{k-1}, \bar{\mathbf{y}}_{k-1}), \mathrm{val}(\varnothing)\}. \end{aligned} \tag{12}$$

Since $\mathcal{H}$ does not have an infinite DSL tree; i.e., P_A does not have a winning strategy, we have that $\mathrm{val}(\varnothing) < \Omega$ and further by Bousquet et al. (2021, Lemma B.7), $\mathrm{val}(\varnothing) < \omega_1$. Here, we claim that $\mathrm{val}(\varnothing) \geqslant 0$. Consider the sequence $\mathbf{w} = (\mathbf{x}_1, C_1, \mathbf{y}_1, \ldots) \in \mathsf{P}_\infty$ constructed as follows. First, fix a hypothesis $h \in \mathcal{H}$. For each $s \geqslant 1$, pick arbitrary $(\mathbf{x}_s, C_s) \in \mathsf{XPC}_s$, set $\mathbf{y}_s = h(\mathbf{x}_s)$. Then, it is obvious that $\mathbf{w} \notin \mathsf{W}_{\mathfrak{B}}$. Thus, we have $\mathrm{val}(\varnothing) \geqslant 0$.

Suppose that for some $j \in \{0, 1, \ldots, k-2\}$, we have that $\bar{\mathbf{y}}_s \in \bar{C}_s$ and

$$0 \leqslant \mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_s, \bar{C}_s, \bar{\mathbf{y}}_s) < \mathrm{val}(\varnothing)$$

for all $s \in \{1, \ldots, j\}$. Then, by the definition of $\widehat{\mathbf{y}}_t$, we have

$$\mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}) < \mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_j, \bar{C}_j, \bar{\mathbf{y}}_j) < \mathrm{val}(\varnothing).$$

If $\bar{\mathbf{y}}_{j+1} \notin \bar{C}_{j+1}$, then, for any $\mathbf{w}' \in \prod_{s=j+2}^{\infty} (\mathsf{XPC}_s \times \mathcal{Y}^s)$, we claim that

$$\mathbf{w} := (\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}, \mathbf{w}') \notin \mathsf{W}_{\mathfrak{B}}.$$

This is because for any $\tau \in [j]$, we must have $\mathcal{H}|_{\bar{\mathbf{x}}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_\tau, \bar{\mathbf{y}}_\tau} \neq \varnothing$ since $\mathrm{val}(\bar{\mathbf{x}}_1, \bar{\mathbf{y}}_1, \bar{C}_1, \ldots, \bar{\mathbf{x}}_\tau, \bar{C}_\tau, \bar{\mathbf{y}}_\tau) \geqslant 0$ and $\bar{\mathbf{y}}_s \in C_s$ for any $s \in [j]$ by the induction hypothesis. Then, if $\bar{\mathbf{y}}_{j+1} \notin \bar{C}_{j+1}$, we must have $\mathbf{w} \notin \mathsf{W}_{\mathfrak{B}}$.

Since $(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}, \mathbf{w}') \notin \mathsf{W}_{\mathfrak{B}}$ for any $\mathbf{w}' \in \prod_{s=j+2}^{\infty} (\mathsf{XPC}_s \times \mathcal{Y}^s)$, there is an infinite active tree starting from $\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}$, which implies that

$$\mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}) = \Omega.$$

However, this cannot happen because we have shown that $\mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}) < \mathrm{val}(\varnothing) < \omega_1$. Thus, it must hold that $\bar{\mathbf{y}}_{j+1} \in \bar{C}_{j+1}$ by contradiction.

Then, we claim that

$$\mathrm{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}) \geqslant 0.$$

If on the contrary $\text{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}) = -1$, we will have $\mathcal{H}|_{\bar{\mathbf{x}}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{\mathbf{y}}_{j+1}} = \varnothing$ because we have shown that $\bar{\mathbf{y}}_s \in \bar{C}_s$ for any $s \in [j+1]$ and $\mathcal{H}|_{\bar{\mathbf{x}}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_j, \bar{\mathbf{x}}_j} \neq \varnothing$. However, since $(x_1, y_1, \ldots, x_t, y_t)$ is a consistent sequence with $\mathcal{H}$, we must have $\mathcal{H}|_{\bar{\mathbf{x}}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{\mathbf{y}}_{j+1}} \neq \varnothing$. Thus, there is a contradiction and we must have $\text{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{j+1}, \bar{C}_{j+1}, \bar{\mathbf{y}}_{j+1}) \geqslant 0$.

Now, by induction, we can conclude that that $\bar{\mathbf{y}}_s \in \bar{C}_s$ and

$$0 \leqslant \text{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_s, \bar{C}_s, \bar{\mathbf{y}}_s) < \text{val}(\varnothing) < \omega_1$$

for all $s \in \{1, \ldots, k-1\}$. For any $\mathbf{y}'' \in \mathcal{Y}^k \backslash C$, we must have

$$\text{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{k-1}, \bar{C}_{k-1}, \bar{\mathbf{y}}_{k-1}, \bar{\mathbf{x}}_k, C, \mathbf{y}'') = \Omega$$

according to the same arguments we used for the proof that $\bar{\mathbf{y}}_{j+1} \in \bar{C}_{j+1}$ in the induction step. Then, by (12), we have that

$$\begin{aligned} &\text{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{k-1}, \bar{C}_{k-1}, \bar{\mathbf{y}}_{k-1}, \bar{\mathbf{x}}_k, C, \mathbf{y}) \\ \geqslant &\min\{\text{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{k-1}, \bar{C}_{k-1}, \bar{\mathbf{y}}_{k-1}), \text{val}(\varnothing)\} \end{aligned}$$

for any $\mathbf{y} \in \mathcal{Y}^k$. However, this contradicts Bousquet et al. (2021, Propostion B.8) since

$$0 \leqslant \text{val}(\bar{\mathbf{x}}_1, \bar{C}_1, \bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{x}}_{k-1}, \bar{C}_{k-1}, \bar{\mathbf{y}}_{k-1}) < \omega_1.$$

Thus, we have $\dim(\mathcal{H}(S, \hat{\mathbf{y}}_t)) < k$. ■

Let us fix a $\mathcal{H}$-realizable distribution P on $\mathcal{X} \times \mathcal{Y}$. Let $(\Omega_P, \mathcal{F}_P, \mathbf{P})$ denote the underlying probability space. Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be i.i.d. random variables on $(\Omega_P, \mathcal{F}_P, \mathbf{P})$ with $(X_1, Y_1) \sim P$. We have the following result regarding the consistency of the random sequence $((X_t, Y_t))_{t \geqslant 1}$ with $\mathcal{H}$.

Lemma 59 *If P is $\mathcal{H}$-realizable and $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d. random variables with distribution P, then, with probability one, for any $t \geqslant 1$, there exists some $h \in \mathcal{H}$ such that $h(X_s) = Y_s$ for all $s \in [t]$.*

The proof of Lemma 59 can be found in the proof of Bousquet et al. (2021, Lemma 4.3.).

For any $k \in \mathbb{N}$ and function $g : \mathcal{X}^k \to 2^{\mathcal{Y}^k}$ where 2^S denotes the power set of the set S, we define

$$\text{per}(g) = \mathbf{P}[(Y_1, \ldots, Y_k) \in g(X_1, \ldots, X_k)].$$

Now, let

$$\begin{aligned} &\tau_t := T_t(X_1, Y_1, \ldots, X_t, Y_t), \\ &\hat{\mathbf{y}}_t(x_1, \ldots, x_{\tau_t}) := \hat{\mathbf{Y}}_t(X_1, Y_1, \ldots, X_t, Y_t, x_1, \ldots, x_{\tau_t}). \end{aligned}$$

We first prove the following result when $\text{per}(g) = 0$.

Lemma 60 *For any $k, n \in \mathbb{N}$ with $n \geqslant k$, any function $g : \mathcal{X}^k \to 2^{\mathcal{Y}^k}$, and any sequence $S = ((X_i, Y_i))_{i=1}^n \sim P^n$, if $\text{per}(g) = 0$, then $((i, Y_i))_{i=1}^n$ is consistent with $\mathcal{H}(S|_{\mathcal{X}}, g)$ and $\mathcal{D}$ is $\mathcal{H}(S|_{\mathcal{X}}, g)$-realizable a.s., where $S|_{\mathcal{X}} := (X_1, X_2 \ldots, X_n)$ and $\mathcal{D}$ denotes the uniform distribution over $\{(i, Y_i)\}_{i=1}^n$, i.e., $\mathcal{D}(\{(i, Y_i)\}) = \frac{1}{n}$ for any $i \in [n]$.*

Proof Since $S \sim P^n$, according to Lemma 59, there exists a random variable $H : \Omega \to \mathcal{H}$ such that for **P**-a.e. $\omega \in \Omega$, $h = H(\omega) \in \mathcal{H}$ satisfies that $h(X_i(\omega)) = Y_i(\omega)$ for any $i \in [n]$. Since per$(g) = 0$, we have that for **P**-a.e. $\omega \in \Omega$, $(Y_{i_1}(\omega), \ldots, Y_{i_k}(\omega)) \notin g(X_{i_1}(\omega), \ldots, X_{i_k}(\omega))$ for all distinct $1 \leqslant t_1, \ldots, t_k \leqslant n$. Thus, for **P**-a.e. $\omega \in \Omega$, $h = H(\omega)$ satisfies that $(h(X_{i_1}(\omega)), \ldots, h(X_{i_k}(\omega))) \notin g(X_{i_1}(\omega), \ldots, X_{i_k}(\omega))$ for all distinct $1 \leqslant t_1, \ldots, t_k \leqslant n$.

Define the random variable $\bar{H} : \Omega \to \mathcal{Y}^{[n]}$ by $\bar{H}(\omega)(i) := H(\omega)(X_i(\omega)) = h(X_i(\omega))$ where $h = H(\omega)$. Then, by the definition of $\mathcal{H}(S|_{\mathcal{X}}, g)$ and $\bar{h}$, we have that for **P**-a.e. $\omega \in \Omega$, $\bar{h} = \bar{H}(\omega) \in \mathcal{H}(S(\omega)|_{\mathcal{X}}, g)$ and $\bar{h}(i) = Y_i(\omega)$ for any $i \in [n]$. Thus,

$$\mathbf{P}[\bar{h}(I) \neq Y_I|S] = \mathbf{P}[h(X_I) \neq Y_I|S] = \frac{1}{n}\mathbb{1}\{h(X_i) \neq Y_i\} = 0$$

where I is a random variable uniformly distributed over $[n]$ and is independent of S. Then, we know that (I, Y_I) follows distribution $\mathcal{D}$ conditional on S. Therefore,

$$\text{er}(\bar{h}) = \mathbf{P}[\bar{h}(I) \neq Y_I|S] = 0, \mathbf{P}\text{-a.s.}$$

Thus, $\inf_{h' \in \mathcal{H}(S|_{\mathcal{X}}, g)} \text{er}(h') = 0$ a.s., which implies that $\mathcal{D}$ is $\mathcal{H}(S|_{\mathcal{X}}, g)$-realizable a.s. ■

Similar to Bousquet et al. (2021, Lemma 5.7), we have the following lemma.

Lemma 61 $\mathbf{P}[\text{per}(\hat{\mathbf{y}}_t) > 0] \to 0$ *as* $t \to \infty$.

Proof According to Lemma 59, we have that $((X_i, Y_i))_{i \in \mathbb{N}}$ is consistent with $\mathcal{H}$ a.s.. Then, by Proposition 51, we have that

$$T := \sup\left\{s \geqslant 1 : (Y_{s-\tau_{s-1}+1}, \ldots, Y_s) \in \hat{\mathbf{y}}_{s-1}(X_{s-\tau_{s-1}+1}, \ldots, X_s)\right\}$$

is finite a.s.. Since $((X_i, Y_i))_{i \in \mathbb{N}}$ is an i.i.d. sequence of random variables, we have that for any $t \geqslant 1$, $\hat{\mathbf{y}}_{t-1}$ is independent of $((X_i, Y_i))_{i \geqslant t}$. Then, by the strong laws of large number, we have that with probability one,

$$\begin{aligned}
&\lim_{m\to\infty} \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}\left\{(Y_{t+(k-1)\tau_{t-1}}, \ldots, Y_{t+k\tau_{t-1}-1}) \in \hat{\mathbf{y}}_{t-1}(X_{t+(k-1)\tau_{t-1}}, \ldots, X_{t+k\tau_{t-1}-1})\right\} \\
&= \mathbf{E}\left[\mathbb{1}\left\{(Y_t, \ldots, Y_{t+\tau_{t-1}-1}) \in \hat{\mathbf{y}}_{t-1}(X_t, \ldots, X_{t+\tau_{t-1}-1})\right\}\right] \\
&= \text{per}(\hat{\mathbf{y}}_{t-1}).
\end{aligned}$$

Since $T < \infty$ implies that $\tau_{s-1} = \tau_{t-1} < \infty$ and $\hat{\mathbf{y}}_{s-1} = \hat{\mathbf{y}}_{t-1}$ for any $T < s, t < \infty$, it follows that for any $t \in \mathbb{N}$ such that $T < t < \infty$,

$$(Y_{t+(k-1)\tau_{t-1}}, \ldots, Y_{t+k\tau_{t-1}-1}) \notin \hat{\mathbf{y}}_{t-1}(X_{t+(k-1)\tau_{t-1}}, \ldots, X_{t+k\tau_{t-1}-1}), \ \forall k \in \mathbb{N}$$

and thus,

$$\lim_{m\to\infty} \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}\left\{(Y_{t+(k-1)\tau_{t-1}}, \ldots, Y_{t+k\tau_{t-1}-1}) \in \hat{\mathbf{y}}_{t-1}(X_{t+(k-1)\tau_{t-1}}, \ldots, X_{t+k\tau_{t-1}-1})\right\} = 0.$$

Therefore, we can conclude that for any $t \in \mathbb{N}$,

$$\{T < t\} \subseteq \Big\{ \lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{(Y_{t+(k-1)\tau_{t-1}}, \ldots, Y_{t+k\tau_{t-1}-1}) \in \hat{\mathbf{y}}_{t-1}(X_{t+(k-1)\tau_{t-1}}, \ldots, X_{t+k\tau_{t-1}-1})\} = 0 \Big\}.$$

Given the above results, we have

$$\begin{aligned} &\mathbf{P}[\mathrm{per}(\hat{\mathbf{y}}_t) = 0] \\ =&\mathbf{P}\left[\lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{(Y_{t+(k-1)\tau_{t-1}}, \ldots, Y_{t+k\tau_{t-1}-1}) \in \hat{\mathbf{y}}_{t-1}(X_{t+(k-1)\tau_{t-1}}, \ldots, X_{t+k\tau_{t-1}-1})\} = 0 \right] \\ \geqslant&\mathbf{P}\left[T < t\right]. \end{aligned}$$

and

$$\mathbf{P}[\mathrm{per}(\hat{\mathbf{y}}_t) > 0] = 1 - \mathbf{P}[\mathrm{per}(\hat{\mathbf{y}}_t) = 0] \leqslant [T \geqslant t].$$

Since we have proved $T < \infty$ a.s., we have

$$\limsup_{t\to\infty} \mathbf{P}[\mathrm{per}(\hat{\mathbf{y}}_t) > 0] \leqslant \lim_{t\to\infty} \mathbf{P}[T \geqslant t] = 0$$

Therefore, $\lim_{t\to\infty} \mathbf{P}[\mathrm{per}(\hat{\mathbf{y}}_t) > 0] = 0$. ■

Analogous to Bousquet et al. (2021, Lemma 5.10), we have

Lemma 62 *Given any $t^* \in \mathbb{N}$ such that $\mathbf{P}[\mathrm{per}(\hat{\mathbf{y}}_{t^*}) > 0] \leqslant \frac{1}{8}$, if $n \geqslant \max\{4(t^*+1), N\}$ for some $N \geqslant 1$ dependent on the adversarial algorithm defined in Section C.2, P, and t^*, then there exists a universally measurable function $\hat{t}_n = \hat{t}_n(X_1, Y_1, \ldots, X_{\lfloor n/2 \rfloor}, Y_{\lfloor n/2 \rfloor}) \in [\lfloor n/4 \rfloor - 1]$ whose definition does not depend on the data distribution P and some constants C and c independent of n (but dependent on the algorithm, P, and t^*) such that*

$$\mathbf{P}[\hat{t}_n \in \mathcal{T}_{\mathrm{good}}] \geqslant 1 - Ce^{-cn},$$

where $\mathcal{T}_{\mathrm{good}} := \big\{1 \leqslant t \leqslant t^* : \mathbf{P}[\mathrm{per}(\hat{\mathbf{y}}_t) > 0] \leqslant \frac{3}{8}\big\}$.

Proof For each $1 \leqslant t \leqslant \lfloor n/4 \rfloor - 1$ and $1 \leqslant i \leqslant \lfloor n/(4t) \rfloor$, define

$$\tau_t^i := T_t(X_{(i-1)t+1}, Y_{(i-1)t+1}, \ldots, X_{it}, Y_{it}) \leqslant t + 1 \leqslant \lfloor n/4 \rfloor,$$

$$\hat{\mathbf{y}}_t^i(x_1, \ldots, x_{\tau_t^i}) := \hat{\mathbf{Y}}_t(X_{(i-1)t+1}, Y_{(i-1)t+1}, \ldots, X_{it}, Y_{it}, x_1, \ldots, x_{\tau_t^i})$$

for $\forall (x_1, \ldots, x_{\tau_t^i}) \in \mathcal{X}^{\tau_t^i}$, and

$$\hat{e}_t := \frac{1}{\lfloor n/(4t) \rfloor} \sum_{i=1}^{\lfloor n/(4t) \rfloor} \mathbb{1}\{(Y_{s+1,\ldots,Y_{s+\tau_t^i}}) \in \hat{\mathbf{y}}_t^i(X_{s+1,\ldots,X_{s+\tau_t^i}}) \text{ for some } \lfloor n/4 \rfloor \leqslant s \leqslant \lfloor n/2 \rfloor - \tau_t^i\}.$$

Since $\mathrm{per}(\widehat{\mathbf{y}}_t^i) = 0$ implies that $(Y_{s+1,\ldots,Y_{s+\tau_t^i}}) \notin \widehat{\mathbf{y}}_t^i(X_{s+1,\ldots,X_{s+\tau_t^i}})$ for all $\lfloor n/4 \rfloor \leqslant s \leqslant \lfloor n/2 \rfloor - \tau_t^i$ a.s., we have that

$$\widehat{e}_t \leqslant e_t := \frac{1}{\lfloor n/(4t) \rfloor} \sum_{i=1}^{\lfloor n/(4t) \rfloor} \mathbb{1}\big\{\mathrm{per}(\widehat{\mathbf{y}}_t^i) > 0\big\} \text{ a.s.}$$

Define

$$\widehat{t}_n := \min\big\{ \inf\big\{1 \leqslant t \leqslant \lfloor n/4 \rfloor - 1 : \widehat{e}_t < 1/4\big\},\ \lfloor n/4 \rfloor - 1\big\}$$

with the convention that $\inf \varnothing = +\infty$. Since $t^* \leqslant \lfloor n/4 \rfloor - 1$, we can conclude that $\widehat{e}_{t^*} < 1/4$ implies $\widehat{t}_n \leqslant t^*$. Then, by Hoeffding's inequality, we have

$$\mathbf{P}[\widehat{t}_n > t^*] \leqslant \mathbf{P}[\widehat{e}_{t^*} \geqslant 1/4] \leqslant \mathbf{P}[e_{t^*} - \mathbf{E}[e_{t^*}] \geqslant 1/4] \leqslant e^{-\lfloor n/(4t^*) \rfloor/32}.$$

For any $t \in \mathbb{N}$ such that $t \leqslant t^*$ and $\mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_t) > 0] > 3/8$, since

$$\lim_{z \to 0} \mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_t) > z] = \mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_t) > 0] > 3/8$$

by the continuity of probability measures, there exists some $\varepsilon_t > 0$ such that $\mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_t) > \varepsilon_t] > 1/4 + 1/16$. Because $t^* < \infty$, there exists $\varepsilon > 0$ such that $\mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_t) > \varepsilon] > 1/4 + 1/16$ for all $1 \leqslant t \leqslant t^*$ such that $\mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_t) > 0] > 3/8$.

Fixing an arbitrary $1 \leqslant t \leqslant t^*$ such that $\mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_t) > 0] > 3/8$, by Hoeffding's inequality, we have

$$\mathbf{P}\Big[\frac{1}{\lfloor n/(4t) \rfloor} \sum_{i=1}^{\lfloor n/(4t) \rfloor} \mathbb{1}\left\{\mathrm{per}(\widehat{\mathbf{y}}_t) > \varepsilon\right\} < \frac{1}{4}\Big] \leqslant e^{-\lfloor n/(4t^*) \rfloor/128}.$$

For any $2 \leqslant \tau \leqslant \lfloor n/4 \rfloor$ and any $g : \mathcal{X}^\tau \to \{0,1\}^\tau$ with $\mathrm{per}(g) > \varepsilon$, we have that

$$\begin{aligned}
&\mathbf{P}\big[(Y_{s+1}, \ldots, Y_{s+\tau}) \notin g(X_{s+1}, \ldots, X_{s+\tau}) \text{ for all } \lfloor n/4 \rfloor \leqslant s \leqslant \lfloor n/2 \rfloor - \tau\big] \\
\leqslant&\mathbf{P}\big[(Y_{\lfloor n/4 \rfloor + (k-1)\tau + 1}, \ldots, Y_{\lfloor n/4 \rfloor + k\tau}) \notin g(X_{\lfloor n/4 \rfloor + (k-1)\tau + 1}, \ldots, X_{\lfloor n/4 \rfloor + k\tau}) \\
&\qquad \text{for all } 1 \leqslant k \leqslant \lfloor n/(4\tau) \rfloor\big] \\
=&(1 - \mathrm{per}(g))^{\lfloor n/(4\tau) \rfloor} \\
\leqslant&(1 - \varepsilon)^{\lfloor n/(4\tau) \rfloor}.
\end{aligned}$$

Thus, by union bound and the fact that $\tau_t^i \leqslant t + 1 \leqslant t^* + 1$, we have

$$\begin{aligned}
&\mathbf{P}\Big[\mathbb{1}\big\{\mathrm{per}(\widehat{\mathbf{y}}_t^i) > \varepsilon\big\} > \mathbb{1}\big\{(Y_{s+1,\ldots,Y_{s+\tau_t^i}}) \in \widehat{\mathbf{y}}_t^i(X_{s+1,\ldots,X_{s+\tau_t^i}}) \text{ for some } \lfloor n/4 \rfloor \leqslant s \leqslant \lfloor n/2 \rfloor - \tau_t^i\big\} \\
&\qquad \text{for some } 1 \leqslant i \leqslant \lfloor n/(4t) \rfloor\Big] \\
\leqslant&\mathbf{P}\Big[\mathbb{1}\big\{(Y_{s+1,\ldots,Y_{s+\tau_t^i}}) \in \widehat{\mathbf{y}}_t^i(X_{s+1,\ldots,X_{s+\tau_t^i}}) \text{ for some } \lfloor n/4 \rfloor \leqslant s \leqslant \lfloor n/2 \rfloor - \tau_t^i\big\} = 0 \\
&\qquad \text{for some } 1 \leqslant i \leqslant \lfloor n/(4t) \rfloor\Big] \\
=&\mathbf{P}\Big[\exists 1 \leqslant i \leqslant \lfloor n/(4t) \rfloor \text{ s.t. } (Y_{s+1,\ldots,Y_{s+\tau_t^i}}) \notin \widehat{\mathbf{y}}_t^i(X_{s+1,\ldots,X_{s+\tau_t^i}}) \text{ for all } \lfloor n/4 \rfloor \leqslant s \leqslant \lfloor n/2 \rfloor - \tau_t^i\Big] \\
\leqslant&\lfloor n/(4t) \rfloor (1 - \varepsilon)^{\lfloor n/(4(t^*+1)) \rfloor}.
\end{aligned}$$

Then, we can conclude that

$$
\begin{aligned}
&\mathbf{P}[\widehat{t}_n = t] \\
\leqslant&\mathbf{P}[\widehat{e}_t < 1/4] \\
\leqslant&\mathbf{P}\Big[\frac{1}{\lfloor n/(4t)\rfloor}\sum_{i=1}^{\lfloor n/(4t)\rfloor} \mathbb{1}\left\{\mathrm{per}(\widehat{\mathbf{y}}_t) > \varepsilon\right\} < \frac{1}{4}\Big] \\
&+\mathbf{P}\Big[\mathbb{1}\big\{\mathrm{per}(\widehat{\mathbf{y}}_t^i) > \varepsilon\big\} > \mathbb{1}\big\{(Y_{s+1,\dots,Y_{s+\tau_t^i}}) \in \widehat{\mathbf{y}}_t^i(X_{s+1,\dots,X_{s+\tau_t^i}}) \text{ for some } \lfloor n/4\rfloor \leqslant s \leqslant \lfloor n/2\rfloor - \tau_t^i\big\} \\
&\qquad\qquad \text{for some } 1 \leqslant i \leqslant \lfloor n/(4t)\rfloor\Big] \\
\leqslant& e^{-\lfloor n/(4t^*)\rfloor/128} + \lfloor n/(4t)\rfloor(1-\varepsilon)^{\lfloor n/(4(t^*+1))\rfloor}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{P}[\widehat{t}_n \notin \mathcal{T}_{\text{good}}] \leqslant&\mathbf{P}[\widehat{t}_n > t^*] + \mathbf{P}[\widehat{t}_n \leqslant t^* \text{ and } \mathrm{per}(\widehat{\mathbf{y}}_{\widehat{t}_n}) > 3/8] \\
=&\mathbf{P}[\widehat{t}_n > t^*] + \mathbf{P}[\widehat{t}_n = t \text{ for some } t \text{ s.t. } 1 \leqslant t \leqslant t^* \text{ and } \mathrm{per}(\widehat{\mathbf{y}}_t) > 3/8] \\
\leqslant& e^{-\lfloor n/(4t^*)\rfloor/32} + t^* e^{-\lfloor n/(4t^*)\rfloor/128} + t^*\lfloor n/4\rfloor(1-\varepsilon)^{\lfloor n/(4(t^*+1))\rfloor}
\end{aligned}
$$

Note that $e^{-\lfloor n/(4t^*)\rfloor/32} \leqslant e^{\frac{1}{32}}e^{-\frac{n}{128t^*}}$, $e^{-\lfloor n/(4t^*)\rfloor/32} \leqslant t^* e^{\frac{1}{128}}e^{-\frac{n}{512t^*}}$, and

$$
-\log\left(t^*\lfloor n/4\rfloor(1-\varepsilon)^{\lfloor n/(4(t^*+1))\rfloor}\right) \geqslant \frac{\log\left(\frac{1}{1-\varepsilon}\right)}{8t^*+2}n - \log\left(\frac{t^*}{1-\varepsilon}\right) + \frac{\log\left(\frac{1}{1-\varepsilon}\right)}{8t^*+2}n - \log(n/4),
$$

Since $\log(n/4) \leqslant \sqrt{n}$ for all $n \geqslant 4$, we have that $\frac{\log\left(\frac{1}{1-\varepsilon}\right)}{8t^*+2}n \geqslant \log(n/4)$ and

$$
t^*\lfloor n/4\rfloor(1-\varepsilon)^{\lfloor n/(4(t^*+1))\rfloor} \leqslant \frac{t^*}{1-\varepsilon}\exp\left(-\frac{\log\left(\frac{1}{1-\varepsilon}\right)}{8t^*+2}n\right)
$$

for all $n \geqslant \max\{4, \left(\frac{8t^*+2}{\log(1-\varepsilon)}\right)^2\}$.

Thus, for any $n \geqslant \max\left\{4(t^*+1), \left(\frac{8t^*+2}{\log(1-\varepsilon)}\right)^2\right\}$, we have $\mathbf{P}[\widehat{t}_n \notin \mathcal{T}_{\text{good}}] \leqslant Ce^{-cn}$ for $c := \min\left\{\frac{1}{512t^*}, \frac{\log\left(\frac{1}{1-\varepsilon}\right)}{8t^*+2}\right\}$ and $C := e^{\frac{1}{32}} + t^* e^{\frac{1}{128}} + \frac{t^*}{1-\varepsilon}$. Since ε depends on t^*, the data distribution P, and the algorithm, but does not depend on n, the lemma is proved. ■

According to Brukhim et al. (2022, Theorem 1) and its proof in Brukhim et al. (2022, Section 4.5), we have the following theorem.

Theorem 63 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be an hypothesis class with DS dimension $d < \infty$. There is a learning algorithm $\mathcal{A} : \cup_{n=1}^{\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}$ with the following guarantee. For any $\mathcal{H}$-realizable distribution $\mathcal{D}$, any $\delta \in (0,1)$, any integer $n \geqslant 1$, any sample $(S,(X,Y)) \sim \mathcal{D}^{n+1}$ where $S \in (\mathcal{X}\times\mathcal{Y})^n$ and $(X,Y) \in \mathcal{X}\times\mathcal{Y}$, the output hypothesis $\mathcal{A}(\mathcal{H},S)$ satisfies that*

$$
\mathbf{P}[\mathcal{A}(\mathcal{H},S)(X) \neq Y|S] \leqslant O\left(\frac{d^{3/2}\log^2(n) + \log(1/\delta)}{n}\right).
$$

with probability at least $1-\delta$.

We immediately have the following corollary from Theorem 63.

Corollary 64 *For the hypothesis class $\mathcal{H}$, learning algorithm $\mathcal{A}$, distribution $\mathcal{D}$, and any integer $n \geqslant 1$ described in Theorem 63, we have*

$$\mathbf{P}[\mathcal{A}(\mathcal{H}, S)(X) \neq Y] \leqslant \frac{Cd^{3/2}\log^2(n)}{n}.$$

for some constant $C > 0$.

Proof Define $R = \mathbf{P}[A(S)(X) \neq Y|S]$. Then, by Theorem 63, there exists some constant $C_1 > 0$ such that

$$\mathbf{P}\left[R \geqslant \frac{C_1 d^{3/2}\log^2(n) + C_1\log(1/\delta)}{n}\right] \leqslant \delta.$$

Define $t = \frac{C_1 d^{3/2}\log^2(n)+C_1\log(1/\delta)}{n} \in (\frac{C_1 d^{3/2}\log^2(n)}{n}, \infty)$. Then, we have $\delta = \exp(d^{3/2}\log^2(n) - nt/C_1)$. It follows that

$$\begin{aligned}\mathbf{P}[\mathcal{A}(S)(X) \neq Y] =& \mathbf{E}[R]\\ =& \mathbf{E}\left[R\mathbb{1}\Big\{R \leqslant \frac{C_1 d^{3/2}\log^2(n)}{n}\Big\}\right] + \mathbf{E}\left[R\mathbb{1}\Big\{R > \frac{C_1 d^{3/2}\log^2(n)}{n}\Big\}\right]\\ \leqslant& \frac{C_1 d^{3/2}\log^2(n)}{n} + \int_{\frac{C_1 d^{3/2}\log^2(n)}{n}}^{\infty} \mathbf{P}[R \geqslant t]dt\\ \leqslant& \frac{C_1 d^{3/2}\log^2(n)}{n} + \int_{\frac{C_1 d^{3/2}\log^2(n)}{n}}^{\infty} \exp(d^{3/2}\log^2(n) - nt/C_1)dt\\ =& \frac{C_1 d^{3/2}\log^2(n) + C_1}{n}.\end{aligned}$$

Thus, there exists some constant $C > 0$ such that $\mathbf{P}[\mathcal{A}(S)(X) \neq Y] \leqslant \frac{Cd^{3/2}\log^2(n)}{n}$. ■

The following lemma is very important in upper bounding the error probability for learning algorithms with access to leave-one-out samples using their guarantees on all samples.

Lemma 65 *Suppose that A is an algorithm that for any positive integer n, any feature space $\mathcal{Z}$, and any label space $\mathcal{W}$, given a hypothesis class $\mathcal{H} \subseteq \mathcal{W}^{\mathcal{Z}}$ and a sequence of samples $((z_i, w_i))_{i=1}^n$ consistent with $\mathcal{H}$, outputs a hypothesis $h \in \mathcal{W}^{\mathcal{Z}}$.*

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the feature space and label space of the samples. Suppose that $H : \cup_{n=1}^{\infty}\mathcal{X}^n \to \cup_{n=1}^{\infty}2^{\mathcal{Y}^{[n]}}$ is a function that for any positive integer n, given a sequence $(x_1, \dots, x_n) \in \mathcal{X}^n$, constructs a hypothesis class $H((x_1, \dots, x_n)) \subseteq \mathcal{Y}^{[n]}$ such that $((1, x_1), \dots, (n, x_n))$ is consistent with $H((x_1, \dots, x_n))$.

For any positive integer n and any sequence $S' = ((x_1, y_1), \dots, (x_{n+1}, y_{n+1})) \in (\mathcal{X} \times \mathcal{Y})^{n+1}$, define $S'|_{\mathcal{X}} := (x_1, x_2, \dots, x_{n+1})$. Let $\mathcal{D}'$ denote the uniform distribution over $\{(i, y_i)\}_{i=1}^{n+1}$ and $\mathcal{D}$ denote the uniform distribution over $\{(i, y_i)\}_{i=1}^{n}$ (i.e., $\mathcal{D}'(\{(i, y_i)\}) = \frac{1}{n+1}$ for any $i \in [n+1]$ and $\mathcal{D}(\{(i, y_i)\}) = \frac{1}{n}$ for any $i \in [n]$).

Then, for any $T \sim \mathcal{D}^{\lceil n/2 \rceil}$ *and* $(T', (I, y_I)) \sim (\mathcal{D}')^{\lceil n/2 \rceil + 1}$ *with* $T' \in \{(i, y_i) : 1 \leqslant i \leqslant n+1\}^{\lceil n/2 \rceil}$, $I \in [n+1]$, *and* $\lceil x \rceil := \min\{n \in \mathbb{Z} : n \geqslant x\}$ *for any* $x \in \mathbb{R}$, *we have*

$$\mathbf{P}[A(H(S'|_{\mathcal{X}}), T)(n+1) \neq y_{n+1}] \leqslant 2\mathbf{P}[A(H(S'|_{\mathcal{X}}), T')(I) \neq y_I].$$

Proof Since $(T', (I, y_I)) \sim (\mathcal{D}')^{\lceil n/2 \rceil + 1}$, we have that

$$\mathbf{P}[(n+1, y_{n+1}) \in T'] \leqslant |T'| \frac{1}{n+1} = \frac{\lceil n/2 \rceil}{n+1} \leqslant \frac{1}{2}.$$

By the assumption on H, we know that T' is consistent with $H(S'|_{\mathcal{X}})$. Thus, by the assumption on A, we have that $A(H(S'|_{\mathcal{X}}), T')(i) = y_i$ for any $(i, y_i) \in T'$. It follows that

$$\begin{aligned}
&\mathbf{P}[A(H(S'|_{\mathcal{X}}), T')(I) \neq y_I] \\
=&\mathbf{E}[\mathbf{P}[A(H(S'|_{\mathcal{X}}), T')(I) \neq y_I | T']] \\
=&\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{P}[A(H(S'|_{\mathcal{X}}), T')(i) \neq y_i] \\
\geqslant&\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{E}[\mathbb{1}\{(i, y_i) \notin T'\} \mathbb{1}\{A(H(S'|_{\mathcal{X}}), T')(i) \neq y_i\}] \\
=&\mathbf{E}[\mathbb{1}\{(n+1, y_{n+1}) \notin T'\} \mathbb{1}\{A(H(S'|_{\mathcal{X}}), T')(n+1) \neq y_{n+1}\}] \\
=&\mathbf{P}[(n+1, y_{n+1}) \notin T'] \mathbf{E}[\mathbb{1}\{A(H(S'|_{\mathcal{X}}), T')(x_{n+1}) \neq y_{n+1}\} | (n+1, y_{n+1}) \notin T'] \\
\geqslant&\frac{1}{2} \mathbf{P}[A(H(S'|_{\mathcal{X}}), T')(n+1) \neq y_{n+1} | (n+1, y_{n+1}) \notin T'] \\
=&\frac{1}{2} \mathbf{P}[A(H(S'|_{\mathcal{X}}), T)(n+1) \neq y_{n+1}],
\end{aligned}$$

where the last inequality follows from the fact that $\mathbf{P}[T' \in B | x_{n+1} \notin T'] = \mathbf{P}[T \in B]$ for any $B \subseteq \{((i_1, y_{i_1}), \ldots, (i_{\lceil n/2 \rceil}, y_{i_{\lceil n/2 \rceil}})) : 1 \leqslant i_1, \ldots, i_{\lceil n/2 \rceil} \leqslant n+1\}$. ■

Now, we are ready to prove the main theorem that relates guarantees of learning algorithms on uniform rates to universal rates.

Theorem 66 *Suppose that* A *is a learning algorithm which for any hypothesis class* H *with DS dimension at most* $d < \infty$, *any* H*-realizable distribution* $\mathcal{D}$, *any number* $n \in \mathbb{N}$, *and any sample* $S \sim \mathcal{D}^n$, *outputs a hypothesis* $A(H, S)$ *with* $\mathbf{E}[\mathrm{er}(A(H, S))] \leqslant r(n, d)$, *where* $r : \mathbb{N} \times \mathbb{N} \to [0, 1]$ *is some rate function non-increasing for any* $d \in \mathbb{N}$. *Then, there is an algorithm* A' *satisfying that for any hypothesis class* $\mathcal{H}$ *that does not have an infinite DSL tree and any* $\mathcal{H}$*-realizable distribution* P, *there exist some constants* $C, c > 0$ *and* $d_0 \in \mathbb{N}$ *such that for all* $n \in \mathbb{N}$ *and* $S \sim P^n$, A' *outputs a hypothesis* $A'(\mathcal{H}, S) \in \mathcal{H}$ *with*

$$\mathbf{E}[\mathrm{er}(A'(\mathcal{H}, S))] \leqslant Ce^{-cn} + 32r(\lceil n/4 \rceil, d_0).$$

Proof According to Lemma 61, there exists $t^* \in \mathbb{N}$ such that $\mathbf{P}[\mathrm{per}(\widehat{\mathbf{y}}_{t^*}) > 0] \leqslant \frac{1}{8}$. Then, for any $n \geqslant \max\{4(t^*+1), N\}$ with N specified in Lemma 62, let $\widehat{t}_n \in [\lfloor n/4 \rfloor - 1]$ be the random time constructed in Lemma 62. For any $t \in [\lfloor n/4 \rfloor - 1]$ and any $i \in [n/(4\widehat{t}_n)]$, define

$$\tau_t^i := T_t(X_{(i-1)t+1}, Y_{(i-1)t+1}, \ldots, X_{it}, Y_{it}) \leqslant t+1 \leqslant \lfloor n/4 \rfloor,$$

and

$$\widehat{\mathbf{y}}_t^i : \mathcal{X}^{\tau_t^i} \to \mathcal{Y}^{\tau_t^i},\ (x_1,\dots,x_{\tau_t^i}) \mapsto \widehat{\mathbf{Y}}_t(X_{(i-1)t+1}, Y_{(i-1)t+1},\dots,X_{it},Y_{it},x_1,\dots,x_{\tau_t^i})$$

as in the proof of Lemma 62.

For any $t \in \mathcal{T}_{\text{good}}$, since $\mathbf{P}[\text{per}(\widehat{\mathbf{y}}_t) > 0] \leqslant \frac{3}{8}$, by a Chernoff bound, we have

$$\mathbf{P}\left[\frac{1}{\lfloor n/(4t)\rfloor}\sum_{i=1}^{\lfloor n/(4t)\rfloor}\mathbb{1}\{\text{per}(\widehat{\mathbf{y}}_t^i) > 0\} > \frac{7}{16}\right] \leqslant e^{-\lfloor n/(4t)\rfloor/128} \leqslant e^{-\lfloor n/(4t^*)\rfloor/128}.$$

Using union bound, we have

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{\lfloor n/(4\widehat{t}_n)\rfloor}\sum_{i=1}^{\lfloor n/(4\widehat{t}_n)\rfloor}\mathbb{1}\{\text{per}(\widehat{\mathbf{y}}_{\widehat{t}_n}^i) > 0\} > \frac{7}{16}, \widehat{t}_n \in \mathcal{T}_{\text{good}}\right]\\
\leqslant &\sum_{t\in\mathcal{T}_{\text{good}}}\mathbf{P}\left[\frac{1}{\lfloor n/(4t)\rfloor}\sum_{i=1}^{\lfloor n/(4t)\rfloor}\mathbb{1}\{\text{per}(\widehat{\mathbf{y}}_t^i) > 0\} > \frac{7}{16}\right]\\
\leqslant & t^* e^{-\lfloor n/(4t^*)\rfloor/128}.
\end{aligned} \tag{13}$$

Define the sequence $S := ((1, Y_{\lfloor n/2\rfloor+1}), (2, Y_{\lfloor n/2\rfloor+2}), \dots, (n - \lfloor n/2\rfloor, Y_n))$. Let $\mathcal{D}$ denote the uniform distribution over the elements in S (i.e., $\mathcal{D}(\{(i, Y_{\lfloor n/2\rfloor+i})\}) = \frac{1}{n-\lfloor n/2\rfloor}$ for any $i \in [n - \lfloor n/2\rfloor]$). Let $T^1,\dots,T^{\lfloor n/(4\widehat{t}_n)\rfloor}$ denote an i.i.d. sequence of random variables with $T^1 \sim \mathcal{D}^{\lceil (n-\lfloor n/2\rfloor)/2\rceil}$. For any $i \in [\lfloor n/(4\widehat{t}_n)\rfloor]$ and any $x \in \mathcal{X}$, define the hypothesis class $\mathcal{H}^i(x) := \mathcal{H}((X_{\lfloor n/2\rfloor+1},\dots,X_n,x), \widehat{\mathbf{y}}_{\widehat{t}_n}^i)$. Then, for any $i \in [\lfloor n/(4\widehat{t}_n)\rfloor]$, we can define the following prediction function

$$\widehat{y}^i : \mathcal{X} \to \mathcal{Y},\ x \mapsto A(\mathcal{H}^i(x), T^i)(n - \lfloor n/2\rfloor + 1).$$

Let $\widehat{h}_n$ be the majority vote of $\widehat{y}^i$ for $i \in [\lfloor n/(4\widehat{t}_n)\rfloor]$. $\widehat{h}_n$ will be the final output of our learning algorithm. Now, we need to upper bound the error rate $\mathbf{E}[\text{er}(\widehat{h}_n)]$.

Recall that P denotes the underlying data distribution that is $\mathcal{H}$-realizable. Suppose that $(X,Y) \sim P$ and is independent of $\{(X_i,Y_i)\}_{i=1}^n$. Then, we have

$$\begin{aligned}
&\mathbf{E}[\text{er}(\widehat{h}_n)]\\
=&\mathbf{P}[\widehat{h}_n(X) \neq Y]\\
\leqslant&\mathbf{P}\left[\frac{1}{\lfloor n/(4\widehat{t}_n)\rfloor}\sum_{i=1}^{\lfloor n/(4\widehat{t}_n)\rfloor}\mathbb{1}\{\widehat{y}^i(X) \neq Y\} \geqslant \frac{1}{2}\right]\\
\leqslant&\mathbf{P}[\widehat{t}_n \notin \mathcal{T}_{\text{good}}] + \mathbf{P}\left[\frac{1}{\lfloor n/(4\widehat{t}_n)\rfloor}\sum_{i=1}^{\lfloor n/(4\widehat{t}_n)\rfloor}\mathbb{1}\{\text{per}(\widehat{\mathbf{y}}_{\widehat{t}_n}^i) > 0\} > \frac{7}{16}, \widehat{t}_n \in \mathcal{T}_{\text{good}}\right]\\
&+\mathbf{P}\left[\widehat{t}_n \in \mathcal{T}_{\text{good}},\ \frac{1}{\lfloor n/(4\widehat{t}_n)\rfloor}\sum_{i=1}^{\lfloor n/(4\widehat{t}_n)\rfloor}\mathbb{1}\{\text{per}(\widehat{\mathbf{y}}_{\widehat{t}_n}^i) = 0\} \geqslant \frac{9}{16},\ \frac{1}{\lfloor n/(4\widehat{t}_n)\rfloor}\sum_{i=1}^{\lfloor n/(4\widehat{t}_n)\rfloor}\mathbb{1}\{\widehat{y}^i(X) \neq Y\} \geqslant \frac{1}{2}\right].
\end{aligned} \tag{14}$$

Define the sequence $S' := ((1, Y_{\lfloor n/2 \rfloor+1}), \ldots, (n - \lfloor n/2 \rfloor, Y_n), (n - \lfloor n/2 \rfloor + 1, Y))$ and conditional on S', let $\mathcal{D}'$ denote the uniform distribution over the elements in S' (i.e., $\mathcal{D}'(\{(i, Y_{\lfloor n/2 \rfloor+i})\}) = \frac{1}{n-\lfloor n/2 \rfloor+1}$ for any $i \in [n - \lfloor n/2 \rfloor]$ and $\mathcal{D}'(\{(n - \lfloor n/2 \rfloor + 1, Y)\}) = \frac{1}{n-\lfloor n/2 \rfloor+1}$). Let $T' \sim (\mathcal{D}')^{\lceil (n-\lfloor n/2 \rfloor)/2 \rceil}$ and $(I, Y') \sim \mathcal{D}'$ be two independent samples from S' conditional on S'.

For any $i \in [\lfloor n/(4\widehat{t}_n) \rfloor]$, by Lemma 59, $(X_{(i-1)\widehat{t}_n+1}, Y_{(i-1)\widehat{t}_n+1}, \ldots, X_{i\widehat{t}_n}, Y_{i\widehat{t}_n})$ is consistent with $\mathcal{H}$ a.s. Then, by Lemma 58, we have that with probability 1, $\dim(\mathcal{H}^i(X)) < \tau^i_{\widehat{t}_n}$ and therefore,

$$\mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \dim(\mathcal{H}^i(X)) < t^*.$$

Moreover, if $\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0$, by Lemma 60, we have that S' is consistent with $\mathcal{H}^i(X)$ and $\mathcal{D}'$ is $\mathcal{H}^i(X)$-realizable a.s. Then, it follows from Lemma 65 and the property of A that

$$\begin{aligned}
&\mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbf{P}[\widehat{y}^i(X) \neq Y | ((X_j, Y_j))_{j=1}^n, X, Y] \\
=& \mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbf{P}[A(\mathcal{H}^i(X), T^i)(n - \lfloor n/2 \rfloor + 1) \neq Y | ((X_j, Y_j))_{j=1}^n, X, Y] \\
\leqslant& 2 \mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbf{P}[A(\mathcal{H}^i(X), T')(I) \neq Y' | ((X_j, Y_j))_{j=1}^n, X, Y] \\
\leqslant& 2r(\lceil (n - \lfloor n/2 \rfloor)/2 \rceil, t^*).
\end{aligned}$$

By the properties of conditional expectation, we have that

$$\begin{aligned}
&\mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbf{P}\left[\widehat{y}^i(X) \neq Y \middle| ((X_j, Y_j))_{j=1}^{\lfloor n/2 \rfloor}\right] \\
=& \mathbf{E}\left[\mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbf{P}[\widehat{y}^i(X) \neq Y | ((X_j, Y_j))_{j=1}^n, X, Y] \middle| ((X_j, Y_j))_{j=1}^{\lfloor n/2 \rfloor}\right] \\
\leqslant& 2r(\lceil (n - \lfloor n/2 \rfloor)/2 \rceil, t^*).
\end{aligned}$$

Since $\frac{9}{16} + \frac{1}{2} = 1 + \frac{1}{16}$, by Markov's inequality and the above inequality, we have

$$\begin{aligned}
&\mathbf{P}\left[\widehat{t}_n \in \mathcal{T}_{\text{good}}, \frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \geqslant \frac{9}{16}, \frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbb{1}\{\widehat{y}^i(X) \neq Y\} \geqslant \frac{1}{2}\right] \\
\leqslant& \mathbf{P}\left[\mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbb{1}\{\widehat{y}^i(X) \neq Y\} \geqslant \frac{1}{16}\right] \\
\leqslant& 16 \mathbf{E}\left[\mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbb{1}\{\widehat{y}^i(X) \neq Y\}\right] \\
=& 16 \mathbf{E}\left[\frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbb{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\} \mathbb{1}\{\text{per}(\widehat{\mathbf{y}}^i_{\widehat{t}_n}) = 0\} \mathbf{P}\left[\widehat{y}^i(X) \neq Y \middle| ((X_j, Y_j))_{j=1}^{\lfloor n/2 \rfloor}\right]\right] \\
\leqslant& 32 \mathbf{E}\left[\frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} r(\lceil (n - \lfloor n/2 \rfloor)/2 \rceil, t^*)\right] \\
\leqslant& 32 r(\lceil (n - \lfloor n/2 \rfloor)/2 \rceil, t^*) \\
\leqslant& 32 r(\lceil n/4 \rceil, t^*).
\end{aligned} \tag{15}$$

By (14), (13), Lemma 62, and (15), we have

$$\mathbf{E}[\mathrm{er}(\widehat{h})] \leqslant C_1 e^{-c_1 n} + t^* e^{-\lfloor n/(4t^*) \rfloor/128} + 32r(\lceil n/4 \rceil, t^*).$$

■

Then, we immediately have the following result.

Theorem 67 *If $\mathcal{H}$ does not have an infinite DSL tree, then $\mathcal{H}$ is learnable at rate $\frac{\log^2(n)}{n}$.*

Proof According to Corollary 64 and Theorem 66, we know that there exists an algorithm A satisfies that for any $\mathcal{H}$-realizable distribution P, there exists some constants $C_0, C_1, c_0 > 0$ and $d_0 \in \mathbb{N}$ such that for all $n \in \mathbb{N}$ large enough and $S \sim P^n$, the output hypothesis $A(\mathcal{H}, S)$ of A has the error rate

$$\mathbf{E}[A(\mathcal{H}, S)] \leqslant C_0 e^{-c_0 n} + C_1 \frac{d_0^{3/2} \log^2(n)}{n}.$$

Thus, there exists some constants $C > 0$ such that

$$\mathbf{E}[A(\mathcal{H}, S)] \leqslant C \frac{\log^2(n)}{n},$$

which implies that $\mathcal{H}$ is learnable at rate $\frac{\log^2(n)}{n}$. ■

C.5. Concluding proof

We conclude with the proof of Theorem 14.
Proof of Theorem 14 Theorem 14 follows directly from Theorem 44 and Theorem 67. ■

Appendix D. Arbitrary Slow Rates

In this section, we provide the complete proof of Theorem 15. First, we show two lemmas regarding the properties of pseudo-cubes.

Lemma 68 *For any positive integer d, any label class $\mathcal{Y}$, any pseudo-cube $H \subseteq \mathcal{Y}^d$ of dimension d, any $j \in [d]$, and any label $y \in \mathcal{Y}$, define $N := |H|$ and $H_y^j := \{h \in H : h(j) = y\}$. Then, we have*

$$|H_y^j| \leqslant \frac{1}{2} N.$$

Proof We prove by contradiction. Suppose on the contrary that there exist some $j \in [d]$ and $y \in \mathcal{Y}$ such that $|H_y^j| > \frac{1}{2} n$. The definition of pseudo-cube implies that $|H| \geqslant 2$. Then, there exist $h, h' \in H_y^j$ with $h \neq h'$. Let f and f' denote the j-neighbors of h and h' in H; i.e., there exists $f, f' \in H$ such that $f(j) \neq h(j) = y$, $f'(j) \neq h'(j) = y$, $f(i) = h(i)$, and $f'(i) = h'(i)$ for any $i \in [d] \backslash \{j\}$. Since $h \neq h'$ and $h(j) = y = h'(j)$, there exists some $j' \in [d] \backslash \{j\}$ such that $h(j') \neq h'(j')$. It follows that $f(j') = h(j') \neq h'(j') = f'(j')$ and $f' \neq f$. Then, we have

$$|\{h \in H : h(j) \neq y\}| \geqslant |H_y^j| > \frac{1}{2}n$$

and

$$n = |\{h \in H : h(j) \neq y\}| + |H_y^j| > n,$$

which is a contradiction. Thus, we must have $|H_y^j| \leqslant \frac{1}{2}n$. ■

Lemma 69 *For any integer $d \geqslant 2$, $n \in [d-1]$, and $1 \leqslant j_1 < \cdots < j_n \leqslant d$, any label class $\mathcal{Y}$, any pseudo-cube $H \subseteq \mathcal{Y}^d$ of dimension d, and any hypothesis $g \in H$, define $\mathsf{J} := (j_1, \ldots, j_n)$ and $\mathsf{K} = (k_1, \ldots, k_{d-n})$ such that $1 \leqslant k_1 < \cdots < k_{n-d} \leqslant d$ and $\{j_1, \ldots, j_n, k_1, \ldots, k_{n-d}\} = [d]$. Then, $H_{g,\mathsf{J}} := \{h|_{\mathsf{K}} : h \in H, h(j_i) = g(j_i), \forall i \in [n]\}$ is a pseudo-cube of dimension $n - d$.*

Proof For any $f \in H_{g,J}$, there exists some $f' \in H$ such that $f = f'|_{\mathsf{K}}$. Then, for any $i \in [n-d]$, there exists some $h' \in H$ such that $h'(k_i) \neq f'(k_i)$ and $h'(l) = f'(l)$ for all $l \in [d] \backslash \{k_i\}$. Since $k_i \notin \{j_1, \ldots, j_n\}$, we have $h'|_{\mathsf{J}} = f'|_{\mathsf{J}} = g|_{\mathsf{J}}$ and $h := h'|_{\mathsf{K}} \in H_{g,\mathsf{J}}$. Then, we have $h(i) = h'(k_i) \neq f'(k_i) = f(i)$ and $h(m) = h'(k_m) = f'(k_m) = f(m)$ for any $m \in [n-d] \backslash \{i\}$, which implies that $H_{g,\mathsf{J}}$ is a pseudo-cube. ■

Now, we present the proof of Theorem 15

Proof of Theorem 15 Suppose that $\mathcal{H}$ has an infinite DSL tree. Fix an arbitrary rate R with $\lim_{n\to\infty} R(n) = 0$ and an arbitrary learning algorithm A. According to Bousquet et al. (2021, Lemma 5.12), there exist a sequence of non-negative numbers $(p_i)_{i\in\mathbb{N}}$ for which $\sum_{i=1}^{\infty} p_i = 1$, two strictly increasing sequences of positive integers $(n_i)_{i\in\mathbb{N}}$ and $(k_i)_{i\in\mathbb{N}}$, and a constant $\frac{1}{2} \leqslant C \leqslant 1$ such that for all $i \in \mathbb{N}$, we have $\sum_{k>k_i} p_k \leqslant \frac{1}{n_i}$, $n_i p_{k_i} \leqslant k_i$, and $p_{k_i} = CR(n_i)$.

For the infinite DSL tree, let $v_\varnothing \in \mathcal{X}$ denote the root node and $c_\varnothing \in \mathbb{N}$ denote the number of children of $v_\varnothing$. For any $i \in [c_\varnothing]$, let v_i denote the i-th child of $v_\varnothing$ and c_i denote the number of the children of v_i. Suppose that for some $k \in \mathbb{N}$, $v_{i_1,\ldots,i_k}$ and $c_{i_1,\ldots,i_k}$ has been defined for any $i_1 \in [c_\varnothing], \ldots, i_k \in [c_{i_1,\ldots,i_{k-1}}]$. For any $i \in [c_{i_1,\ldots,i_k}]$, let $v_{i_1,\ldots,i_k,i}$ denote the i-th child of $v_{i_1,\ldots,i_k}$ and $c_{i_1,\cdots_k,i}$, denote the number of children of $v_{i_1,\ldots,i_k,i}$. Then, by induction, $\{c_{i_1,\ldots,i_k}\}_{i_1\in[c_\varnothing],\ldots,i_k\in[c_{i_1,\ldots,i_{k-1}}]}$ and $\{v_{i_1,\ldots,i_k}\}_{i_1\in[c_\varnothing],\ldots,i_k\in[c_{i_1,\ldots,i_{k-1}}]}$ are defined for all $k \geqslant 0$. Thus, every node in the infinite DSL tree has been denoted and the tree can be denoted with $\mathbf{t} := \{v_{i_1,\ldots,i_k} : i_1 \in [c_\varnothing], \ldots, i_k \in [c_{i_1,\ldots,i_{k-1}}], k \geqslant 0\}$. For any $k \geqslant 0$ and any $i_1 \in [c_\varnothing], \ldots, i_k \in [c_{i_1,\ldots,i_{k-1}}], i_{k+1} \in [c_{i_1,\ldots,i_k}]$, define $\mathbf{x}_{i_1,\ldots,i_k} \in \mathcal{X}^{k+1}$ to be the label of $v_{i_1,\ldots,i_k}$ and $\mathbf{y}_{i_1,\ldots,i_k,i_{k+1}}$ to be the label of the edge connecting $v_{i_1,\ldots,i_k}$ and $v_{i_1,\ldots,i_k,i_{k+1}}$.

Let I_1 be a random variable following the uniform distribution over $[c_\varnothing]$ (i.e., $\mathbf{P}(I_1 = i) = \frac{1}{c_\varnothing}$ for any $i \in [c_\varnothing]$). For any $k \geqslant 1$, suppose that I_j has been defined for all $j \in [k]$. Define I_{k+1} to be a random variable such that conditional on $I_1, \ldots, I_k$, I_{k+1} follows the uniform distribution over $c_{I_1,\ldots,I_k}$ (i.e., $\mathbf{P}(I_{k+1} = i | I_1, \ldots, I_k) = \frac{1}{c_{I_1,\ldots,I_k}}$ for any $i \in [c_{I_1,\ldots,I_k}]$). Define $\mathbf{I} := (I_1, I_2, \ldots)$. Then, the support of $\mathbf{I}$ is

$$\mathcal{I} := \{(i_1, i_2, \ldots, i_k, \ldots) : i_1 \in [c_\varnothing], i_2 \in [c_{i_1}], \ldots, i_k \in [c_{i_1,\ldots,i_{k-1}}], \ldots\}.$$

For any $\mathbf{i} = (i_1, i_2, \dots) \in \mathcal{I}$, define the distribution $P_{\mathbf{i}}$ on $\mathcal{X} \times \mathcal{Y}$ as

$$P_{\mathbf{i}}(\{(x^j_{i_1,\dots,i_{k-1}}, y^j_{i_1,\dots,i_k})\}) = \frac{p_k}{k} \text{ for } j \in [k], k \in \mathbb{N},$$

where $x^j_{i_1,\dots,i_{k-1}}$ and $y^j_{i_1,\dots,i_k}$ denote the j-th element in $\mathbf{x}_{i_1,\dots,i_{k-1}}$ and $\mathbf{y}_{i_1,\dots,i_k}$ respectively. Note that as in the proof of Theorem 44, the mapping $\mathbf{i} \mapsto P_{\mathbf{i}}$, $\mathbf{i} \in \mathcal{I}$ is measurable. By the definition of DSL tree, for any $n \in \mathbb{N}$, there exists $h_n \in \mathcal{H}$ such that $h_n(\mathbf{x}_{i_1,\dots,i_{k-1}}) = \mathbf{y}_{i_1,\dots,i_k}$ for all $k \in [n]$. Hence,

$$\text{er}_{\mathbf{i}}(h_n) := P_{\mathbf{i}}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : h_n(x) \neq y\}) \leqslant \sum_{k>n} p_k,$$

which, together with the fact that $\sum_{k=1}^{\infty} p_k = 1$ and $p_k \geqslant 0$ for all $k \in \mathbb{N}$, implies that $\inf_{h \in \mathcal{X}} \text{er}_{\mathbf{i}}(h) = 0$. Thus, $P_{\mathbf{i}}$ is $\mathcal{H}$-realizable for any $\mathbf{i} \in \mathcal{I}$.

Let $(T, J), (T_1, J_1), (T_2, J_2), \dots$ be a sequence of i.i.d. random variables, independent of $\mathbf{I}$, with distribution

$$\mathbf{P}(T = k, J = j) = \frac{p_k}{k} \quad \text{for } j \in [k],\ k \in \mathbb{N}.$$

Define

$$X = x^J_{I_1,\dots,I_{T-1}},\ Y = y^J_{I_1,\dots,I_T},\ X_i = x^{J_i}_{I_1,\dots,I_{T_i-1}},\ Y_i = y^{J_i}_{I_1,\dots,I_{T_i}} \text{ for } i \in \mathbb{N}.$$

Then, we know that conditional on $\mathbf{I}$, $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ is a sequence of i.i.d. random variables with distribution $\mathbf{P}_{\mathbf{I}}$.

For any $k \in \mathbb{N}$ and any $(i_1, i_2, \dots, i_{k-1}) \in \mathbb{N}^{k-1}$ such that $i_1 \in [c_\varnothing], i_2 \in [c_{i_1}], \dots, i_{k-1} \in [c_{i_1,\dots,i_{k-2}}]$, we know that $C_{i_1,\dots,i_{k-1}} := \{\mathbf{y}_{i_1,\dots,i_{k-1},i} : i \in [c_{i_1,\dots,i_{k-1}}]\} \subseteq \mathcal{Y}^{[k]}$ is a pseudo-cube of dimension k by the definition of DSL trees.

For any $n \in \mathbb{N}$ and $j \in [k]$, define the sequence family

$$\mathcal{J}_{j,k,n} := \{(j_1, \dots, j_m) \in ([k] \backslash \{j\})^m : j_1 < j_2 < \cdots < j_m, m \in [\min\{k-1, n\}]\}.$$

For any $i_k \in [c_{i_1,\dots,i_{k-1}}]$ and $\mathsf{J} = (j_1, \dots, j_m) \in \mathcal{J}_{j,k,n}$ with $m \in [\min\{k-1, n\}]$, by Lemma 69, we know that

$$(C_{i_1,\dots,i_{k-1}})_{\mathbf{y}_{i_1,\dots,i_k},\mathsf{J}}$$

is a pseudo-cube of dimension $k - m$ following the notation given in Lemma 69. Then, by Lemma 68, for any $y \in \mathcal{Y}$, we have

$$|((C_{i_1,\dots,i_{k-1}})_{\mathbf{y}_{i_1,\dots,i_k},\mathsf{J}})^{j'}_y| \leqslant \frac{1}{2}|(C_{i_1,\dots,i_{k-1}})_{\mathbf{y}_{i_1,\dots,i_k},\mathsf{J}}|, \tag{16}$$

where $j' := j - \max\{l \in [m] : j_l < j\}$ with the convention that $\max \varnothing := 0$.

By the definition of $\mathbf{I}$, I_k follows the uniform distribution over $c_{I_1,\dots,I_{k-1}}$ conditional on $I_1, \dots, I_{k-1}$. Note that for any $i, i' \in [c_{i_1,\dots,i_{k-1}}]$ such that $i \neq i'$ and $\mathbf{y}_{i_1,\dots,i_{k-1},i}|_{\mathsf{J}} = \mathbf{y}_{i_1,\dots,i_{k-1},i'}|_{\mathsf{J}}$, we must have

$$\mathbf{y}_{i_1,\dots,i_{k-1},i}|_{\mathsf{J}'} \neq \mathbf{y}_{i_1,\dots,i_{k-1},i'}|_{\mathsf{J}'}, \tag{17}$$

where $\mathsf{J}' := (j_1', \ldots, j_{k-m}')$ with $1 \leqslant j_1' < \cdots < j_{k-m}' \leqslant k$ and $\{j_1, \ldots, j_m, j_1', \ldots, j_{k-m}'\} = [k]$. Therefore, conditional on $I_1, \ldots, I_{k-1}$ and $\mathbf{y}_{I_1,\ldots,I_k}|_{\mathsf{J}}$, $\mathbf{y}_{I_1,\ldots,I_k}|_{\mathsf{J}'}$ distributes uniformly over the set $(C_{I_1,\ldots,I_{k-1}})_{\mathbf{y}_{I_1,\ldots,I_k},\mathsf{J}}$. Then, by (16), we have

$$\mathbf{P}\left(y^j_{I_1,\ldots,I_k} \neq y \Big| I_1, \ldots, I_{k-1}, (\mathsf{J}, \mathbf{y}_{I_1,\ldots,I_k}|_{\mathsf{J}})\right) \geqslant \frac{1}{2}.$$

By Lemma 68, we also have

$$|(C_{i_1,\ldots,i_{k-1}})^j_y| \leqslant \frac{1}{2}|C_{i_1,\ldots,i_{k-1}}|,$$

which implies that

$$\mathbf{P}\left(y^j_{I_1,\ldots,I_k} \neq y \Big| I_1, \ldots, I_{k-1}\right) \geqslant \frac{1}{2}. \tag{18}$$

Now, define $\widehat{h}_n := A(\mathcal{H}, ((X_1, Y_1), \ldots, (X_n, Y_n)))$ and the random sequence $\mathfrak{J} := \mathrm{seq}(\{J_i : T_i = k, i \in [n]\})$, where $\mathrm{seq}(\varnothing) := \varnothing$ and for a finite set of integers $\{a_1, \ldots, a_q\}$ with $q \in \mathbb{N}$, $\mathrm{seq}(\{a_1, \ldots, a_q\}) := (a_{(1)}, \ldots, a_{(q)})$ where $a_{(i)}$ denotes the i-th smallest element among $(a_1, \ldots, a_q)$ for any $i \in [q]$. Then, by (17) and (18), we have

$$\begin{aligned}
&\mathbf{P}\left(\widehat{h}_n(X) \neq Y, T = k\right) \\
\geqslant& \sum_{j=1}^{k} \mathbf{P}\left(\widehat{h}_n(x^j_{I_1,\ldots,I_{k-1}}) \neq y^j_{I_1,\ldots,I_k}, T = k, J = j, T_1, \ldots, T_n \leqslant k, (T_1, J_1), \ldots, (T_n, J_n) \neq (k, j)\right) \\
=& \sum_{j=1}^{k} \mathbf{E}\Big[\mathbb{1}\{T = k, J = j, T_1, \ldots, T_n \leqslant k, (T_1, J_1), \ldots, (T_n, J_n) \neq (k, j)\} \\
&\qquad \cdot \mathbf{P}\Big(\widehat{h}_n(x^j_{I_1,\ldots,I_{k-1}}) \neq y^j_{I_1,\ldots,I_k} \Big| I_1, \ldots, I_{k-1}, T_1, \ldots, T_n, J_1, \ldots, J_n, (\mathfrak{J}, \mathbf{y}_{I_1,\ldots,I_k}|_{\mathfrak{J}})\Big)\Big] \\
\geqslant& \sum_{j=1}^{k} \mathbf{E}\Big[\frac{1}{2}\mathbb{1}\{T = k, J = j, T_1, \ldots, T_n \leqslant k, (T_1, J_1), \ldots, (T_n, J_n) \neq (k, j)\}\Big] \\
=& \frac{1}{2}\sum_{j=1}^{k} \mathbf{P}\left(T = k, J = j, T_1, \ldots, T_n \leqslant k, (T_1, J_1), \ldots, (T_n, J_n) \neq (k, j)\right) \\
=& \frac{p_k}{2}\left(1 - \sum_{l>k} p_l - \frac{p_k}{k}\right)^n.
\end{aligned}$$

Then, for any $i \geqslant 3$, by Bousquet et al. (2021, Lemma 5.12), we have

$$\begin{aligned}
\mathbf{P}\left(\widehat{h}_{n_i}(X) \neq Y, T = k_i\right) \geqslant& \frac{p_{k_i}}{2}\left(1 - \sum_{l>k_i} p_l - \frac{p_{k_i}}{k_i}\right)^{n_i} \\
\geqslant& \frac{p_{k_i}}{2}\left(1 - \frac{2}{n_i}\right)^{n_i} \\
\geqslant& \frac{CR(n_i)}{54}.
\end{aligned}$$

Since

$$\frac{1}{R(n_i)}\mathbf{P}\left(\widehat{h}_{n_i}(X) \neq Y, T = k_i|\mathbf{I}\right) \leqslant \frac{1}{R(n_i)}\mathbf{P}\left(T = k_i|\mathbf{I}\right) = \frac{1}{R(n_i)}\mathbf{P}\left(T = k_i\right) = \frac{p_{k_i}}{R(n_i)} = C \text{ a.s.},$$

by Fatou's lemma, we have

$$\mathbf{E}\left[\limsup_{i\to\infty}\frac{1}{R(n_i)}\mathbf{P}\left(\widehat{h}_{n_i}(X) \neq Y, T = k_i|\mathbf{I}\right)\right] \geqslant \limsup_{i\to\infty}\frac{1}{R(n_i)}\mathbf{P}\left(\widehat{h}_{n_i}(X) \neq Y, T = k_i\right) \geqslant \frac{C}{54}.$$

Because

$$\mathbf{E}[\mathrm{er}_{\mathbf{I}}(\widehat{h}_n)|\mathbf{I}] = \mathbf{P}(\widehat{h}_n(X) \neq Y|\mathbf{I}) \geqslant \mathbf{P}(\widehat{h}_n(X) \neq Y, T = k|\mathbf{I}) \text{ a.s.},$$

we have $\mathbf{E}[\limsup_{i\to\infty}\frac{1}{R(n_i)}\mathbf{E}[\mathrm{er}_{\mathbf{I}}(\widehat{h}_{n_i})|\mathbf{I}]] \geqslant \frac{C}{54} > \frac{C}{55}$, which implies that there exists $\mathbf{i} \in \mathcal{I}$ such that $\mathbf{E}[\mathrm{er}_{\mathbf{i}}(\widehat{h}_n)] \geqslant \frac{C}{55}R(n)$ for infinitely many n. By choosing $P = P_{\mathbf{i}}$, we see that $\mathcal{H}$ requires arbitrarily slow rates.

Since $\mathcal{X}$ is Polish and $\mathcal{Y}$ is countable, there exists a learning algorithm with $\mathbf{E}[\mathrm{er}(\widehat{h}_n)] \to 0$ for all realizable distributions P (Hanneke et al., 2021). It follows that $\mathcal{H}$ is learnable but requires arbitrarily slow rates. ■

Appendix E. Proof of Theorem 9

In this section, we provide the complete proof of Theorem 9 below.

Proof According to the proof of Brukhim et al. (2022, Theorem 2), for any $d \in \mathbb{N}$, there exists a d-dimensional pseudo-cube $B_d \subseteq Y_d^{X_d}$ for some spaces X_d and Y_d with $|X_d| = d$ and $|Y_d| < \infty$.

Therefore, for $B_1 \subseteq Y_1^{X_1}$, we can pick $c_1 := |B_1|$ feature spaces $X_{1,1}, \ldots, X_{1,d_1}$ of size 2, label spaces $Y_{1,1}, \ldots, Y_{1,d_1}$, and pseudo-cubes $B_{1,1} \subseteq Y_{1,1}^{X_{1,1}}, \ldots, B_{1,c_1} \subseteq Y_{1,c_1}^{X_{1,c_1}}$ of dimension 2 such that $X_1, X_{1,1}, \ldots, X_{1,c_1}$ are pairwise disjoint and $Y_1, Y_{1,1}, \ldots, Y_{1,c_1}$ are also pairwise disjoint. Define $c_{1,i} := |B_{1,i}|$ for any $i \in [c_1]$. Now, suppose that for some $d \in \mathbb{N}$, $c_{\mathbf{i}_k}$, $X_{\mathbf{i}_k}$, $Y_{\mathbf{i}_k}$ and $B_{\mathbf{i}_k} \subseteq Y_{\mathbf{i}_k}^{X_{\mathbf{i}_k}}$ have been defined such that $|X_{\mathbf{i}_k}| = k$ and $B_{\mathbf{i}_k}$ is a pseudo-cube of dimension k for any $k \in [d]$, $\mathbf{i}_k \in \mathcal{I}_k := \{(i_1, \ldots, i_k) : i_1 \in [1], i_2 \in [c_{i_1}], \ldots, i_k \in [c_{i_1,\ldots,i_{k-1}}]\}$, $\{X_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}_k, k \in [d]\}$ are pairwise disjoint, and $\{Y_{i_1,\ldots,i_k} : \mathbf{i} \in \mathcal{I}_k, k \in [d]\}$ are also pairwise disjoint. Then, for any $\mathbf{i}_d \in \mathcal{I}_d$, pick $c_{\mathbf{i}_d}$ feature spaces $X_{\mathbf{i}_d,1}, \ldots, X_{\mathbf{i}_d,c_{\mathbf{i}_d}}$ of size $d+1$, label spaces $Y_{\mathbf{i}_d,1}, \ldots, Y_{\mathbf{i}_d,c_{\mathbf{i}_d}}$, and pseudo-cubes $B_{\mathbf{i}_d,1} \subseteq Y_{\mathbf{i}_d,1}^{X_{\mathbf{i}_d,1}}, \ldots, B_{\mathbf{i}_d,c_{\mathbf{i}_d}} \subseteq Y_{\mathbf{i}_d,c_{\mathbf{i}_d}}^{X_{\mathbf{i}_d,c_{\mathbf{i}_d}}}$ of dimension $d+1$ such that $\{X_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}_k, k \in [d+1]\}$ are pairwise disjoint and $\{Y_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}_k, k \in [d+1]\}$ are also pairwise disjoint where $\mathcal{I}_{d+1} := \{(\mathbf{i}_d, i) : i \in [c_{\mathbf{i}}], \mathbf{i}_d \in \mathcal{I}_d\}$. Then, we define $c_{i_1,\ldots,i_d,i} := |B_{i_1,\ldots,i_d,i}|$ for any $i \in [c_{i_1,\ldots,i_d}]$.

By induction, for any $k \in \mathbb{N}$ and any $\mathbf{i}_k \in \mathcal{I}_k$, $c_{\mathbf{i}_k}$, $X_{\mathbf{i}_k}$, $Y_{\mathbf{i}_k}$, and $B_{\mathbf{i}_k} \subseteq Y_{\mathbf{i}_k}^{X_{\mathbf{i}_k}}$ have been defined such that $|X_{\mathbf{i}_k}| = k$, $B_{\mathbf{i}_k}$ is a pseudo-cube of dimension k, $\{X_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}_k, k \in \mathbb{N}\}$ are pairwise disjoint, and $\{Y_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}_k, k \in \mathbb{N}\}$ are also pairwise disjoint.

Now, we define $\mathcal{I} := \cup_{k\in\mathbb{N}}\mathcal{I}_k$, $\mathcal{X} := \cup_{\mathbf{i}\in\mathcal{I}}X_{\mathbf{i}}$, and $\mathcal{Y} := \cup_{\mathbf{i}\in\mathcal{I}}Y_{\mathbf{i}} \cup \{\star\}$ where $\star \notin \cup_{\mathbf{i}\in\mathcal{I}}Y_{\mathbf{i}}$ is a new label. Note that $\mathcal{X}$ and $\mathcal{Y}$ are countable. Now, for any $d \in \mathbb{N}$ and $\mathbf{i} = (i_1, \ldots, i_d) \in \mathcal{I}_d$, since $|B_{\mathbf{i}}| = c_{\mathbf{i}} \in \mathbb{N}$, we use $h_{\mathbf{i}}^{(i)}$ to denote the i-th hypothesis in $B_{\mathbf{i}}$ for any $i \in [c_{\mathbf{i}}]$ and extend the domain of $h_{\mathbf{i}}^{(i)}$ to $\mathcal{X}$ by defining $h_{\mathbf{i}}^{(i)}|_{X_{i_1,\ldots,i_k}} := h_{i_1,\ldots,i_k}^{(i_{k+1})}|_{X_{i_1,\ldots,i_k}}$ for any $k \in [d-1]$ and $h_{\mathbf{i}}^{(i)}(x) := \star$

for any $x \in \mathcal{X}\backslash(\cup_{k\in[d]} X_{i_1,\ldots,i_k})$. Letting $H_{\mathbf{i}}$ denote the extended hypotheses in $B_{\mathbf{i}}$, we define the following hypothesis class

$$\mathcal{H} := \cup_{\mathbf{i}\in\mathcal{I}} H_{\mathbf{i}}.$$

By setting $\{X_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}\}$ to be the set of nodes and $\{H_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}\}$ to be the set of edges, we obtain an infinite DSL tree of $\mathcal{H}$. To prove that $\mathcal{H}$ has no NL tree of depth 2, it suffices to show that the Natarajan dimension of $\mathcal{H}$ is 1. For any $k_1, k_2 \in \mathbb{N}$ with $k_1 \leqslant k_2$, $\mathbf{i}_1 \in \mathcal{I}_{k_1}$, $\mathbf{i}_2 \in \mathcal{I}_{k_2}$, and $x_1 \in X_{\mathbf{i}_1}$ and $x_2 \in X_{\mathbf{i}_2}$ with $x_1 \neq x_2$, if $\mathbf{i}_2|_{1:k_1} \neq \mathbf{i}_1$, then, $\star \in \{h(x_1), h(x_2)\}$ for any $h \in \mathcal{H}$, which implies that $\{x_1, x_2\}$ is not N-shattered (see Brukhim et al. 2022, Definition 4 for the definition of "N-shattered") by $\mathcal{H}$. If $k_1 < k_2$ and $\mathbf{i}_2|_{1:k_1} = \mathbf{i}_1$, then, for any $h_1, h_2 \in \mathcal{H}$, in order to have $h_1(x_1) \neq h_2(x_1)$ and $h_1(x_2) \neq h_2(x_2)$, we must have either $\{h_1(x_1), h_2(x_1)\} \times \{h_1(x_2), h_2(x_2)\} = \{\star, y_1\} \times \{\star, y_2\}$ or $\{h_1(x_1), h_2(x_1)\} \times \{h_1(x_2), h_2(x_2)\} = \{y_1', y_1''\} \times \{\star, y_2'\}$ for some $y_1, y_1', y_1'', y_2, y_2' \in \mathcal{H}\backslash\{\star\}$ with $y_1' \neq y_1''$. For $\{h_1(x_1), h_2(x_1)\} \times \{h_1(x_2), h_2(x_2)\} = \{\star, y_1\} \times \{\star, y_2\}$, there is no $h \in \mathcal{H}$ such that $(h(x_1), h(x_2)) = (\star, y_2)$ by our construction. For $\{h_1(x_1), h_2(x_1)\} \times \{h_1(x_2), h_2(x_2)\} = \{y_1', y_1''\} \times \{\star, y_2'\}$, WOLG, we may assume that $(h_1(x_1), h_1(x_2)) = (y_1', y_2')$. Then, there is no $h \in \mathcal{H}$ such that $(h(x_1), h(x_2)) = (y_1'', y_2')$ because any h such that $h(x_2) = h_1(x_2) \neq \star$ must have $h(x_1) = h_1(x_1) = y_1'$. Thus, $\{x_1, x_2\}$ is not N-shattered by $\mathcal{H}$. Finally, if $\mathbf{i}_1 = \mathbf{i}_2 = \mathbf{i}$, for any $h_1 \in \mathcal{H}\backslash\bar{H}_{\mathbf{i}}$ where $\bar{H}_{\mathbf{i}} := \{h \in H_{\mathbf{i}'} : \mathbf{i}' \in \mathcal{I}_k, \mathbf{i}'|_{1:k_1} = \mathbf{i}, k \geqslant k_1\}$, we have $h_1(x_1) = h_1(x_2) = \star$. However, there is no $h \in \mathcal{H}$ such that $(h(x_1), h(x_2)) = (\star, y_2)$ for $y_2 \neq \star$. On ther other hand, for any $h_1, h_2 \in \bar{H}_{\mathbf{i}}$, we have $\bar{H}_{\mathbf{i}}|_{(x_1,x_2)} = B_{\mathbf{i}}|_{(x_1,x_2)}$. Since the Natarajan dimension of $B_{\mathbf{i}}$ is 1, (x_1, x_2) is also not N-shattered by $\bar{H}_{\mathbf{i}'}$. Thus, (x_1, x_2) is also not N-shattered by $\mathcal{H}$. In conclusion, any $(x_1, x_2) \in \mathcal{X}^2$ is not N-shattered by $\mathcal{H}$, the Natarajan dimension of $\mathcal{H}$ is 1, and $\mathcal{H}$ has no NL tree of depth 2.

■

Appendix F. Proof of Theorem 10

In this section, we prove Theorem 10. We first prove the following general lemma.

Lemma 70 *Suppose that $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\mathcal{Y} := [K]$ for some $K \in \mathbb{N}\backslash\{1\}$ has an infinite GL tree $\mathsf{T} = \cup_{n=0}^{\infty}\{(\mathbf{x_u}, \mathbf{s_u}) \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ with its associated hypothesis set $\{h_{\mathbf{u}} \in \mathcal{H} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l, n \in \mathbb{N}\}$. For any $d \in \mathbb{N}_0$, $\boldsymbol{\eta} \in \prod_{l=1}^{d}\{0,1\}^l$, and $\mathbf{w} \in \{0,1\}^{d+1}\backslash\{0\}^{d+1}$, there exist a sequence $\mathbf{y}_{\boldsymbol{\eta},\mathbf{w}} \in \mathcal{Y}^{d+1}$ and an infinite GL tree $\cup_{n=0}^{\infty}\{(\tilde{\mathbf{x}}_{\mathbf{u}}, \tilde{\mathbf{s}}_{\mathbf{u}}) \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ with its associated hypothesis set $\{\tilde{h}_{\mathbf{u}} \in \mathcal{H} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l, n \in \mathbb{N}\}$ such that $(\tilde{x}_{\varnothing}, \tilde{s}_{\varnothing}) = (x_{\varnothing}, s_{\varnothing})$, $(\tilde{\mathbf{x}}_{\mathbf{u}}, \tilde{\mathbf{s}}_{\mathbf{u}}, \tilde{h}_{\mathbf{u}}) = (\mathbf{x_u}, \mathbf{s_u}, h_{\mathbf{u}})$ for all $\mathbf{u} \in \big(\cup_{n\in\mathbb{N}}\prod_{l=1}^{n}\{0,1\}^l\big)\backslash\big(\cup_{n=d+1}^{\infty}(\{\boldsymbol{\eta},\mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l)\big)$, $\{\tilde{h}_{\mathbf{u}} : \mathbf{u} \in \{\boldsymbol{\eta},\mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l, n \geqslant d+1\} \subseteq \{h_{\mathbf{u}} : \mathbf{u} \in \{\boldsymbol{\eta},\mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l, n \geqslant d+1\}$, and for all $0 \leqslant i \leqslant d$ and $\mathbf{u} \in \cup_{n=d+1}^{\infty}\big(\{\boldsymbol{\eta},\mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l\big)$, we have $\tilde{h}_{\mathbf{u}}(\tilde{x}_{\boldsymbol{\eta}}^i) = y_{\boldsymbol{\eta},\mathbf{w}}^i = s_{\boldsymbol{\eta}}^i$ if $w^i = 0$ and $\tilde{h}_{\mathbf{u}}(\tilde{x}_{\boldsymbol{\eta}}^i) = y_{\boldsymbol{\eta},\mathbf{w}}^i \neq s_{\boldsymbol{\eta}}^i$ if $w^i = 1$.*

Proof For any $\mathbf{u} \in \cup_{n=d+1}^{\infty}(\{\boldsymbol{\eta},\mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l)$, we color $v_{\mathbf{u}} := (\mathbf{x_u}, \mathbf{s_u}, h_{\mathbf{u}})$ with

$$(h_{\mathbf{u}}(x_{\boldsymbol{\eta}}^0), \ldots, h_{\mathbf{u}}(x_{\boldsymbol{\eta}}^d)) \in \mathcal{Y}^{d+1}.$$

Since $|\mathcal{Y}| = K < \infty$, by the Milliken's tree theorem (Milliken, 1979), for the colored infinite tree $\mathsf{T}_{\boldsymbol{\eta},\mathbf{w}} := \{v_{\mathbf{u}} : \mathbf{u} \in \{\boldsymbol{\eta},\mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l, n \geqslant d+1\}$, there exists some color $\mathbf{y}_{\boldsymbol{\eta},\mathbf{w}} \in \mathcal{Y}^{d+1}$

and a strongly embedded infinite subtree $\breve{\mathsf{T}}_{\boldsymbol{\eta},\mathbf{w}}$ of $\mathsf{T}_{\boldsymbol{\eta},\mathbf{w}}$ such that all the nodes in $\breve{\mathsf{T}}_{\boldsymbol{\eta},\mathbf{w}}$ have the same color $\mathbf{y}_{\boldsymbol{\eta},\mathbf{w}}$. Since $\breve{\mathsf{T}}_{\boldsymbol{\eta},\mathbf{w}}$ is a strongly embedded subtree of $\mathsf{T}_{\boldsymbol{\eta},\mathbf{w}}$, there exists some sequence $(n_l)_{l\in\mathbb{N}_0} \in \mathbb{N}^{\mathbb{N}_0}$ such that $n_{l+1} > n_l \geqslant d+1$ for any $l \in \mathbb{N}_0$,

$$\breve{\mathsf{T}}_{\boldsymbol{\eta},\mathbf{w}} = \cup_{\tau=0}^{\infty}\Big\{(\breve{\mathbf{x}}_{\mathbf{b}}, \breve{\mathbf{s}}_{\mathbf{b}}, \breve{h}_{\mathbf{b}}) \in \mathcal{X}^{n_\tau+1} \times \mathcal{Y}^{n_\tau+1} \times \mathcal{H} : \mathbf{b} \in \prod_{l=0}^{\tau-1}\{0,1\}^{n_l+1}\Big\},$$

and $(\breve{\mathbf{x}}_{\mathbf{b}}, \breve{\mathbf{s}}_{\mathbf{b}}, \breve{h}_{\mathbf{b}})$ is a node in level n_τ of T for all $\mathbf{b} \in \prod_{l=0}^{\tau-1}\{0,1\}^{n_l+1}$ and $\tau \in \mathbb{N}_0$. For any $t \in \mathbb{N}$ with $t \geqslant d+1$ and $\mathbf{u} = (u_1^0, (u_2^0, u_2^1), \ldots, (u_t^0, \ldots, u_t^{t-1})) \in \{\boldsymbol{\eta}, \mathbf{w}\} \times \prod_{l=d+2}^{t}\{0,1\}^l$, define

$$\begin{aligned}\boldsymbol{\beta}(\mathbf{u}) :=& ((\beta(\mathbf{u})_1^0, \ldots, \beta(\mathbf{u})_1^{n_0}), (\beta(\mathbf{u})_2^0, \ldots, \beta(\mathbf{u})_2^{n_1}), \ldots, (\beta(\mathbf{u})_{t-d-1}^0, \ldots, \beta(\mathbf{u})_{t-d-1}^{n_{t-d-2}}))\\ &\in \prod_{l=0}^{t-d-2}\{0,1\}^{n_l+1}\end{aligned}$$

by

$$\beta(\mathbf{u})_l^i := \begin{cases} u_{l+d+1}^i, & \text{if } 0 \leqslant i \leqslant l+d, \\ 0, & \text{if } l+d+1 \leqslant i \leqslant n_{l-1} \end{cases}$$

for all $0 \leqslant i \leqslant n_{l-1}$ and $1 \leqslant l \leqslant t-d-1$ and

$$\widetilde{v}_{\mathbf{u}} := (\widetilde{\mathbf{x}}_{\mathbf{u}}, \widetilde{\mathbf{s}}_{\mathbf{u}}, \widetilde{h}_{\mathbf{u}})$$

with

$$\widetilde{\mathbf{x}}_{\mathbf{u}} := (\breve{\mathbf{x}}^0_{\boldsymbol{\beta}(\mathbf{u})}, \ldots, \breve{\mathbf{s}}^n_{\boldsymbol{\beta}(\mathbf{u})}) \in \mathcal{X}^{n+1},\ \widetilde{\mathbf{s}}_{\mathbf{u}} := (\breve{\mathbf{x}}^0_{\boldsymbol{\beta}(\mathbf{u})}, \ldots, \breve{\mathbf{s}}^n_{\boldsymbol{\beta}(\mathbf{u})}) \in \mathcal{Y}^{n+1}, \text{and } \widetilde{h}_{\mathbf{u}} := \breve{h}_{\boldsymbol{\beta}(\mathbf{u})} \in \mathcal{H}.$$

Define $\widetilde{x}_{\varnothing} := x_{\varnothing}$, $\widetilde{s}_{\varnothing} := s_{\varnothing}$, and $\widetilde{\mathbf{x}}_{\mathbf{u}} := \mathbf{x}_{\mathbf{u}}, \widetilde{\mathbf{s}}_{\mathbf{u}} := \mathbf{s}_{\mathbf{u}}, \widetilde{h}_{\mathbf{u}} := h_{\mathbf{u}}$ for any

$$\mathbf{u} \in \Big(\cup_{n\in\mathbb{N}} \prod_{l=1}^{n}\{0,1\}^l\Big)\setminus\Big(\cup_{n=d+1}^{\infty} \Big(\{\boldsymbol{\eta}, \mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l\Big)\Big).$$

Then, we obtain the following infinite tree

$$\widetilde{\mathsf{T}} := \cup_{n=0}^{\infty}\Big\{(\widetilde{\mathbf{x}}_{\mathbf{u}}, \widetilde{\mathbf{s}}_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\Big\}.$$

Since T is an infinite GL tree and $\breve{\mathsf{T}}_{\boldsymbol{\eta},\mathbf{w}}$ is a strongly embedded infinite subtree of $\mathsf{T}_{\boldsymbol{\eta},\mathbf{w}}$, we have that $\widetilde{\mathsf{T}}$ is an infinite GL tree with the associated hypothesis set $\{h_{\mathbf{u}} \in \mathcal{H} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l, n \in \mathbb{N}_0\}$. Since all the nodes in $\breve{\mathsf{T}}_{\boldsymbol{\eta},\mathbf{w}}$ have the smae color $\mathbf{y}_{\boldsymbol{\eta},\mathbf{w}}$, by the construction of coloring and $\widetilde{\mathsf{T}}$, for all $0 \leqslant i \leqslant d$ and $\mathbf{u} \in \cup_{n=d+1}^{\infty}\big(\{\boldsymbol{\eta}, \mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l\big)$, we have $\widetilde{h}_{\mathbf{u}}(\widetilde{x}^i_{\boldsymbol{\eta}}) = y^i_{\boldsymbol{\eta},\mathbf{w}} = s^i_{\boldsymbol{\eta}}$ if $w^i = 0$ and $\widetilde{h}_{\mathbf{u}}(\widetilde{x}^i_{\boldsymbol{\eta}}) = y^i_{\boldsymbol{\eta},\mathbf{w}} \neq s^i_{\boldsymbol{\eta}}$ if $w^i = 1$. Finally, we have $(\widetilde{x}_{\varnothing}, \widetilde{s}_{\varnothing}) = (x_{\varnothing}, s_{\varnothing})$ and $(\mathbf{x}_{\mathbf{u}}, \mathbf{s}_{\mathbf{u}}, h_{\mathbf{u}}) = (\widetilde{\mathbf{x}}_{\mathbf{u}}, \widetilde{\mathbf{s}}_{\mathbf{u}}, \widetilde{h}_{\mathbf{u}})$ for all $\mathbf{u} \in \big(\cup_{n\in\mathbb{N}} \prod_{l=1}^{n}\{0,1\}^l\big)\setminus\big(\cup_{n=d+1}^{\infty} (\{\boldsymbol{\eta}, \mathbf{w}\} \times \prod_{l=d+2}^{n}\{0,1\}^l)\big)$ by our definition.

■

Now, we are ready to carry out the proof of Theorem 10.

Proof of Theorem 10 For any NL tree $\cup_{n=0}^{d-1}\{(\mathbf{x_u}, \mathbf{s_u}^{(0)}, \mathbf{s_u}^{(1)}) \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} \times \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ of $\mathcal{H}$ of depth $1 \leqslant d \leqslant \infty$, $\cup_{n=0}^{d-1}\{(\mathbf{x_u}, \mathbf{s_u}^{(0)}) \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ is a GL tree of $\mathcal{H}$ of the same depth d. Thus, an infinite NL tree of $\mathcal{H}$ implies an infinite GL tree of $\mathcal{H}$.

Now, suppose that $\cup_{n=0}^{\infty}\{(\mathbf{x_u}, \mathbf{s_u}) \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ is an infinite GL tree of $\mathcal{H}$. For any $n \in \mathbb{N}$ and $\mathbf{u} = (u_1^0, (u_2^0, u_2^1), \ldots, (u_n^0, \ldots, u_n^{n-1})) \in \prod_{l=1}^{n}\{0,1\}^l$, there exists some $h_{\mathbf{u}} \in \mathcal{H}$ such that $h_{\mathbf{u}}(x^i_{\mathbf{u}_{\leqslant l}}) = s^i_{\mathbf{u}_{\leqslant l}}$ if $u^i_{l+1} = 0$ and $h_{\mathbf{u}}(x^i_{\mathbf{u}_{\leqslant l}}) \neq s^i_{\mathbf{u}_{\leqslant l}}$ otherwise for all $0 \leqslant i \leqslant l$ and $0 \leqslant l < n$, where

$$\mathbf{u}_{\leqslant l} := (u_1^0, (u_2^0, u_2^1), \ldots, (u_l^0, \ldots, u_l^{l-1})),\ x_{\mathbf{u}_{\leqslant l}} := (x^0_{\mathbf{u}_{\leqslant 1}}, \ldots, x^l_{\mathbf{u}_{\leqslant 1}}).$$

Then, we define $\mathsf{T}_G = \{v_\varnothing = (x_\varnothing, s_\varnothing)\} \cup \{v_{\mathbf{u}} = (\mathbf{x_u}, \mathbf{s_u}, h_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l, 1 \leqslant n < \infty\}$ which is the infinite GL tree with the associated hypotheses. Next, we use induction to show that $\mathcal{H}$ has an infinite NL tree.

Applying Lemma 70 to T_G for $d = 0$, $\boldsymbol{\eta} = \varnothing$, and $\mathbf{w} = 1$, we obtain a label $\bar{s}^0_\varnothing \in \mathcal{Y}\backslash\{s^0_\varnothing\}$ and an infinite GL tree with the associated hypotheses $\breve{\mathsf{T}}_G = \{\breve{v}_\varnothing = (\breve{x}_\varnothing, \breve{s}_\varnothing)\} \cup \big(\cup_{n=1}^{\infty}\{\breve{v}_{\mathbf{u}} = (\breve{\mathbf{x}}_{\mathbf{u}}, \breve{\mathbf{s}}_{\mathbf{u}}, \breve{h}_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}\big)$ such that for all $\mathbf{u} \in \cup_{n=1}^{\infty}\big(\{1\} \times \prod_{l=2}^{n}\{0,1\}^l\big)$, we have $\breve{h}_{\mathbf{u}}(\breve{x}_\varnothing) = \bar{s}^0_\varnothing$. Then, we replace T_G with $\breve{\mathsf{T}}_G$. With abuse of notation, we still use T_G to denote $\breve{\mathsf{T}}_G$, use $v_\varnothing = (x_\varnothing, s_\varnothing)$ to denote $\breve{v}_\varnothing = (\breve{x}_\varnothing, \breve{s}_\varnothing)$, and use $v_{\mathbf{u}} = (\mathbf{x_u}, \mathbf{s_u}, h_{\mathbf{u}})$ to denote $\breve{v}_{\mathbf{u}} = (\breve{\mathbf{x}}_{\mathbf{u}}, \breve{\mathbf{s}}_{\mathbf{u}}, \breve{h}_{\mathbf{u}})$ for all $\mathbf{u} \in \cup_{n=1}^{\infty}(\prod_{l=1}^{n}\{0,1\}^l)$.

Suppose that for some $d \in \mathbb{N}$, there exists a set $\cup_{n=0}^{d-1}\{\bar{\mathbf{s}}_{\mathbf{u}} \in \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ and an infinite GL tree with the associated hypotheses $\mathsf{T}_G = \{v_\varnothing = (x_\varnothing, s_\varnothing)\} \cup \big(\cup_{n=1}^{\infty}\{v_{\mathbf{u}} = (\mathbf{x_u}, \mathbf{s_u}, h_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}\big)$ of $\mathcal{H}$ such that $\cup_{n=0}^{d-1}\{(\mathbf{x_u}, \mathbf{s_u}, \bar{\mathbf{s}}_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ is a NL of $\mathcal{H}$ of depth d and for any $\mathbf{u} \in \cup_{n=d}^{\infty}(\prod_{l=1}^{n}\{0,1\}^l)$, we have $h_{\mathbf{u}}(x^i_{\mathbf{u}_{\leqslant l}}) = s^i_{\mathbf{u}_{\leqslant l}}$ if $u^i_{l+1} = 0$ and $h_{\mathbf{u}}(x^i_{\mathbf{u}_{\leqslant l}}) = \bar{s}^i_{\mathbf{u}_{\leqslant l}}$ otherwise for all $0 \leqslant i \leqslant l$ and $0 \leqslant l < d$. Define $r := \lceil 5\log_2 K \rceil \in \mathbb{N}$. For any $\boldsymbol{\kappa} \in \prod_{l=1}^{d}\{0,1\}^l$, consider $\boldsymbol{\eta} \in \{\boldsymbol{\kappa}\} \times \prod_{l=d+1}^{r(d+1)-1}\{0\}^l$. Applying Lemma 70 for $\boldsymbol{\eta}$ and each $\mathbf{w} \in \{0,1\}^{r(d+1)}\backslash\{0\}^{r(d+1)}$ iteratively and defining $\mathbf{y}_{\boldsymbol{\eta},\mathbf{w}} := \mathbf{s}_{\boldsymbol{\eta}}$ for $\mathbf{w} \in \{0\}^{r(d+1)}$, we obtain a class $H_{\boldsymbol{\eta}} = \{\mathbf{y}_{\boldsymbol{\eta},\mathbf{w}} \in \mathcal{Y}^{r(d+1)} : \mathbf{w} \in \{0,1\}^{r(d+1)}\}$ and an infinite GL tree with the associated hypotheses $\widetilde{\mathsf{T}}_{G,\boldsymbol{\eta}} = \{\widetilde{v}_\varnothing = (x_\varnothing, s_\varnothing)\} \cup \big(\cup_{n=1}^{\infty}\{\widetilde{v}_{\mathbf{u}} = (\widetilde{\mathbf{x}}_{\mathbf{u}}, \widetilde{\mathbf{s}}_{\mathbf{u}}, \widetilde{h}_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}\big)$ such that $\widetilde{\mathsf{T}}_{G,\boldsymbol{\eta}}$ satisfies the induction hypothesis for d and $\cup_{n=0}^{d-1}\{\bar{\mathbf{s}}_{\mathbf{u}} \in \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ and for all $\mathbf{w} \in \{0,1\}^{r(d+1)}\backslash\{0\}^{r(d+1)}$, $0 \leqslant i \leqslant r(d+1)-1$, and $\mathbf{u} \in \cup_{n=r(d+1)}^{\infty}\big(\{\boldsymbol{\eta}, \mathbf{w}\} \times \prod_{l=r(d+1)+1}^{n}\{0,1\}^l\big)$, we have $\widetilde{h}_{\mathbf{u}}(\widetilde{x}^i_{\boldsymbol{\eta}}) = y^i_{\boldsymbol{\eta},\mathbf{w}} = s^i_{\boldsymbol{\eta}}$ if $w^i = 0$ and $\widetilde{h}_{\mathbf{u}}(\widetilde{x}^i_{\boldsymbol{\eta}}) = y^i_{\boldsymbol{\eta},\mathbf{w}} \neq s^i_{\boldsymbol{\eta}}$ if $w^i = 1$. Then, we replace T_G with $\widetilde{\mathsf{T}}_{G,\boldsymbol{\eta}}$. With abuse of notation, we still use T_G to denote $\widetilde{\mathsf{T}}_{G,\boldsymbol{\eta}}$, use $v_\varnothing = (x_\varnothing, s_\varnothing)$ to denote $\widetilde{v}_\varnothing = (\widetilde{x}_\varnothing, \widetilde{s}_\varnothing)$, and use $v_{\mathbf{u}} = (\mathbf{x_u}, \mathbf{s_u}, h_{\mathbf{u}})$ to denote $\widetilde{v}_{\mathbf{u}} = (\widetilde{\mathbf{x}}_{\mathbf{u}}, \widetilde{\mathbf{s}}_{\mathbf{u}}, \widetilde{h}_{\mathbf{u}})$ for all $\mathbf{u} \in \cup_{n=1}^{\infty}(\prod_{l=1}^{n}\{0,1\}^l)$.

Since we have shown that $y^i_{\boldsymbol{\eta},\mathbf{w}} \neq s^i_{\boldsymbol{\eta}}$ if $w^i = 1$ and $y^i_{\boldsymbol{\eta},\mathbf{w}} = s^i_{\boldsymbol{\eta}}$ if $w^i = 0$ for any $\mathbf{w} \in \{0,1\}^{r(d+1)}$ and $0 \leqslant i \leqslant r(d+1) - 1$, we have $\dim_G(H_{\boldsymbol{\eta}}) = r(d+1)$ and by Bendavid et al. (1995), $\dim_N(H_{\boldsymbol{\eta}}) > \frac{\dim_G(H_{\boldsymbol{\eta}})}{5\log_2 K} \geqslant d+1$. Thus, there exists a subset $\{i_0, \ldots, i_d\} \subseteq [r(d+1)]$ and two sequences $\mathbf{f}_0 = (f_0^0, \ldots, f_0^d)$, $\mathbf{f}_1 = (f_1^0, \ldots, f_1^d) \in \mathcal{Y}^{d+1}$ such that $f_0^i \neq f_1^i$ for all $0 \leqslant i \leqslant d$ and $H_{\boldsymbol{\eta}}|_{(i_0,\ldots,i_d)} \supseteq \{f_0^0, f_1^0\} \times \cdots \times \{f_0^d, f_1^d\}$. Then, we must have either $\mathbf{f}_0 = \mathbf{s}_{\boldsymbol{\eta}}|_{(i_0,\ldots,i_d)}$ or $\mathbf{f}_1 = \mathbf{s}_{\boldsymbol{\eta}}|_{(i_0,\ldots,i_d)}$. If $\mathbf{f}_0 = \mathbf{s}_{\boldsymbol{\eta}}|_{(i_0,\ldots,i_d)}$, we define $\bar{\mathbf{s}}_{\boldsymbol{\kappa}} := \mathbf{f}_1$ and $\mathbf{s}_{\boldsymbol{\kappa}} := \mathbf{f}_0$. If $\mathbf{f}_1 = \mathbf{s}_{\boldsymbol{\eta}}|_{(i_0,\ldots,i_d)}$, we define $\bar{\mathbf{s}}_{\boldsymbol{\kappa}} := \mathbf{f}_0$ and $\mathbf{s}_{\boldsymbol{\kappa}} := \mathbf{f}_1$. Then, we have $\bar{s}^i_{\boldsymbol{\kappa}} \neq s^i_{\boldsymbol{\kappa}}$ for all $0 \leqslant i \leqslant d$ and $H_{\boldsymbol{\eta}}|_{(i_0,\ldots,i_d)} \supseteq \{s^0_{\boldsymbol{\kappa}}, \bar{s}^0_{\boldsymbol{\kappa}}\} \times \cdots \times \{s^d_{\boldsymbol{\kappa}}, \bar{s}^d_{\boldsymbol{\kappa}}\}$.

Define the set $W_{\boldsymbol{\eta}} := \{\mathbf{w} \in \{0,1\}^{r(d+1)} : y^{i_j}_{\boldsymbol{\eta},\mathbf{w}} = \bar{s}^{i_j}_{\boldsymbol{\kappa}}$ for all $j \in [d+1]$ s.t. $w^{i_j} = 1\}$ and $H'_{\boldsymbol{\eta}} := \{\mathbf{y}_{\boldsymbol{\eta},\mathbf{w}}|_{(i_0,\ldots,i_d)} : \mathbf{w} \in W_{\boldsymbol{\eta}}\}$. We have $H'_{\boldsymbol{\eta}} = \{s^0_{\boldsymbol{\kappa}}, \bar{s}^0_{\boldsymbol{\kappa}}\} \times \cdots \times \{s^d_{\boldsymbol{\kappa}}, \bar{s}^d_{\boldsymbol{\kappa}}\}$ and $|H'_{\boldsymbol{\eta}}| = 2^{d+1}$. For any $\mathbf{g} = (g_1, \ldots, g_{d+1}) \in \{0,1\}^{d+1}$, define the set

$$W_{\boldsymbol{\eta},\mathbf{g}} := W_{\boldsymbol{\eta}} \cap \{\mathbf{w} \in \{0,1\}^{r(d+1)} : w^{i_j} = g^j \text{ for all } j \in [d+1]\}.$$

Since $H_{\boldsymbol{\eta}}|_{(i_1,\ldots,i_d)} \supseteq \{s^0_{\boldsymbol{\kappa}}, \bar{s}^0_{\boldsymbol{\kappa}}\} \times \cdots \times \{s^d_{\boldsymbol{\kappa}}, \bar{s}^d_{\boldsymbol{\kappa}}\}$, we have $W_{\boldsymbol{\eta},\mathbf{g}} \neq \varnothing$. Then, we pick a sequence $\mathbf{w}_{\mathbf{g}} \in W_{\boldsymbol{\eta},\mathbf{g}}$ for any $\mathbf{g} \in \{0,1\}^{d+1}$. For any $n \in \mathbb{N}$ with $n \geqslant d+1$ and any $\mathbf{u} = (\mathbf{u}_{d+1}, \ldots, \mathbf{u}_n) \in \prod_{l=d+1}^n \{0,1\}^l$, we define

$$\begin{aligned}\boldsymbol{\alpha}(\mathbf{u}) := \big(&(\mathbf{w}_{\mathbf{u}_{d+1}}, (\alpha(\mathbf{u})^0_{r(d+1)+1}, \ldots, \alpha(\mathbf{u})^{r(d+1)}_{r(d+1)+1}), \ldots, \\ &(\alpha(\mathbf{u})^0_{n+(r-1)(d+1)}, \ldots, \alpha(\mathbf{u})^{n+(r-1)(d+1)-1}_{n+(r-1)(d+1)})\big) \\ \in &\prod_{l=r(d+1)}^{n+(r-1)(d+1)} \{0,1\}^l\end{aligned}$$

with

$$\alpha(\mathbf{u})^i_l := \begin{cases} u^i_{l-(r-1)(d+1)}, & \text{if } 0 \leqslant i \leqslant l-(r-1)(d+1)-1, \\ 0, & \text{if } l-(r-1)(d+1) \leqslant i \leqslant l \end{cases}$$

for any $r(d+1)+1 \leqslant l \leqslant n+(r-1)(d+1)$ and $0 \leqslant i \leqslant l-1$.

Next, for any $n \in \mathbb{N}$ with $n \geqslant d$ and any $\mathbf{u} \in \prod_{l=d+1}^n \{0,1\}^l$, define

$$\tilde{v}_{\boldsymbol{\kappa},\mathbf{u}} := \begin{cases} ((\mathbf{x}^{i_0}_{\boldsymbol{\eta}}, \ldots, \mathbf{x}^{i_d}_{\boldsymbol{\eta}}), (\mathbf{s}^{i_0}_{\boldsymbol{\eta}}, \ldots, \mathbf{s}^{i_d}_{\boldsymbol{\eta}}), h_{\boldsymbol{\eta}}) & \text{if } n = d, \\ ((\mathbf{x}^0_{\boldsymbol{\eta},\boldsymbol{\alpha}(\mathbf{u})}, \ldots, \mathbf{x}^n_{\boldsymbol{\eta},\boldsymbol{\alpha}(\mathbf{u})}), (\mathbf{s}^0_{\boldsymbol{\eta},\boldsymbol{\alpha}(\mathbf{u})}, \ldots, \mathbf{s}^n_{\boldsymbol{\eta},\boldsymbol{\alpha}(\mathbf{u})}), h_{\boldsymbol{\eta},\boldsymbol{\alpha}(\mathbf{u})}), & \text{if } n \geqslant d+1. \end{cases}$$

Then, we obtain the following tree

$$\tilde{\mathsf{T}}_{G,\boldsymbol{\kappa}} := \cup_{n=d}^\infty \Big\{\tilde{v}_{\boldsymbol{\kappa},\mathbf{u}} : \mathbf{u} \in \prod_{l=d+1}^n \{0,1\}^d\Big\} = \Big\{\tilde{v}_{\mathbf{u}} : \mathbf{u} \in \cup_{n=d}^\infty \Big(\{\boldsymbol{\kappa}\} \times \prod_{l=d+1}^n \{0,1\}^d\Big)\Big\}.$$

We replace $\mathsf{T}_{G,\boldsymbol{\kappa}}$ with $\tilde{\mathsf{T}}_{G,\boldsymbol{\kappa}}$ in T_G by replacing $v_{\mathbf{u}} = (\mathbf{x}_{\mathbf{u}}, \mathbf{s}_{\mathbf{u}}, h_{\mathbf{u}})$ with $\tilde{v}_{\mathbf{u}}$ and still use $v_{\mathbf{u}} = (\mathbf{x}_{\mathbf{u}}, \mathbf{s}_{\mathbf{u}}, h_{\mathbf{u}})$ to denote $\tilde{v}_{\mathbf{u}}$ in T_G after the replacement for all $\mathbf{u} \in \cup_{n=d}^\infty(\{\boldsymbol{\kappa}\} \times \prod_{l=d+1}^n \{0,1\}^l)$. Now, we have $h_{\mathbf{u}}(x^i_{\boldsymbol{\kappa}}) = s^i_{\boldsymbol{\kappa}}$ if $u^i_{d+1} = 0$ and $h_{\mathbf{u}}(x^i_{\boldsymbol{\kappa}}) = \bar{s}^i_{\boldsymbol{\kappa}}$ if $u^i_{d+1} = 1$ for all $0 \leqslant i \leqslant d$ and $\mathbf{u} \in \cup_{n=d}^\infty(\{\boldsymbol{\kappa}\} \times \prod_{l=d+1}^n \{0,1\}^l)$ and T_G is still an infinite GL tree with the associated hypotheses after the replacement.

After the above procedure for all $\boldsymbol{\kappa} \in \prod_{l=1}^d \{0,1\}^l$, we obtain a set $\cup_{n=0}^d \{\bar{\mathbf{s}}_{\mathbf{u}} \in \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^n \{0,1\}^l\}$ and an infinite GL tree with the associated hypotheses $\{(x_\varnothing, s_\varnothing)\} \cup \big(\cup_{n=1}^\infty \{(\mathbf{x}_{\mathbf{u}}, \mathbf{s}_{\mathbf{u}}, h_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^n \{0,1\}^l\}\big)$ of $\mathcal{H}$ such that $\cup_{n=0}^d \{(\mathbf{x}_{\mathbf{u}}, \mathbf{s}_{\mathbf{u}}, \bar{\mathbf{s}}_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^n \{0,1\}^l\}$ is a NL of $\mathcal{H}$ of depth $d+1$ and for any $\mathbf{u} \in \cup_{n=d+1}^\infty (\prod_{l=1}^n \{0,1\}^l)$, we have $h_{\mathbf{u}}(x^i_{\mathbf{u}_{\leqslant l}}) = s^i_{\mathbf{u}_{\leqslant l}}$ if $u^i_{l+1} = 0$ and $h_{\mathbf{u}}(x^i_{\mathbf{u}_{\leqslant l}}) = \bar{s}^i_{\mathbf{u}_{\leqslant l}}$ otherwise for all $0 \leqslant i \leqslant l$ and $0 \leqslant l \leqslant d$. Thus, the induction hypothesis has been shown for $d+1$.

By induction, there exists an infinite set $\cup_{n=0}^{\infty}\{\bar{\mathbf{s}}_{\mathbf{u}} \in \mathcal{Y}^{n+1} : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ and an infinite GL tree $\cup_{n=0}^{\infty}\{(\mathbf{x}_{\mathbf{u}}, \mathbf{s}_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ of $\mathcal{H}$ such that $\cup_{n=0}^{\infty}\{(\mathbf{x}_{\mathbf{u}}, \mathbf{s}_{\mathbf{u}}, \bar{\mathbf{s}}_{\mathbf{u}}) : \mathbf{u} \in \prod_{l=1}^{n}\{0,1\}^l\}$ is an infinite NL of $\mathcal{H}$. It follows that an infinite GL tree of $\mathcal{H}$ implies an infinite NL tree of $\mathcal{H}$.

Finally, we can conclude that $\mathcal{H}$ has an infinite NL tree if and only if it has an infinite GL tree. ■

Appendix G. Proof of Proposition 12

Proof For any $n \in \mathbb{N}$ and $\mathbf{x} \in \mathcal{X}$, if $\widehat{h}_n(\mathbf{x}) = k$ for some $k \in [K]\backslash\{1\}$, there exists some $\mathbf{z}' \in [0,\infty)^d$, $1 \leqslant i_1 < \cdots < i_t \leqslant n$ and $(\alpha_1, \ldots \alpha_t) \in [0,1]^t$ such that $\sum_{\tau=1}^{t} \alpha_\tau = 1$ and

$$\mathbf{x} = \mathbf{z}' + \sum_{\tau=1}^{t} \alpha_\tau \mathbf{x}_{i_\tau}.$$

Then, for any $k' < k$, we have

$$\mathbf{w}_{k'} \cdot \mathbf{x} - b_{k'} = \mathbf{w}_{k'} \cdot \mathbf{z}' + \sum_{\tau=1}^{t} \alpha_\tau (\mathbf{w}_{k'} \cdot \mathbf{x}_{i_\tau} - b_{k'}) \leqslant \mathbf{w}_k \cdot \mathbf{z}' + \sum_{\tau=1}^{t} \alpha_\tau (\mathbf{w}_k \cdot \mathbf{x}_{i_\tau} - b_k),$$

which implies that $h_{n+1}(\mathbf{x}) \geqslant k = \widehat{h}_n(\mathbf{x})$. Then, for any $n \in \mathbb{N}$ such that $\widehat{h}_n(\mathbf{x}_{n+1}) \neq y_{n+1} = h_{n+1}(\mathbf{x}_{n+1})$, we must have $\widehat{h}_n(\mathbf{x}_{n+1}) < y_{n+1}$. It follows from the definition of $\widehat{h}_n$ that for every $i \leqslant n$ such that $y_i \geqslant y_{n+1}$, there exists some $j \in [d]$ such that $(\mathbf{x}_i)_j > (\mathbf{x}_{n+1})_j$.

Suppose on the contrary that there exists a strictly increasing infinite sequence $(n_t)_{t\in\mathbb{N}}$ such that $\widehat{h}_{n_t}(\mathbf{x}_{n_{t+1}}) \neq y_{n_{t+1}}$ for all $t \in \mathbb{N}$. Now, define an infinite complete graph with vertex set $\{\mathbf{x}_{n_t}\}_{t\in\mathbb{N}}$ and color each edge $\{\mathbf{x}_{n_t}, \mathbf{x}_{n_{t'}}\}$ with $t < t'$ to be $\min\{j \in [d] : (\mathbf{x}_{n_t})_j > (\mathbf{x}_{n_{t'}})_j\} \in [d]$. Then, by the infinite Ramsey theory, there exist some $j \in [d]$ a strictly increasing infinite sequence $(t_i)_{i\in\mathbb{N}}$ such that the edge $\{\mathbf{x}_{n_{t_i}}, \mathbf{x}_{n_{t_{i'}}}\}$ is colored with j for all $i \neq i'$. Thus, by the rule of coloring, $(\mathbf{x}_{n_{t_i}})_j$ is a strictly decreasing infinite sequence in i, which contradicts the fact that $(\mathbf{x}_{n_{t_i}})_j \in \mathbb{N}$ for all $i \in \mathbb{N}$. Therefore, $(\widehat{h}_n)_{n\in\mathbb{N}}$ only makes finitely many mistakes for any consistent sequence $((\mathbf{x}_n, \mathbf{y}_n))_{n\in\mathbb{N}}$.

Moreover, if $\widehat{h}_n(\mathbf{x}_{n+1}) = y_{n+1}$, we claim that $\widehat{h}_{n+1} = \widehat{h}_n$. Indeed, for any $\mathbf{x} \in \mathcal{X}$, if $\mathsf{Y}_{S_n,\mathbf{x}} \neq \varnothing$, we have $\mathsf{Y}_{S_{n+1},\mathbf{x}} = \mathsf{Y}_{S_n,\mathbf{x}}$. Thus, $\widehat{h}_{n+1}(\mathbf{x}) = \widehat{h}_n(\mathbf{x})$. If $\mathsf{Y}_{S_n,\mathbf{x}} = \varnothing$, we must have $\widehat{h}_n(\mathbf{x}) = 1$, which implies that we have $k \notin \mathsf{Y}_{S_{n+1},\mathbf{x}}$ for any $k > 1$. Thus, $\widehat{h}_{n+1}(\mathbf{x}) = 1 = \widehat{h}_n(\mathbf{x})$. ■