Universal Rates for Regression: Separations between Cut-Off and Absolute Loss

Idan Attias IDANATTI@POST.BGU.AC
Ben-Gurion University

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Alkis Kalavasis ALKIS.KALAVASIS@YALE.EDU
Yale University

Amin Karbasi AMIN.KARBASI@YALE.EDU
Yale University

Grigoris Velegkas GRIGORIS.VELEGKAS@YALE.EDU
Yale University

Editors: Shipra Agrawal and Aaron Roth

Abstract

In this work we initiate the study of regression in the *universal rates* framework of (Bousquet et al., 2021). Unlike the traditional *uniform* learning setting, we are interested in obtaining learning guarantees that hold for all fixed data-generating distributions, but do not hold uniformly across them. We focus on the realizable setting and we consider two different well-studied loss functions: the cut-off loss at scale $\gamma > 0$, which asks for predictions that are γ-close to the correct one, and the absolute loss, which measures how far away the prediction is from the correct one. Our results show that the landscape of the achievable rates in the two cases is completely different. First we give a trichotomic characterization of the optimal learning rates under the cut-off loss: each class is learnable either at an exponential rate, a (nearly) linear rate or requires arbitrarily slow rates. Moving to the absolute loss, we show that the achievable learning rates are significantly more involved by illustrating that an *infinite* number of different learning rates is achievable. This is the first time that such a rich landscape of rates is obtained in the universal rates literature.

Keywords: Regression, Universal Rates, Statistical Learning Theory

1. Introduction

Regression stands as a cornerstone problem in statistical analysis and data science, extensively investigated in Machine Learning (ML) literature (Vapnik, 1999; Goodfellow et al., 2016; Bach, 2021), with wide-ranging applications spanning domains like Economics and Medicine (Dua and Graff, 2017). Despite its practical importance, a complete theoretical comprehension of the topic remains elusive. Consequently, the study of regression and associated error rates has remained a pivotal focus within learning theory (Alon et al., 1997; Bartlett et al., 1994; Simon, 1997; Bartlett and Long, 1998; Mendelson, 2002; Aden-Ali et al., 2023; Attias et al., 2023).

In the foundational Probably Approximately Correct (PAC) learning paradigm (Valiant, 1984), a recent work by Attias et al. (2023) characterized learnability in the setting of real-valued *realizable* regression. This characterization is based on a scaled variant of the One-Inclusion Graph algorithm of Haussler et al. (1994) and builds on an extensive line of research concerning binary and multiclass classification (Haussler et al., 1994; Blumer et al., 1989; Rubinstein et al., 2009; Hanneke, 2016; Daniely and Shalev-Shwartz, 2014; Brukhim et al., 2022; Daniely et al., 2015).

Although the PAC model offers a clean and elegant theoretical framework, it falls short in capturing the real-world dynamics of ML problems. The main drawback of this model is that it is worst-case: for any fixed learning algorithm, one seeks to establish bounds on its error rate that hold *uniformly* over all distributions. In particular, this means that as the size n of the dataset that the algorithm has access to increases, this worst-case distribution that witnesses the performance of the algorithm changes. Nonetheless, practical scenarios often involve measuring the error rate of algorithms as a function of the size of the dataset (or other resources) for *fixed* data distributions, prompting the exploration of the *learning curves* under such fixed distributions.

These thoughts, further empirically motivated by Cohn and Tesauro (1990, 1992); Schuurmans (1997), led to the theoretical framework of *universal learning* of Bousquet et al. (2021). This framework aims to understand the best possible asymptotic error rate that can hold for every data distribution, but without requiring an upper bound which applies uniformly to all of these distributions. In other words, this error rate is allowed to depend on distribution-specific constants. In the setting of realizable binary classification, Bousquet et al. (2021) showed the surprising result that the following trichotomy of rates exists for any concept class $\mathcal{H}$: given n samples, $\mathcal{H}$ is either universally learnable at an optimal exponential rate e^{-n}, or universally learnable with optimal linear rate $1/n$, or requires arbitrarily slow rates. This is in contrast to the standard dichotomy of PAC binary classification where any class is either learnable at an optimal linear rate or is not learnable at all. This result, which validated existing empirical evidence (Cohn and Tesauro, 1990) and demonstrated the importance of the universal learning framework, inspired a series of follow-up works in that area (Bousquet et al., 2022; Kalavasis et al., 2022; Hanneke et al., 2022a, 2023).

In our work, we revisit the fundamental problem of realizable regression establishing results in the *universal rates* setting.

1.1. The Regression Learning Task and Universal Rates

In this section we define the learning setting we consider in our work. There is a domain $\mathcal{X}$, which we assume to be a Polish space, a label space $\mathcal{Y} = [0,1] \cap \mathbb{Q}$[1] and a concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, which satisfies standard measurability assumptions. There is also a data generating process which is modeled as a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We define a *regressor* $h : \mathcal{X} \to [0,1]$ to be a universally measurable function and we consider two different loss functions for this regressor.

$$\text{Expected Absolute Loss:} \quad \mathrm{er}_{\mathcal{D}}(h) = \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}[|h(x) - y|]\,, \tag{1}$$

$$\text{Expected Cut-Off Loss:} \quad \mathrm{er}^{\gamma}_{\mathcal{D}}(h) = \mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}}[|h(x) - y| > \gamma],\ \text{for some fixed } \gamma > 0\,. \tag{2}$$

We call $\mathcal{D}$ *realizable* with respect to the hypothesis class $\mathcal{H}$ if $\inf_{h\in\mathcal{H}} \left\{\mathbf{Pr}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]\right\} = 0$. We denote it by $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$. We sometimes consider the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$ which we denote by $\mathcal{D}_{\mathcal{X}}$. We study the *universal* setting (Bousquet et al., 2021), in which a class $\mathcal{H}$ is learnable at rate $R(n)$ under the expected absolute loss if there exists a learning rule $\widehat{h}_n$ such that

$$(\forall \mathcal{D} \in \mathrm{RE}(\mathcal{H}))\,(\exists C, c)\ \text{such that}\ L_{\mathcal{D}}(\widehat{h}_n) \leq CR(cn), \forall n \in \mathbb{N}\,,$$

[1] Our results hold for label space $\mathcal{Y} = \mathbb{Q} \cap [0,1]$. They also extend to countable subspace of $[0,1]$. Extending our results to uncountable spaces is out of the scope of this work since it requires non-trivial measure-theoretic tools.

where $L_{\mathcal{D}}(\widehat{h}_n) \triangleq \mathbf{E}[\mathrm{er}_{\mathcal{D}}(\widehat{h}_n)]$, and the expectation is over the training set of $\widehat{h}_n$. Similarly, for some fixed $\gamma > 0$, a class $\mathcal{H}$ is learnable at rate $R(n)$ under the expected cut-off loss if there exists a learning rule $\widehat{h}_n$ whose expected loss $L^{\gamma}_{\mathcal{D}}(\widehat{h}_n) \triangleq \mathbf{E}[\mathrm{er}^{\gamma}_{\mathcal{D}}(\widehat{h}_n)]$ satisfies

$$(\forall \mathcal{D} \in \mathrm{RE}(\mathcal{H}))\,(\exists C_\gamma, c_\gamma) \text{ such that } L^{\gamma}_{\mathcal{D}}(\widehat{h}_n) \leq C_\gamma R(c_\gamma n), \forall n \in \mathbb{N}\,.$$

For the latter case, we have fixed $\gamma > 0$ a priori and so the constants may depend on γ (and the distribution $\mathcal{D}$). Notice that in the well-studied PAC learning setting (Valiant, 1984) the order of the quantifiers is flipped, i.e., the constants do not depend on the data-generating distribution and the rates are *uniform* over all realizable distributions.

Next, we define precisely what it means to be learnable at some rate $R(n)$ in the universal setting. The definition comes from the work of Bousquet et al. (2021).

Definition 1 (Learning Rates (Bousquet et al., 2021)) *Fix a concept class $\mathcal{H}$, and let $R : \mathbb{N} \to [0,1], R(n) \overset{n\to\infty}{\longrightarrow} 0$ be a rate function, where n is the number of i.i.d. samples from $\mathcal{D}$.*

- *$\mathcal{H}$ is learnable at rate R under the expected absolute loss if there is an algorithm $\hat{h}_n$ such that for every distribution $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$, there exist c, C for which $L_{\mathcal{D}}(\widehat{h}_n) \leq CR(cn), \forall n \in \mathbb{N}$. Similarly, for the γ-cut-off loss, if there exist c_γ, C_γ for which $L^{\gamma}_{\mathcal{D}}(\widehat{h}_n) \leq C_\gamma R(c_\gamma n), \forall n \in \mathbb{N}$.*

- *$\mathcal{H}$ is not learnable at rate faster than R under the expected absolute loss if for all algorithms $\widehat{h}_n$ there exists a distribution $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$ and c, C for which $L_{\mathcal{D}}(\widehat{h}_n) \geq CR(cn)$, for infinitely many $n \in \mathbb{N}$. Similarly, for the γ-cut-off loss, if there exist c_γ, C_γ for which $L^{\gamma}_{\mathcal{D}}(\widehat{h}_n) \geq C_\gamma R(c_\gamma n)$, for infinitely many $n \in \mathbb{N}$.*

- *$\mathcal{H}$ is learnable with optimal rate R if it is learnable at rate R and it is not learnable at rate faster than R.*

- *$\mathcal{H}$ requires arbitrarily slow rates if for all rates R, it is not learnable at rate faster than R.*

1.2. Main Results

We are now ready to present the main results of our work. Our first main result gives a characterization[2] of the optimal universal rates for the *cut-off* loss function. In particular, for every $\gamma > 0$, we show a trichotomic characterization of the optimal achievable rates which is reminiscent of the landscape for binary (Bousquet et al., 2021) and multiclass (Kalavasis et al., 2022; Hanneke et al., 2023) classification. In our results, we exclude trivially learnable classes $\mathcal{H}$, see Remark 6 for details.

Theorem 1 (Universal Regression for Cut-Off Loss) *Fix $\gamma \in (0,1)$. For any non-trivial hypothesis class $\mathcal{H}$, exactly one of the following holds for the expected γ-cut-off loss:*

- *$\mathcal{H}$ is learnable at an optimal rate e^{-n}.*

- *$\mathcal{H}$ is learnable at an optimal rate $\widetilde{\Theta}(1/n)$.*

- *$\mathcal{H}$ is learnable but requires arbitrarily slow rates.*

[2] Up to poly-logarithmic factors, which appear due to the fact that the optimal rate in the *uniform* setting is not known.

Next, we move on to the case of the *absolute* loss function, where we show a qualitatively different and more complicated landscape. In particular, we prove that for *all* rates between $o(1/n)$ and arbitrarily slow, there is a hypothesis class for which such a rate is (almost) optimal. This landscape provides an *infinite* collection of possible optimal rates and is much richer compared to what prior work on universal rates has provided and dealt with.

Theorem 2 (Universal Regression for Absolute Loss) *For any non-trivial hypothesis class $\mathcal{H}$, one of the following holds for the expected absolute loss:*

- *$\mathcal{H}$ is learnable at an optimal rate e^{-n} or*
- *$\mathcal{H}$ is learnable an optimal rate that is not faster than $o(1/n)$.*

Moreover,

- *There exists a hypothesis class $\mathcal{H}$ that is learnable at an optimal rate $o(1/n)$, but requires rates arbitrarily close to $1/n$.*
- *There exists a hypothesis class $\mathcal{H}$ that is learnable at an optimal rate $1/n$.*
- *For every rate $R(n)$ such that $n \cdot R(n)$ is non-decreasing, there exists a hypothesis class $\mathcal{H}$ such that no algorithm can learn $\mathcal{H}$ at a rate faster than $o(R(n))$, and there exists an algorithm that learns $\mathcal{H}$ at a rate $R(n)$.*[3]

Let us provide some further explanation for the $o(1/n)$ rate. To say the optimal rate is $o(1/n)$ means that there is a learner that, for every realizable distribution, achieves *some* $o(1/n)$ rate (e.g., for one distribution it might be $1/(n\log(n))$ and for another it may be $1/(n\log\log(n))$), but for any learner and any fixed rate $R(n) = o(1/n)$ there exists a distribution where the learner's rate is at least $R(n)$. This is not to be confused with saying that there is an optimal rate $R^*(n)$ which satisfies $R^*(n) = o(1/n)$; that would be a very different claim. This is also illustrated in the construction of our lower bound (see Section B.3), where for every target rate $R(n) = o(1/n)$ we can construct a "hard" distribution based on that particular $R(n)$.

Our result demonstrates that the landscape is more complicated than all the other problems that have been considered in the universal rates literature (Bousquet et al., 2021; Kalavasis et al., 2022; Hanneke et al., 2022a, 2023). In particular we show a clear separation between the universal rates obtained by using cut-off and absolute loss. While in the uniform PAC setting, the rates achieved by the two losses coincide (Attias et al., 2023), in the universal framework, the landscape of the expected absolute loss is highly more complicated. For instance, our result implies that in the absolute loss case, there is an infinite collection of possible optimal rates, while in the cut-off loss, there are only three. In contrast to the cut-off loss landscape, our result does not provide a complete characterization of the optimal rates with respect to the absolute loss. This is in large due to the fact that we are lacking a *quantitative* characterization of the optimal learning rates under this loss in the *uniform* case (Attias et al., 2023). The main intuition for this separation is that, taking γ to be fixed for the cut-off loss makes the regression problem similar to a classification task (and hence this is the reason why we get a trichotomy in Theorem 1). On the other hand, the case of the absolute loss is more "continuous" since the scale γ is not fixed and reveals the true difficulty of regression,

[3] For technical reasons, when we consider rates between $1/n$ and arbitrarily slow (e.g., $R(n) = 1/\sqrt{n}$), we need the function $n \cdot R(n)$ to be non-decreasing. We will elaborate on these details shortly in Section 3.4.

which was not captured by any prior work on universal rates. Interestingly we can give examples of classes realizing the rates of Theorem 2. Finite classes and thresholds over $\mathbb{N}$ are learnable at an exponential rate and thresholds over $\mathbb{R}$ at linear. We also provide a class that is learnable at an optimal rate arbitrarily close to $o(1/n)$ (see Section 3.3). Finally, for each rate slower than linear we provide an example class in Section 3.4.

Remark 3 *We mention that while our results are presented for the absolute loss, most of our techniques generalize directly to pseudo-metrics, as in Attias et al. (2023), including ℓ_p losses.*

Combinatorial Dimensions and Universal Rates. Our next set of results provides some necessary and sufficient combinatorial conditions that give rise to the (optimal) learning rates we mentioned before. We stress that our results regarding the combinatorial characterizations of the achievable rates, as in the prior work on universal rates (Bousquet et al., 2021, 2022; Kalavasis et al., 2022; Hanneke et al., 2023), are based on some *tree* structures. In our work, we introduce some novel tree structures, namely the γ-Littlestone tree for $\gamma \geq 0$ (a scaled version of the Littlestone tree from Littlestone (1988); Bousquet et al. (2021)) and a tree we call γ-OIG-Littlestone tree, which is based on a combination of the scaled One-Inclusion Graph (OIG) dimension of Attias et al. (2023) and the Littlestone tree. Since the definitions of these combinatorial measures are complicated, we first state our results regarding necessary and sufficient conditions to achieve the optimal learning rates for universal regression and then explain how these trees are defined.

Theorem 4 (Combinatorial Characterization for Cut-Off Loss) *Fix $\gamma \in (0,1)$. For any non-trivial class $\mathcal{H}$, exactly one of the following holds for the expected γ-cut-off loss*[4]*:*

- *$\mathcal{H}$ is learnable at an optimal rate e^{-n} if and only if it does not have an infinite γ-Littlestone tree.*
- *$\mathcal{H}$ is learnable at an optimal rate $\widetilde{\Theta}(1/n)$ if and only if it has an infinite γ-Littlestone tree, but does not have an infinite γ-OIG-Littlestone tree.*
- *$\mathcal{H}$ requires arbitrarily slow rates if and only if it has an infinite γ-OIG-Littlestone tree.*

Shifting our attention to the absolute loss, we obtain the following result.

Theorem 5 (Combinatorial Implications for Absolute Loss) *Let $\mathcal{H}$ be a non-trivial class. Then, the following hold for the expected absolute loss:*

- *$\mathcal{H}$ is learnable at an optimal rate e^{-n} if and only if it does not have an infinite 0-Littlestone tree.*
- *If $\mathcal{H}$ has an infinite 0-Littlestone tree, then it is not learnable at an optimal rate which is faster than $o(1/n)$.*
- *If there exists $\gamma > 0$ such that $\mathcal{H}$ has an infinite γ-Littlestone tree, then it is not learnable at an optimal rate which is faster than $1/n$.*

[4] For the ease of exposition we ignore a gap of a factor of 2 on the dependence on γ of the combinatorial measures we state. We will elaborate on these details shortly in Section 2.

- *There exists some $\mathcal{H}$ that has an infinite 0-Littlestone tree and is learnable at an optimal rate $o(1/n)$.*
- *If there exists some $\gamma > 0$ such that $\mathcal{H}$ has an infinite γ-OIG-Littlestone tree, then $\mathcal{H}$ requires arbitrarily slow rates.*

We can now introduce the combinatorial measures of interest, which we use in the above statements. First, we introduce a scaled variant of a Littlestone tree.

Definition 2 (Scaled-Littlestone Tree (Informal, see Definition 9)) *A **scaled-Littlestone tree** of depth $\mathrm{d} \leq \infty$ for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a complete binary tree of depth d whose internal nodes are labeled by $\mathcal{X}$, and whose two edges connecting a node of level $i \leq \mathrm{d}$ to its children are labeled by two different elements in $\mathcal{Y}$, such that every path of length at most d emanating from the root is consistent with a concept $h \in \mathcal{H}$. We say that $\mathcal{H}$ has an **infinite scaled-Littlestone tree** if there exists a scaled-Littlestone tree for $\mathcal{H}$ with depth $\mathrm{d} = \infty$. Moreover, if all the gaps (absolute difference) of the labels of the edges of each node of this tree are lower bounded by $\gamma > 0$, we call it γ-**Littlestone tree**, whereas if there is not a non-zero lower bound, we call it 0-**Littlestone tree**.*

Next, we describe a combinatorial measure that is based on the scaled *One-Inclusion Graph* (OIG) dimension which was shown recently to characterize *uniform* rates for realizable regression (Attias et al., 2023). We will only provide an informal description of this definition here and refer the reader to Section C. First, it is important to describe the real-valued version of the OIG.

Definition 3 (OIG (Informal, see Definition 10)) *Consider the domain $[n]$ and a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{[n]}$. The OIG induced by $\mathcal{H}$ is a hypergraph where the set of vertices is the set of all hypotheses and, for each direction $i \in [n]$, every non-empty hyperedge in that direction corresponds to the set of all hypotheses in $\mathcal{H}$ that agree with the labelings of all the points except for that of point i.*

Subsequently, we give the description of an *orientation* of the OIG hypergraph as well as the definition of the out-degree of a node in the OIG induced by a given orientation.

Definition 4 (Orientation and Scaled Out-Degree (Informal, see Definition 11)) *An orientation σ of the OIG (V, E) is a mapping from every hyperedge to vertices of that particular hyperedge. Given $\gamma \in (0,1)$, the γ-out-degree of every node $v \in V$ under σ is defined to be $|\{i \in [n] : |\sigma(e_{i,v}) - v(i)| > \gamma\}|$, where, given $i \in [n]$, $v(i)$ is the value of v in direction i and $e_{i,v} \in E$ is the hyperedge $\{h \in V : h(j) = v(j), \forall j \in [n] \setminus \{i\}\}$.*

Let us give some intuition about Definition 3 and Definition 4. We have a set of $[n]$ unlabeled points. All the hypotheses that agree with the labels of the $[n]$ points form an equivalence class, and each of these classes is a vertex of the OIG graph. In other words, the set of vertices of the graph is a set of equivalence classes within $\mathcal{H}$. For every direction $i \in [n]$, we form a hyperedge e_i by considering all the vertices that agree in the labels of the $[n] \setminus \{i\}$ points. By doing that, we can think of the orientation of e_i as the prediction that the algorithm would make on the datapoint i if it had only observed the points $[n] \setminus \{i\}$.

Equipped with these definitions, we are now ready to describe the scaled OIG dimension as defined by Attias et al. (2023). For $S = (x_1, ..., x_n) \in \mathcal{X}^n$, recall that $\mathcal{H}|_S = \{(h(x_1), ..., h(x_n)) : h \in \mathcal{H}\}$ is the projection of $\mathcal{H}$ onto the unlabeled dataset S.

Definition 5 (OIG Dimension (Attias et al., 2023) (Informal, see Definition 12)) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $\gamma \in (0,1)$. We define the scaled OIG dimension of $\mathcal{H}$ to be the largest $n \in \mathbb{N}$ for which there exists some $S \in \mathcal{X}^n$ such that there exists a finite subgraph $G = (V, E)$ of the OIG induced by $\mathcal{H}|_S$ so that for all orientations, there exists a node in V with γ-out-degree at least $n/3$.*

Intuitively, this definition says that for some unlabeled dataset S one can find a *finite* subgraph of $\mathcal{H}_{|S}$ so that, no matter how we orient the hyperedges, there exists one node with large out-degree. Recall that orientations of this graph correspond to predictions of the algorithm trained on S, so, intuitively, subgraphs with large out-degree witness the difficulty of the learning task.

We are now ready to describe the structure of the scaled-OIG-Littlestone tree, a novel combinatorial measure that combines the structure of the Littlestone tree and the scaled OIG dimension.

Definition 6 (OIG-Littlestone Tree (Informal, see Definition 13) *Fix some $\gamma \in (0,1)$. A **γ-OIG-Littlestone tree** of depth $d \leq \infty$ for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a tree of depth d whose internal nodes in every level i are labeled by an element $S \in \mathcal{X}^{i+1}$ and a finite set of labelings $L \subseteq \mathcal{Y}^{i+1}$ of these elements with the property that for all the orientations of the OIG induced by S, L, there exists a node that has γ-out-degree at least $(i+1)/3$, and whose edges connecting a node of level $i \leq d$ to its children are labeled by elements of all the labelings L that are contained in that node. Moreover, the labels that appear on the edges of every path of length at most d from the root to a leaf must be consistent with a concept $h \in \mathcal{H}$. We say that $\mathcal{H}$ has an **infinite γ-OIG-Littlestone tree** if there exists a γ-OIG-Littlestone tree for $\mathcal{H}$ with depth $d = \infty$.*

The intuition is that this structure is a tree whose nodes have size that increases linearly with the depth and witness a finite subgraph of the OIG that has “large” out-degree for all possible orientations. Along each path, the hypothesis class gets more and more restricted since it has to be consistent with the chosen path, so it is getting increasingly more difficult to “shatter” a set of points, in the sense of the OIG-shattering described in Definition 5.

Remark 6 (Non-Trivial Classes) *In our main results, we state the requirement that $\mathcal{H}$ is non-trivial. This is to exclude some classes that are learnable by a single sample. For the formal description of these classes, we refer the reader to Definition 7 and Definition 8.*

Remark 7 (General Loss Functions) *In Theorem 4 and Theorem 5 we provide combinatorial structures characterizing the achievable rates. The designed tree are based on the absolute loss. If one would like to study some other loss (e.g., squared loss), then these tree structures would be defined on a notion of distance based on the new loss but the proofs do not require further modification.*

1.3. Novel Challenges Compared to Prior Work

In this section, we briefly discuss some challenging points behind our results that have not appeared in prior works, neither in the universal rates nor in the uniform regression literature.

Scaled Trees. Compared to prior universal papers, since we study regression problems, our combinatorial measures, i.e., the trees that we consider, depend on some scaling factor γ, which intuitively corresponds to the gap between the two labels of each node. However, this adds a complication to our results since different resolutions potentially yield different rates. For instance, one can consider a (γ_n)-Littlestone tree (see Definition 9) where the gaps of each level $n \in \mathbb{N}$ change according to the sequence γ_n. The different behaviors of this sequence could yield qualitatively different rates in the learning problem. A manifestation of this phenomenon is that, for the absolute loss, (i) infinite γ-Littlestone trees for fixed γ imply linear lower bounds on the rate while (ii) (the weaker) infinite 0-Littlestone trees give a lower bound of order $o(1/n)$. These changes in the rate based on the different resolutions of the scale is a novel aspect of our work.

Collection of Rates. While the cut-off loss enjoys the standard trichotomic characterization that appeared in prior works, the landscape of the absolute loss is significantly more involved. Our Theorem 2 implies an infinite collection of possible optimal rates: the rate can be $e^{-n}, o(1/n), 1/n$ and (roughly) any function that decreases slower than linearly. We find the case that the rate $o(1/n)$ (i.e., faster than linear but arbitrarily close to it) is optimal for some universal regression tasks quite interesting, since this rate was only observed in prior works in active learning tasks (Hanneke, 2012) and not supervised ones. More to that, this work is the first in the universal rates literature that shows that infinitely many rates are possible for some task. While characterizing when each rate occurs for the absolute loss seems to be a far-reaching goal, even showing that all these rates are achievable is non-trivial and requires various new ideas (see Section 3.3 and Section 3.4).

1.4. Related Work

PAC Regression. The problem of learning real-valued functions under various losses, such as the cut-off loss and the absolute loss, has received a lot of attention in the PAC learning theory literature. Simon (1997) showed that the finiteness of the scaled Natarajan dimension is a necessary condition for realizable uniform regression. Subsequently, Bartlett and Long (1998) used the OIG algorithm (Haussler et al., 1994) to get a real-valued predictor whose expected error is upper bounded by the V_γ-dimension. We refer the interested reader to the work of Kleer and Simon (2023) for details about this dimension which was introduced by Alon et al. (1997). A series of works (Alon et al., 1997; Bartlett et al., 1994) showed that the finiteness of the fat shattering dimension at all scales is a necessary and sufficient condition for uniform learnability, learnability in the agnostic setting and some noise models but does not characterize learnability in the *realizable* setting. Recently, the work of Aden-Ali et al. (2023) provided high-probability bounds for the OIG algorithm in the realizable uniform regression setting using the V_γ-dimension. Subsequently, Attias et al. (2023) proposed a dimension which is based on the structure of the OIG algorithm and showed that it characterizes learnability of uniform realizable regression. For a discussion about more general loss functions, we refer the reader to Mendelson (2002); Bartlett et al. (2005); Attias et al. (2023). Prasadan and Neykov (2024) study minimax rates of realizable regression of some concept class $\mathcal{H}$ for some fixed marginal $\mathcal{D}_X$ and the rate depends on it, where realizability means zero-mean noise to the labels; in our case, we have no noise and this is crucial since, as mentioned in Attias et al. (2023), our notion of realizability heavily changes realizability (there are classes learnable with 1 sample in our noiseless case, that require infinitely many samples in the presense of noise). As in Prasadan and Neykov (2024), we also get distribution-dependent rates $CR(cn)$, but only the constants C, c depend on $\mathcal{D}$ (not just $\mathcal{D}_X$). This is the difference with standard PAC learning where the constants are distribution independent. More to that Prasadan and Neykov (2024), $\mathcal{H}$ is convex and bounded, but we handle general classes. One of the key tools in Prasadan and Neykov (2024) and other related works is packing numbers. Packing numbers fail to capture the gap between realizable and agnostic PAC due to "discretization" error, so some notion of "packing trees" could not characterize our setting. Our dimensions capture the universal rates for realizable regression and are $\mathcal{D}_X$-independent.

Universal Rates. The study of universal learning rates was initiated by the seminal work of Bousquet et al. (2021) who derived a complete characterization of the optimal rates in the binary classification setting. Subsequently, Bousquet et al. (2022) derived more fine-grained results for binary classification. Later, Kalavasis et al. (2022) extended the result of Bousquet et al. (2021) to the multiclass setting, with the restriction that the number of different classes is bounded. Recently, Hanneke et al. (2023) improved upon this result by characterizing multiclass classification with an infinite number of labels. Diverging from the supervised learning line of work, Hanneke et al. (2022b) derived a complete characterization of the optimal learning rates for binary classification in a general interactive learning setting.

1.5. Further Remarks and Open Problems

We conclude this introductory section with some remarks and open questions. In this work we have initiated the study of regression problems in the universal rates framework, both under the cut-off loss and the absolute loss. Our results show that this problem exhibits a much richer landscape than other learning tasks that have been studied in this framework.

Beyond Realizability. Ideally, the literature should provide a complete understanding of the regression problem in the universal setting, including learning (i) without noise, (ii) with well-structured noise (e.g., zero mean Gaussian noise added to the true label) or (iii) with arbitrary corruptions. In our work, we are handling one case of this: realizable learning, i.e., without noise. Recent work by Attias et al. (2023) shows a separation between the noise-free setting and noisy setting for the PAC framework (uniform rates), and we expect that this distinction will remain valid under a universal analysis. We emphasize that extending the universal rates results to noisy settings is a very interesting direction and even for the much easier task of binary classification, no such result is known. We expect that our techniques would be useful but they would rely on a different tree structure, like a fat-shattering type of tree.

Other Open Questions. One concrete open problem that follows from our results is whether all classes that are uniformly learnable at some optimal rate $R(n)$, which is faster than $\mathrm{poly}(\log n)/n$, can be learned at a rate $o(R(n))$ in the universal setting, under the absolute loss. Furthermore, it is very interesting to study this problem in the *agnostic* setting, i.e., where there is no restriction on the data-generating distribution. Another direction is to resolve the PAC optimal rates and then translate this to universal rates. Since the OIGL tree is complicated, we believe that a simpler tree based on the scaled DS dimension (see Attias et al. (2023)) could yield similar results. Finally it is interesting to study the case of uncountable label space or some specific function classes such as Hölder or Sobolev.

2. Universal Rates Landscape for Cut-Off Loss

The first step in order to show Theorem 1 is the exponential rates case. This proof builds on the tools provided by Bousquet et al. (2021), carefully adapted to the notion of γ-Littlestone trees. For the proof, we refer to Section A.1 and Section A.2.

2.1. Near Linear Rates for Cut-Off Loss

We now move on to the second part of the trichotomy, the case of (near) linear rates. We sketch how to prove the next result, which characterizes the concept classes that are learnable at a near linear rate $\widetilde{\Theta}(1/n)$ with respect to the γ-cut-off loss.

Theorem 8 *Let $\gamma > 0$. For any non-trivial class $\mathcal{H}$, if $\mathcal{H}$ has an infinite 2γ-Littlestone tree, but does not have an infinite γ-OIG-Littlestone tree, then it is learnable at a rate $\log^2(n)/n$, but is not learnable faster than $1/n$ with respect to the expected γ-cut-off loss.*

The proof of the linear lower bound appears in Section A.3. It essentially extends the techniques of Bousquet et al. (2021); Kalavasis et al. (2022); Hanneke et al. (2023) to the regression setting. The lower bound construction uses the probabilistic method: it first picks a branch of the tree $\boldsymbol{y} \in \{0,1\}^\infty$ uniformly at random and then designs a distribution $\mathcal{D}_{\boldsymbol{y}}$ associated with this path which is realizable with respect to $\mathcal{H}$. A careful design of $\mathcal{D}_{\boldsymbol{y}}$ implies that some $(1/n)$-fraction of the distribution will lie deeper in the tree than the deepest point which appears in the dataset, and for this fraction the learner will make a mistake, in the sense of the 2γ-cut-off loss, with probability $1/2$. The challenging part of the proof is the upper bound (cf. Section A.4), i.e., the design of an algorithm that has near linear universal rate, when $\mathcal{H}$ does not have an infinite γ-OIG-Littlestone (OIGL) tree. Similarly to Hanneke et al. (2023), our main result is a general reduction from uniform to universal rates.

Lemma 1 (Informal, see Theorem 18) *Consider any learning algorithm $\widehat{h}_n$ that, in the uniform setting, learns at rate $R(n;d)$ a concept class $\mathcal{H}$ with γ-OIG dimension at most d, using n samples from a distribution $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$ with respect to the γ-cut-off loss, i.e., $L^{\gamma}_{\mathcal{D}}(h_n) \le R(n;d)$.*

Then if $\mathcal{H}$ does not have an infinite γ-OIG-Littlestone tree, there is a learning algorithm h'_n that, in the universal setting, achieves the same $R(n;d')$ rate for some constant $d' \in \mathbb{N}$.

This lemma allows us to use a recent result of Attias et al. (2023) that learns classes of γ-OIG dimension equal to d at a rate $d\log^2(n)/n$. We mention that a positive aspect of this black-box reduction is that any improvement in the uniform learning rate would directly imply an improvement in our universal rate. The proof of this key lemma is complicated and requires various tools from Bousquet et al. (2021); Hanneke et al. (2023) together with an extension of the results of Attias et al. (2023). As a starting point, we consider the adversarial online setting and design a game between an adversary P_A and a learner P_L (as in the exponential rates case). This can be found in Section A.4.1. The game is convoluted and relies on the structure of an OIGL tree. In particular, the strategy space of the adversary will be the set of "OIG-type" of graphs that have large out-degree for every orientation of the edges (see Definition 14). The game in every round $\tau \in \mathbb{N}$ is defined as follows. First, the adversary P_A chooses a sequence $x_\tau = (x^0_\tau, \ldots, t^{\tau-1}_\tau) \in \mathcal{X}^\tau$ and a finite set of labelings $L_\tau \subseteq \mathrm{LG}_{\tau,\gamma}$, where $\mathrm{LG}_{\tau,\gamma}$ appears in Definition 14 [5] and then P_L chooses an element $y_\tau \in L_\tau$. The winning condition of the game for player P_L is the following: player P_L wins if for some round τ we have $\mathcal{H}_{x_1,y_1,\ldots,x_\tau,y_\tau} = \emptyset$, where $\mathcal{H}_{x_1,y_1,\ldots,x_\tau,y_\tau} := \left\{h \in \mathcal{H} : h(x^i_s) = y^i_s, \forall\, 0 \le i \le s, 1 \le s \le \tau\right\}$.

Crucially, one can prove that $\mathcal{H}$ does not have an infinite γ-OIG-Littlestone tree if and only if P_L has a universally measurable winning strategy in the above OIG-Littlestone game. We now have to

[5] It is the set of all finite subsets of $\mathcal{Y}^\tau$ that have the property that the graph whose nodes are all the elements of that particular finite subset of $\mathcal{Y}^\tau$ and whose hyperedges are defined as in the OIG, has the property that all its orientations have a node with γ-out-degree at least $n/3$.

turn this winning strategy to an algorithm that makes predictions about the labels. In this adversarial setting, we assume access to an infinite sequence of labeled data. Informally, our approach is to use a constant fraction of the labeled data sequence that the learner has access to in order to "simulate" the OIG-Littlestone (Gale-Stewart) game between the adversary P_A and the learner P_L. Once this game has "converged", this will give rise to a pattern-avoidance function which will define some constraints that all realizable datasets need to satisfy.

This pattern-avoidance function can then be used to define a *partial* concept class (Alon et al., 2022). This partial concept class[6] is defined based on the constraints generated by the pattern-avoidance function. The crucial observation is that the scaled OIG dimension of this partial concept class is finite. We now have to extend the results of Attias et al. (2023), which PAC learn total concept classes with finite scaled OIG dimension, to partial concept classes. This is done in Theorem 26 and the uniform rate is $d\log^2(n)/n$, where d is the bound on the γ-OIG dimension. As a final step, given a collection of samples generated by some unknown $\mathcal{H}$-realizable distribution, we split the data into batches and, roughly, use our pattern avoidance function (obtained by the winning strategy of P_L) to design a partial concept class of scaled OIG dimension d that we then learn at a rate $d\log^2(n)/n$. Thus every batch gives rise to a regressor and we can show that the majority of them have small γ-cut-off loss. The final result follows by aggregating their predictions and outputting the median.

2.2. Arbitrarily Slow Rates for Cut-Off Loss

In this subsection, we discuss the following result, whose proof appears in Section A.5.

Theorem 9 *Let $\gamma \in (0,1)$. For any non-trivial class $\mathcal{H}$, if $\mathcal{H}$ has an infinite 2γ-OIG-Littlestone tree, then it requires arbitrarily slow rates with respect to the expected γ-cut-off loss.*

The proofs of all the lower bounds appearing in Bousquet et al. (2021); Kalavasis et al. (2022); Hanneke et al. (2023) and the proof of our linear lower bound for the cut-off loss (cf. Section 2.1) share a common structure: they use the probabilistic method and the construction is independent of the learning algorithm. However, for the proof of Theorem 9 we take a different route. In particular, due to the structure of the scaled OIGL tree, it is not clear how to obtain a hard distribution using the probabilistic method. To this end, we design a hard instance that is *algorithm-dependent*, and we choose the distribution that witnesses the lower bound in a *deterministic* way.

Let us now sketch the main idea of the lower bound. Assume an infinite 2γ-OIGL tree for $\mathcal{H}$. Fix an arbitrary rate R that is slower than linear and consider a learning algorithm $\mathbb{A}$. We would like to show that there exists a realizable distribution for which the algorithm $\mathbb{A}$ has expected γ-cut-off loss at least R. In particular, the hard distribution will crucially depend on $\mathbb{A}$.

Let us first recall the structure of the OIGL tree. We know that, for any level $k \geq 1$, the node of the tree contains a tuple: it contains an element S in $\mathcal{X}^{k+1}$ and it also contains a finite set of labelings $L \subseteq \mathcal{Y}^{k+1}$ of S. The node of the k-th level of the scaled OIGL tree has the special property that for all the orientations of the scaled OIG induced by (S, L), there exists a node, i.e., a particular labeling of S, which has 2γ-out-degree at least $(k+1)/3$. The next step is to utilize the conceptual connection between learning algorithms and orientations of the OIG, i.e., the fact

[6] Inspired by Alon et al. (2022), instead of dealing with concept classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ where each concept $h \in \mathcal{H}$ is a **total function** $h : \mathcal{X} \to \mathcal{Y}$, we study **partial concept classes** $\mathcal{H} \subseteq (\mathcal{Y} \cup \{\star\})^{\mathcal{X}}$, where each concept h is now a **partial function** and $h(z) = \star$ means that the function h is **undefined** at $z \in \mathcal{X}$. For further details, we refer to Section D.

that every algorithm induces some orientation of each hyperedge (Haussler et al., 1994; Attias et al., 2023). Hence, given the learner $\mathbb{A}$, we can identify an orientation $\sigma_{\mathbb{A}}$ and then based on the infinite scaled OIGL tree, find the node that has large scaled out-degree *with respect to* $\sigma_{\mathbb{A}}$. Hence, instead of randomly picking a branch in the infinite tree, as the previous constructions do, we have to deterministically pick the node that has large out-degree. In order to complete the proof, it remains to (i) argue how to assign mass in different levels of the tree and inside the large out-degree nodes, (ii) verify that the designed distribution is $\mathcal{H}$-realizable, and, (iii) show that the expected cut-off loss of the algorithm A is at least $R(n)$ for infinitely many values of n. This is done in Section A.5.

3. Universal Rates Landscape for Absolute Loss

3.1. Exponential Rates for the Absolute Loss

In this subsection, we sketch the following theorem.

Theorem 10 *For any non-trivial class $\mathcal{H}$, $\mathcal{H}$ is learnable at an optimal rate e^{-n} with respect to the expected absolute loss if and only if it does not have an infinite 0-Littlestone tree.*

The lower bound follows from the cut-off loss. For the upper bound, we reduce the problem to multiclass classification. Essentially, we can show that if $\mathcal{H}$ does not have an infinite 0-Littlestone tree, then it does not have an infinite multiclass Littlestone tree. This is because at any level n of the 0-Littlestone tree, there exists a non-zero positive gap value γ_n such that any pair of labels having the same parent node differ by at least γ_n. But this tree, corresponds to a multiclass Littlestone tree, which by assumption, is not infinite. Since (i) the label space we are working on is countable and (ii) any class $\mathcal{H}$ without an infinite multiclass Littlestone tree is learnable at exponential universal rates (Hanneke et al., 2023), we get the desired rate. For the formal argument, see Section B.2.

3.2. Infinite Scaled Littlestone Trees and Lower Bounds

In this section, the landscape of absolute regression rates starts to become more complicated. We show two lower bounds on the possible rate depending on the scaled-Littlestone trees of different resolution being infinite.

Theorem 11 *For any non-trivial class $\mathcal{H}$:*

- *If $\mathcal{H}$ has an infinite γ-Littlestone tree for some $\gamma > 0$, then it is not learnable at a rate faster than $1/n$ with respect to the expected absolute loss.*
- *If $\mathcal{H}$ has an infinite 0-Littlestone tree, then it is not learnable at a rate faster than $o(1/n)$ with respect to the expected absolute loss.*

The fist part of this result follows immediately from the lower bound regarding the γ-cut-off loss. In particular, this lower bound shows that after observing n samples the learner will be making predictions that are $\gamma/2$ far of the correct one for a $(1/n)$-fraction of the population, hence its absolute loss will be at least $\gamma/(2n)$. The proof of the second part of the result uses the probabilistic method to construct “hard” distributions for the learning algorithms. Given the description of any 0-Littlestone and any sublinear rate $R(n)$, we will construct a realizable distribution that is supported on a random branch of the tree. We define the support of the distribution inductively, by skipping sufficiently many levels of the tree as dictated by the rate function and the gap on each level of the tree. Moreover, the mass on each level of the support is inversely proportional to the level. Under this choice of the probability distribution, we can show that there exists an infinite sequence $\{n_\ell\}_{\ell\in\mathbb{N}}$, so that when the learner observes $O(n_\ell)$ many samples, with some constant probability, it will not observe any elements of the tree that are in a deeper level than n_ℓ. Thus, its prediction for the element of this level will be $\gamma_\ell/2$-away from the correct one, where γ_ℓ is the minimum gap size on level n_ℓ. Since the gap sizes are decreasing, the mistakes that the learner makes become less significant as the sample size increases. This is why we can merely show an $o(1/n)$ lower bound instead of a linear $1/n$ lower bound, as in the cut-off loss case. There are some technical details of the proof which make it more complicated than the one in the case of the cut-off loss and they are handled in Section B.3.

A natural question is whether there exist concept classes that realize these rates. First, note that the class of thresholds enjoys a linear rate and so there is a class that is learnable at an optimal $1/n$ rate in the universal regression setting with respect to the absolute loss. The more interesting question is whether there exists a class with optimal $o(1/n)$. This rate is essentially sublinear but arbitrarily close to linear and appears in active learning problems (Hanneke, 2012). Perhaps surprisingly, we show that they also appear in the universal regression setting.

3.3. Sublinear (But Arbitrarily Close to Linear) Rates are Achievable

In this section, we provide a concept class with the following property.

Theorem 12 *There exists a hypothesis class $\mathcal{H}$ that is learnable at an optimal rate $o(1/n)$ with respect to the expected absolute loss.*

For the construction we start with a countable domain $\mathcal{X}$ and we assume that all these elements appear on an infinite Littlestone tree whose gap is decreasing exponentially across deeper levels. We can construct a hypothesis class $\mathcal{H}$ by uniquely identifying each hypothesis with a branch of the tree. To be more precise, for every path there exists a hypothesis which perfectly labels this path and returns a default label (e.g., $1/2$) on elements outside of that path.

On the one side, by the construction of the class we know that $\mathcal{H}$ has an infinite 0-Littlestone tree, so, by the previous section, it cannot be learned at a rate faster than $o(1/n)$. It remains to argue about the upper bound. Every realizable distribution $\mathcal{D}$ can be essentially decomposed into two parts: one that is supported on some path, potentially skipping some levels of it, and one outside of the path. We construct the following classifier: given a training set S, let d^* denote the deepest level of the tree for which there exists some $(X^*, Y^*) \in S$ with X^* being on the d^*-th level of the tree and $Y^* \neq 1/2$. Then, given some test point X, if X is an ancestor of X^* the algorithm can predict its label correctly (since we have a tree), with probability 1. Otherwise, the algorithm predicts the default value $1/2$. The idea is that this classifier can only make mistakes on the part of the distribution that is supported on the path and lies deeper than d^*, and after observing n samples this part of the distribution has mass $O(1/n)$. Moreover, as it observes more and more samples, the magnitude of the errors that it makes decreases. It can hence be shown that the designed classifier learns $\mathcal{H}$ with respect to the absolute loss at a rate $o(1/n)$. The details of the construction are handled in Section B.4.

3.4. Infinitely Many Slower than Linear Rates are Achievable

A further indication that the landscape of absolute loss is very rich and complex is the following result, which shows that infinitely many rates (slower than linear) are admissible as optimal rates.

Theorem 13 *For every rate $R(n)$ such that $R(n)$ is non-increasing and $n \cdot R(n)$ is non-decreasing, there exists a hypothesis class $\mathcal{H}$ such that no algorithm can learn $\mathcal{H}$ at a rate faster than $o(R(n))$, and there exists an algorithm that learns $\mathcal{H}$ at a rate $R(n)$.*

The reason why we need this assumption on $nR(n)$ is in order to avoid pathological cases where $R(n)$ is "flat" for a very large interval and then exhibits a large drop. Intuitively, we should not expect to achieve such a learning rate because it would mean that for some $n_0 \in \mathbb{N}$, when we increase the sample size by a few points, the error drops significantly. Instead, our assumption, which we believe is mild, requires that the rate function behaves more smoothly. The construction of this result is provided in Section B.5 and is sketched below.

We consider a countable domain $\mathcal{X}$ and an infinite sequence of blocks of size k_i, $i \in \mathbb{N}$. Each block consists of unique elements of $\mathcal{X}$. Within each block i, we consider gaps of size $\epsilon_i \in [0, 1]$. For the i-th block, consider the 2^{k_i} possible labelings $L_i = \{1/2 - \epsilon_i, 1/2 + \epsilon_i\}^{k_i}$ of its elements. We define the hypothesis class $\mathcal{H}$ in such a way that, for each block i, there is one hypothesis that realizes each element of L_i. The exact choice of $\{k_i\}_{i\in\mathbb{N}}, \{\epsilon_i\}_{i\in\mathbb{N}}$ is related to the target rate $R(n)$, but $\{k_i\}_{i\in\mathbb{N}}$ is always increasing sufficiently fast and $\{\epsilon_i\}_{i\in\mathbb{N}}$ is decreasing sufficiently fast. For the lower bound, we pick the target label for each example x in the i-th block uniformly at random from $\{1/2 - \epsilon_i, 1/2 + \epsilon_i\}$ and we can show that there is a sequence of sample sizes $\{n_i\}_{i\in\mathbb{N}}$ so that whenever the learner observes only n_i points, it will not have observed at least half of the points of some block with index ℓ_i, so for half of the elements of that block it will make a mistake of magnitude ϵ_{ℓ_i}. Because the magnitude of the mistakes is decreasing, we can only show a $o(R(n))$ lower bound instead of a $R(n)$ lower bound. There are several technical details, such as the application of Fatou's lemma to "derandomize" the choice of the target function that can be found in Section B.5.

Let us now shift our attention to the upper bound. For each $n \in \mathbb{N}$, let k_{i_n} be the block size such that $k_{i_n} \leq n \leq k_{i_n+1}$. The idea is that upon observing n samples, there are three types of mistakes the learner can make: (i) mistakes on blocks with index $\ell < i_n$, (ii) mistakes on the block i_n, and, (iii) mistakes on blocks indexed by $\ell > i_n$. The choices of the gap size and the block size guarantee that the dominant term in the loss will be the one coming from mistakes of type (ii). Then, the choice of $\{k_i\}_{i\in\mathbb{N}}$ guarantees that the learner will not observe at most a (k_{i_n}/n)-fraction of the total mass of the i_n-th block. Thus, the total loss will be of the order of $\epsilon_{i_n} \cdot k_{i_n}/n$. This is one aspect of the construction that dictates the choice of the block size and gap size. Similarly, as in the lower bound, there are several technical details that are missing in this discussion, such as the choice of the parameters in a way that balances the three different loss terms, and are handled in Section B.5.

Acknowledgments

We would like to thank the anonymous COLT reviewers for their comments and suggestions. Amin Karbasi acknowledges funding in direct support of this work from NSF (IIS-1845032), ONR (N00014- 19-1-2406), and the AI Institute for Learning-Enabled Optimization at Scale (TILOS). Grigoris Velegkas is supported by TILOS, the Onassis Foundation, and the Bodossaki Foundation. Idan Attias is supported by the Vatat Scholarship from the Israeli Council for Higher Education.

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. Optimal pac bounds without uniform convergence. *arXiv preprint arXiv:2304.09167*, 2023.

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

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of pac learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671. IEEE, 2022.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: Pac learning and online learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Francis Bach. Learning theory from first principles. *Online version*, 2021.

Peter Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett and Philip M Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.

Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310, 1994.

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

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*, pages 532–541, 2021.

Olivier Bousquet, Steve Hanneke, Shay Moran, Jonathan Shafer, and Ilya Tolstikhin. Fine-grained distribution-dependent learning curves. *arXiv preprint arXiv:2208.14615*, 2022.

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.

David Cohn and Gerald Tesauro. Can neural networks do better than the vapnik-chervonenkis bounds? *Advances in Neural Information Processing Systems*, 3, 1990.

David Cohn and Gerald Tesauro. How tight are the vapnik-chervonenkis bounds? *Neural Computation*, 4(2):249–269, 1992.

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. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

Dheeru Dua and Casey Graff. UCI machine learning repository. 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Steve Hanneke. Activized learning: Transforming passive to active with improved label complexity. *The Journal of Machine Learning Research*, 13(1):1469–1587, 2012.

Steve Hanneke. The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1):1319–1333, 2016.

Steve Hanneke, Amin Karbasi, Shay Moran, and Grigoris Velegkas. Universal rates for interactive learning. *Advances in Neural Information Processing Systems*, 35:28657–28669, 2022a.

Steve Hanneke, Amin Karbasi, Shay Moran, and Grigoris Velegkas. Universal rates for interactive learning. In *Advances in Neural Information Processing Systems*, 2022b.

Steve Hanneke, Shay Moran, and Qian Zhang. Universal rates for multiclass learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5615–5681. PMLR, 2023.

David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Alkis Kalavasis, Grigoris Velegkas, and Amin Karbasi. Multiclass learnability beyond the pac framework: Universal rates and partial concept classes. *Advances in Neural Information Processing Systems*, 35:20809–20822, 2022.

Pieter Kleer and Hans Simon. Primal and dual combinatorial dimensions. *Discrete Applied Mathematics*, 327:185–196, 2023.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

Shahar Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002.

Akshay Prasadan and Matey Neykov. Characterizing the minimax rate of nonparametric regression under bounded convex constraints. *arXiv preprint arXiv:2401.07968*, 2024.

Benjamin IP Rubinstein, Peter L Bartlett, and J Hyam Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59, 2009.

Dale Schuurmans. Characterizing rational versus exponential learning curves. *journal of computer and system sciences*, 55(1):140–160, 1997.

Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, 26(3):751–763, 1997.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Appendix A. Universal Rates Landscape for Cut-Off Loss

In this section we prove our results regarding the characterization of the optimal rates under the cut-off loss. We first start by providing a definition of non-trivial classes and showing that such classes cannot be learned at a rate faster than exponential.

Definition 7 (Non-Trivial Class for Cut-Off Loss) *A hypothesis class $\mathcal{H}$ is non-trivial with respect to the γ-cut-off loss if $|\mathcal{H}| \geq 2$ and there exist $x_1, x_2 \in \mathcal{X}$ and $h_1, h_2 \in \mathcal{H}$ such that $h_1(x_1) = h_2(x_1), |h_1(x_2) - h_2(x_2)| > 2\gamma$.*

It is easy to see that if a class $\mathcal{H}$ does not satisfy the non-triviality definition with respect to the γ-cut-off loss, then there is an algorithm that achieves zero loss using just one sample since that sample.

A.1. Exponential Rates for Cut-Off Loss (Lower Bound)

In this section, we show that any non-trivial class cannot be learned at a rate faster than exponential with respect to the cut-off loss.

Proposition 1 (Cut-Off Loss - Exponential Rates - Lower Bound) *Fix $\gamma \in (0,1)$. Assume that $\mathcal{H}$ is non-trivial with respect to the γ-cut-off loss. Then $\mathcal{H}$ cannot be learned at a rate faster than exponential under the expected γ-cut-off loss.*

Proof Let $\mathcal{H}$ be a non-trivial class for the γ-cut-off loss. Hence, there are $h_0, h_1 \in \mathcal{H}$ and $x, x' \in \mathcal{X}$ such that $h_0(x) = h_1(x) = y \in [0,1]$ and $h_0(x') = y_0, h_1(x') = y_1$ with $|y_0 - y_1| > 2\gamma$. We fix some learning algorithm $\widehat{h}_n$ and two distributions $\mathcal{D}_0, \mathcal{D}_1$ where $\mathcal{D}_i\left\{(x,y)\right\} = \frac{1}{2}, \mathcal{D}_i\left\{(x', y_i)\right\} = \frac{1}{2}, i \in \{0,1\}$. We let $I \sim \mathrm{Bernoulli}(1/2)$ and given I, we let $(X_1, Y_1), (X_2, Y_2), \ldots$ be i.i.d. samples from $\mathcal{D}_I$. In particular, $(X_1, Y_1), \ldots, (X_n, Y_n)$ are the training samples for $\widehat{h}_n$ and (X_{n+1}, Y_{n+1}) is the test point for the learner. Then, we have that

$$\mathbf{E}\left[\mathbf{Pr}\left[\,|\widehat{h}_n(X_{n+1}) - Y_{n+1}| > \gamma \mid \{X_t, Y_y\}_{t=1}^n, I\,\right]\right] \geq \frac{1}{2}\,\mathbf{Pr}(X_1 = \ldots = X_n = x, X_{n+1} = x') = 2^{-n-2}\,.$$

We also have that

$$\begin{aligned}&\mathbf{E}\left[\mathbf{Pr}\left[\,|\widehat{h}_n(X_{n+1}) - Y_{n+1}| > \gamma \mid \{X_t, Y_y\}_{t=1}^n, I\,\right]\right] \\ &= \frac{1}{2}\sum_{i \in \{0,1\}} \mathbf{E}\left[\mathbf{Pr}\left[|\widehat{h}_n(X_{n+1}) - Y_{n+1}| > \gamma \mid \{X_t, Y_y\}_{t=1}^n, I = i\right] | I = i\right]\,.\end{aligned}$$

Thus, for every n, there exists some $i_n \in \{0,1\}$ such that for $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. from P_{i_n} it holds that

$$\mathbf{E}[\mathrm{er}^{\gamma}_{\mathcal{D}_{i_n}}(\widehat{h}_n)] = \Omega(2^{-n-2})\,,$$

where $\mathrm{er}^{\gamma}_{\mathcal{D}}(h)$ is the expected γ-cut-off loss under $\mathcal{D}$. Hence, there exists some fixed $i \in \{0,1\}$ such that $\mathbf{E}[\mathrm{er}^{\gamma}_{\mathcal{D}_i}(\widehat{h}_n)] = \Omega(2^{-n-2})$ for infinitely many n. ■

A.2. Exponential Rates for Cut-Off Loss (Upper Bound)

We now move on to proving that if a class does not contain a γ-Littlestone tree then it is learnable at exponential rates with respect to the cut-off loss.

Theorem 14 (Cut-Off Loss - Exponential Rates - Upper Bound) *Fix $\gamma \in (0,1)$. Assume that $\mathcal{H}$ does not admit an infinite γ-Littlestone tree. Then for any $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$, there exist constants C_γ, c_γ:*

$$L^{\gamma}_{\mathcal{D}}(\widehat{h}_n) \leq C_\gamma e^{-c_\gamma n}\,,$$

where $\widehat{h}_n$ is the output of Algorithm 1.

Algorithm 1 Exponential Rates Algorithm for Universal Regression with Cut-Off Loss

```
Exponential Rates (Input $\gamma$)
Let $g_t$ be an eventually $\gamma$-correct regressor.
Let $(X_1, Y_1, \ldots, X_n, Y_n)$ be the training set.
Estimate $\hat{t}_n$ such that $\mathbf{Pr}[\mathrm{er}(g_{\hat{t}_n})] \leq 3/8$.
Break the training set into $N = n/\widehat{t}_n$ batches.
Create $N$ copies of $g$: $g^1, ..., g^N$ where the $i$-th copy is trained on the $i$-th batch.
To predict the label of some $x \in \mathcal{X}$, take the median over all $g^i_{\hat{t}_n}$.
```

```
Exponential GS Game
For any $t \in \mathbb{N}$ :
  $\mathtt{P}_A$ picks $\kappa_t = (\xi_t, y_t^{(0)}, y_t^{(1)}) \in \mathcal{X} \times [0,1] \times [0,1]$.
  $\mathtt{P}_A$ reveals $\kappa_t$ to the learner $\mathtt{P}_L$.
  $\mathtt{P}_L$ chooses $\eta_t \in \{0,1\}$.
$\mathtt{P}_L$ wins the game if for some $t \in \mathbb{N}$
$\Big\{h \in \mathcal{H} : |h(\xi_\ell) - y_\ell^{(\eta_\ell)}| \leq \gamma \;\forall \ell \in [1..t]\Big\} = \emptyset\,.$
```

Our goal is to design an algorithm that achieves the exponential universal rate if $\mathcal{H}$ does not have an infinite γ-Littlestone tree. In short, our approach builds on the framework initiated by Bousquet et al. (2021). We consider an adversarial online learning game $\mathcal{G}$ played in rounds between an adversary P_A and a learner P_L, defined in Figure 1 and Figure 2. Our main result is that $\mathcal{H}$ that does not have an infinite scaled Littlestone tree if and only if there exists a universally measurable strategy for the learning player P_L in the game that only makes finitely many mistakes in terms of the γ-cut-off loss against any adversary P_A (see Appendix A.2.1). This result can then be employed in the probabilistic setting (see Appendix A.2.2) and yield a learning algorithm that achieves the exponential universal rate for the regression task with respect to the cut-off loss.

Let us sketch how the structure of a γ-Littlestone tree allows us to obtain an algorithm. To this end, we have to shortly introduce the two-player game between P_A and P_L. In each round $t \geq 1$, the following interaction takes place:

- P_A picks a three-tuple $\xi_t = (x_t, y_t^0, y_t^1) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ and reveals it to P_L.
- P_L picks $\eta_t \in \{0,1\}$.

The learner P_L wins the game if for some $t \in \mathbb{N}$, the set $\Big\{h \in \mathcal{H} : |h(\xi_\ell) - y_\ell^{(\eta_\ell)}| \leq \gamma,\ \forall \ell \in [1..t]\Big\} = \emptyset$. Hence, one can show that a winning strategy for the adversary P_A is equivalent to the existence of an infinite γ-Littlestone tree. This implies that, since P_A and P_L are playing a Gale-Stewart game (Bousquet et al., 2021), the non-existence of such a tree, gives a winning strategy g_t for the learning player, which can then be used to obtain an algorithm with exponential rates. We proceed with the formal proof.

A.2.1. ADVERSARIAL SETTING VIA GALE-STEWART GAMES

In order to design our algorithms, we consider the following setting. We introduce the following online learning game (Figure 1). In this game, there are two players, the adversary who chooses features and reveals them to the second player, the learner whose goal is to guess a real-valued label for the given example. **The learner makes a mistake in round t whenever the guess $\widehat{y}_t$ differs from the true label y_t by at least** γ. The goal of the learner is to minimize her loss and the adversary's intention is to provoke many errors to the learner.

On each round $t \geq 1$:

1. The adversary picks a point $x_t \in \mathcal{X}$.
2. The learner guesses a value $\widehat{y}_t \in [0,1]$.
3. The adversary chooses the value y_t as true label so that $y_t = h(x_t)$ for some $h \in \mathcal{H}$ that is consistent with the previous examples (x_p, y_p) for any $p \leq t$.

Figure 1: Realizable Online Setting

We say that the concept class $\mathcal{H}$ is γ-online learnable if there exists a strategy $\widehat{y}_t = \widehat{y}_t(x_1, y_1, ..., x_{t-1}, y_{t-1}, x_t)$ that makes a mistake only *finitely* many times, regardless of what realizable sequence is presented by the adversary. The main result in this setting is the following.

Theorem 15 (Strategies in the Adversarial Setting) *Fix $\gamma \in (0,1)$. We say that a learner (that predicts x) makes a mistake against an adversary (that reveals y) if $|x - y| > \gamma$. For any concept class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, the following dichotomy occurs.*

1. *If $\mathcal{H}$ does not have an infinite γ-Littlestone tree, then there is a strategy for the learner that makes only finitely many mistakes against any adversary.*
2. *If $\mathcal{H}$ has an infinite γ-Littlestone tree, then there is a strategy for the adversary that forces any learner to make a mistake in every round.*

Proof [Proof of Theorem 15] Let us fix γ for the proof. We first introduce a two-player game $\mathcal{G}$ that is played in discrete timesteps $t = 1, 2, \ldots$ between the adversary and the learner.

1. The adversary picks a point $\kappa_t = \left(\xi_t, y_t^{(0)}, y_t^{(1)}\right) \in \mathcal{X} \times [0,1] \times [0,1]$ and reveals it to the learner.
2. The learner chooses a point $\eta_t \in \{0,1\}$.

Figure 2: Adversarial Setting - 2-Player Game

The learning player wins in some finite round t if $\mathcal{H}_{\xi_1, y_1^{(\eta_1)}, \ldots, \xi_t, y_t^{(\eta_t)}} = \emptyset$. The adversary wins if the game continues indefinitely (i.e., the class of consistent hypotheses from $\mathcal{H}$ never gets empty) .

Clearly, the set of winning strategies for the learning player is

$$\mathcal{W} = \{(\boldsymbol{\kappa}, \boldsymbol{\eta}) \in (\mathcal{X}\times[0,1]\times[0,1]\times\{0,1\})^\infty : \exists\, 1 \leq t^\star < \infty \text{ such that } \mathcal{H}_{\xi_1, y_1^{(\eta_1)}, \ldots, \xi_{t^\star}, y_{t^\star}^{(\eta_{t^\star})}} = \emptyset\}\,.$$

We now recall an important theorem about Gale-Stewart games:

Proposition 2 *In any Gale-Stewart game, either P_A or P_L has a winning strategy.*

Equipped with Proposition 2, we can show that the adversary has a winning strategy if and only if $\mathcal{H}$ has an infinite γ-scaled Littlestone tree (provided that $\mathcal{G}$ is a Gale-Stewart game). This is summarized in the next claim.

Claim 1 *The game $\mathcal{G}$ is a Gale-Stewart game and the adversary has a winning strategy in $\mathcal{G}$ if and only if the hypothesis class $\mathcal{H}$ has an infinite γ-Littlestone tree.*

Proof It is clear from the definition of $\mathcal{W}$ that every winning strategy of the learner is finitely decidable, hence $\mathcal{G}$ is a Gale-Stewart game. For the other part of the claim, notice that if $\mathcal{H}$ has an infinite γ-Littlestone tree, then the adversary's strategy is to present the learner at step t the point of the tree at depth t that is consistent with the execution of the game so far along with the labels of the edges that connect it with its children. By the definition of the tree, this strategy ensures that the game will keep going on forever. For the other direction, assume that the adversary has a winning strategy $\kappa_\tau(\eta_1, \ldots, \eta_{\tau-1}) \in \mathcal{X} \times [0,1] \times [0,1]$. Then, define the γ-Littlestone tree $\mathcal{T} = \{x_{\boldsymbol{u}} : 0 \leq k < \infty, \boldsymbol{u} \in \{0,1\}^k\}$ where $x_{\eta_1,\ldots,\eta_{\tau-1}} = \xi_\tau(\eta_1, \ldots, \eta_{\tau-1})$ where the labels that connect $x_{\eta_1,\ldots,\eta_{\tau-1}}$ with its left, right children are $y_\tau^{(0)}(\eta_1, \ldots, \eta_{\tau-1}), y_\tau^{(1)}(\eta_1, \ldots, \eta_{\tau-1})$, respectively. We can see that $\mathcal{T}$ is infinite since this is a winning strategy for the adversary. ■

Having shown the above statement, we are ready to establish the desired dichotomy in the online game. Assume first that $\mathcal{H}$ has an infinite γ-Littlestone tree $\{x_u\}$. The adversary's strategy is defined inductively based on the path followed so far in the game: in round t, set $\boldsymbol{b}_t = (b_1, ..., b_{t-1}) \in \{0,1\}^{t-1}$ denote the path parsed so far in the tree by the two players. Then, the adversary picks $x_t = x_{\boldsymbol{b}_t}$. After the learner reveals her choice $\widehat{y}_t$, the worst case adversary chooses as a response the branch of the γ-Littlestone tree which will incur loss at least γ given the learner's choice (the adversary may even have two choices). By the definition of the tree, this chosen label is valid since there exists some $h \in \mathcal{H}$ that realizes the path $(x_{b_1}, ..., x_{b_{t-1}}, x_t)$. Moreover, this choice provokes a mistake (in the sense of γ gap) to the learning player and this is true for any round. Hence, there is a strategy for the adversary that forces any learner to make a mistake in every round.

For the other direction, assume that the class $\mathcal{H}$ does not have an infinite γ-Littlestone tree. Before we describe the winning strategy of the learner, we need to introduce the notion of ordinal γ-Littlestone dimension. We will assign an ordinal to every finite γ-Littlestone tree. For some preliminaries on ordinals and transfinite recursion, we refer to Bousquet et al. (2021). The rank is defined by a partial order $\prec$. We set $t' \prec t$ if t' is a γ-Littlestone tree that extends t by one level, i.e., t is obtained from t' by removing its leaves. A γ-Littlestone tree t is minimal if it cannot be extended to a γ-Littlestone tree of larger depth. For such a tree, we set $\mathrm{rank}(t) = 0$. If the tree t is non-minimal, then it can be extended and this is quantified using transfinite recursion by

$$\mathrm{rank}(t) = \sup\{\mathrm{rank}(t') + 1 : t' \prec t\}\,.$$

The rank is well-defined as long as $\mathcal{H}$ has no infinite γ-Littlestone tree (since $\prec$ is well-founded). In particular, we define

$$\overline{\mathrm{Ldim}}_\gamma(\mathcal{H}) = \begin{cases} -1 & \text{if } \mathcal{H} \text{ is empty}\,, \\ \Omega & \text{if } \mathcal{H} \text{ has an infinite } \gamma\text{-Littlestone tree}\,, \\ \mathrm{rank}(\emptyset) & \text{otherwise}\,. \end{cases}$$

The strategy is chosen so that $\overline{\mathrm{Ldim}}_\gamma(\mathcal{H}_{x_1,y_1,\ldots,x_t,y_t})$ decreases in every round and the learner that follows this strategy will win the game, since the ordinals do not admit an infinite decreasing chain. We note that this statement at first is purely existential via the theory of Gale-Stewart games. We next shortly provide a "constructive" way to compute the winning strategy of the learning player in the set of games we consider. Let us describe the winning strategy: the learner invokes the scaled version of the (ordinal) Standard Optimal Algorithm and chooses the label y_t (given x_t) that maximizes the ordinal γ-Littlestone dimension, i.e., $y_t = \mathrm{argmax}_{y \in [k]} \overline{\mathrm{Ldim}}_\gamma(V_t^y)$, where = $V_t^y = \{h \in \mathcal{H}_{x_1,y_1,\ldots,x_{t-1},y_{t-1}} : h(x_t) = y\}$. The ordinal SOA at round $t = 1, 2, \ldots$ with initial set $V_0 = \mathcal{H}$ works as follows:

1. Receive x_t.
2. For any $y \in [k]$, let $V_t^y = \{h \in V_{t-1} : h(x_t) = y\}$.
3. Predict $\widehat{y}_t \in \mathrm{argmax}_{y \in [0,1]} \overline{\mathrm{Ldim}}_\gamma(V_t^y)$, where $\overline{\mathrm{Ldim}}_\gamma$ is the ordinal γ-Littlestone dimension.
4. Receive true answer y_t and set $V_t = V_t^{y_t}$.

This algorithm drives the game in a win-win phenomenon for the learner in every round: if the adversary forces the learner to a mistake, then she will "prune" the tree and set the learner closer to winning the game. Otherwise, the learner will be correct and will not incur any loss.

In order to show that the scaled ordinal SOA makes a finite number of mistakes, we couple the online game with a Gale-Stewart game. The idea is that every time the learner makes a mistake in the online game on point x_t, we advance the Gale-Stewart game by one round where we pretend that $\xi_\tau = x_t, y_\tau^{(0)} = \widehat{y}_t, y_\tau^{(1)} = y_t, \eta_\tau = y_t$. Notice that if the learner makes an infinite number of mistakes in the online game using the ordinal SOA, then the Gale-Stewart game can proceed infinitely. Hence, to conclude the proof, we need to show that in this coupled game, there is some finite point τ^* such that $\mathcal{H}_{\xi_1, y_1^{(\eta_1)}, \ldots, \xi_{\tau^\star}, y_{\tau^\star}^{(\eta_{\tau^\star})}} = \emptyset$. The following result helps us establish that. In fact, the next lemma follows from Bousquet et al. (2021)(Proposition B.8) by choosing the value of the game being the ordinal γ-Littlestone dimension.

Lemma 2 (See Proposition B.8 of Bousquet et al. (2021)) *Assume that $\mathcal{H}$ does not contain an infinite γLittlestone tree. Then, for any choices of the adversary $\kappa_1, \ldots, \kappa_{t-1}$ up to round t and for any choice $\kappa_t = (\xi_t, y_t^{(0)}, y_t^{(1)})$ in round t there is a choice η_t of the learner such that*

$$\overline{\mathrm{Ldim}}_\gamma\left(\mathcal{H}_{\xi_1, y_1^{(\eta_1)}, \ldots, \xi_{t-1}, y_{t-1}^{(\eta_{t-1})}, \xi_t, y_t^{(\eta_t)}}\right) < \overline{\mathrm{Ldim}}_\gamma\left(\mathcal{H}_{\xi_1, y_1^{(\eta_1)}, \ldots, \xi_{t-1}, y_{t-1}^{(\eta_{t-1})}}\right)\,.$$

The previous result shows that for every ξ_t there is at most one label $\ell_t \in [0,1]$ such that

$$\overline{\mathrm{Ldim}}_\gamma\left(\mathcal{H}_{\xi_1, y_1^{(\eta_1)}, \ldots, \xi_{t-1}, y_{t-1}^{(\eta_{t-1})}, \xi_t, \ell_t}\right) = \overline{\mathrm{Ldim}}_\gamma\left(\mathcal{H}_{\xi_1, y_1^{(\eta_1)}, \ldots, \xi_{t-1}, y_{t-1}^{(\eta_{t-1})}}\right)\,.$$

Indeed, assume that there are two such labels ℓ_t, ℓ_t' for some ξ_t. Then, if the adversary proposes the point (ξ_t, ℓ_t, ℓ_t'), there is no choice η_t of the learner that decreases that ordinal Littlestone dimension in this round, which leads to a contradiction. Hence, the learner can pick any label as long as it is not the one that maximizes the ordinal Littlestone dimension. This is exactly how the coupled Gale-Stewart game proceeds, so we know that every time the learner makes a mistake in the online game the ordinal Littlestone dimension of the coupled game decreases. Since ordinals that are less than Ω do not admit infinitely decreasing chains, we get the desired result.

■

The measurability of the winning strategies and of the learning algorithm developed in the previous section constitutes an important detail, extensively discussed in Bousquet et al. (2021), in order to move from the adversarial setting to the probabilistic one. We provide the next useful result.

Lemma 3 *Let $\mathcal{X}$ be Polish and $\mathcal{H} \subseteq ([0,1] \cap \mathbb{Q})^{\mathcal{X}}$ be measurable. Then, the Gale-Stewart game $\mathcal{G}$ of Figure 2 has a universally measurable winning strategy.*

Crucially the above result states that the winning strategy η_t of the learning player is measurable. However, the previous proof made use of the scaled ordinal SOA algorithm, whose measurability is not directly implied. To this end, we modify the adversarial algorithm to handle the measurability issue. The modification follows:

```
1. $\mathcal{Y} \triangleq \mathbb{Q} \cap [0,1]$.
2. Initialize $\tau \leftarrow 1, G = \texttt{Clique}(V = \mathcal{Y}), f(\cdot,\cdot,\cdot) \leftarrow \eta_1(\cdot,\cdot,\cdot)$ $\triangleright$ $\tau$ is the mistake counter
3. For every round $t \geq 1$ :
   (a) Observe $x_t$
   (b) For any $y \neq y'$ with $y, y' \in \mathcal{Y}$, orient the edge $(y, y')$ of $G$ according to $f(x_t, y, y')$
   (c) Let $G'$ the directed clique
   (d) Predict $\widehat{y}_t \leftarrow \mathrm{argmax}_{y \in \mathcal{Y}} \mathrm{outdeg}(y; G')$
   (e) If $|\widehat{y}_t - y_t| > \gamma$, let $\xi_\tau \leftarrow x_t, f(\cdot,\cdot,\cdot) \leftarrow \eta_{\tau+1}(x_1, y_1, \ldots, x_\tau, y_\tau, \cdot,\cdot,\cdot), \tau \leftarrow \tau + 1$
```

Figure 3: Measurable Modification of Online Learning Algorithm for Exponential Rates

The above algorithm makes use of a tournament procedure. The algorithm is a measurable function since (i) the winning strategy of the learner is measurable and (ii) the countable maximum of measurable functions is measurable. This algorithm can be used in order to show that if $\mathcal{H}$ does not have an infinite γ-Littlestone tree, then the above algorithm makes only a finite number of mistakes (in the sense of γ gaps) against any adversary. Essentially, this is due to the fact that when the winning strategy has converged to a zero-mistake prediction rule (which occurs after a finite number of mistakes), the tournament procedure will always output the correct label for the observed example. Hence, the algorithm will eventually make a finite number of mistakes in the adversarial setting.

A.2.2. FROM ADVERSARIAL TO PROBABILISTIC LEARNING

The algorithm of Figure 3 works in the adversarial setting. We first show that it also applies to the probabilistic setting (and this is why we require the above measurability discussion).

Lemma 4 (From Adversarial to Probabilistic) *Fix $\gamma \in (0,1)$. For any distribution $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$ and for the learning algorithm $\widehat{y}_t : \mathcal{X} \to [0,1]$ of Theorem 15, we have*

$$\mathop{\mathbf{Pr}}_{S_t}\left[\mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}}\left[|\widehat{y}_t(x) - y| > \gamma\right] > 0\right] \to 0 \; as \;\; t \to \infty\,,$$

where S_t is the training set $(x_1, y_1, ..., x_{t-1}, y_{t-1})$ of the algorithm.

Proof Since the distribution $\mathcal{D}$ is realizable, there exists a sequence of functions $h_k \in \mathcal{H}$ so that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}}[|h_k(x) - y| > \gamma] < \frac{1}{2^k}\,.$$

Let us fix $t \geq 1$. We have that

$$\sum_{k=1}^{\infty} \mathbf{Pr}[\exists s \leq t : |h_k(X_s) - Y_s| > \gamma] \leq t \sum_{k=1}^{\infty} \mathop{\mathbf{Pr}}_{(X,Y)\sim\mathcal{D}}[|h_k(X) - Y| > \gamma] < \infty\,,$$

where the first inequality is due to union bound. By Borel-Cantelli, with probability one, there exists for every $t \geq 1$ a hypothesis $h \in \mathcal{H}$ so that $h(X_s) = Y_s$ for all $s \leq t$. Hence, the sequence $X_1, Y_1, X_2, Y_2, ...$ is a valid input for the online learning game with probability one. In particular, we make use of the following statement: If $\mathcal{H}$ does not have an infinite γ-Littlestone tree, then there is a strategy for the learner that makes only finitely many mistakes against any adversary. This is proved in Theorem 15. The existence of a winning strategy $\widehat{y}_t$ for the learning player implies that the time T where the player makes a mistake is

$$T = \sup\{s \in \mathbb{N} : |\widehat{y}_{s-1}(X_s) - Y_s| > \gamma\}$$

is a random variable that is finite with probability one. Moreover, the online learner is selected so that it is changed only when a loss is observed. This means that $\widehat{y}_s = \widehat{y}_t$ for all rounds $s \geq t \geq T$.

We now employ the law of large numbers in order to understand the asymptotic behavior of the online learner:

$$\mathbf{Pr}\left[\mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}}[|\widehat{y}_t(x) - y| > \gamma] = 0\right] = \mathbf{Pr}\left[\lim_{S\to\infty} \frac{1}{S} \sum_{s=t+1}^{t+S} 1\{|\widehat{y}_t(X_s) - Y_s| > \gamma\} = 0\right]$$

and this probability is at least the probability of this event and of the event that $T \leq t$, i.e.,

$$\mathbf{Pr}\left[\mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}}[|\widehat{y}_t(x) - y| > \gamma] = 0\right] \geq \mathbf{Pr}\left[\lim_{S\to\infty} \frac{1}{S} \sum_{s=t+1}^{t+S} 1\{|\widehat{y}_t(X_s) - Y_s| > \gamma\} = 0, T \leq t\right] = \mathbf{Pr}[T \leq t]\,,$$

where the last inequality follows from the observation that since $s \geq t$ and t is greater than the critical time T then the first event occurs with probability one. This implies that

$$\mathbf{Pr}\left[\mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}}[|\widehat{y}_t(x) - y| > \gamma] = 0\right] \geq \mathbf{Pr}[T \leq t]$$

and so

$$\mathbf{Pr}\left[\underset{(x,y)\sim\mathcal{D}}{\mathbf{Pr}}[|\widehat{y}_t(x)-y|>\gamma]>0\right]\leq\lim_{t\to\infty}1-\mathbf{Pr}[T\leq t]=0\,.$$

■

A.2.3. Concluding the proof

The above result guarantees that the expected γ cut-off error of the learning algorithm tends to zero as t goes to infinity, i.e., we have established that that $\mathbf{E}[\mathrm{er}_\gamma(\widehat{y}_t)]\to 0$ as $t\to\infty$, where $\mathrm{er}_\gamma(\widehat{y}_t)=\mathbf{Pr}_{(x,y)\sim\mathcal{D}}[|\widehat{y}_t(x)-y|>\gamma]$ for $\mathcal{D}\in\mathrm{RE}(\mathcal{H})$. This means that the scaled ordinal SOA is a consistent algorithm in the statistical setting. However, this fact is not enough to establish the exponential convergence rate.

We can easily show that a modification of this algorithm actually leads to exponential rates. The modification is standard and appears in all the previous papers in the universal rates literature (Bousquet et al., 2021; Kalavasis et al., 2022; Hanneke et al., 2023). In particular, we can apply Lemma 4 together with simple adaptations of Lemma 4.4 and Corollary 4.5 of Bousquet et al. (2021) in order to obtain a learning algorithm that achieves exponential rate. In more detail, the algorithm splits the data into two parts: the first is used to obtain an estimator $\widehat{t_n}$ of $t^\star$, which is defined to be a critical time such that if we run the game for t^* many rounds, then we will obtain a function $\widehat{y}_{t^*}$ that does not make γ-mistakes, with at least some constant probability, i.e., $\mathbf{Pr}[\mathrm{er}_\gamma(\widehat{y}_{t^*})>0]<1/4$. Standard arguments (Bousquet et al., 2021; Kalavasis et al., 2022; Hanneke et al., 2023) show that this estimator will be accurate with probability at least $1-e^{-\Omega(n)}$. The second part of the dataset is used as follows: we create roughly $n/\widehat{t_n}$ different batches and compute the classifier $\widehat{y}^i_{\widehat{t_n}}$ separately for each batch $i\in[n/\widehat{t_n}]$. Again, applying the same ideas as in Bousquet et al. (2021); Kalavasis et al. (2022); Hanneke et al. (2023) we can show that with probability at least $1-e^{-\Omega(n)}$ the majority of $\widehat{y}_{\widehat{t_n}}$ will not be making any γ-mistakes on any points. Finally, we choose our guess $\widehat{h}_n$ to be the median among these classifiers.

A.3. Linear Rates for Cut-Off Loss (Lower Bound)

In this section we show that for the cut-off loss, a rate that is slower than exponential cannot be faster than linear. We prove that when the class has an infinite 2γ-Littlestone tree then the fastest rate one can hope for is linear.

Theorem 16 (Cut-Off Loss - Linear Rates - Lower Bound) *Fix $\gamma\in(0,1)$. Assume that $\mathcal{H}$ admits an infinite 2γ-Littlestone tree. Then there exists $\mathcal{D}\in\mathrm{RE}(\mathcal{H})$ such that there exists a constant C_γ:*

$$L^\gamma_{\mathcal{D}}(\widehat{h}_n)\geq\frac{C_\gamma}{n}\,,$$

for infinitely many $n\in\mathbb{N}$.

Proof Let us first fix $\gamma\in(0,1)$. Fix any learning algorithm $\widehat{h}_n$ and an infinite 2γ-Littlestone tree for $\mathcal{H}$. Let $\boldsymbol{y}=(y_1,y_2,...)\in\{0,1\}^\infty$ be a random branch of this tree, where the sequence $\{y_i\}_{i\in[n]}$ is an i.i.d. sequence of unbiased Bernoulli coins. We introduce the random distribution over $\mathcal{X}\times\mathcal{Y}$ as

$$P_{\boldsymbol{y}}((x_{\boldsymbol{y}_{\leq\ell}},z_{\ell+1}))=\frac{1}{2^{\ell+1}},\ \ell\geq 0\,,$$

where $z_{\ell+1} \in \mathcal{Y}$ is the label of the edge connecting $x_{\boldsymbol{y}_{\leq \ell}}$ to its child according to the chosen path $\boldsymbol{y}$. For any $n < \infty$, there exists a hypothesis $h \in \mathcal{H}$ so that

$$h(x_{\boldsymbol{y}_{\leq \ell}}) = z_{\ell+1}$$

for all $0 \leq \ell \leq n$. This is due to the construction of the scaled Littlestone tree. We have that

$$L^{\gamma}_{\boldsymbol{y}}(h) = \underset{(x,z)\sim\mathcal{D}_{\boldsymbol{y}}}{\mathbf{Pr}}[|h(x) - z| > \gamma] \leq \sum_{\ell > n} 2^{-\ell-1},$$

which goes to 0 as $n \to \infty$. This implies that $\mathcal{D}_{\boldsymbol{y}}$ is realizable for every infinite branch $\boldsymbol{y} \in \{0,1\}^{\infty}$. Moreover, the mapping $y \to \mathcal{D}_y$ is measurable. Let us draw $(X, Z), (X_1, Z_1), (X_2, Z_2), ...$ i.i.d. samples from $\mathcal{D}_{\boldsymbol{y}}$. The first sample corresponds to the test sample and the other samples are associated with the training phase. Moreover, let $T, T_1, T_2, ...$ be i.i.d. Geometric random variables with success probability $1/2$ starting at 0. We can set

1. $X = x_{\boldsymbol{y}\leq T}, Z = z_{T+1}$ and
2. $X_i = x_{\boldsymbol{y}\leq T_i}, Z_i = z_{T_i+1}$.

Crucially, on the event that $\{T = \ell, \max\{T_1, ..., T_n\} < \ell\}$, the value of $\widehat{h}_n(X)$ is conditionally independent of $z_{\ell+1}$ given $X, (X_1, Z_1), ..., (X_n, Z_n)$. We next have that

$$\mathbf{Pr}[|\widehat{h}_n(X) - Z| > \gamma, T = \ell, \max\{T_1, ..., T_n\} < \ell] = \mathbf{Pr}[|\widehat{h}_n(X) - Z_{\ell+1}| > \gamma, T = \ell, \max\{T_1, ..., T_n\} < \ell].$$

This is equal to

$$\mathbf{E}[\underset{Z}{\mathbf{Pr}}[|\widehat{h}_n(X) - Z| > \gamma | X, (X_1, Z_1), ..., (X_n, Z_n)]\mathbf{1}\{T = \ell, \max\{T_1, ..., T_n\} < \ell\}]$$

Now conditional on this event, any algorithm will incur a loss of at least γ with probability $1/2$, since the realization of the true label is independent of the guess of the algorithm. Hence, this quantity is lower bounded by

$$\frac{1}{2}\mathbf{Pr}[T = \ell, \max\{T_1, ..., T_n\} < \ell] = 2^{-\ell-2}(1 - 2^{-\ell})^n.$$

We are free now to pick ℓ. Choosing $\ell = \ell_n := \lceil 1 + \log(n) \rceil$, we have that $1/2^{\ell} > 1/(4n)$ and $(1 - 2^{-\ell})^n \geq 1/2$. Our goal is to apply the reverse Fatou lemma. This can be done since almost surely, we have that

$$n\,\mathbf{Pr}[|\widehat{h}_n(X) - Z| > \gamma, T = \ell_n | \boldsymbol{y}] \leq n\,\mathbf{Pr}[T = \ell_n] = n2^{-\ell_n - 1} \leq 1/4.$$

Hence, we can apply the reverse Fatou lemma and get

$$\mathbf{E}\left[\limsup_{n\to\infty} n\,\mathbf{Pr}[|\widehat{h}_n(X) - Z| > \gamma, T = \ell_n | \boldsymbol{y}]\right] \geq \limsup_{n\to\infty} n\,\mathbf{Pr}[|\widehat{h}_n(X) - Z| > \gamma, T = \ell_n] > 1/32.$$

But, almost surely, it holds that

$$\mathbf{E}[L^{\gamma}_{\boldsymbol{y}}(\widehat{h}_n)|\boldsymbol{y}] = \mathbf{Pr}[|\widehat{h}_n(X) - Z| > \gamma | \boldsymbol{y}] \geq \mathbf{Pr}[|\widehat{h}_n(X) - Z| > \gamma, T = \ell_n | \boldsymbol{y}].$$

So, combining the above inequalities

$$\mathbf{E}\left[\limsup_{n\to\infty} n\,\mathbf{E}[L^{\gamma}_{\boldsymbol{y}}(\widehat{h}_n)]\right] > \Omega(1).$$

Hence, there must exist a realization of $\boldsymbol{y}$ so that $\mathbf{E}[L^{\gamma}_{\boldsymbol{y}}(\widehat{h}_n)] = \Omega(1/n)$ infinitely often. Choosing $\mathcal{D} = \mathcal{D}_{\boldsymbol{y}}$ completes the proof. ■

A.4. Near Linear Rates for Cut-Off Loss (Upper Bound)

We then move on to the setting where the class has an infinite 2γ-Littlestone tree but not an infinite γ-OIG-Littlestone tree. The main result we will show in this setting is the following.

Theorem 17 (Cut-Off Loss - Near Linear Rates - Upper Bound) *Fix $\gamma \in (0,1)$. Assume that $\mathcal{H}$ does not admit an infinite γ-One-Inclusion Graph Littlestone tree. Then for any $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$ it holds that:*

$$L_{\mathcal{D}}^{\gamma}(\widehat{h}_n) \leq \widetilde{O}\left(\frac{1}{n}\right), \forall n \in \mathbb{N},$$

where this notation suppresses constants that depend on $\mathcal{D}, \gamma$.

Our approach to show this result consists of several steps and follows the template introduced in Bousquet et al. (2021) that has also been extensively used in other works related to universal learning rates. However, there are several non-trivial issues we need to overcome which we will comment on.

An important technical result we need for our proof is an extension of the scaled OIG algorithm of Attias et al. (2023) to the partial concepts setting (see Section D). We believe that this could be of independent interest.

Now, we proceed with the proof of the nearly-linear rate upper bound. To this end, we first give an algorithm for the adversarial setting and then study the probabilistic setting (as we did in the exponential case).

A.4.1. ADVERSARIAL SETTING VIA GALE-STEWART GAMES AND PATTERN AVOIDANCE

The first step in our approach is to consider a *Gale-Stewart* game between a learner P_L and an adversary P_A, in which the learner has a winning strategy if and only if $\mathcal{H}$ does not admit an infinite γ-OIG-Littlestone tree. The game in every round $\tau \in \mathbb{N}$ is defined as follows:

- Player P_A chooses a sequence $x_\tau = (x_\tau^0, \ldots, t_\tau^{\tau-1}) \in \mathcal{X}^\tau$ and a finite set of labelings $L_\tau \in \mathrm{LG}_{\gamma,\tau}$.
- Player P_L chooses an element $y_\tau \in L_\tau$.

The winning condition of the game for player P_L is the following:

- Player P_L wins if there exists some $\tau \in \mathbb{N}$ such that $\mathcal{H}_{x_1,y_1,\ldots,x_\tau,y_\tau} = \emptyset$, where

$$\mathcal{H}_{x_1,y_1,\ldots,x_\tau,y_\tau} := \left\{h \in \mathcal{H} : h(x_s^i) = y_s^i, \forall\, 0 \leq i \leq s, 1 \leq s \leq \tau\right\}.$$

As usual in Gale-Stewart games, player P_A wins the game if the game continues indefinitely. We first show that the winning condition of this game for the learning player P_L is tightly captured by the finiteness of the OIG-Littlestone tree of the underlying concept class.

Lemma 5 *The class $\mathcal{H}$ does not have an infinite γ-OIG-Littlestone tree if and only if P_L has a universally measurable winning strategy in the OIG-Littlestone game.*

Proof It is clear from the definition of the game that every winning strategy of the learner P_L is finitely decidable, hence the OIG-Littlestone game is a Gale-Stewart game. Thus, either P_A or P_L has a winning strategy for the scaled OIGL game.

Notice that if $\mathcal{H}$ has an infinite γ-OIG-Littlestone tree, then the adversary's strategy is to present the learner at step τ the point of the tree at depth τ that is consistent with the execution of the game so far along with the labels of the edges that connect it with its children. By the definition of the tree, this strategy ensures that the game will keep going on forever and so P_A has a winning strategy.

For the other direction, let us now consider the case where the adversary has a winning strategy in the scaled OIGL game. Denote the τ-th decision of the adversary as $\kappa_\tau(\eta_1, ..., \eta_{\tau-1})$, where η_i are the decisions of P_A and P_L in round $i < \tau$. Then, P_A can construct an infinite scaled OIGL tree by setting $(x_\tau, L_\tau) = \kappa_\tau(\eta_1, ..., \eta_{\tau-1})$. Note that the class $\mathcal{H}_{\kappa_1,\eta_1,...,\kappa_\tau,\eta_\tau}$ is non-empty for any round τ and this implies that there exists an infinite scaled OIGL tree for $\mathcal{H}$.

Finally, since the game is GS, we have that the learning player has a winning strategy in the game if the class $\mathcal{H}$ has no infinite scaled OIGL tree. The measurability of the learner's winning strategy follows from the analysis of Hanneke et al. (2023) (Proposition 50). ■

Winning Strategy for $P_L \implies$ Pattern Avoidance. From now on we focus on the case where $\mathcal{H}$ does not have an infinite OIG-Littlestone tree. In the adversarial setting, we assume access to an infinite sequence of labeled data $(\mathcal{X} \times \mathcal{Y})^\infty$. Our high-level approach is to use this data sequence that the learner has access to in order to "simulate" the OIG-Littlestone Gale-Stewart game between P_A and P_L. Once this game has "converged", this will give rise to a *pattern-avoidance function* which will define some constraints that all realizable datasets need to satisfy. The last step is to use this pattern-avoidance function in order to define a *partial* concept class whose OIG dimension is *finite*.

We first start with the description of the execution of the game on the data. This is an online algorithm that is executed on an *infinite* sequence of labeled data $S \in (\mathcal{X} \times \mathcal{Y})^\infty$ and will serve as an important building block of our algorithm in the probabilistic setting.

- Let $g_1 : \mathcal{X} \times \mathrm{LG}_{\gamma,1} \to \mathcal{Y}$ be the function that corresponds to the strategy of P_L in the first round of the game.
- Initialize $\tau_0 \leftarrow 1$.
- For every $t \geq 1$:
  - If there exists $L \in \mathrm{LG}_{\gamma,\tau_{t-1}}$ such that $g_{\tau_{t-1}}(x_{t-\tau_{t-1}+1}, \ldots, x_t, L) = y_{t-\tau_{t-1}+1}, \ldots, y_t$: # proceed to the next round in the game.
    - $\tau_t \leftarrow \tau_{t-1} + 1$.
    - $c_{\tau_{t-1}} \leftarrow (x_{t-\tau_{t-1}+1}, \ldots, x_t, L)$
    - $g_{\tau_t}(\cdot,\cdot) := \eta_{\tau_t}(c_1, \ldots, c_{\tau_{t-1}} \cdot, \cdot)$, where $\eta_{\tau_t}(c_1, \ldots, c_{\tau_{t-1}} \cdot, \cdot)$ is the strategy of P_L in round τ_t of the game, given its history.
  - Else: $\tau_t \leftarrow \tau_{t-1}$.

Based on the above game, we define a pattern-avoidance function as

$$\hat{y}_t(x'_1,\ldots,x'_{\tau_t}) = \bigcup_{L\in \mathrm{LG}_{\gamma,\tau_{t-1}}(\mathcal{H}|_{(x'_1,\ldots,x'_{\tau_t})})} \{(y_1,\ldots,y_{\tau_t})\in L : g_{\tau_t}(x'_1,\ldots,x'_{\tau_t},L) = y_1,\ldots,y_{\tau_t}\}\,. \tag{3}$$

We also define the functions:

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

and

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

We first show that τ_t can only be increased for a finite number of times.

Proposition 3 *If $\mathcal{H}$ does not have an infinite γ-OIGL tree, for any sequence $x_1,y_1,x_2,y_2,\ldots,$ that is consistent with $\mathcal{H}$, there exists some finite number $t^* \in \mathbb{N}$, such that for all $t > t^\star$*

$$(y_{t-\tau_{t-1}+1},\ldots,,y_t) \notin \hat{y}_{t_i-1}(x_{t-\tau_{t-1}+1},\ldots,x_t), \tau_t = \tau_{t-1}, \hat{y}_t = \hat{y}_{t-1}\,.$$

Assume that this happens infinitely many times. Then, since in this Gale-Stewart game the learner is using a winning strategy there exists some $t^\star$ such that the version space is empty after the first k rounds of the game. But this contradicts the fact that the sequence is consistent with $\mathcal{H}$. We next make this sketch more formal.

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

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

for $i \in \mathbb{N}$. Take g_t to be the winning strategy of P_L, which exists since $\mathcal{H}$ does not have an infinite scaled OIGL tree. Hence there is some finite index k so that $\mathcal{H}_{\xi_1,\eta_1,\ldots,\xi_k,\eta_k} = \emptyset$, where $\xi_i = (x_{t_i-\tau_{t_i-1}+1},...,x_{t_i},L_{\tau_{t_i}-1})$ and $\eta_i = (y_{t_i-\tau_{t_i-1}+1},\ldots,,y_{t_i})$. This contradicts the fact that the sequence $x_1,y_1,...,$ is $\mathcal{H}$-consistent. ■

Now we focus on the measurability of $T_t,\hat{Y}_t$. The next result follows directly from Proposition 52 Hanneke et al. (2023).

Proposition 4 (Proposition 52 in Hanneke et al. (2023)) *For any $t \ge 1$ the functions $T_t,\hat{Y}_t$ are universally measurable.*

Pattern Avoidance $\implies$ Finite OIG Dimension. Pattern avoidance functions essentially identify constraints that any realizable data sequence must satisfy. Since we have now identified a constraint that the data need to satisfy we will express it through the following partial concept class

$$\mathcal{F} = \left\{f:\mathcal{X}\to\{0,1,\star\} : \forall(x_1,\ldots,x_{\tau_{t^\star}}) \in \mathcal{X}^{\tau_{t^\star}}, (f(x_1),\ldots,f(x_{\tau_{t^\star}})) \notin \hat{y}_{t^\star}(x_1,\ldots,x_{\tau_{t^\star}})\right\}\,.$$

Importantly, we can show that this is a partial concept class whose OIG-dimension is bounded by $\tau_{t^\star} - 1$.

We proceed formally as follows. For any $k, n \in \mathbb{N}$, any $g : \mathcal{X}^k \to 2^{\mathcal{Y}^k}$ and any $S = (x_1, ..., x_n)$, define

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

Moreover, for any $t \ge 0, n \ge \tau_t$, and sequence $S = (x_1, ..., x_n)$, let

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

Lemma 6 *Fix $\gamma > 0$ and assume that $\mathcal{H}$ does not have an infinite γ-OIGL tree. For any $t > 0$ and any sequence $(x_1, y_1, ..., x_t, y_t) \in (\mathcal{X} \times \mathcal{Y})^t$ that is consistent with $\mathcal{H}$, any $n \ge \tau_t$, and any $S = (x'_1, ..., x'_n) \in \mathcal{X}^n$, we have that*

$$\mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H}(S, \widehat{y}_t)) < \tau_t ,$$

where $\widehat{y}_t$ is the pattern avoidance function defined in (3).

Proof This follows immediately from the definition of $\widehat{y}_t$ and the class $\mathcal{H}(S, \widehat{y}_t)$. ■

A.4.2. PROBABILISTIC SETTING AND UNIFORM-TO-UNIVERSAL REDUCTION

We now move to the probabilistic setting where the data are generated by some unknown realizable distribution $\mathcal{D}$. It is in that step that we have to use our results regarding learnability of partial concept classes with finite OIG dimension (cf. Section D). Another important challenge we will need to handle is that we can only show that the game converges with high probability, so we need to consider multiple batches of it that will induce different classifiers, similarly as in Bousquet et al. (2021). When we are aggregating these different classifiers, instead of using their majority vote over the labels we use the *median* prediction. The details of our approach follow.

Let us fix a $\mathcal{H}$-realizable distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. Let $(X_1, Y_1), (X_2, Y_2), ...$ be i.i.d. random variables drawn from $\mathcal{D}$. We have the following result regarding the consistency of the random sequence:

Lemma 7 (Lemma 4.3 in Bousquet et al. (2021)) *If $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$ and $(X_1, Y_1), (X_2, Y_2), ...$ are i.i.d. random variables drawn from $\mathcal{D}$, then, with probability one, for any $t \ge 1$, there exists some $h \in \mathcal{H}$ such that $h(X_s) = Y_s$ for any $s \le t$.*

For any $k \in \mathbb{N}$ and set function $g : \mathcal{X}^k \to 2^{\mathcal{Y}^k}$, define

$$\mathrm{per}(g) = \mathbf{Pr}[(Y_1, ..., Y_k) \in g(X_1, ..., X_k)]$$

We now connect the mappings T_t and $\widehat{Y}_t$ of the adversarial setting with the probabilistic setting by defining:

$$\tau_t = T_t(X_1, Y_1, ..., X_t, Y_t)$$

and

$$\widehat{y}_t(x_1, ..., x_{\tau_t}) = \widehat{Y}_t(X_1, Y_1, ..., X_t, Y_t, x_1, ..., x_{\tau_t}) .$$

Lemma 8 (Zero Pattern Error Implies Consistency (Lemma 60 in Hanneke et al. (2023))) *For any $k, n \in \mathbb{N}$ with $n \geq k$, any function $g : \mathcal{X}^k \to 2^{\mathcal{Y}^k}$, and any sequence $S = ((X_i, Y_i))_{i=1}^n \sim \mathcal{D}^n$, if $\mathrm{per}(g) = 0$, 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 with probability one, where $S|_\mathcal{X} = (X_1, ..., X_n)$ and $\mathcal{D}'$ denotes the uniform distribution over $\{(i, Y_i)\}_{i=1}^n$, i.e., $\mathcal{D}'((i, Y_i)) = 1/n$ for any $i \in [n]$.*

Essentially, this result states that whenever the games have converged the pattern avoidance function we obtain from them gives rise to partial concept classes which can perfectly label all the data that we have seen so far. An identical result was proven by Hanneke et al. (2023) for pattern avoidance functions that are related to *pseudo-cubes* rather than scaled OIGs, but an adaptation to accommodate the modified pattern avoidance function is straightforward.

Lemma 9 (Eventual Convergence of Pattern Error (Lemma 61 in Hanneke et al. (2023))) *It holds that $\mathbf{Pr}[\mathrm{per}(\widehat{y}_t) > 0] \to 0$ as $t \to \infty$.*

Again, this is a result that appears in all the universal rates literature (Bousquet et al., 2021; Kalavasis et al., 2022; Hanneke et al., 2022a, 2023) and formalizes the intuitive fact that as the size of the training set increases, the probability that the pattern avoidance function makes a mistake goes to zero.

The previous result is asymptotic and cannot be utilized directly to design an algorithm for our problem. Nevertheless, it is a standard result in the universal rates literature how to move from the asymptotic setting to the finite sample size setting. The learner can simply estimate some time $\widehat{t}_n$ so that, with constant probability over the generated dataset, the Gale-Stewart game will terminate with $\widehat{t}_n$. The proof of this result is standard and appears in all the works in the universal rates literature (Bousquet et al., 2021; Kalavasis et al., 2022; Hanneke et al., 2022a, 2023).

Lemma 10 (Rate of Convergence of Pattern Error (Lemma 62 in Hanneke et al. (2023))) *For any $n \in \mathbb{N}$, consider a training set $\{(X_i, Y_i)\}$ consisting of n points i.i.d. drawn from $\mathcal{D}$. Then there exists a universally measurable $\widehat{t}_n = \widehat{t}_n(X_1, Y_1, ..., X_{\lfloor n/2 \rfloor}, Y_{\lfloor n/2 \rfloor})$ whose definition does not depend on $\mathcal{D}$ so that the following holds. Set the critical time $t^\star \in \mathbb{N}$ be such that*

$$\mathbf{Pr}[\mathrm{per}(\widehat{y}_{t^\star}) > 0] \leq 1/8\,,$$

where the probability is over the training set of the algorithm $\widehat{y}_t$. Then, there exist $C, c > 0$ that depend on $\mathcal{D}, t^\star$ but not n so that

$$\mathbf{Pr}[\widehat{t}_n \in T^\star] \geq 1 - Ce^{-cn}\,.$$

where the probability is over the training of the estimator $\widehat{t}_n$ and $T^\star$ is the set

$$T^\star = \{1 \leq t \leq t^\star : \mathbf{Pr}[\mathrm{per}(\widehat{y}_{t^\star}) > 0] \leq 3/8\}\ ,$$

where the probability is over the training of $\widehat{y}_t$.

Uniform Rates $\implies$ Universal Rates. Now, we can apply the previous collection of lemmas concerning probabilistic aspects of pattern avoidance functions in order to design a template for building learning algorithms in the probabilistic setting. We show the following reduction. Any learning algorithm with some guaranteed uniform rate for finite scaled OIG 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 scaled OIG-Littlestone tree.

Theorem 18 (Cut-Off Loss - Reduction from Uniform) *Fix $\gamma \in (0,1)$. Suppose that $\mathbb{A}$ is a learning algorithm which for any hypothesis class $\mathcal{H}$ with γ-OIG dimension at most d, any distribution $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$, any number $n \in \mathbb{N}$, and any sample $S \sim \mathcal{D}^n$, outputs a hypothesis $h_n \sim \mathbb{A}(S)$ such that $L^{\gamma}_{\mathcal{D}}(h_n) \leq R(n,d)$ for some rate function $R : \mathbb{N} \times \mathbb{N} \to [0,1]$ which is non-increasing for any $d \in \mathbb{N}$.*

Then, there is an algorithm $\mathbb{A}'$ satisfying that for any hypothesis class $\mathcal{H}$ that does not have an infinite γ-OIGL tree and any distribution $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$, there exist some constants $C, c > 0$ and $d_0 \in \mathbb{N}$, such that such that for all $n \in \mathbb{N}$ and $S' \sim \mathcal{D}^n$, $\mathbb{A}'$ outputs a hypothesis $h'_n \sim \mathbb{A}'(S')$ with

$$L^{\gamma}_{\mathcal{D}}(h'_n) \leq Ce^{-cn} + 32R(\lceil n/4 \rceil, d_0)\,.$$

Proof Lemma 9 implies that there exists $t^* \in \mathbb{N}$ such that $\mathbf{Pr}[\mathrm{per}(\widehat{y}_{t^\star}) > 0] \leq \frac{1}{8}$. Then, for any $n \in \mathbb{N}$, let us define $\widehat{t}_n \in [\lfloor n/4 \rfloor - 1]$ to be the random time constructed in Lemma 10. For any $t \in [\lfloor n/4 \rfloor - 1]$ and any $i \in [n/(4\widehat{t}_n)]$, define also

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

and pattern avoidance mappings

$$\widehat{y}^i_t : \mathcal{X}^{\tau^i_t} \to 2^{\mathcal{Y}^{\tau^i_t}},\ (x_1, \ldots, x_{\tau^i_t}) \mapsto \widehat{Y}_t(X_{(i-1)t+1}, Y_{(i-1)t+1}, \ldots, X_{it}, Y_{it}, x_1, \ldots, x_{\tau^i_t})\,.$$

For any $t \in \mathcal{T}_{\mathrm{good}}$, since $\mathbf{E}[\mathbf{1}\{\mathrm{per}(\widehat{y}_t) > 0\}] \leq \frac{3}{8}$, by a standard Chernoff bound, we have

$$\mathbf{Pr}\left[\frac{1}{\lfloor n/(4t) \rfloor} \sum_{i=1}^{\lfloor n/(4t) \rfloor} \mathbf{1}\{\mathrm{per}(\widehat{y}^i_t) > 0\} > \frac{7}{16}\right] \leq e^{-\lfloor n/(4t) \rfloor/128} \leq e^{-\lfloor n/(4t^\star) \rfloor/128}\,.$$

This implies that our estimate $\widehat{t}_n$ satisfies

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

The first inequality follows from a union bound.

Define the sequence $S := ((1, Y_{\lfloor n/2 \rfloor+1}), (2, Y_{\lfloor n/2 \rfloor+2}), \ldots, (n - \lfloor n/2 \rfloor, Y_n))$. 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 $\mathcal{H}^i(x) := \mathcal{H}((X_{\lfloor n/2 \rfloor + 1}, \dots, X_n, x), \widehat{y}^i_{\widehat{t}_n})$. Then, for any $i \in [\lfloor n/(4\widehat{t}_n) \rfloor]$, we can define the following prediction function

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

where $\mathbb{A}(\mathcal{H}^i(x), T^i)$ is the hypothesis returned by the uniform PAC learner $\mathbb{A}$.

Let $\widehat{h}_n$ be the median 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.

Let $\mathcal{D}$ denote the labeled data distribution that is $\mathcal{H}$-realizable. Now, we need to upper bound the error rate

$$L^{\gamma}_{\mathcal{D}}(\widehat{h}_n) = \mathop{\mathbf{Pr}}_{(X,Y) \sim P}[|\widehat{h}_n(X) - Y| > \gamma] \le \mathbf{Pr}\left[\frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbf{1}\{|\widehat{y}^i(X) - Y| > \gamma\} \ge \frac{1}{2}\right].$$

This probability is at most $A + B + C$, where

$$A = \mathbf{Pr}[\widehat{t}_n \notin \mathcal{T}_{\text{good}}]\,,$$

$$B = \mathbf{Pr}\left[\frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbf{1}\{\mathrm{per}(\widehat{y}^i_{\widehat{t}_n}) > 0\} > \frac{7}{16}, \widehat{t}_n \in \mathcal{T}_{\text{good}}\right],$$

and

$$C = \mathbf{Pr}\left[\frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbf{1}\{\mathrm{per}(\widehat{y}^i_{\widehat{t}_n}) = 0\} > \frac{9}{16}, \widehat{t}_n \in \mathcal{T}_{\text{good}}, \frac{1}{\lfloor n/(4\widehat{t}_n) \rfloor} \sum_{i=1}^{\lfloor n/(4\widehat{t}_n) \rfloor} \mathbf{1}\{|\widehat{y}^i(X) - Y| > \gamma\} \ge \frac{1}{2}\right].$$

We know how to control A and B, it remains to argue about C. This is the context of the rest of the proof.

Define the sequence $S' := ((1, Y_{\lfloor n/2 \rfloor + 1}), \dots, (n - \lfloor n/2 \rfloor, Y_n), (n - \lfloor n/2 \rfloor + 1, Y))$ 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 7, $(X_{(i-1)\widehat{t}_n + 1}, Y_{(i-1)\widehat{t}_n + 1}, \dots, X_{i\widehat{t}_n}, Y_{i\widehat{t}_n})$ is consistent with $\mathcal{H}$ a.s. Then, by Lemma 6, we have that with probability 1, $\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H}^i(X)) < \tau^i_{\widehat{t}_n}$ and therefore,

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

Moreover, if $\mathrm{per}(\widehat{y}^i_{\widehat{t}_n}) = 0$, by Lemma 8, 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 Hanneke et al. (2023)[Lemma 65] and the property of $\mathbb{A}$ that

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

Standard properties of conditional expectation and the above bound imply that

$$\mathbf{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\}\mathbf{1}\{\text{per}(\widehat{y}^i_{\widehat{t}_n}) = 0\}\,\mathbf{Pr}[|\widehat{y}^i(X) - Y| > \gamma|((X_j, Y_j))_{j=1}^{\lfloor n/2 \rfloor}] \leq 2R(\lceil (n - \lfloor n/2 \rfloor)/2 \rceil, t^*)\,. \tag{5}$$

Hence, we have that

$$C \leq \mathbf{Pr}\left[\mathbf{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\}\frac{1}{\lfloor n/(4\widehat{t}_n)\rfloor}\sum_{i=1}^{\lfloor n/(4\widehat{t}_n)\rfloor}\mathbf{1}\{\text{per}(\widehat{y}^i_{\widehat{t}_n}) = 0\}\mathbf{1}\{\widehat{y}^i(X) \neq Y\} \geq \frac{1}{16}\right].$$

Now by Markov's inequality, this is at most

$$16\,\mathbf{E}\left[\mathbf{1}\{\widehat{t}_n \in \mathcal{T}_{\text{good}}\}\frac{1}{\lfloor n/(4\widehat{t}_n)\rfloor}\sum_{i=1}^{\lfloor n/(4\widehat{t}_n)\rfloor}\mathbf{1}\{\text{per}(\widehat{y}^i_{\widehat{t}_n}) = 0\}\mathbf{1}\{\widehat{y}^i(X) \neq Y\} \geq \frac{1}{16}\right].$$

This can be upper bounded by $32R(\lceil (n - \lfloor n/2 \rfloor)/2 \rceil, t^\star) \leq 32R(\lceil n/4 \rceil, t^\star)$ using (5).

Hence we have that

$$L^{\gamma}_{\mathcal{D}}(\widehat{h}_n) \leq A + B + C \leq C_1 e^{-c_1 n} + t^\star e^{-\lfloor n/(4t^*)\rfloor/128} + 32R(\lceil n/4 \rceil, t^\star)\,.$$

■

By combining Theorem 18 and Theorem 26, we conclude that

Corollary 1 *Fix $\gamma \in (0, 1)$. Assume that $\mathcal{H}$ does not have an infinite γ-OIGL tree. Then $\mathcal{H}$ is learnable with respect to the expected γ-cut-off loss at rate $\frac{\log^2 n}{n}$.*

A.5. Arbitrarily Slow Rates for Cut-Off Loss

In this section, we show that whenever $\mathcal{H}$ has an infinite OIG-Littlestone tree it requires arbitrarily slow rates. The following result from Bousquet et al. (2021) is useful for our construction.

Lemma 11 (Lemma 5.12 from Bousquet et al. (2021)) *Let $R(\cdot)$ be a rate function. There exist probabilities $p_1, p_2, \ldots \geq 0$ so that $\sum_{k\geq 1} p_k = 1$, two increasing sequences of integers $(n_i)_{i\geq 1}$ and $(k_i)_{i\geq 1}$, and a constant $1/2 \leq C \leq 1$ such that for all $i > 1$: (i) $\sum_{k>k_i} p_k \leq 1/n_i$, (ii) $n_i p_{k_i} \leq k_i$, and (iii) $p_{k_i} = CR(n_i)$.*

We also state the following result from Attias et al. (2023) that our construction relies on.

Lemma 12 (Lemma 6 from Attias et al. (2023)) *Let $\mathbb{A}$ be any learning algorithm and $\epsilon, \delta, \gamma \in (0, 1)^3$ such that $\delta < \varepsilon$. Then, the algorithm $\mathbb{A}$ requires at least*

$$\Omega\left(\frac{\mathbb{D}^{\text{OIG}}_{2\gamma}(\mathcal{H})}{\varepsilon}\right),$$

many samples to achieve expected γ-cut-off loss at most ϵ with probability $1 - \delta$ in the uniform setting, where $\mathbb{D}^{\text{OIG}}_{2\gamma}(\mathcal{H})$ is the 2γ-OIG dimension of $\mathcal{H}$.

Remark 19 (Hard Distribution for OIG) *Let us explain the structure of the construction of Attias et al. (2023) that we will utilize in our arbitrarily slow lower bound. Essentially, they show that for any learning algorithm $\mathbb{A}$ if there are n_0 elements of $\mathcal{X}$, which we denote by S, such that the restriction of $\mathcal{H}$ on S, which we denote by $\mathcal{H}_{|S}$, induces a OIG type of graph where for every orientation there exists a node of the graph that has 2γ-out-degree at least $n_0/3$, there is a way to define a realizable distribution $\mathcal{D}'$ with respect to $\mathcal{H}'$, where $\mathcal{X}' = S, \mathcal{H}' = \mathcal{H}_{|S}$, so that upon receiving n samples, the algorithm $\mathbb{A}$ will make $\Omega(n_0/n)$ mistakes that are of magnitude γ, in expectation over the random draws of the samples of $\mathcal{D}'$.*

Equipped with the previous two results, we can give the high-level idea of our construction in more detail. We will define a distribution that is supported on a single path of the infinite OIGL tree, potentially skipping some levels of it. Unlike the previous arbitrarily slow rates constructions in the universal rates literature (Bousquet et al., 2021; Kalavasis et al., 2022; Hanneke et al., 2023), we will choose the target path in a *deterministic* way. Given any target rate $R(\cdot)$, we will use the sequence $\{p_k\}_{k\in\mathbb{N}}$ to assign total mass p_k on some node of the k-th level of the tree. We define the path branch inductively. Starting from the root of the tree, we choose the edge that is indicated by Remark 19. This is something that can be done since every node of the OIGL tree is an instance of a graph described in Remark 19. We keep following the constructed path and picking the appropriate edge on every level of the tree. So far we have described (i) the construction of the path, and, (ii) the total mass on each level. What remains to be described is how the mass within each node is distributed. Again, this follows by the construction of Remark 19. Conditional on some node of the tree, the distribution within the node is exactly the one that Attias et al. (2023) define. The idea to prove the result is that there is an infinite sequence $\{n_i\}_{i\in\mathbb{N}}$ so that when the learner takes as input n i.i.d. samples from our constructed distribution, with some constant probability, it will only see elements up to level k_i. Thus, roughly speaking, for the points that lie on the node of level k_i it will have the error rate indicated in Lemma 12.

We are now ready to state and prove our result.

Theorem 20 (Cut-Off Loss - Arbitrarily Slow Rates) *Fix $\gamma \in (0,1)$. Assume that $\mathcal{H}$ admits an infinite 2γ-One-Inclusion Graph Littlestone tree. Then $\mathcal{H}$ requires arbitrarily slow rates.*

Proof We proceed with the proof of the lower bound. Fix an arbitrary rate $R : \mathbb{N} \to [0,1]$ (slower than linear), so that $R(n) \overset{n\to\infty}{\longrightarrow} 0$, a learning algorithm $\mathbb{A}$, an infinite 2γ-One-Inclusion Graph Littlestone tree $\{x_u\}$ for $\mathcal{H}$, and $C, \{p_k\}_{k\in\mathbb{N}}, \{k_i\}_{i\in\mathbb{N}}, \{n_i\}_{i\in\mathbb{N}}$ as in Lemma 11. We would like to show that there exists a realizable distribution for which the algorithm has expected cut-off loss R. Our goal is to design this distribution based on the input learning algorithm.

Let us first recall the definition of the scaled OIGL tree, with a simplified notation. We know that, for any level $k \geq 1$, the node of the tree contains a tuple: it contains $S(y_{\leq k}) \in \mathcal{X}^{k+1}$ (where $y_{\leq k}$ is the path one has to follow to reach that node) and it also contains a finite set of labelings $L(y_{\leq k}) \subseteq \mathcal{Y}^{k+1}$ of the node $S(y_{\leq k})$. Any edge connecting this node to its children is labeled by one element in $L(y_{\leq k})$. The node of the k-th level of the scaled OIGL tree has the special property that for all the orientations of the scaled OIG induced by $(S(y_{\leq k}), L(y_{\leq k}))$, there exists a node of that OIG, i.e., a particular labeling of $S(y_{\leq k})$, which has 2γ-out-degree at least $(k+1)/3$. Thus, for any learning algorithm $\mathbb{A}$, the node admits a "hard" distribution as described in Lemma 12, Remark 19, which is defined by the marginal distribution on $S(y_{\leq k})$ and the target labeling $L(y_{\leq k})_*$. We denote the marginal distribution on $S(y_{\leq k})$ as $\mathcal{D}^*_{S(y_{\leq k})}$.

In order to define the construction, we need to (i) described how to choose the target path ii) argue how to assign mass on different levels of the tree, (iii) verify that the designed distribution is realizable by $\mathcal{H}$ and (iv) show that the expected cut-off loss of the algorithm is at least $R(n)$ for infinitely many values of n.

We start from the root of the tree and choose the edge $L(y_{\leq 0})_*$. We use this edge to move on to the next level of the tree and continue inductively in the same way. We let $y^*_{\leq k}$ to denote the node of the k-the level of the tree we have chosen in our path.

The total mass on the node of the $k-$th level is p_k, and conditional on the node, the mass is distributed among its elements as indicated by $\mathcal{D}^*_{S(y^*_{\leq k})}$. The labels of the elements $S(y^* \leq k)$ of that node are given by $L(y^*_{\leq k})_*$. This completes the description of the data-generating distribution $\mathcal{D}^*$.

Let us now move on to arguing about the realizability of the distribution. By the definition of the tree, for every level $n \in \mathbb{N}$ and every level $k \leq n$, there is some $h_{y^*_{\leq n}} \in \mathcal{H}$ that perfectly labels all the elements that appear on the path $y^*_{\leq n}$. For this classifier, we can bound its cut-off loss by

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}^*}[h_{y^*_{\leq n}}(x) \neq y] \leq \sum_{k>n+1} p_k \,,$$

which goes to zero as $n \to \infty$. Hence, the distribution is indeed realizable.

Consider the sequences $C, \{p_k\}_{k\in\mathbb{N}}, \{k_i\}_{i\in\mathbb{N}}, \{n_i\}_{i\in\mathbb{N}}$ as in Lemma 11. For each such $n_i, i \in \mathbb{N}$, our goal is to show a lower bound of $O(1/n_i)$. Notice that

$$\sum_{k>k_i} p_k \leq \frac{1}{n_i} \,,$$

so with probability at least $(1-1/n_i)^{n_i} \geq 1/4$, the learner will not observe any samples from levels deeper than k_i upon receiving n_i samples from $\mathcal{D}^*$. Let us call this event E_1^i and condition on it. Moreover, notice that $n_i \cdot p_{k_i} \leq k_i$. Also, notice that since the node on level k_i has mass p_{k_i}, the expected number of samples the learner observes from that node is at most $n_i \cdot p_{k_i} \leq k_i$. Thus, Markov's inequality shows us that with probability at least $4/5$, the learner will observe at most $5 \cdot k_i$ points from the node on the path that lies on level k_i. Let us call this event E_2^i and condition on it. By a union bound, $\mathbf{Pr}[E_1^i \cap E_2^i] \geq 1/20$. Let us denote by $\widehat{h}_{n_i}$ the output of the algorithm when it receives n_i i.i.d. samples from $\mathcal{D}^*$. Let us also condition on the event E_3^i that the test point (X, Y) is coming from the node of the target path that lies on level k_i. Under these events, the construction described in Lemma 12, Remark 19, shows that

$$\mathbf{E}\left[\mathbf{Pr}[|\widehat{h}_{n_i}(X) - Y| > \gamma]|E_1^i \cap E_2^i \cap E_3^i\right] \geq C' \cdot \frac{k_i}{5k_i} \,,$$

for some absolute constant $C' > 0$. Moreover, since E_3^i is independent from $E_1^i \cap E_2^i$ we have that

$$\begin{aligned}\mathbf{Pr}[E_1^i \cap E_2^i \cap E_3^i] &= \mathbf{Pr}[E_1^i \cap E_2^i] \cdot \mathbf{Pr}[E_3^i] \\ &\geq \frac{1}{20} \cdot p_{k_i} \\ &\geq \frac{1}{20} \cdot C \cdot R(n_i) \,.\end{aligned}$$

Putting it together, we see that

$$\mathbf{E}\left[\mathbf{Pr}[|\widehat{h}_{n_i}(X) - Y| > \gamma]\right] \geq C' \cdot \frac{k_i}{5k_i} \cdot \frac{1}{20} \cdot C \cdot R(n_i) \geq \widetilde{C} \cdot R(n_i)\,,$$

where $\widetilde{C} > 0$ is some absolute constant. This concludes the proof.

■

Appendix B. Universal Rates Landscape for Absolute Loss

As in the cut-off case, we start by providing a definition of non-trivial classes.

Definition 8 (Non-Trivial Class for Absolute Loss) *A hypothesis class $\mathcal{H}$ is non-trivial with respect to the expected absolute loss if $|\mathcal{H}| \geq 2$ and there exists $x_1, x_2 \in \mathcal{X}$ and $h_1, h_2 \in \mathcal{H}$ such that $h_1(x_1) = h_2(x_1), h_1(x) \neq h_2(x)$.*

Similar to the case of the cut-off loss, if $\mathcal{H}$ is trivial then there is an algorithm that learns this class using just one sample, since it can exactly learn the target hypothesis.

B.1. Exponential Rates for Absolute Loss (Lower Bound)

The exponential rates lower bound follows from Proposition 5, which is an adaptation of Bousquet et al. (2021).

Proposition 5 (Absolute Loss - Exponential Rates - Lower Bound) *Assume that $\mathcal{H}$ is non-trivial with respect to the absolute loss. Then $\mathcal{H}$ cannot be learned at a rate faster than exponential under the expected absolute loss.*

The proof of this result follows directly from the exponential rates lower bound of the cut-off case.

B.2. Exponential Rates for Absolute Loss (Upper Bound)

We will next design a learning algorithm that achieves exponential rates in the case where $\mathcal{H}$ does not have a 0-Littlestone tree. Interestingly, we will reduce the regression problem with expected absolute loss to multiclass classification.

Lemma 13 (Regression to Classification) *Assume that $\mathcal{H}$ does not have an infinite 0-Littlestone tree. Then $\mathcal{H}$ does not have an infinite multiclass Littlestone tree.*

Proof Consider an arbitrary multiclass Littlestone tree with labels coming from the label space L. At any level n of the tree, there exists a gap $\gamma_n > 0$ such that any pair of labels to the same parent node differ by at least γ_n in absolute value. Hence this multiclass tree induces a (γ_n)-Littlestone tree for the regression problem which is not infinite by assumption. The result follows. ■

This implies the following.

Theorem 21 (Absolute Loss - Exponential Rates - Upper Bound) *Assume that $\mathcal{H}$ does not admit an infinite 0-Littlestone tree. Then $\mathcal{H}$ is learnable at an optimal exponential rate.*

Proof Since $\mathcal{H}$ does not have an infinite multiclass Littlestone tree, the main result of Hanneke et al. (2023) shows that it is learnable in exponential rates, under the 0-1 loss. The result follows by noticing that the 0-1 loss upper bounds our loss. ■

B.3. Sublinear Rates for Absolute Loss (Lower Bound)

In this section we show that whenever $\mathcal{H}$ has a 0-Littlestone tree, no algorithm can achieve rate faster than $o(1/n)$.

Theorem 22 (Absolute Loss - Sublinear Rates - Lower Bound) *Assume that $\mathcal{H}$ admits an infinite 0-Littlestone tree. Fix any rate $R(n) = o(1/n)$. Then for any algorithm $\widehat{h}_n$, there exists a $\mathcal{D} \in \mathrm{RE}(\mathcal{H})$ and there exist constants C, c such that*

$$L_{\mathcal{D}}(\widehat{h}_n) \geq C \cdot R(c \cdot n)\,,$$

for infinitely many $n \in \mathbb{N}$.

Proof Our goal is to construct a realizable distribution $\mathcal{D}$ so that given some rate function $R(n) = o(1/n)$, there exist constants C, c, such that for any learning algorithm it holds $L_{\mathcal{D}}(\widehat{h}_n) \geq C \cdot R(c \cdot n)$ for infinitely many $n \in \mathbb{N}$. We will construct this distribution using the probabilistic method in the following manner: we first pick a branch of the tree uniformly at random. Then, on each level i of the tree we put small enough mass p_i that will lead to the $o(1/n)$ rates. The details follow.

Fix any learner $\widehat{h}_n$ and an infinite 0-Littlestone tree for $\mathcal{H}$. We let $\widetilde{\gamma}_0 = \gamma_\emptyset$ and for any $\ell \geq 1$ we let $\widetilde{\gamma}_\ell = \min\{\min_{\boldsymbol{y}\leq\ell} \gamma_{\boldsymbol{y}\leq\ell}, \widetilde{\gamma}_{\ell-1}\}$, i.e., the minimum gap across all nodes of level ℓ. We also let $n_1 = \inf\{n \in \mathbb{N} : \widetilde{\gamma}_1/n \geq R(n)\}$ and for all $\ell \geq 2$ we let $n_\ell = \inf\{n \in \mathbb{N}, n > 4 \cdot n_{\ell-1} : \widetilde{\gamma}_\ell/n \geq R(n)\}$. Finally, we let $p_1 = 1/n_1$, for all $\ell \geq 1$ we let $p_\ell = 1/n_\ell$, and $p_0 = 1 - \sum_{\ell>0} p_\ell$. Notice that this distribution is well-defined sine $R(n)$ is sublinear. Let $\boldsymbol{y} = (y_1, y_2, ...)$ be an i.i.d. sequence of fair Bernoulli coins. We introduce the random distribution over $\mathcal{X} \times \{0, 1, ..., k\}$ as

$$\mathcal{D}_{\boldsymbol{y}}((x_{\boldsymbol{y}_{\leq\ell}}, z_{\ell+1})) = p_\ell, \ell \geq 0\,,$$

where $z_{\ell+1} \in [0, 1]$ is the label of the edge connecting $x_{\boldsymbol{y}_{\leq\ell}}$ to its child according to the chosen path $\boldsymbol{y}$. For any $n < \infty$, there exists a hypothesis $h \in \mathcal{H}$ so that

$$h(x_{\boldsymbol{y}_{\leq\ell}}) = z_{\ell+1}$$

for $0 \leq \ell \leq n$. This is due to the construction of a 0-Littlestone tree. We have that

$$\mathrm{er}_{\boldsymbol{y}}(h) \leq \underset{(x,z)\sim\mathcal{D}_{\boldsymbol{y}}}{\mathbf{Pr}}[h(x) \neq z] \leq \sum_{\ell>n} p_\ell\,,$$

which goes to 0 as $n \to \infty$. This implies that $\mathcal{D}_{\boldsymbol{y}}$ is realizable for every infinite branch $\boldsymbol{y} \in \{0, 1\}^\infty$. Moreover, the mapping $y \to \mathcal{D}_y$ is measurable. Let us draw $(X, Z), (X_1, Z_1), (X_2, Z_2), ...$ i.i.d. samples from $\mathcal{D}_{\boldsymbol{y}}$. The first sample corresponds to the test sample and the other samples deal with the training phase. Moreover, let $T, T_1, T_2, ...$ be i.i.d. Geometric random variables with success probability $1/2$ starting at 0. We can set

1. $X = x_{\boldsymbol{y}\leq T}, Z = z_{T+1}$ and

2. $X_i = x_{\boldsymbol{y}\leq T_i}, Z_i = z_{T_i+1}$.

We consider the infinite sequence $\{n_\ell\}_{\ell\geq 1}$. Observe that on the event that $\{T = \ell, \max\{T_1, ..., T_{n_\ell}\} < \ell\}$, the value of $\widehat{h}_{n_\ell}(X)$ is conditionally independent of $z_{\ell+1}$ given $X, (X_1, Z_1), ..., (X_{n_\ell}, Z_{n_\ell})$. We next have that

$$\begin{aligned}\mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell/2, T = \ell] &\geq \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell/2, T = \ell, \max\{T_1, ..., T_{n_\ell}\} < \ell] \\ &= \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z_{\ell+1}| \geq \widetilde{\gamma}_\ell/2, T = \ell, \max\{T_1, ..., T_{n_\ell}\} < \ell] \,.\end{aligned}$$

This is equal to

$$\mathbf{E}[\mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z_{\ell+1}| \geq \widetilde{\gamma}_\ell/2|X, (X_1, Z_1), ..., (X_{n_\ell}, Z_{n_\ell})]\mathbf{1}\{T = \ell, \max\{T_1, ..., T_{n_\ell}\} < \ell\}]$$

Now conditional on this event, any algorithm will predict $\widetilde{\gamma}_\ell/2$ far from the true label with probability at least 1/2. Thus, the previous quantity is lower bounded by

$$\frac{1}{2}\mathbf{Pr}[T = \ell, \max\{T_1, ..., T_{n_\ell}\} < \ell] = \frac{p_\ell}{2}\left(1 - \sum_{i>\ell} p_i\right)^{n_\ell} \geq \frac{p_\ell}{2}\left(1 - \frac{p_\ell}{3}\right)^{n_\ell} = \frac{1}{2n_\ell}\left(1 - \frac{1}{3n_\ell}\right)^{n_\ell} \geq \frac{1}{3n_\ell} \,.$$

Our goal is to apply the reverse Fatou lemma. This can be done since almost surely, we have that

$$n_\ell \, \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell, T = \ell|\boldsymbol{y}] \leq n_\ell \, \mathbf{Pr}[T = \ell|\boldsymbol{y}] = n_\ell \, \mathbf{Pr}[T = \ell] \leq n_\ell \cdot \frac{1}{n_\ell} \leq 1 \,.$$

Hence, we can apply the reverse Fatou lemma and get

$$\mathbf{E}\left[\limsup_{\ell\to\infty} n_\ell \, \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell/2, T = \ell|\boldsymbol{y}]\right] \geq \limsup_{\ell\to\infty} n_\ell \, \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell, T = \ell] \geq 1/3 \,.$$

But, almost surely, it holds that

$$\begin{aligned}\frac{1}{R(n_\ell)} \cdot \mathbf{E}[\mathrm{er}_{\boldsymbol{y}}(\widehat{h}_{n_\ell})|\boldsymbol{y}] &\geq \frac{1}{R(n_\ell)} \cdot \widetilde{\gamma}_\ell \cdot \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell|\boldsymbol{y}] \\ &\geq \frac{1}{R(n_\ell)} \cdot \widetilde{\gamma}_\ell \cdot \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell, T = \ell|\boldsymbol{y}] \\ &\geq n_\ell \cdot \mathbf{Pr}[|\widehat{h}_{n_\ell}(X) - Z| \geq \widetilde{\gamma}_\ell, T = \ell|\boldsymbol{y}] \,.\end{aligned}$$

So, combining the above inequalities

$$\mathbf{E}\left[\limsup_{\ell\to\infty} \frac{1}{R(n_\ell)} \, \mathbf{E}[\mathrm{er}_{\boldsymbol{y}}(\widehat{h}_{n_\ell})]\right] \geq \frac{1}{3} \,.$$

Hence, there must exist a realization of $\boldsymbol{y}$ and constant $C > 0$ so that $\mathbf{E}[\mathrm{er}_{\boldsymbol{y}}(\widehat{h}_n)] \geq C \cdot R(n)$ infinitely often. Choosing $\mathcal{D} = \mathcal{D}_{\boldsymbol{y}}$ completes the proof. ■

B.4. Sublinear Rates for Absolute Loss (Achievable)

We now construct a hypothesis class $\mathcal{H}$ that contains an infinite 0-Littlestone tree and we show that for any sublinear rate $R(n)$, there exists an algorithm that learns the class at this rate.

Theorem 23 (Absolute Loss - Sublinear Rates - Achievable) *There exists a class $\mathcal{H}$ that (i) admits an infinite 0-Littlestone tree but (ii) there exists an algorithm $\widehat{h}_n$ so that the loss $L_{\mathcal{D}}(\widehat{h}_n)$ has an optimal rate $o(1/n)$ (arbitrarily close to $1/n$ but not linear).*

Proof Let $\mathcal{X}$ be a countable instance space and $\left\{\gamma_i := \frac{1}{3^{i+1}}\right\}_{i\in\mathbb{N}}$. Let T be a complete infinite binary tree whose nodes are labeled by unique elements of $\mathcal{X}$ and for every level $\ell_i \in \mathbb{N}$ and every node of level ℓ_i, the left edge of that node is labeled by $1/2 - \gamma_i$ and the right edge of that node is labeled by $1/2 + \gamma_i$. Moreover, assume that all the elements of $\mathcal{X}$ appear in T. We now define a hypothesis class $\mathcal{H}$ based on T. For every infinite path $\boldsymbol{y} \in \{0,1\}^{\mathbb{N}}$ we let $h_{\boldsymbol{y}}$ be as follows: for every $x \in \mathcal{X}$ that is not on the path $\boldsymbol{y}$, we let $h_{\boldsymbol{y}}(x) = 1/2$, otherwise if x is on the level ℓ of path $\boldsymbol{y}$ we let $h_{\boldsymbol{y}}(x) = 1/2 + (2\boldsymbol{y}_{\ell+1} - 1)\gamma_\ell$, i.e., $h_{\boldsymbol{y}}(x)$ agrees with the label of x along the path $\boldsymbol{y}$. Notice that $h_{\boldsymbol{y}}$ is well-defined on all of $\mathcal{X}$. We let $\mathcal{H} = \left\{h_{\boldsymbol{y}} : \boldsymbol{y} \in \{0,1\}^{\mathbb{N}}\right\}$. Notice that, by construction, $\mathcal{H}$ admits a 0-Littlestone tree so the fastest rate we can get is $o(1/n)$ (cf. Theorem 22). Let $\widehat{h}_n$ be defined as follows: define $\widehat{\ell}_n$ to be the deepest level so that there exists some $x_{\boldsymbol{y}\leq\widehat{\ell}_n}$ in the training set whose label is different from $1/2$. Let $\widehat{h} \in \mathcal{H}$ be a function that labels $x_{\boldsymbol{y}\leq\widehat{\ell}_n}$ correctly. For every $x \in \mathcal{X}$ that is an ancestor of $x_{\boldsymbol{y}\leq\widehat{\ell}_n}$ in the tree T, define $\widehat{h}_n(x) = \widehat{h}(x)$ and for every other point $x \in \mathcal{X}$, let $\widehat{h}_n(x) = 1/2$. We will show that $\widehat{h}_n$ achieves sublinear rates for every realizable distribution.

Let $\mathcal{D}$ be a distribution that is realizable with respect to $\mathcal{H}$. Let S_1 be the points (x,y) in the support of $\mathcal{D}$ such that $y = 1/2$, and S_2 be the points (x,y) in the support of $\mathcal{D}$ such that $y \neq 1/2$. Let also p be the total mass in S_1 and $1-p$ be the total mass in S_2. Because of the realizability assumption, the labels of all the points in S_2 must be consistent with some $h^\star \in \mathcal{H}$. Formally, let ℓ_1, ℓ_2 with $\ell_1 \leq \ell_2$ be the depth of two points $x_{\boldsymbol{y}\leq\ell_1}, x_{\boldsymbol{y}\leq\ell_2}$, such that $(x_{\boldsymbol{y}\leq\ell_1}, \widehat{y}_1), (x_{\boldsymbol{y}\leq\ell_2}, \widehat{y}_2) \in S_2$ but there is no $h \in \mathcal{H}$ such that $h(x_{\boldsymbol{y}\leq\ell_1}) = \widehat{y}_1, h(x_{\boldsymbol{y}\leq\ell_2}) = \widehat{y}_2$. Then, the realizability assumption is violated. Thus, all the points of this set belong to a single path of T and there exists (at least) one $h^\star \in \mathcal{H}$ that perfectly labels all the elements in S_2.[7] In order to bound the loss of $\widehat{h}_n$ we consider two cases: if the test point (x,y) comes from set S_1, then $\widehat{h}_n(x) = y$. This is because there cannot be a point that is an ancestor of $x_{\boldsymbol{y}\leq\widehat{\ell}_n}$ but is in S_1. Otherwise, if (x,y) comes from the set S_2 let ℓ be the level of the tree that x appears on. Then, the classifier $\widehat{h}_n$ will make a mistake of magnitude at most $\gamma_{\widehat{\ell}_n}$ if and only if the $\ell > \widehat{\ell}_n$ (recall that $\widehat{\ell}_n$ is defined to be the deepest level of a point whose label is different from $1/2$ that appears in the training set).

First, let us condition on the event E_n that the training sample contains at least $(1-p)n/2$ many points from the part of the distribution that is supported on the path. By Chernoff, this happens with probability at least $1 - e^{\Omega(n)}$, where we are hiding some distribution dependent constant. If the test point comes from the target path, the probability that it lies deeper than $\widehat{\ell}_n$ is at most $2/\left((1-p)n\right)$

[7] If there are finitely many points in S_2 there could be more than one such functions, but they all agree on these finitely many labels.

and the loss of the algorithm is bounded by $\gamma_{\widehat{\ell}_n}$. Putting it together

$$\begin{aligned}\mathbf{E}[L_{\mathcal{D}}(\widehat{h}_n)|E_n] &\leq \mathbf{E}[\Pr_{(x,y)\sim\mathcal{D}}[y \neq 1/2, \text{ depth of } x \text{ in } T \ \geq \widehat{\ell}_n] \cdot \gamma_{\widehat{\ell}_n}|E] \\ &\leq \frac{C}{n} \cdot \mathbf{E}[\gamma_{\widehat{\ell}_n}|E] \,,\end{aligned}$$

where C is some absolute numerical constant. Since $\mathbf{E}[\gamma_{\widehat{\ell}_n}|E_n]$ is non-increasing in n, $\mathbf{E}[\gamma_{\widehat{\ell}_n}|E_n] \to 0$, as $n \to \infty$, and the probability of E_n is at least $1 - e^{\Omega(n)}$ we can see that $\widehat{h}_n$ achieves rate $o(1/n)$. ■

B.5. Slower than Linear Rates for Absolute Loss

In this section we construct a family of hypothesis classes that witness rate functions between $1/n$ and arbitrarily slow as optimal rates. In particular, given any rate $R(n)$ such that $R(n)$ is non-increasing and $nR(n)$ is non-decreasing, we can construct some $\mathcal{H}$ for which no algorithm can achieve rate faster than $o(R(n))$ and there is an algorithm that achieves rate $R(n)$. This is formalized in the following result.

Theorem 24 *Given any rate function $R(n)$ such that $\lim_{n\to\infty} R(n) = 0$, and $n \cdot R(n)$ is non-decreasing, there is hypothesis class $\mathcal{H}$ which is not learnable at a rate faster than $o(R(n))$ and for which there exists a learning algorithm that achieves rate $R(n)$.*

Proof Let us first describe the high-level idea of our construction. We consider an infinite sequence of blocks of different elements,[8] where each block has size $\{k_i\}_{i\in\mathbb{N}}$, and k_i is increasing sufficiently fast at the rate which we will specify later. Moreover, we consider a sequence $\{\epsilon\}_{i\in\mathbb{N}}$, where ϵ_i is decreasing at a rate that we will specify later. Intuitively, ϵ_i indicates the gap size within each block. We identify the domain $\mathcal{X}$ with all the elements that appear in these blocks, we denote by $\mathcal{X}_i$ the set of elements that appear in the i-th block, and for $j \in [k_i]$ we refer to the $j-$th element in this block by x^i_j. The hypothesis class $\mathcal{H}$ is defined to be the one that realizes every unique pattern $1/2 + \epsilon_i, 1/2 - \epsilon_i$ for all the elements that appear in the block k_i. Formally,

$$\mathcal{H} = \left\{h : \mathcal{X} \to \{0,1\} : \forall p \in \{-1,1\}^{k_1} \times \{-1,1\}^{k_2} \times \ldots, \forall i \in \mathbb{N}, \forall j \in [k_i], \exists h \text{ so that } h(x^i_j) = 1/2 + p^i_j \cdot \epsilon_i\right\} .$$

The intuition is that $\mathcal{H}$ is rich enough to shatter every block, while the size of each block is increasing and the gap between the elements is decreasing. Let us start by describing the approach to achieve the upper bound. We define the sequence $\{k_i\}_{i\in\mathbb{N}}$ inductively, starting with $k_1 = 4$, and for all $i \in \mathbb{N}, i \geq 1$, we let $k_{i+1} \in \mathbb{N}$ to be the smallest number such that $k_{i+1} \cdot R(k_{i+1}) \geq 2k_i$. Notice that since $n \cdot R(n)$ is non-decreasing in n and $\lim_{n\to\infty} R(n) = 0$, this number k_{i+1} is well-defined. Moreover, for each $i \in \mathbb{N}$ we set $\epsilon_i = R(k_i)/2$. Notice that for all $i \in \mathbb{N}$ we have that $\sum_{j<i+1} k_j \leq k_{i+1}$. Let n denote the number of i.i.d. samples from some realizable distribution $\mathcal{D}$ that the learner observes and let i^* be such that $k_{i^*} \leq n \leq k_{i^*+1}$. Consider the learner $\widehat{h}_n : \mathcal{X} \to \{0,1\}$ that works as follows: if the test point X has appeared in the dataset, then the learner predicts the correct label, otherwise it outputs $1/2$. Our goal is to show that for every $n \in \mathbb{N}$ the expected absolute loss of of

[8] The elements are different both within the same block and across all the different blocks.

the learner is bounded by $R(n)$. Let us consider the following cases. If $X \in \mathcal{X}_j, j \geq i^* + 1$, then the loss of the learner is $R(k_{i^*+1})/2 \leq R(n)/2$, so we have shown the desired bound. Now notice that the total number of points that appear in blocks $k_1, \ldots, k_{i^*-1}$ is at most $2 \cdot k_{i^*-1}$. Moreover, notice that, by construction $2 \cdot k_{i^*-1} \leq k_{i^*} \cdot R(k_{i^*})$. For the elements that appear in these blocks, we can use the $0-1$ loss, i.e., if we do not predict the label correctly we pay loss 1, otherwise we pay loss 0. Then, we have that

$$\begin{aligned}
\mathbf{E}[|\widehat{h}_n(X) - Y| \mid X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j] &\leq \mathbf{E}[\mathbb{1}_{\widehat{h}_n(X)\neq Y} \mid X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j] \\
&= \sum_{X' \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j} \frac{p_{X'}}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j]} \cdot (1-p_{X'})^n \\
&\leq \frac{2 \cdot \left(\sum_{1\leq j\leq i^*-1} k_j\right)/n}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j]} \\
&\leq \frac{4k_{i^*-1}/n}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j]} \\
&\leq \frac{2 \cdot k_{i^*}R(k_{i^*})/n}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j]} \\
&\leq \frac{2 \cdot n \cdot R(n)/n}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j]} \\
&\leq \frac{2 \cdot R(n)}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j]}
\end{aligned}$$

where the first inequality follows from the definition of the $0-1$ loss, $p_{X'}$ is the probability mass placed on X under $\mathcal{D}$, the second inequality from Lemma 14, and the rest from the definition of k_{i^*}. Finally, let us consider the case where $X \in \mathcal{X}_{i^*}$. A similar analysis gives that

$$\begin{aligned}
\mathbf{E}[|\widehat{h}_n(X) - Y| \mid X \in \mathcal{X}_{i^*}] &\leq R(k_i) \cdot \sum_{X' \in \mathcal{X}_{i^*}} \frac{p_{X'}}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \mathcal{X}_{i^*}]} \cdot (1-p_{X'})^n \\
&\leq R(k_{i^*}) \cdot \frac{2 \cdot k_{i^*}/n}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \mathcal{X}_{i^*}]} \\
&\leq R(n) \cdot \frac{2 \cdot n/n}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \mathcal{X}_{i^*}]} \\
&\leq \frac{2R(n)}{\mathbf{Pr}_{(X,Y)\sim\mathcal{D}}[X \in \mathcal{X}_{i^*}]} \,,
\end{aligned}$$

where again the first inequality follows from the definition of $\mathcal{X}_{i^*}$, the second inequality from Lemma 14 and the rest from the definition of k_{i^*} and the fact that $n'R(n')$ is non-decreasing. Putting it all together, we see that

$$\begin{aligned}
\mathbf{E}[|\widehat{h}_n(X) - Y|] = &\Pr_{(X,Y)\sim\mathcal{D}}[X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j] \cdot \mathbf{E}[|\widehat{h}_n(X) - Y| \mid X \in \cup_{1\leq j\leq i^*-1}\mathcal{X}_j] \\
&+ \Pr_{(X,Y)\sim\mathcal{D}}[X \in \mathcal{X}_{i^*}] \cdot \mathbf{E}[|\widehat{h}_n(X) - Y| \mid X \in \mathcal{X}_{i^*}] \\
&+ \Pr_{(X,Y)\sim\mathcal{D}}[X \in \cup_{j>i^*}\mathcal{X}_j] \cdot \mathbf{E}[|\widehat{h}_n(X) - Y| \mid X \in \cup_{j>i^*}\mathcal{X}_j] \\
\leq\ & 4 \cdot R(n) + R(n)/2 \,,
\end{aligned}$$

which shows that indeed the class is learnable at a rate $R(n)$.

We now move on to proving the lower bound. For that, it suffices to construct a realizable distribution $\mathcal{D}$ and an infinite sequence $\{n_j\}_{j\in\mathbb{N}}$ so that when the learner receives n_j datapoints generated i.i.d. by $\mathcal{D}$ its expected error is at least $o(R(n_j))$. We will construct this distribution in a randomized way. First we define the marginal distribution on $\mathcal{X}$. We choose a sequence $\{i_j\}_{j\in\mathbb{N}}$ and we put total mass 2^{-j} on the block i_j which is distributed uniformly among the elements of the block. For the target function h^*, we label the points within each block i_j uniformly at random between the choices $1/2+\epsilon_{i_j}, 1/2-\epsilon_{i_j}$. We will define the choice of $\{i_j\}_{j\in\mathbb{N}}$ shortly. Let $\{n_j = k_{i_j}/2\}_{j\in\mathbb{N}}$. When the learner receives n_j samples it does not see at least half of the points of the block k_{i_j}, so its expected absolute loss over the random choice of the labels is at least $2^{-j}\cdot R(k_{i_j})/4 = 2^{-j}\cdot R(2n_j)/4$. Consider any function $R'(n) = o(R(n))$. We need to show that for any constant $C>1$ there exists a $\mathcal{D}$ and $\{n_j\}_{j\in\mathbb{N}}$ such that

$$\mathbf{E}[|\widehat{h}_{n_j}(X)-Y|] \geq R'(n_j/C)\,.$$

So far, we have shown that

$$\mathbf{E}[|\widehat{h}_{n_j}(X)-Y|] \geq 2^{-j}R(2n_j)/4\,.$$

Hence, we need to show that $2^{-j}R(2n_j)/4 \geq R'(n_j/C)$. We have that $2n_jR(2n_j) \geq n_j/CR(n_j/C)$, which follows from the fact that $nR(n)$ is non-decreasing, which means that $R(2n_j)/4 \geq \frac{1}{8C}R(n_j/C)$. Combining it with the previous result, we have that

$$\mathbf{E}[|\widehat{h}_{n_j}(X)-Y|] \geq \frac{2^{-j}}{8C}R(n_j/C)\,.$$

Thus, we let $n_j = \min\{n\in\mathbb{N} : \frac{2^{-j}}{8C}R(n_j/C) \geq R'(n_j/C)\}$, which is always well defined since $R' = o(R(n))$. This also defines $k_{i_j} = 2n_j, \forall j\in\mathbb{N}$. Notice that so far we have shown the lower bound the particular n_j. The last step for the proof is to apply Fatou's lemma to get the result for all the infinitely many $\{n_j\}_{j\in\mathbb{N}}$. Thus, so far we have shown that

$$\mathbf{E}[|\widehat{h}_{n_j}(X)-Y|] \geq R'(n_j/C)\,.$$

Moreover a similar argument using Chernofff's bound shows that with probability at least $1-e^{-C'n_j}$ we have that

$$\mathbf{E}[\widehat{h}_{n_j}(X)-Y|] \geq R'(n_j/C)/2\,.$$

Let $\boldsymbol{y}$ denote the random choices of the labels. Using Fatou's lemma we have that

$$
\begin{aligned}
\mathbf{E}\left[\limsup_{j\to\infty}\frac{1}{R'(n_j/C)}\mathop{\mathbf{E}}_{(X,Y)\sim\mathcal{D}_{\boldsymbol{y}}}[|\widehat{h}_{n_j}(X)-Y|]\right] &\geq \mathbf{E}\left[\limsup_{j\to\infty}\frac{1}{R'(n_j/C)}\cdot\right.\\
&\qquad\left.\min\left\{\mathop{\mathbf{E}}_{(X,Y)\sim\mathcal{D}_{\boldsymbol{y}}}[|\widehat{h}_{n_j}(X)-Y|], 1/2R'(n_j/C)\right\}\right]\\
&\geq \limsup_{j\to\infty}\frac{1}{R'(n_j/C)}\cdot\\
&\qquad\mathbf{E}\left[\min\left\{\mathop{\mathbf{E}}_{(X,Y)\sim\mathcal{D}_{\boldsymbol{y}}}[|\widehat{h}_{n_j}(X)-Y|], 1/2R'(n_j/C)\right\}\right]\\
&\geq \limsup_{j\to\infty}\frac{\mathbf{Pr}_{\boldsymbol{y}}\left[\mathbf{E}_{(X,Y)\sim\mathcal{D}_{\boldsymbol{y}}}[|\widehat{h}_{n_j}(X)-Y|]\geq 1/2\cdot R'(n_j/C)\right]}{R'(n_j/C)}\cdot\\
&\qquad\frac{R'(n_j/C)}{2}\\
&\geq \limsup_{j\to\infty}\frac{1}{2}\cdot(1-e^{-C'n_j})\\
&=\frac{1}{2},
\end{aligned}
$$

where the first inequality follows by definition of the $\min$, the second inequality follows from Fatou's lemma, the third inequality follows from Markov's inequality, and the fourth inequality follows from the Chernoff bound argument we discussed above. Notice that Fatou's lemma applies since

$$
\frac{1}{R'(n_j/C)}\cdot\min\left\{\mathop{\mathbf{E}}_{(X,Y)\sim\mathcal{D}_{\boldsymbol{y}}}[|\widehat{h}_{n_j}(X)-Y|], 1/2R'(n_j/C)\right\}\leq\frac{1}{2}.
$$

This concludes the proof.

■

The following result is useful for the derivation of the upper bound of our algorithm.

Lemma 14 *Let $n, K \in \mathbb{N}$. Let $\{p_i \in [0,1]\}_{i\in[K]}$ be a sequence of numbers. Then,*

$$
\sum_{i\in[K]} p_i(1-p_i)^n \leq \frac{2K}{n}.
$$

Proof We can write

$$\begin{aligned}\sum_{i\in[K]} p_i \cdot (1-p_i)^n &\leq \frac{K}{n} + \sum_{i\in[K]:p_i\geq 1/n} p_i \cdot (1-p_i)^n \\ &\leq \frac{K}{n} + \sum_{i\in[K]:p_i\geq 1/n} p_i \cdot e^{-p_i \cdot n} \\ &\leq \frac{K}{n} + \sum_{i\in[K]:p_i\geq 1/n} \frac{1}{n} \cdot e^{-1} \\ &= \left(1+\frac{1}{e}\right)\frac{K}{n}\,,\end{aligned}$$

where the second inequality follows from the fact that $xe^{-x\cdot n}$ is decreasing for $x \geq 1/n$. ■

Appendix C. Ommitted Definitions

For a sequence $\boldsymbol{y} = (y_1, y_2, ...)$, we denote $\boldsymbol{y}_{\leq k} = (y_1, ..., y_k)$. We may also usually identify elements of $\{0,1\}^d$ with strings or a prefix of a sequence of length d. We begin with a formal definition of a crucial combinatorial measure, namely the (γ_n)-Littlestone tree of a class $\mathcal{H}$.

Definition 9 ((γ_n)-Littlestone tree) *Fix some non-increasing sequence of scales $(\gamma_n) \in [0,1]^{\mathbb{N}}$. An (γ_n)-Littlestone tree of depth $d \leq \infty$ for $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ is a complete binary tree of depth d whose internal nodes are labeled by $\mathcal{X}$, and whose two edges connecting a node of level $i \leq d$ to its children are labeled by two elements in $[0,1]$ that differ by at least γ_i, such that every path of length at most d emanating from the root is consistent with a concept $h \in \mathcal{H}$. More formally, the tree consists of a set of nodes*

$$\bigcup_{0\leq \ell < d} \left\{x_u \in \mathcal{X} : u \in \{0,1\}^\ell\right\} = \{x_\emptyset\} \cup \{x_0, x_1\} \cup \{x_{00}, x_{01}, x_{10}, x_{11}\} \cup ... \subseteq \mathcal{X},$$

and real-valued scales

$$\bigcup_{0\leq \ell < d} \left\{\gamma_u \in [0,1] : \boldsymbol{u} \in \{0,1\}^\ell\right\} = \{\gamma_\emptyset\} \cup \{\gamma_0, \gamma_1\} \cup \{\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}\} \cup ... \subseteq [0,1]^{\mathbb{N}},$$

such that for every path $\boldsymbol{y} \in \{0,1\}^d$ and finite $n < d$, there exists $h \in \mathcal{H}$ so that $h(x_{\boldsymbol{y}_{\leq \ell}}) = s_{\boldsymbol{y}_{\leq \ell+1}}$ for $0 \leq \ell \leq n$, where $s_{\boldsymbol{y}_{\leq \ell+1}} \in [0,1]$ is the label of the edge connecting the nodes $x_{\boldsymbol{y}_{\leq \ell}}$ and $x_{\boldsymbol{y}_{\leq \ell+1}}$ and $|s_{\boldsymbol{y}_{\leq \ell},0} - s_{\boldsymbol{y}_{\leq \ell},1}] \geq \gamma_{\boldsymbol{y}_{\leq \ell}}$. We say that $\mathcal{H}$ has an infinite (γ_n)-Littlestone tree if there exists an (γ_n)-Littlestone tree for $\mathcal{H}$ with depth $d = \infty$. As a special case, we have a fixed-scale γ-Littlestone tree, for $\gamma \in [0,1]$.

Definition 10 (One-Inclusion Hypergraph (Rubinstein et al., 2009; Brukhim et al., 2022)) *Consider the set $[n]$ and a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{[n]}$. We define a graph $G_{\mathcal{H}}^{\mathrm{OIG}} = (V, E)$ such that $V = \mathcal{H}$. Consider a direction $i \in [n]$ and a mapping $f : [n] \setminus \{i\} \to \mathcal{Y}$. We introduce the hyperedge $e_{i,f} = \{h \in V : h(j) = f(j),\ \forall j \in [n] \setminus \{i\}\}$. We define the edge set of $G_{\mathcal{H}}^{\mathrm{OIG}}$ to be the collection*

$$E = \{e_{i,f} : i \in [n], f : [n] \setminus \{i\} \to \mathcal{Y}, e_{i,f} \neq \emptyset\}\,.$$

Definition 11 (Orientation and Scaled Out-Degree Attias et al. (2023)) *Let $\gamma \in [0,1], n \in \mathbb{N}, \mathcal{H} \subseteq [0,1]^{[n]}$. An orientation of the one-inclusion graph $G^{\mathrm{OIG}}_{\mathcal{H}} = (V,E)$ is a mapping $\sigma : E \to V$ so that $\sigma(e) \in e$ for any $e \in E$. Let $\sigma_i(e) \in [0,1]$ denote the i-th entry of the orientation.*

For a vertex $v \in V$, corresponding to some hypothesis $h \in \mathcal{H}$ (see Definition 10), let v_i be the i-th entry of v, which corresponds to $h(i)$. The (scaled) out-degree of a vertex v under σ is $\mathrm{outdeg}(v;\sigma,\gamma) = |\{i \in [n] : |\sigma_i(e_{i,v}) - v_i| > \gamma\}|$. The maximum (scaled) out-degree of σ is $\mathrm{outdeg}(\sigma,\gamma) = \max_{v\in V} \mathrm{outdeg}(v;\sigma,\gamma)$.

Definition 12 (γ-OIG Dimension Attias et al. (2023)) *Consider a class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and let $\gamma \in [0,1]$. We define the γ-one-inclusion graph dimension $\mathbb{D}^{\mathrm{OIG}}_{\gamma}$ of $\mathcal{H}$ as follows:*

$$\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H}) = \sup\{n \in \mathbb{N} : \exists S \in \mathcal{X}^n \textit{ such that } \exists \textit{ finite subgraph } G=(V,E) \textit{ of } G^{\mathrm{OIG}}_{\mathcal{H}|_S} = (V_n,E_n) \\ \textit{such that } \forall \textit{ orientations } \sigma, \exists v \in V, \textit{ where } \mathrm{outdeg}(v;\sigma,\gamma) > n/3\}\,.$$

We define the dimension to be infinite if the supremum is not attained by a finite n.

Definition 13 (Scaled OIG-Littlestone Tree) *Fix some non-increasing sequence of scales $(\gamma_n) \in [0,1]^{\mathbb{N}}$. An (γ_n)-OIG-Littlestone tree of depth $d \le \infty$ for $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ is a complete binary tree of depth d whose internal nodes at every level $i \le d$ are labeled by some element in $S \in \mathcal{X}^{i+1}$ and a finite collection L of labelings in $\mathcal{Y}^{i+1}$ so that the graph whose vertices are the elements of L and hyperedges are defined as in Definition 10 has an element with γ-out-degree at least $i/3$ for every orientation of the hyperedges (Definition 11). The edges connecting a node of level $i \le d$ to its children are labeled by the elements of L such that every path of length at most d emanating from the root is consistent with a concept $h \in \mathcal{H}$. More formally:*

- *For every $0 \le i < d$ and each node v of level i of the tree (defining the root to be at level 0) node v is labeled by some element S_v of $\mathcal{X}^{i+1}$ and a* finite *collection L_v of elements of $\mathcal{Y}^{i+1}$, where L_v can be identified with a hypothesis class defined on S_v. The requirement is that the OIG defined on S_v, L_v has the property that for every orientation of the hyperedges there exists a vertex that has γ_i-out-degree at least $i/3$. Moreover, node v has exactly $|L_v|$ children and each one is labeled by a different element of L_v.*

- *Consider any root-to-leaf path, let $x_i \in \mathcal{X}^{i+1}, y_i \in \mathcal{Y}^{i+1}$ be the node, edge of level $0 \le i < d$ that appears in the path. Let us index the elements of x_i as $x_i^0, \ldots, x_i^i, y_i^0, \ldots, y_i^i$. Then, there exists some $h \in \mathcal{H}$ such that $h(x_i^j) = y_i^j, 0 \le i < d, 0 \le j < i+1$.*

We say that $\mathcal{H}$ has an infinite (γ_n)-OIG-Littlestone tree if there exists an (γ_n)-OIG-Littlestone tree for $\mathcal{H}$ with depth $d = \infty$. As a special case, we have a fixed-scale γ-OIG-Littlestone tree, for $\gamma \in [0,1]$.

Definition 14 (Finite OIGs with Large Out-Degree (informal, see Definition 15) *Let $\gamma \in (0,1), n \in \mathbb{N}$. We define the set $\mathrm{LG}_{n,\gamma}$ to be the set of all finite subsets of $\mathcal{Y}^n$ that have the property that the graph whose nodes are all the elements of that particular finite subset of $\mathcal{Y}^n$ and whose hyperedges are defined as in the OIG, has the property that all its orientations have a node with γ-out-degree at least $n/3$.*

Definition 15 *Let $\gamma \in (0,1), n \in \mathbb{N}, \mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}, S = (x_1, \ldots, x_n) \in \mathcal{X}^n$. Let $\mathcal{H}|_S = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$. We define the set $\mathrm{LG}_{n,\gamma}, \mathrm{LG}_{n,\gamma}(\mathcal{H}|_S)$ (resp.) to be the set $\mathcal{V}$ that contains all finite subsets $V \subseteq \mathcal{Y}^n, V \subseteq \mathcal{H}|_S$ (resp.) that have the following property: the hypergraph $G = (V, E)$ where a hyperedge $e_{i,f} = \{h \in V : h(j) = f(j),\ \forall j \in [n] \setminus \{i\}\}$ and $E = \{e_{i,f} : i \in [n], f : [n] \setminus \{i\} \to \mathcal{Y}, e_{i,f} \neq \emptyset\}$, has the property that for every orientation $\sigma : E \to V$ where $\sigma(e) \in e$, there exists some $h \in V$ such that* $\mathrm{outdeg}(v; \sigma, \gamma) > n/3$.

Appendix D. PAC Realizable Regression for Partial Concepts

Inspired by Alon et al. (2022), instead of dealing with concept classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ where each concept $h \in \mathcal{H}$ is a **total function** $h : \mathcal{X} \to \mathcal{Y}$, we study **partial concept classes** $\mathcal{H} \subseteq (\mathcal{Y} \cup \{\star\})^{\mathcal{X}}$, where each concept h is now a **partial function** and $h(x) = \star$ means that the function h is **undefined** at x. We define the support of h as the set $\mathrm{supp}(h) = \{x \in \mathcal{X} : h(x) \neq \star\}$.

In this section, we will characterize PAC regression of partial concepts in the realizable setting. A distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ is **realizable** by $\mathcal{H}$ if, almost surely, for any n, a training set $(x_i, y_i)_{i \in [n]} \sim \mathcal{D}^n$ is realizable by some partial concept $h \in \mathcal{H}$, i.e., $\{x_i\}_{i \in [n]} \subseteq \mathrm{supp}(h)$ and $h(x_i) = y_i$ for all $i \leq n$. For a partial concept h and a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, whenever h outputs $\star$ it is counted as a mistake.

Attias et al. (2023) has established the following result for total concepts.

Theorem 25 (OIG Upper Bound for PAC Regression - Cut-Off, Lemma 11 in Attias et al. (2023)) *Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and $\varepsilon, \delta, \gamma \in (0,1)^3$. Then, the sample complexity of (ϵ, δ)-PAC learning $\mathcal{H}$ under the expected γ-cut-off loss is*

$$\mathcal{M}(\mathcal{H}; \varepsilon, \delta, \gamma) = O\left(\frac{\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H})}{\varepsilon} \log^2\left(\frac{\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H})}{\varepsilon}\right) + \frac{1}{\varepsilon} \log \frac{1}{\delta}\right).$$

We now show that this can be extended to partial concepts. Similar results for classification are established in Alon et al. (2022); Kalavasis et al. (2022); Hanneke et al. (2023).

Theorem 26 *Let $\epsilon, \delta, \gamma \in (0,1)^3$. For any partial concept class $\mathcal{H} \subseteq ([0,1] \cup \{\star\}\}^{\mathcal{X}}$ with $\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H}) \leq \infty$, the sample complexity of (ϵ, δ)-PAC learning $\mathcal{H}$ under the expected γ-cut-off loss is*

$$\mathcal{M}(\mathcal{H}; \epsilon, \delta, \gamma) = O\left(\frac{\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H})}{\epsilon} \log^2\left(\frac{\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H})}{\epsilon}\right) + \frac{1}{\epsilon} \log(1/\delta)\right).$$

In particular, if $\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H}) = \infty$ for some $\gamma \in [0,1]$, then $\mathcal{H}$ is not PAC learnable.

Proof Our algorithm will make use of the scaled one-inclusion graph algorithm, introduced in Attias et al. (2023) whose utility is provided by Theorem 25 for total concepts. We first show the next lemma for the scaled one-inclusion hypergraph predictor for partial concepts.

Lemma 15 *Fix $\gamma \in (0,1)$. For any partial concept class $\mathcal{H} \subseteq ([0,1] \cup \{\star\})^{\mathcal{X}}$ with $\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H}) < \infty$, there exists an algorithm $\mathbb{A} : (\mathcal{X} \times [0,1])^* \times \mathcal{X} \to [0,1]$ such that, for any $n \in \mathbb{N}$ and any sequence $\{(x_1, y_1), ..., (x_n, y_n)\} \in (\mathcal{X} \times [0,1])^n$ that is realizable with respect to $\mathcal{H}$,*

$$\mathop{\mathbf{Pr}}_{\sigma \sim \mathcal{U}(\mathbb{S}_n)}[|\mathbb{A}(x_\sigma(1), y_\sigma(1), ..., x_\sigma(n-1), y_\sigma(n-1), x_\sigma(n)) - y_\sigma(n)| > \gamma] = \widetilde{O}\left(\frac{\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H})}{n}\right).$$

Proof Fix $n \in \mathbb{N}$. Consider a set of points $S = \{x_1, ..., x_n\}$ and let S_d be the set of distinct elements of the sequence S. Define the hypothesis class $\mathcal{H}_{S_d}$ that contains all the total functions $h : S_d \to [0, 1]$ such that the sequence $\{(x, h(x)) : x \in S_d\}$ is realizable with respect to $\mathcal{H}$.
CASE A: Assume that $\mathcal{H}_{S_d} \neq \emptyset$. This is a total concept class and so let $\mathbb{A}_{S_d}$ be the algorithm guaranteed to exist by Theorem 25 with $\mathcal{X} = S_d$ and $\mathcal{H} = \mathcal{H}_{S_d}$. For any $y_1, ..., y_n \in [0, 1]$ so that the training sequence $(x_1, y_1), ..., (x_n, y_n)$ is realizable with respect to $\mathcal{H}$ (and so realizable with respect to $\mathcal{H}_{S_d}$), define

$$\mathbb{A}(x_1, y_1, ..., x_{n-1}, y_{n-1}, x_n) \triangleq \mathbb{A}_{S_d}(\mathcal{H}_{S_d}, x_1, y_1, ..., x_{n-1}, y_{n-1}, x_n)\,.$$

Moreover, we can consider any permutation of the sequence $x_1, ..., x_n$ and let the feature space S_d and the hypothesis class $\mathcal{H}_{S_d}$ the same. Finally, we have that $\mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H}_{S_d}) \leq \mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H})$. The guarantees of the total algorithm of Theorem 25 give the desired bound.

CASE B: Assume that $\mathcal{H}_{S_d}$ is empty. In this case, set $\mathbb{A}(x_1, y_1, ..., x_{n-1}, y_{n-1}, x_n) = 0$ for all sequences $(x_1, ..., x_n) \in \mathcal{X}^n$ and $(y_1, ..., y_{n-1}) \in [0, 1]^{n-1}$ that satisfy $\{h \in \mathcal{H} : h(x_i) = y_i \text{ with } i < n \text{ and } h(x_n) \in [0, 1]\} = \emptyset$. ■

Let us now focus on the upper bound given that $\mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H}) < \infty$. For any distribution $\mathcal{D}$ realizable with respect to $\mathcal{H}$ and for a sequence of n labeled i.i.d. examples from $\mathcal{D}$, we define the strategy $\widehat{h}_n(\cdot) = \mathbb{A}(X_1, Y_1, ..., X_n, Y_n, \cdot)$ and so the expected cut-off loss is

$$\mathbf{E}[\mathrm{er}_{\mathcal{D}}^{\gamma}(\widehat{h}_n)] = \mathop{\mathbf{E}}_{(X_i,Y_i)_{i\leq n}} \left[\mathop{\mathbf{Pr}}_{(X_{n+1},Y_{n+1})} [|\mathbb{A}(X_1, Y_1, ..., X_n, Y_n, X_{n+1}) - Y_{n+1}| > \gamma] \right] \leq O\left(\frac{\mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H})}{n+1} \right).$$

Essentially now we have to boost our predictor. In particular, we have to convert this algorithm which guarantees an expected error bounded of $\mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H})/(n+1)$ into an algorithm that guarantees a bound on the error with probability at least $1 - \delta$. In order to boost the algorithm, we use a standard median boosting algorithm by decomposing the dataset into $\log(1/\delta)$ parts and using Chernoff bounds. For the details we refer to Attias et al. (2023).

Let $\mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H}) = \infty$. We will show that $\mathcal{H}$ is not PAC learnable. For any $\ell \leq \mathbb{D}_\gamma^{\mathrm{OIG}}(\mathcal{H})$, let $\mathcal{X}_\ell = \{x_1, ..., x_\ell\}$ be a set OIG-shattered by $\mathcal{H}$. Let $\mathcal{H}_\ell$ be the class of all total functions $\mathcal{X}_\ell \to [0, 1]$, any distribution $\mathcal{D}$ on $\mathcal{X}_\ell \times [0, 1]$ realizable with respect to $\mathcal{H}_\ell$ can be extended to a distribution on $\mathcal{X} \times [0, 1]$ realizable with respect to $\mathcal{H}$ with $\mathcal{D}((\mathcal{X} \setminus \mathcal{X}_k) \times [0, 1]) = 0$. Thus, any lower bound on the sample complexity of PAC learning the total concept class $\mathcal{H}_\ell$ is also a lower bound on the sample complexity of learning the partial class $\mathcal{H}$. This gives the desired lower bound. ■