Universal rates of ERM for agnostic learning

Steve Hanneke   STEVE.HANNEKE@GMAIL.COM
Purdue University

Mingyue Xu   XU1864@PURDUE.EDU
Purdue University

Editors: Nika Haghtalab and Ankur Moitra

Abstract

The universal learning framework has been developed to obtain guarantees on the learning rates that hold for any fixed distribution, which can be much faster than the ones uniformly hold over all the distributions. Given that the *Empirical Risk Minimization* (ERM) principle being fundamental in the PAC theory and ubiquitous in practical machine learning, the recent work of Hanneke and Xu (2024) studied the universal rates of ERM for binary classification under the realizable setting. However, the assumption of realizability is too restrictive to hold in practice. Indeed, the majority of the literature on universal learning has focused on the realizable case, leaving the non-realizable case barely explored.

In this paper, we consider the problem of universal learning by ERM for binary classification under the agnostic setting, where the "learning curve" reflects the decay of the excess risk as the sample size increases. We explore the possibilities of agnostic universal rates and reveal a compact *trichotomy*: there are three possible agnostic universal rates of ERM, being either e^{-n}, $o(n^{-1/2})$, or arbitrarily slow. We provide a complete characterization of which concept classes fall into each of these categories. Moreover, we also establish complete characterizations for the target-dependent universal rates as well as the Bayes-dependent universal rates.

Keywords: PAC learning, Agnostic learning, Universal rates, Empirical Risk Minimization

1. Introduction

The classic PAC (Probably Approximately Correct) theory (Vapnik and Chervonenkis, 1974; Valiant, 1984) focuses on understanding the best worst-case (*uniform*) learning rate by a learning algorithm over all data distributions. Due to its distribution-free nature, the PAC framework fails to capture the distribution-dependent rates of learning hypothesis classes, which are possibly faster than the uniform learning rates (Cohn and Tesauro, 1990, 1992). From a practical perspective, the distribution for data generation is typically fixed in real-world learning problem and the collected data is rarely worst-case, the PAC framework is therefore too pessimistic to explain practical machine learning performance. Universal learning, a distribution-dependent framework that helps to understand machine learning beyond the classic PAC setting, has been proposed by Bousquet et al. (2021) and actively studied recently (Bousquet et al., 2023; Hanneke et al., 2022, 2023; Attias et al., 2024; Hanneke and Xu, 2024; Hanneke et al., 2024a). The universal learning model adopts a setting where the data distribution is fixed and the performance of a learning algorithm is measured by its "learning curve", i.e., the decay of the expected error as a function of the input sample size, and such rate of decay is called a *universal* rate. Indeed, in the work of Bousquet et al. (2021), they showed that for binary classification in the realizable setting, the optimal universal rates are captured by a *trichotomy*: every concept class $\mathcal{H}$ has a universal rate being either exponential, linear or arbitrarily slow. Compared to a well-known *dichotomy* of the optimal uniform rates: every concept class

$\mathcal{H}$ has a uniform rate being either linear $\text{VC}(\mathcal{H})/n$ or "bounded away from zero", this makes an impression that universal rates may differ a lot from the uniform rates.

In supervised learning, the celebrated *empirical risk minimization* (ERM) principle (Vapnik, 1998) stands at the center of many successful learning algorithms that seek to minimize the average error over the training data. In practice, ERM-based algorithms have been innovatively designed and widely applied in different areas of machine learning. For example, most successful applications of deep neural networks in fields such as computer vision (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012), and reinforcement learning (Mnih et al., 2015) have their models trained to minimize the empirical error, leveraging those renowned optimization algorithms such as GD, SGD and Adam (Kingma and Ba, 2015). In learning theory, the ERM principle has also been shown to have fundamental importance in understanding the PAC learnability: a concept class is learnable if and only if it can be learned by any ERM algorithm.

While the role of ERM in the classic PAC theory has been very well understood, the topics of universal learning by ERM have remained underexplored. The recent work of (Hanneke and Xu, 2024) studied the universal rates of ERM for binary classification problem in the realizable setting. They showed that the universal rates of ERM are captured by a *tetrachotomy*: every concept class that is learnable by ERM has a universal rate being either e^{-n}, $1/n$, $\log(n)/n$, or arbitrarily slow. The realizable case is indeed an idealistic scenario where a perfect hypothesis is assumed to exist, i.e., $\inf_{h\in\mathcal{H}} \text{er}_P(h) = 0$. However, in real-world machine learning applications, the ground-truth models are often complicated and unknown to practitioners. These considerations motivate us to study the universal rates of ERM in the agnostic setting, a more realistic and applicable situation where the true concept may not be in the hypothesis class, i.e., $\inf_{h\in\mathcal{H}} \text{er}_P(h) > 0$, and the goal is to find a hypothesis being competitive with the best hypothesis (in class). In this paper, we aim to answer the following fundamental question:

Question 1 *Given a concept class $\mathcal{H}$, what are the possible rates at which $\mathcal{H}$ can be agnostically universally learned by ERM?*

1.1. Notations and preliminaries

Following the classical setup of statistical learning, we consider a binary classification problem with an instance space $\mathcal{X}$ and a concept class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. Let $h : \mathcal{X} \to \{0,1\}$ be a classifier. Given a probability distribution P on $\mathcal{X} \times \{0,1\}$, the *error rate* of h is defined as $\text{er}_P(h) := P((x,y) \in \mathcal{X} \times \{0,1\} : h(x) \neq y)$. Given a dataset $S_n := \{(x_i,y_i)\}_{i=1}^n \in (\mathcal{X} \times \{0,1\})^n$, the *empirical error rate* of h is defined as $\hat{\text{er}}_{S_n}(h) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}(h(x_i) \neq y_i)$. Let P be a data distribution, for an integer n, we denote by $S_n := \{(x_i,y_i)\}_{i=1}^n \sim P^n$ an i.i.d. P-distributed dataset. Recall that a distribution P is called realizable with respect to $\mathcal{H}$ if it satisfies $\inf_{h\in\mathcal{H}} \text{er}_P(h) = 0$. Note that for realizable learning, an ERM learner is any learning algorithm that outputs a sample-consistent classifier, that is, a classifier in the sample-induced version space (Mitchell, 1977). In this paper, we consider instead the often more realistic setting of agnostic learning, where $\inf_{h\in\mathcal{H}} \text{er}_P(h) > 0$. In the agnostic setting, an ERM algorithm is any learning algorithm outputs the hypothesis achieving the best performance on the training data (breaking ties arbitrarily), i.e., $\hat{h}_n = \text{ERM}(S_n) := \arg\min_{h\in\mathcal{H}} \hat{\text{er}}_{S_n}(h)$. For simplicity, we conflate the ERM learner $\hat{h}_n$ with the hypothesis it returns throughout the paper.

In the realizable setting, the PAC learning aims to achieve $\text{er}_P(\hat{h}_n) \leq \epsilon$ for the error ϵ going to 0 as fast as possible with n, and the universal learning focuses on the rate of decay of the so-called *learning curve*, that is, the decay of the expected error rate $\mathbb{E}[\text{er}_P(\hat{h}_n)]$ as a function of sample

size n. In the agnostic setting, the goal of PAC learning is instead to guarantee that the excess risk $er_P(\hat{h}_n) - \inf_{h \in \mathcal{H}} er_P(h) \leq \epsilon$ for ϵ going to 0 as fast as possible with n. Therefore, for universal learning, it is natural to extend the notion of learning curve as the decay of the expected excess risk as a function of sample size n. Concretely, we define the expected excess risk as follow.

Definition 1 (Excess risk) *Let $\mathcal{H}$ be a concept class, and let $\{\hat{h}_n\}_{n \in \mathbb{N}}$ be the output of an ERM algorithm. For any distribution P over $\mathcal{X} \times \{0, 1\}$ and data $S_n := \{(x_i, y_i)\}_{i=1}^n \sim P^n$, we define its (expected) <u>excess risk</u> as*

$$\mathcal{E}(n, P) := \mathbb{E}\left[er_P(\hat{h}_n) - \inf_{h \in \mathcal{H}} er_P(h)\right].$$

Moreover, we say that a distribution <u>P is centered at h^*</u> for some $h^* : \mathcal{X} \rightarrow \{0, 1\}$ if it satisfies $er_P(h^*) = \inf_{h \in \mathcal{H}} er_P(h)$ and also $\inf_{h \in \mathcal{H}} P_{\mathcal{X}}(x : h(x) \neq h^*(x)) = 0$, and then h^* is called a <u>target function</u> of the learning problem.

We underline that a target function may not be in the concept class which is standard for agnostic learning, and a distribution P can have multiple target functions (Example 1). With these settings settled, we are now able to define the problem of agnostic universal learning by ERM. Following Hanneke and Xu (2024), we extend the definition from the realizable case.

Definition 2 (Agnostic universal learning by ERM) *Let $\mathcal{H}$ be a concept class and $R(n) \rightarrow 0$ be a rate function. We say*
- *$\mathcal{H}$ is <u>agnostically universally learnable at rate R by ERM</u>, if for every distribution P, there exist parameters $C, c > 0$ such that for every ERM algorithm, its excess risk satisfies $\mathcal{E}(n, P) \leq CR(cn)$, for all $n \in \mathbb{N}$.*
- *$\mathcal{H}$ is <u>not agnostically universally learnable at rate faster than R by ERM</u>, if there exists a distribution P and parameters $C, c > 0$ such that there is an ERM algorithm satisfying $\mathcal{E}(n, P) \geq CR(cn)$, for infinitely many $n \in \mathbb{N}$.*
- *$\mathcal{H}$ is <u>agnostically universally learnable with exact rate R by ERM</u>, if $\mathcal{H}$ is agnostically universally learnable at rate R by ERM, and is not agnostically universally learnable at rate faster than R by ERM.*
- *$\mathcal{H}$ <u>requires at least arbitrarily slow rates to be agnostically universally learned by ERM</u>, if for any $R(n) \rightarrow 0$, $\mathcal{H}$ is not agnostically universally learnable at rate faster than R by ERM.*

We emphasize that a crucial difference between Definition 2 and the PAC learning is that here the constants $C, c > 0$ are allowed to depend on the distribution P. Moreover, it is straightforward from the definition that we are basically considering the worst-case ERM here. The following extensions are required for presenting our results.

Definition 3 *Following the notations in Definition 2, we say that $\mathcal{H}$ is*
- *<u>agnostically universally learnable at rate $o(n^{-1/2})$ by ERM</u>, if for every distribution P and every ERM algorithm, $\mathcal{E}(n, P) = o(n^{-1/2})$, for all $n \in \mathbb{N}$.*
- *<u>not agnostically universally learnable at rate faster than $o(n^{-1/2})$ by ERM</u>, if for any $T(n) = o(n^{-1/2})$, there exists a distribution P such that there is an ERM algorithm satisfying $\mathcal{E}(n, P) \geq T(n)$, for infinitely many $n \in \mathbb{N}$.*
- *<u>agnostically universally learnable with exact rate $o(n^{-1/2})$ by ERM</u>, if the above two hold.*

Here, o is the standard asymptotic notation that can be distribution-dependent when $n \to \infty$.

Definition 4 *For a class of distributions $\mathcal{P}$, the agnostic universal learning of $\mathcal{H}$ under $\mathcal{P}$ by ERM is defined as the same as Definition 2 except considering only distributions $P \in \mathcal{P}$ instead of all the probability distributions P over $\mathcal{X} \times \{0,1\}$.*

1.2. Related works

PAC learning by ERM. The performance of ERM algorithms in the PAC framework has been well understood. For the realizable case, the optimal sample complexity of ERM learners (Blumer et al., 1989; Vapnik and Chervonenkis, 1974) is $\mathcal{M}_{\mathrm{ERM}}^{\mathcal{H}}(\epsilon, \delta) = \Theta((\mathrm{VC}(\mathcal{H}) \log(1/\epsilon) + \log(1/\delta))/\epsilon)$, resulting in the uniform rate $\mathrm{er}_P(\hat{h}_n) = \Theta((\mathrm{VC}(\mathcal{H}) \log(n/\mathrm{VC}(\mathcal{H})) + \log(1/\delta))/n)$, which is sub-optimal due to an unavoidable logarithmic factor (Auer and Ortner, 2007). Indeed, it has been proved that this uniform rate is the best achievable for any proper learner (Haussler et al., 1994; Simon, 2015), whereas there are improper learners can achieve a rate of $\Theta((\mathrm{VC}(\mathcal{H}) + \log(1/\delta))/n)$ (Hanneke, 2016; Aden-Ali et al., 2023; Larsen, 2023; Aden-Ali et al., 2024). However, for the agnostic case, ERM learners have the optimal sample complexity $\mathcal{M}_{\mathrm{ERM}}^{\mathcal{H},\mathrm{AG}}(\epsilon, \delta) = \Theta((\mathrm{VC}(\mathcal{H}) + \log(1/\delta))/\epsilon^2)$, and thus guaranteeing $\mathrm{er}_P(\hat{h}_n) - \inf_{h \in \mathcal{H}} \mathrm{er}_P(h) = \Theta(\sqrt{(\mathrm{VC}(\mathcal{H}) + \log(1/\delta))/n})$ (Haussler, 1992), which is optimal for any learning algorithm including improper learners. It is worth mentioning that this discrepancy of the optimality of the ERM rule between the two settings has been studied recently in the work of Hanneke et al. (2024b), where they showed that ERM is indeed sub-optimal when treating $\inf_{h \in \mathcal{H}} \mathrm{er}_P(h)$ as a parameter of the rates.

Universal learning. While the standard PAC model has been dominating the learning theory, the fact that practical learning rates can be much faster than the one described in the PAC theory, was not only observed from empirical experiments (Cohn and Tesauro, 1990, 1992) but also verified by some early theoretical works (Schuurmans, 1997; Koltchinskii and Beznosova, 2005; Audibert and Tsybakov, 2007; Pillaud-Vivien et al., 2018), where exponentially fast learning rates were guaranteed under specific model assumptions (e.g., for kernel methods and stochastic gradient decent). These findings motivate the development of alternative learning models that can better help to understand the practice of machine learning. The property of *universal consistency* was first provided by Stone (1977) and later generalized by Hanneke (2021), establishing the existence of universally consistent learning algorithms in any separable metric space. The work of Benedek and Itai (1988) considered a relaxation of the PAC model that lies in between the uniform and the universal settings called *nonuniform learning*, where the learning rate may depend on the target concept but still uniform over marginal distributions. The work of van Handel (2013) studied the uniform convergence property from a universal perspective and gave out a combinatorial characterization of the *universal Glivenko-Cantelli* property (Definition 8). Until very recently, the universal learning framework was first formalized by Bousquet et al. (2021), along with a complete theory of the optimal universal rates. After that, Bousquet et al. (2023) carried out a fine-grained analysis on the "distribution-free tail" of the universal *learning curves* by characterizing the optimal constant factor. As generalizations, Kalavasis et al. (2022); Hanneke et al. (2023, 2022, 2024a) studied the universal rates for other settings including multiclass classification, active learning, interactive learning, etc. The most relevant work to ours is Hanneke and Xu (2024), who studied the universal rates of ERM for binary classification problem in the realizable setting.

2. Main Results

In this section, we summarize the main results of this paper as well as the related technical notions of complexity. In brief, we study both target-independent and target-dependent agnostic universal rates by ERM. Moreover, since the target-dependent result relies on certain ad-hoc conditions which are lacking of intuition, we further propose to categorize a data distribution according to its Bayes-optimal classifier (Definition 10) and show that the corresponding Bayes-dependent universal rates are characterized by some simple combinatorial structures. Further details for these results will be discussed in the following Sections 3-5, along with related technical analyses and proof sketches.

We start with target-independent agnostic universal rates. We reveal a fundamental *trichotomy* in the following Theorem 5, namely there are exactly three possibilities for the agnostic universal rates by ERM: being either exponential (e^{-n}), or super-root ($o(n^{-1/2})$), or at least arbitrarily slow. Moreover, the characterization (that determines for each concept class which of the three categories it belongs to) simply consists of the cardinality and the VC dimension of the concept class.

Theorem 5 (Agnostic universal rates for ERM) *For every concept class $\mathcal{H}$ with $|\mathcal{H}| \geq 3$,*
- *$\mathcal{H}$ is agnostically universally learnable by ERM with exact rate e^{-n} if and only if $|\mathcal{H}| < \infty$.*
- *$\mathcal{H}$ is agnostically universally learnable by ERM with exact rate $o(n^{-1/2})$ if and only if $|\mathcal{H}| = \infty$ and $VC(\mathcal{H}) < \infty$.*
- *$\mathcal{H}$ requires at least arbitrarily slow rates to be agnostically universally learned by ERM if and only if $VC(\mathcal{H}) = \infty$.*

It is worthwhile to mention the following technical aspects of the proof. Firstly, to show the upper bound of super-root rates $o(n^{-1/2})$, we prove a refined version of a classic *uniform Bernstein* inequality (Proposition 40) which improves a result of Vapnik and Chervonenkis (1974) by a logarithmic factor. In summary, its proof applies a combination of localization (Bartlett et al., 2004, 2005; Koltchinskii, 2006), a concentration inequality (Bousquet, 2002) and an entropy integral bound on the rate of uniform convergence (van der Vaart and Wellner, 1996; Giné and Koltchinskii, 2006; van der Vaart and Wellner, 2011) accounting for variances of loss differences, together with well-known bounds on the covering numbers of VC classes (Haussler, 1995). Secondly, to get the lower bounds of $o(n^{-1/2})$ and arbitrarily slow rates, the techniques are quite different from proving ERM lower bounds in the classic PAC theory. Concretely, the idea is to use the following equivalences:

Lemma 6 (Hanneke and Xu, 2024, Lemma 8) *Any concept class $\mathcal{H}$ has an infinite eluder sequence (Definition 11) if and only if $|\mathcal{H}| = \infty$.*

Lemma 7 (Hanneke and Xu, 2024, Lemma 9) *Any concept class $\mathcal{H}$ has an infinite VC-eluder sequence (Definition 12) if and only if $VC(\mathcal{H}) = \infty$.*

We construct skillfully-designed distributions on such infinite sequences to support the lower bounds, unlike in the PAC theory such distributions for lower bounds are often constructed on finite sets.

Before proceeding to the target-dependent rates, we first introduce some relevant definitions and technical assumptions. Throughout this paper, we will often assume that a concept class satisfies the universal Glivenko-Cantelli property.

Definition 8 (Glivenko-Cantelli class, van Handel, 2013) *Let $\mathcal{H}$ be a concept class on an instance space $\mathcal{X}$. Given a probability distribution P on $\mathcal{X} \times \{0, 1\}$, let $\{(X_i, Y_i)\}_{i \geq 1}$ be a sequence*

Realizable case	Agnostic case	Categories	Examples
e^{-n}	e^{-n}	$|\mathcal{H}| < \infty$	finite class
$1/n$	$o(n^{-1/2})$	$|\mathcal{H}| = \infty$, $VC(\mathcal{H}) < \infty$, $\nexists$ infinite star-eluder seq	thresholds on $\mathbb{N}$
$\log(n)/n$	$o(n^{-1/2})$	$VC(\mathcal{H}) < \infty$, $\exists$ infinite star-eluder seq	singletons on $\mathbb{N}$
arbitrarily slow	arbitrarily slow	$VC(\mathcal{H}) = \infty$	non-VC class

Table 1: Comparison of the ERM universal rates between the realizable case and the agnostic case. The definition to a star-eluder sequence can be found in Appendix A.

Uniform rates	Universal rates	Categories
$n^{-1/2}$	e^{-n}	$|\mathcal{H}| < \infty$
$n^{-1/2}$	$o(n^{-1/2})$	$|\mathcal{H}| = \infty$, $VC(\mathcal{H}) < \infty$
$\Omega(1)$ (not learnable)	at least arbitrarily slow	$VC(\mathcal{H}) = \infty$

Table 2: Comparison between the agnostic universal rates and the agnostic uniform rates of ERM.

of independently P-distributed random samples. We say that $\mathcal{H}$ is a <u>P-Glivenko-Cantelli</u> class if

$$\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \xrightarrow{p} 0, \;\; as \; n \to \infty,$$

where the convergence rate can be P-dependent. We say $\mathcal{H}$ is a <u>universal Glivenko-Cantelli (UGC)</u> class if it is P-Glivenko-Cantelli for every distribution P.

We remark that while a finite VC dimension is neither sufficient nor necessary to ensure the universal Glivenko-Cantelli property of a hypothesis class, a bunch of works (Vapnik and Chervonenkis, 1971; Dudley et al., 1991; Van Der Vaart and Wellner, 2000) have shown that under weak measurability conditions (e.g., image-admissible Suslin, universal separability), a finite VC dimension is in fact equivalent to the uniform Glivenko-Cantelli property, which of course implies the universal Glivenko–Cantelli property. We then introduce two technical target-dependent conditions.

Condition 1 *For any distribution P centered at h^*, the following holds*

$$\inf_{h \in \mathcal{H}: er_P(h) > er_P(h^*)} \{er_P(h) - er_P(h^*)\} > 0. \;\; (\text{define } \inf_{\emptyset}\{\cdot\} = 1)$$

Condition 2 *For any distribution P centered at h^*, there exists $\epsilon_0 := \epsilon_0(P) > 0$ such that*

$$VC\left(\{h \in \mathcal{H} : 0 < er_P(h) - er_P(h^*) \le \epsilon_0\}\right) < \infty.$$

Further discussions about these conditions can be found in Section 4. Let us present a *trichotomy* capturing the target-dependent agnostic universal rates by ERM.

Theorem 9 (Target-dependent agnostic universal rates) *For every UGC class $\mathcal{H}$ with $|\mathcal{H}| \ge 3$ and every classifier h^*, let $\mathcal{P}_{h^*}$ be the set of all distributions centered at h^*, then the following hold:*
- *$\mathcal{H}$ is agnostically universally learnable under $\mathcal{P}_{h^*}$ by ERM with exact rate e^{-n} if and only if Condition 1 holds for h^*.*
- *$\mathcal{H}$ is agnostically universally learnable under $\mathcal{P}_{h^*}$ by ERM with exact rate $o(n^{-1/2})$ if and only if Condition 1 fails and Condition 2 holds for h^*.*

- $\mathcal{H}$ *requires at least arbitrarily slow rates to be agnostically universally learned under* $\mathcal{P}_{h^*}$ *by ERM if and only if Condition 2 fails for* h^*.

We notice that, though providing an "if and only if" characterization to the target-dependent universal rates, Conditions 1, 2 are not based on simple combinatorial structures, and thus not broadly applicable to general concept classes. A naturally follow-up question is that whether there is a better function-specified universal rates result with a complete characterization based on some combinatorial structures. To address this limitation, we consider to categorize a distribution according to its Bayes-optimal classifier.

Definition 10 (Bayes-optimal classifier) *A Bayes-optimal classifier with respect to a distribution* P, *denoted by* h^*_{Bayes}, *is defined to be a binary function such that* $er_P(h^*_{Bayes}) = \inf_{h:\mathcal{X}\to\{0,1\}} er_P(h)$. *Moreover, let* $\eta(x;P) := P(Y = 1|X = x)$, *then* $h^*_{Bayes}(x) = \mathbb{1}(\eta(x;P) \geq 1/2)$, *for all* $x \in \mathcal{X}$.

The following sequential structures developed by Hanneke and Xu (2024) formalize the characterization of the Bayes-dependent agnostic universal rates.

Definition 11 (Eluder sequence) *A (finite or infinite) data sequence* $\{(x_1,y_1),(x_2,y_2),\ldots\} \in (\mathcal{X} \times \{0,1\})^\infty$ *is called realizable (with respect to* $\mathcal{H}$*) if for every* $n \in \mathbb{N}$, *there exists* $h_n \in \mathcal{H}$ *such that* $h_n(x_i) = y_i$, *for all* $i \in [n]$. *Let* h *be a classifier, we say that* $\mathcal{H}$ *has an infinite eluder sequence* $\{(x_1,y_1),(x_2,y_2),\ldots\}$ *centered at* h *if it is realizable and labelled by* h, *and for every integer* $k \geq 1$, *there exists* $h_k \in \mathcal{H}$ *such that* $h_k(x_i) = y_i$ *for all* $i < k$ *and* $h_k(x_k) \neq y_k$.

Definition 12 (VC-eluder sequence) *Let* $S_n := \{(x_i,y_i)\}_{i=1}^n$ *be a dataset, the version space (induced by* S_n*), denoted by* $V_{S_n}(\mathcal{H})$ *(or* $V_n(\mathcal{H})$*), is defined as* $V_{S_n}(\mathcal{H}) := \{h \in \mathcal{H} : h(x_i) = y_i, \forall i \in [n]\}$. *We say* $\mathcal{H}$ *has an infinite VC-eluder sequence* $\{(x_1,y_1),(x_2,y_2),\ldots\}$ *centered at* h *if it is realizable and labelled by* h, *and for every integer* $k \geq 1$, $\{x_{n_k+1},\ldots,x_{n_k+k}\}$ *is a shattered set of* $V_{n_k}(\mathcal{H})$, *where* $\{n_k\}_{k\in\mathbb{N}}$ *is a sequence of integers defined as* $n_1 = 0$, $n_k := \binom{k}{2}$ *for all* $k > 1$.

Our result is compact and provides a better concept-dependent characterization of universal rates, leveraging simple combinatorial structures rather than complex conditions.

Theorem 13 (Bayes-dependent agnostic universal rates) *For every class* $\mathcal{H}$ *with* $|\mathcal{H}| \geq 3$ *and every classifier* h^*_{Bayes}, *let* $\mathcal{P}_{h^*_{Bayes}}$ *be the set of all distributions* P *such that* h^*_{Bayes} *is a Bayes-optimal classifier with respect to* P, *then the following hold:*

- $\mathcal{H}$ *is agnostically universally learnable under* $\mathcal{P}_{h^*_{Bayes}}$ *by ERM with exact rate* e^{-n} *if and only if* $\mathcal{H}$ *does not have an infinite eluder sequence centered at* h^*_{Bayes}.
- $\mathcal{H}$ *is agnostically universally learnable under* $\mathcal{P}_{h^*_{Bayes}}$ *by ERM with exact rate* $o(n^{-1/2})$ *if and only if* $\mathcal{H}$ *has an infinite eluder sequence but does not have an infinite VC-eluder sequence centered at* h^*_{Bayes}.
- $\mathcal{H}$ *requires at least arbitrarily slow rates to be agnostically universally learned under* $\mathcal{P}_{h^*_{Bayes}}$ *by ERM if and only if* $\mathcal{H}$ *has an infinite VC-eluder sequence centered at* h^*_{Bayes}.

3. Target-independent rates

In this section, we will introduce the relevant results and their proof sketches for the target-independent universal rates. We will prove each bullet of Theorem 5 separately (Theorems 14, 18, 19). We point out that for each bullet, to prove the sufficiency, both a lower bound and an upper bound are required since we are proving an exact rate. All detailed proofs in this section are deferred to Appendix B.

Theorem 14 (Target-independent e^{-n} exact rates) *A concept class $\mathcal{H}$ is agnostically universally learnable with exact rate e^{-n} by ERM if and only if $|\mathcal{H}| < \infty$.*

Proof [Proof sketch of Theorem 14] To prove the sufficiency, if $|\mathcal{H}| < \infty$, we first have an upper bound from the following lemma

Lemma 15 *Any finite concept class $\mathcal{H}$ is agnostically universally learnable at rate e^{-n} by ERM.*

The proof idea of Lemma 15 is, when $|\mathcal{H}| < \infty$, Condition 1 holds with a constant lower bound ϵ_0. We bound the excess risk of the worst-case ERM $\hat{h}_n$ by the probability of $|\hat{\text{er}}_{S_n}(\hat{h}_n) - \text{er}_P(\hat{h}_n)| \geq \epsilon_0$, and then an exponential rate comes from the Hoeffding's inequality.

Moreover, since the realizable setting is a special case of agnostic setting, we can get a lower bound on the rate from the following known result

Lemma 16 (Schuurmans, 1997) *Given a class $\mathcal{H}$, for any learning algorithm $\hat{h}_n$, there exists a realizable distribution P with respect to $\mathcal{H}$ such that $\mathbb{E}[er_P(\hat{h}_n)] \geq 2^{-(n+2)}$ for infinitely many n.*

To prove the necessity, we assume to the contrary that $|\mathcal{H}| = \infty$, then Lemma 6 implies that $\mathcal{H}$ has an infinite eluder sequence. A contradiction follows from the following lemma

Lemma 17 *If $\mathcal{H}$ has an infinite eluder sequence centered at h^*, then $\mathcal{H}$ is not agnostically universally learnable under $\mathcal{P}_{h^*}$ at rate faster than $o(n^{-1/2})$ by ERM.*

We prove Lemma 17 by designing a distribution supported on an existing infinite eluder sequence centered at the target function h^*. In order to have h^* being the target function, the distribution has to have decreasing probability masses along the eluder sequence. We then show that, given data generated from the constructed distribution, the worst-case ERM yields universal rates no faster than $o(n^{-1/2})$ by applying an anti-concentration inequality on bounding the probability of the event that more incorrect labeled examples are observed. ∎

Theorem 18 (Target-independent $o(n^{-1/2})$ exact rates) *A concept class $\mathcal{H}$ is agnostically universally learnable with exact rate $o(n^{-1/2})$ by ERM if and only if $|\mathcal{H}| = \infty$ and $VC(\mathcal{H}) < \infty$.*

Theorem 19 (Target-independent arbitrarily slow rates) *A concept class $\mathcal{H}$ requires at least arbitrarily slow rates to be agnostically universally learned by ERM if and only if $VC(\mathcal{H}) = \infty$.*

Proof [Proof sketches of Theorems 18 and 19] We first prove Theorem 18. To prove the sufficiency, assume that $|\mathcal{H}| = \infty$ and $VC(\mathcal{H}) < \infty$, the upper bound can be derived from the following lemma

Lemma 20 *Any VC class $\mathcal{H}$ is agnostically universally learnable at rate $o(n^{-1/2})$ by ERM.*

For the proof of Lemma 20, we first utilize a refined version of a classic uniform Bernstein inequality (Proposition 40) to bound the excess risk of the worst-case ERM by $O(\sqrt{P_{\mathcal{X}}\{\hat{h}_n(x) \neq h^*(x)\}/n})$. Since $\mathcal{H}$ is totally bounded in $L_1(P_{\mathcal{X}})$ pseudo-metric, we have $P_{\mathcal{X}}\{\hat{h}_n(x) \neq h^*(x)\}$ is also decreasing as n grows (Lemma 41). Finally, an $o(n^{-1/2})$ rate follows from classic localization argument.

Moreover, the lower bound is from the previous Lemma 6 and Lemma 17. To show the necessity, we prove by contradiction. If $|\mathcal{H}| < \infty$, then Lemma 15 yields the contradiction. If $VC(\mathcal{H}) = \infty$, Lemma 7 implies that $\mathcal{H}$ has an infinite VC-eluder sequence, and then a contradiction follows from the following lemma

Lemma 21 *If $\mathcal{H}$ has an infinite VC-eluder sequence centered at h^*, then $\mathcal{H}$ requires at least arbitrarily slow rates to be agnostically universally learned under $\mathcal{P}_{h^*}$ by ERM.*

The proof of Lemma 21 is similar to the proof of Lemma 17, and the key point is to construct a data distribution supported on an existing infinite VC-eluder sequence centered at the target function h^* with decreasing probability masses assigned to the shattered sets along the sequence. Finally, it is straightforward that Theorem 19 holds based on Theorem 18 and Lemma 21. ∎

4. Target-dependent rates

In this section, we give a proof sketch for the target-dependent exact universal rates (Theorem 9). All the missing proofs in this section are deferred to Appendix C.

Target-dependent universal rates have been studied in the work of Hanneke and Xu (2024) under the realizable setting. Therein, the authors stated the results as "h^* is (not) agnostically universally learnable at some rate R" for a target function h^*, which is indeed equivalent to ours "$\mathcal{H}$ is (not) agnostically universally learnable at rate R under every distribution P centered at h^*". We leave the formalized definitions (Definition 32 and 33) to Appendix A due to the space limitation. In light of this, we will write the related lemmas in this section following either of the two forms. For the realizable setting, such a definition yields a perfect *tetrachotomy* characterized by certain well-defined combinatorial structures, namely eluder sequence, star-eluder sequence and VC-eluder sequence (see Theorem 2, Hanneke and Xu (2024)). However, for the agnostic case, the aforementioned sequences do not support such a compact theory. We start with a simple example which will develop some initial intuition for what makes the agnostic case different.

4.1. Condition 1 and infinite eluder sequence

Example 1 (No centered infinite eulder sequence but no faster than $o(n^{-1/2})$ rates) *Let $\mathcal{X} := \mathbb{N}$, let h_1^* be defined as $h_1^*(x) := \mathbb{1}(x = 0)$ and let h_2^* be defined as $h_2^*(x) = 1 - \mathbb{1}(x = 0)$. Furthermore, for any integer $i \geq 1$, we define $h_i := \mathbb{1}(x \in \{0, i\})$. Finally, we define the concept class to be $\mathcal{H} := \{h_1^*, h_2^*\} \cup \{h_i\}_{i \geq 1}$. We construct a distribution P as follow:*

$$P_{\mathcal{X}}(x = 0) = P_{\mathcal{X}}(x \geq 1) = 1/2;$$
$$P(y = 1 | x = 0) = P(y = 1 | x \geq 1);$$
$$P(y = 1 | x = i) = 1/2 - \epsilon_i, \quad P(y = 0 | x = i) = 1/2 + \epsilon_i, \quad \forall i \geq 1;$$
$$\{P_{\mathcal{X}}(x = i)\epsilon_i\}_{i \geq 1} \text{ is a decreasing sequence.}$$

An interesting observation is: $er_P(h_1^*) = er_P(h_2^*) = 1/2 = \inf_{h \in \mathcal{H}} er_P(h)$, and also $\inf_{h \in \mathcal{H}} P_{\mathcal{X}}(x : h(x) \neq h_1^*(x)) = \inf_{h \in \mathcal{H}} P_{\mathcal{X}}(x : h(x) \neq h_2^*(x)) = 0$, which implies that the constructed distribution P is centered at both h_1^* and h_2^*. However, it is clear that $\mathcal{H}$ has an infinite eluder sequence centered at h_1^*, but does not have an infinite eluder sequence centered at h_2^*. In other words, while aiming to learn h_2^*, an ERM algorithm may try to output h_1^* instead (since h_1^* is also an optimal function) and thus has universal rates no faster then $o(n^{-1/2})$ as proved in Lemma 17.

The above Example 1 implies that "the existence/nonexistence of an infinite eluder sequence in $\mathcal{H}$ centered at h^*" is not an "if and only if" characterization to distinguish between e^{-n} and $o(n^{-1/2})$

ERM universal rates. Then, a naturally follow-up question is what is the desired equivalent characterization. Given a target function h^*, we consider the aforementioned target-specified Condition 1, which basically says that there is a "gap" between the error rate of the target and the best concept in the class. We will show that $\mathcal{H}$ is agnostically universally learnable under $\mathcal{P}_{h^*}$ at exponential rates if and only if Condition 1 holds for h^*. Indeed, the sufficiency holds from the following lemma.

Lemma 22 *Let $\mathcal{H}$ be any UGC concept class and h^* be any classifier. If Condition 1 holds for h^*, then h^* is agnostically universally learnable at rate e^{-n} by ERM.*

Before proceeding to the necessity, we provide deep insights into the relation between Condition 1 and "no infinite eluder sequence centered at h^{*}". **First, we can conclude that Condition 1 (holds for h^*) is stronger than "$\mathcal{H}$ has no infinite eluder sequence centered at h^{*}".** On one hand, it guarantees that there is no infinite eluder sequence centered at h^*, since otherwise the distribution we constructed in the proof of Lemma 17 would fail this condition. On the other hand, Example 1 reveals that they are inequivalent assumptions: there is no infinite eluder sequence centered at h_2^*, but Condition 1 fails for h_2^*, i.e., $\inf_{h \in \mathcal{H}: \mathrm{er}_P(h) > \mathrm{er}_P(h_2^*)} \{\mathrm{er}_P(h) - \mathrm{er}_P(h_2^*)\} = \inf_{i \geq 1} \{\mathrm{er}_P(h_i) - \mathrm{er}_P(h_1^*)\} = \inf_{i \geq 1} \{2 P_{\mathcal{X}}(x = i) \epsilon_i\} = 0$. **Moreover, Condition 1 is weaker than "$\mathcal{H}$ does not have any infinite eluder sequence".** This can be verified easily by finding an infinite class $\mathcal{H}$ such that Condition 1 holds for some target function h^*. We give out one of such examples as follow.

Example 2 (Singletons on $\mathbb{N}$) *Let $\mathcal{X} = \mathbb{N}$ and $\mathcal{H}_{singleton,\mathbb{N}} := \{h_t := \mathbb{1}(x = t) | t \in \mathcal{X}\}$ be the class of all singletons on natural numbers. We consider $\mathcal{H} = \mathcal{H}_{singleton,\mathbb{N}} \cup \{h_{all\text{-}1's}\}$ and a target function $h_{all\text{-}1's}$. Since $|\mathcal{H}| = |\mathcal{H}_{singleton,\mathbb{N}}| = \infty$, it must have an infinite eluder sequence (centered at $h_{all\text{-}0's}$) according to Lemma 6. However, it is straightforward that Condition 1 holds for $h_{all\text{-}1's}$.*

This implies that, while $\mathcal{H}$ is not universally learnable at exponential rate under every distribution when $|\mathcal{H}| = \infty$, but it can be learned exponentially fast under a subclass of distributions $\mathcal{P}_{h^*}$ when h^* satisfies some good property. Indeed, if we interpret Condition 1 as a distribution-specific condition, that is, Condition 1 holds for a distribution P if $\inf_{h \in \mathcal{H}: \mathrm{er}_P(h) > \inf_{h' \in \mathcal{H}} \mathrm{er}_P(h')} \{\mathrm{er}_P(h) - \inf_{h' \in \mathcal{H}} \mathrm{er}_P(h')\} > 0$, then **Condition 1 (holds for P) is equivalent to "$\mathcal{H}$ does not have any infinite eluder sequence centered at any target function of P".** Since P can have multiple target functions, Condition 1 can be stronger than "no infinite eluder sequence centered at h^{*}". Finally, the following two Lemmas formalize the above analysis and will be helpful to the proofs.

Lemma 23 *If $\mathcal{H}$ does not have an infinite eluder sequence centered at h^*, then for any distribution P centered at h^*, let $P_{\mathcal{X}}$ be the associated marginal distribution, the following hold:*

(1) There exists $h \in \mathcal{H}$ such that $P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0$.

(2) $\inf_{h \in \mathcal{H}: P_{\mathcal{X}}\{x: h(x) \neq h^*(x)\} > 0} P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} > 0$.

It is easy to check that in Example 1, the above (2) holds for h_2^* while Condition 1 fails. Intuitively, one may think that "no infinite eluder sequence", as a combinatorial assumption, is insufficient for a guarantee on the joint distribution P, but only possible for the marginal distribution $P_{\mathcal{X}}$. This is never a problem for the realizable case since the target is always unique therein.

Another interesting observation is that the (2) in Lemma 23 fails for h_1^*, which is the bad target that fails exponential learning rates. Note that if the failure of Condition 1 guarantees such a bad target, then the necessity follows. Indeed, we can consider Condition 1 as an aforementioned distribution-specific condition and then show that if it fails for some distribution P, then there must exist some target h^* (with respect to P) such that the (2) in Lemma 23 also fails.

Lemma 24 *Let $\mathcal{H}$ be any concept class and P be any distribution such that $\mathcal{H}$ is totally bounded in $L_1(P_{\mathcal{X}})$ pseudo-metric. If Condition 1 fails for P, i.e. the following holds:*

$$\inf_{h \in \mathcal{H}: er_P(h) > \inf_{h' \in \mathcal{H}} er_P(h')} \left\{ er_P(h) - \inf_{h' \in \mathcal{H}} er_P(h') \right\} = 0,$$

then there exists h^* such that $er_P(h^*) = \inf_{h' \in \mathcal{H}} er_P(h')$ and

$$\inf_{h \in \mathcal{H}: P_{\mathcal{X}}\{x:h(x) \neq h^*(x)\} > 0} P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0.$$

In the proof of Lemma 23, we actually show that if the (2) therein fails for some optimal function h^* under distribution P, then there exists an infinite eluder sequence centered at that h^*. Moreover, Lemma 17 yields an $o(n^{-1/2})$ lower bound on a centered infinite eluder sequence. Hence, Lemma 24 tells us that Condition 1 is not only a sufficient condition to a target-dependent e^{-n} upper bound as shown in Lemma 22, but also necessary. Altogether, we have the following theorem.

Theorem 25 (Target-dependent e^{-n} exact rates) *Let $\mathcal{H}$ be any UGC concept class and h^* be any classifier. h^* is agnostically universally learnable at exact rate e^{-n} by ERM if and only if Condition 1 holds for h^*.*

Proof [Proof of Theorem 25] To prove the sufficiency, if Condition 1 holds for h^*, we know that h^* is agnostically universally learnable at rate e^{-n} by ERM according to Lemma 22, and a lower bound e^{-n} by ERM follows simply from Lemma 16. Hence, the sufficiency holds with exponential exact rates. To prove the necessity, we assume to the contrary that Condition 1 fails for h^*. By Lemma 24 and then Lemma 23, we know that there exists an infinite eluder sequence centered at h^*. Then Lemma 17 yields that h^* is not agnostically universally learnable at rate faster than $o(n^{-1/2})$ by ERM. This leads to a contradiction and completes the proof of the necessity. ∎

4.2. Condition 2 and infinite VC-eluder sequence

Next, we will discuss the target-dependent $o(n^{-1/2})$ exact rate. Recall that " $\mathcal{H}$ has no infinite eluder sequence centered at h^*" is weaker than Condition 1 (holds for h^*), which is a fundamental reason for why "no infinite eluder sequence" is not a correct characterization for the target-dependent exponential rates. As an analogue of Example 1, it is not hard to construct an example of $(\mathcal{H}, P)$, where P has two target functions such that $\mathcal{H}$ has no infinite VC-eluder sequence centered at one but has an infinite VC-eluder sequence centered at the other. Such an example indicates that **" $\mathcal{H}$ has no infinite VC-eluder sequence centered at h^*" is weaker than Condition 2 (holds for h^*)**, and is not the correct characterization for the target-dependent super-root rate. Instead, we find out that $\mathcal{H}$ is agnostically universally learnable under $\mathcal{P}_{h^*}$ at $o(n^{-1/2})$ rate by ERM if and only if the target-specified Condition 2 holds for h^*. The following lemma states that Condition 2 stands sufficiently and necessarily for a super-root upper bound.

Lemma 26 *Let $\mathcal{H}$ be any UGC concept class and h^* be any classifier. Then h^* is agnostically universally learnable at rate $o(n^{-1/2})$ by ERM if and only if Condition 2 holds for h^*.*

Given a target function h^* and any distribution P centered at h^*, let us define the "ϵ-ball" (of $\mathcal{H}$ centered at h^*) as $\mathcal{H}(\epsilon; P, h^*) := \{h \in \mathcal{H} : 0 < er_P(h) - er_P(h^*) \leq \epsilon\}$. Condition 2 basically

says that for any distribution P centered at h^*, the ϵ-ball $\mathcal{H}(\epsilon; P, h^*)$ has finite VC dimension for sufficiently small radius $\epsilon(P)$. However, "no infinite VC-eluder sequence centered at h^*" merely provides marginal information, i.e., for any distribution P centered at h^*, there exists $\epsilon := \epsilon(P) > 0$ such that $\mathcal{H}(\epsilon; P_\mathcal{X}, h^*) := \{h \in \mathcal{H} : 0 < P_\mathcal{X}\{x : h(x) \neq h^*(x)\} \leq \epsilon\}$ has finite VC dimension. To see an analogy, we can interpret Condition 1 as that the ϵ-ball $\mathcal{H}(\epsilon; P, h^*)$ is empty for sufficiently small ϵ. However, "no infinite eluder sequence centered at h^*" only implies that for any distribution P centered at h^*, the marginal ϵ-ball $\mathcal{H}(\epsilon; P_\mathcal{X}, h^*)$ is empty for sufficiently small ϵ.

Moreover, it holds similarly that **Condition 2 is weaker than "$\mathcal{H}$ does not have any infinite VC-eluder sequence".** This implies, for a class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = \infty$, while $\mathcal{H}$ requires arbitrarily slow rates to be agnostically universally learned by ERM, it is possible to be learned at $o(n^{-1/2})$ rate under a subclass of distributions $\mathcal{P}_{h^*}$ when h^* satisfies Condition 2. Here is an example.

Example 3 (Hanneke and Xu, 2024, Ex.15) *Let $\mathcal{X} := \bigcup_{k \in \mathbb{N}} \mathcal{X}_k$ be the disjoint union of finite sets with $|\mathcal{X}_k| = k$ and $\mathcal{H} := (\bigcup_{k \geq 1} \mathcal{H}_k) \cup \{h_{all\text{-}1's}\}$, where $\mathcal{H}_k := \{\mathbb{1}_S : S \subseteq \mathcal{X}_k\}$. We consider a target function $h_{all\text{-}1's}$. Now we have $\mathrm{VC}(\mathcal{H}) = \infty$ and Condition 2 holds for $h_{all\text{-}1's}$.*

If we consider Condition 2 as a distribution-specific condition, i.e., $\mathcal{H}(\epsilon; P, h^*) := \mathcal{H}(\epsilon; P) := \{h \in \mathcal{H} : 0 < \mathrm{er}_P(h) - \inf_{h' \in \mathcal{H}} \mathrm{er}_P(h') \leq \epsilon\}$, then for any distribution P, **Condition 2 (holds for P) is equivalent to "$\mathcal{H}$ does not have any infinite VC-eluder sequence centered at any target function of P".** Finally, together with Theorem 25, we can get the following results of the target-dependent super-root exact rates and arbitrarily slow rates.

Theorem 27 (Target-dependent $o(n^{-1/2})$ exact rates) *Let $\mathcal{H}$ be any UGC concept class and h^* be any classifier. Then h^* is agnostically universally learnable at exact rate $o(n^{-1/2})$ by ERM if and only if Condition 2 holds, but Condition 1 fails for h^*.*

Theorem 28 (Target-dependent arbitrarily slow rates) *Let $\mathcal{H}$ be any concept class and h^* be any classifier. Then h^* requires at least arbitrarily slow rates to be agnostically universally learned by ERM if and only if Condition 2 fails for h^*.*

Proof [Proof of Theorems 27 and 28] To prove Theorem 27, note that given Theorem 25, it suffices to prove the part related to Condition 2. We first prove the sufficiency. To prove an exact rate, the upper bound follows from Lemma 26. For the lower bound, note that if Condition 1 fails for h^*, then there exists an infinite eluder sequence centered at h^*. Together with Lemma 17, we establish the lower bound. We next prove the necessity using the method of contradiction. Indeed, in the proof of Lemma 26, we show that if Condition 2 fails for h^*, then there exists an infinite VC-eluder sequence centered at h^*. Then Lemma 21 yields a contradiction and completes the proof of necessity. To prove Theorem 28, note that the sufficiency has already been established in the proof of Theorem 27. Finally, the necessity follows immediately by the method of contradiction. ■

5. Bayes-dependent rates

Recall that in Section 4, we find out that the target-dependent agnostic universal rates are not characterized by simple combinatorial measures but two contrived target-specified conditions. This flaw motivates us to think whether there is a better function-specified (instead of target-specified) categorization of all data distributions. In this section, we consider to categorize a distribution based

on its Bayes-optimal classifier (Definition 10) and give a theory of Bayes-dependent exact universal rates (Theorem 13). Specifically, for every distribution P, we show that the agnostic universal rates for learning $\mathcal{H}$ under P are determined by the existence/non-existence of infinite eluder sequence and VC-eluder sequence centered at the Bayes-optimal classifier h^*_{Bayes} with respect to P. While the Bayes-optimal classifier may not be unique for a distribution P (at any point x satisfying $P(Y = 1|X = x) = P(Y = 0|X = x) = 1/2$, h^*_{Bayes} can output either 0 or 1), we can show that if one of the Bayes-optimal classifiers does (not) have a centered infinite eluder/VC-eluder sequence, so do (not) the others. Therefore, this Bayes-dependent characterization provides a complete concept-dependent characterization of universal rates. All the detailed proofs (if required) in this section are deferred to Appendix D.

Theorem 29 (Bayes-dependent e^{-n} exact rates) *Let $\mathcal{H}$ be any concept class and h be any classifier. For any distribution P such that h is a Bayes-optimal classifier with respect to P, then $\mathcal{H}$ is agnostically universally learnable with exact rate e^{-n} by ERM under P if and only if $\mathcal{H}$ does not have an infinite eluder sequence centered at h.*

Theorem 30 (Bayes-dependent $o(n^{-1/2})$ exact rates) *Let $\mathcal{H}$ be any concept class and h be any classifier. For any distribution P such that h is a Bayes-optimal classifier with respect to P, then $\mathcal{H}$ is agnostically universally learnable at exact rate $o(n^{-1/2})$ by ERM under P if and only if $\mathcal{H}$ does not have an infinite VC-eluder sequence, but has an infinite eluder sequence centered at h.*

Theorem 31 (Bayes-dependent arbitrarily slow rates) *Let $\mathcal{H}$ be any concept class and h be any classifier. For any distribution P such that h is a Bayes-optimal classifier with respect to P, then $\mathcal{H}$ requires at least arbitrarily slow rates to be agnostically universally learned by ERM under P if and only if $\mathcal{H}$ has an infinite VC-eluder sequence centered at h.*

6. Conclusion and Future Work

In this work, we have revealed all the possible universal rates of ERM for binary classification under the agnostic setting and provided a complete characterization based on simple complexity notions. Our theory can be shaped into a compact *trichotomy*: there are three possible universal rates of ERM for agnostic learning, being either e^{-n} for finite concept classes, or $o(n^{-1/2})$ for VC classes, or arbitrarily slow for non-VC classes.

Since the (worst-case) ERM algorithm is a special proper learner, a natural follow-up direction is to study the optimal universal rates of proper learning. Let us consider a simple example of the class of singletons on natural numbers, that is, $\mathcal{H}_{\mathrm{singleton},\mathbb{N}} := \{h_t : t \in \mathbb{N}\}$, where $h_t(x) := \mathbb{1}\{x = t\}$ for all $x \in \mathbb{N}$. For the realizable case, Hanneke and Xu (2024) has proved a $\log(n)/n$ universal rate for ERM. However, a simple proper learner that predicts $h_{\mathrm{all\text{-}0's}}$ unless it sees a 1-labeled example (and then predict the singleton on that example), will yield an exponential universal rate. For the agnostic case, we are able to design a proper learning algorithm that achieves $e^{-o(n)}$ agnostic universal rate, while an $o(n^{-1/2})$ universal rate for ERM has been proved in this paper. Therefore, this example provides some intuitions that the universal rates (and probably also the characterizations) for proper learning can be quite different from ERM's.

Another interesting future direction is to understand the universal rates of ERM for multiclass learning (with an infinite label space). Recall that the characterization for the ERM uniform rates of multiclass learning is the graph dimension of $\mathcal{H}$ (Daniely et al., 2015), which is defined as the VC

dimension of the binary class $\Phi := \{(x,y) \mapsto \mathbb{1}\{h(x) \neq y\} : h \in \mathcal{H}\}$. Let us consider a simple concept class $\mathcal{H} := \{h_t : t \in \mathbb{N}\}$ defined on the instance space $\mathcal{X} := \{x_0, x_1\}$, where $h_t(x_0) = 0$ and $h_t(x_1) = t$ for every $t \in \mathbb{N}$. For this multiclass learning problem, we know that the uniform rates of ERM (for both realizable and agnostic cases) are exponential. However, since Φ is infinite, applying our theory to Φ does not yield a correct characterization for the universal rates of ERM for multiclass learning. Therefore, universal rates of ERM for multiclass learning is not a trivial extension of our results.

Acknowledgments

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. Optimal PAC bounds without uniform convergence. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1203–1223. IEEE, 2023.

Ishaq Aden-Ali, Mikael Møller Høandgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy. Majority-of-three: The simplest optimal learner? In *The 37th Annual Conference on Learning Theory*, pages 22–45. PMLR, 2024.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Universal rates for regression: Separations between cut-off and absolute loss. In *The 37th Annual Conference on Learning Theory*, pages 359–405. PMLR, 2024.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Peter Auer and Ronald Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2):151–163, 2007.

P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In *In The 17th Annual Conference on Learning Theory*, 2004.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Gyora M Benedek and Alon Itai. Nonuniform learnability. In *Automata, Languages and Programming: 15th International Colloquium Tampere, Finland, July 11–15, 1988 Proceedings 15*, pages 82–92. Springer, 1988.

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. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

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. In *The 36th Annual Conference on Learning Theory*, pages 5890–5924. PMLR, 2023.

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, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16(1):2377–2404, 2015.

Richard M Dudley, Evarist Giné, and Joel Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4(3):485–510, 1991.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

Steve Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

Steve Hanneke. Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research*, 22(130):1–116, 2021.

Steve Hanneke and Mingyue Xu. Universal rates of empirical risk minimization. *Advances in Neural Information Processing Systems*, 37:116291–116331, 2024.

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

Steve Hanneke, Shay Moran, and Qian Zhang. Universal rates for multiclass learning. In *Proceedings of the 36th Annual Conference on Learning Theory*, pages 5615–5681. PMLR, 2023.

Steve Hanneke, Amin Karbasi, Shay Moran, and Grigoris Velegkas. Universal rates for active learning. *Advances in Neural Information Processing Systems*, 37:74770–74807, 2024a.

Steve Hanneke, Kasper Green Larsen, and Nikita Zhivotovskiy. Revisiting agnostic PAC learning. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2024b.

D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory A*, 69(2):217–232, 1995.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

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

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

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.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *Proceedings of International Conference on Computational Learning Theory*, pages 295–307. Springer, 2005.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Kasper Green Larsen. Bagging is an optimal PAC learner. In *The 36th Annual Conference on Learning Theory*, pages 450–468. PMLR, 2023.

Tom M Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, volume 1, pages 305–310, 1977.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *The 31st Annual Conference on Learning Theory*, pages 250–296. PMLR, 2018.

Dale Schuurmans. Characterizing rational versus exponential learning curves. *Journal of Computer and System Sciences*, 55(1):140–160, 1997.

Hans U Simon. An almost optimal PAC algorithm. In *The 28th Annual Conference on Learning Theory*, pages 1552–1563. PMLR, 2015.

Eric V Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, pages 404–412, 1977.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22 (1):28–76, 1994.

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

A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Aad Van Der Vaart and Jon A Wellner. Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High dimensional probability II*, pages 115–133. Springer, 2000.

Ramon van Handel. The universal Glivenko–Cantelli property. *Probability Theory and Related Fields*, 155(3):911–934, 2013.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition. *Nauka, Moscow*, 1974.

Vladimir Naumovich Vapnik. *Statistical learning theory*. Wiley-Interscience, 1998.

Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2): 264–280, 1971.

Appendix A. Extra definitions

Definition 32 *Following the notations in Definition 2, let $h^* : \mathcal{X} \to \{0,1\}$, we say*
- h^* *is* <u>*agnostically universally learnable at rate R by ERM,*</u> *if for every distribution P centered at h^* (Definition 1), there exist parameters $C, c > 0$ such that for every ERM algorithm, $\mathcal{E}(n, P) \leq CR(cn)$, for all $n \in \mathbb{N}$.*
- h^* *is* <u>*not agnostically universally learnable at rate faster than R by ERM,*</u> *if there exists a distribution P centered at h^* and parameters $C, c > 0$ such that there is an ERM algorithm satisfying $\mathcal{E}(n, P) \geq CR(cn)$, for infinitely many $n \in \mathbb{N}$.*
- h^* *is* <u>*agnostically universally learnable with exact rate R by ERM,*</u> *if h^* is agnostically universally learnable at rate R by ERM, and is not agnostically universally learnable at rate faster than R by ERM.*

Definition 33 *Following the notations in Definition 2, let $h^* : \mathcal{X} \to \{0,1\}$, we say*
- h^* *is* <u>*agnostically universally learnable at rate $o(R)$ by ERM,*</u> *if for every distribution P centered at h^*, there exists a (distribution-dependent) function $T = o(R)$ such that $\mathcal{E}(n, P) \leq CT(cn)$, for all $n \in \mathbb{N}$, with some distribution-dependent parameters $C, c > 0$.*
- h^* *is* <u>*not agnostically universally learnable at rate faster than $o(R)$ by ERM,*</u> *if for any function $T = o(R)$, there exists a distribution P centered at h^* such that $\mathcal{E}(n, P) \geq CT(cn)$, for infinitely many $n \in \mathbb{N}$, with some distribution-dependent parameters $C, c > 0$.*
- h^* *is* <u>*agnostically universally learnable with exact rate $o(R)$ by ERM,*</u> *if h^* is agnostically universally learnable at rate $o(R)$ by ERM, and is not agnostically universally learnable at rate faster than $o(R)$ by ERM.*

Definition 34 (Region of disagreement) *Let $\mathcal{X}$ be an instance space and $\mathcal{H}$ be a concept class. We define the* <u>*region of disagreement*</u> *of $\mathcal{H}$ as $DIS(\mathcal{H}) := \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\}$.*

Definition 35 (Star number) *Let $\mathcal{X}$ be an instance space and $\mathcal{H}$ be a concept class. Let h be a classifier, the* <u>*star number of h,*</u> *denoted by $\mathfrak{s}_h(\mathcal{H})$ or $\mathfrak{s}_h$ for short, is defined to be the largest integer s such that there exist distinct points $x_1, \ldots, x_s \in \mathcal{X}$ and concepts $h_1, \ldots, h_s \in \mathcal{H}$ satisfying $DIS(\{h, h_i\}) \cap \{x_1, \ldots, x_s\} = \{x_i\}$, for every $1 \leq i \leq s$. (We say $\{x_1, \ldots, x_s\}$ is a star set of $\mathcal{H}$* <u>*centered at h.*</u>*) If no such largest integer s exists, we define $\mathfrak{s}_h = \infty$. The* <u>*star number of $\mathcal{H}$,*</u> *denoted by $\mathfrak{s}(\mathcal{H})$ or $\mathfrak{s}_{\mathcal{H}}$, is defined to be the maximum possible cardinality of a star set of $\mathcal{H}$, or $\mathfrak{s}_{\mathcal{H}} = \infty$ if no such maximum exists.*

Definition 36 (Star-eluder sequence) *Let $\mathcal{H}$ be a concept class and h be a classifier. We say that $\mathcal{H}$ has an infinite* <u>*star-eluder sequence*</u> $\{(x_1, y_1), (x_2, y_2), \ldots\}$ <u>*centered at h*</u> *, if it is realizable and for every integer $k \geq 1$, $\{x_{n_k+1}, \ldots, x_{n_k+k}\}$ is a star set of $V_{n_k}(\mathcal{H})$ centered at h, where $\{n_k\}_{k \in \mathbb{N}}$ is a sequence of integers defined as $n_1 = 0$, $n_k := \binom{k}{2}$ for all $k > 1$.*

Appendix B. Missing proofs from Section 3

Lemma 37 (Lemma 15 restated) *Any finite concept class $\mathcal{H}$ is agnostically universally learnable at rate e^{-n} by ERM.*

Proof [Proof of Lemma 37] Let $\mathcal{H}$ be any finite concept class and P be any distribution, let $h^* := \arg\min_{h \in \mathcal{H}} \mathrm{er}_P(h)$ be a target function. We define the following distribution-dependnet constant

$$\epsilon_0 := \epsilon_0(P) := \min_{h \in \mathcal{H}: \mathrm{er}_P(h) - \mathrm{er}_P(h^*) > 0} \{\mathrm{er}_P(h) - \mathrm{er}_P(h^*)\} > 0.$$

Note that when $|\mathcal{H}| < \infty$, such ϵ_0 always exists. Now we can bound the excess risk by

$$
\begin{aligned}
\mathcal{E}(n, P) =& \mathbb{E}\left[\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*)\right] \\
\leq & \mathbb{P}\left(\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) > 0\right) \\
=& \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{\mathrm{er}}_{S_n}(h) \leq \hat{\mathrm{er}}_{S_n}(h^*), \mathrm{er}_P(h) - \mathrm{er}_P(h^*) > 0\right) \\
=& \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{\mathrm{er}}_{S_n}(h) \leq \hat{\mathrm{er}}_{S_n}(h^*), \mathrm{er}_P(h) - \mathrm{er}_P(h^*) \geq \epsilon_0\right) \\
\leq & \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h) \leq \hat{\mathrm{er}}_{S_n}(h^*) - \mathrm{er}_P(h^*) - \epsilon_0\right). \quad (1)
\end{aligned}
$$

Suppose that the following event holds:

$$\left|\hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h)\right| < \epsilon_0/2, \quad \forall h \in \mathcal{H}.$$

Since $h^* \in \mathcal{H}$, we will have for any $h \neq h^* \in \mathcal{H}$,

$$(\hat{\mathrm{er}}_{S_n}(h^*) - \mathrm{er}_P(h^*)) - (\hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h)) < \epsilon_0/2 - (-\epsilon_0/2) = \epsilon_0,$$

and then the probability in (1) is zero. Therefore, we can continue the inequality (1) by

$$
\begin{aligned}
& \mathbb{P}\left(\exists h \in \mathcal{H} : \hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h) \leq \hat{\mathrm{er}}_{S_n}(h^*) - \mathrm{er}_P(h^*) - \epsilon_0\right) \\
\leq & \ \mathbb{P}\left(\exists h \in \mathcal{H} : \left|\hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h)\right| \geq \epsilon_0/2\right) \\
\overset{\text{Union bound}}{\leq} & \sum_{h \in \mathcal{H}} \mathbb{P}\left(\left|\hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h)\right| \geq \epsilon_0/2\right) \\
\overset{\text{Lemma 53}}{\leq} & \ |\mathcal{H}| \cdot 2e^{-n\epsilon_0^2/2},
\end{aligned}
$$

which completes the proof. $\blacksquare$

Lemma 38 (Lemma 17 restated) *If $\mathcal{H}$ has an infinite eluder sequence centered at h^*, then h^* is not agnostically universally learnable at rate faster than $o(n^{-1/2})$ by ERM.*

Proof [Proof of Lemma 38] Let $\{(x_1, y_1), (x_2, y_2), \ldots\}$ be an infinite eluder sequence centered at h^*, i.e. $h^*(x_i) = y_i$ for all $i \in \mathbb{N}$. We consider the following data distribution P on $\mathcal{X} \times \{0, 1\}$:

$$P_{\mathcal{X}}(x_i) = p_i, \ P(y = y_i | x_i) = 1/2 + \epsilon_i, \ P(y = 1 - y_i | x_i) = 1/2 - \epsilon_i, \ \forall i \in \mathbb{N},$$

where $\{p_i\}_{i \in \mathbb{N}}$ is a sequence of probabilities satisfying $\sum_{i \in \mathbb{N}} p_i \leq 1$ and $\{\epsilon_i\}_{i \in \mathbb{N}}$ is a sequence of positive real values. Further requirements on these two sequences will be specified later. Now according to the definition of an eluder sequence, for any $i \in \mathbb{N}$, there is $h_i \in \mathcal{H}$ such that $h_i(x_j) = y_j$ for all $1 \leq j < i$ and $h_i(x_i) = 1 - y_i$. Suppose that

$$\sum_{j=i+1}^{\infty} p_j \epsilon_j \leq p_i \epsilon_i, \quad \forall i \in \mathbb{N}. \quad (2)$$

Then we have

$$\mathrm{er}_P(h_{i+1}) \leq \sum_{j=1}^{i} p_j \left(\frac{1}{2} - \epsilon_j\right) + \sum_{j=i+1}^{\infty} p_j \left(\frac{1}{2} + \epsilon_j\right) \overset{(2)}{\leq} \sum_{j \neq i} p_j \left(\frac{1}{2} - \epsilon_j\right) + p_i \left(\frac{1}{2} + \epsilon_i\right) \leq \mathrm{er}_P(h_i).$$

It follows that $\inf_{h \in \mathcal{H}} \mathrm{er}_P(h) = \inf_{i \in \mathbb{N}} \mathrm{er}_P(h_i) = \mathrm{er}_P(h^*)$. Furthermore, by choosing $\{p_i\}_{i \in \mathbb{N}}$ to be a decreasing sequence, we also have $\inf_{h \in \mathcal{H}} P_{\mathcal{X}}(x : h(x) \neq h^*(x)) = \inf_{i \in \mathbb{N}} P_{\mathcal{X}}(x : h_i(x) \neq h^*(x)) = \inf_{i \in \mathbb{N}} p_i = 0$, that is, P is centered at h^*. Next, note that if the number of copies of (x_i, y_i) contained in S_n is less than the number of copies of $(x_i, 1 - y_i)$ in S_n, then the worst-case ERM algorithm will output a prediction $\hat{h}_n$ labeling $1 - y_i$ on x_i, which has an excess risk $\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) \geq 2p_i \epsilon_i$. Let n_{x_i} be the number of copies of x_i in S_n. The probability of such an event is lower bounded by

$$\mathbb{P}\left(\sum_{k \in [n_{x_i}]} \mathbb{1}\{Y_k = 1 - y_i\} \geq \frac{n_{x_i}}{2}\right) \overset{\text{Lemma } 56}{\geq} \frac{1}{2}\left(1 - \sqrt{1 - \exp\left(-4n_{x_i}\epsilon_i^2/(1 - 4\epsilon_i^2)\right)}\right). \quad (3)$$

We further assume that

$$n_{x_i} \leq 1/8\epsilon_i^2, \quad \forall i \in \mathbb{N},$$

which implies that $-4n_{x_i}\epsilon_i^2/(1 - 4\epsilon_i^2) \geq -8n_{x_i}\epsilon_i^2 \geq -1$. The RHS of (3) can be further extend to

$$\mathbb{P}\left(\sum_{k \in [n_{x_i}]} \mathbb{1}\{Y_k = 1 - y_i\} \geq \frac{n_{x_i}}{2}\right) \geq \frac{1}{2}\left(1 - \sqrt{1 - e^{-1}}\right) \geq \frac{1}{10}.$$

Now we let $n := n_i = 1/8\epsilon_i^2$, and it follows trivially that $n_{x_i} \leq n_i = 1/8\epsilon_i^2$. Hence, we have

$$\mathcal{E}(n, P) = \mathcal{E}(n_i, P) \geq \frac{p_i \epsilon_i}{5} = \frac{p_i}{5\sqrt{8n_i}}, \quad \forall i \in \mathbb{N}.$$

To show an $o(1/\sqrt{n})$ lower bound, it suffices to show that for any function $R(n) \to 0$, there exists a sequence of probabilities $\{p_i\}_{i \in \mathbb{N}}$ and an increasing sequence of integers $\{n_i\}_{i \in \mathbb{N}}$ such that the following hold with some universal constant $C > 0$:

$$\sum_{i \in \mathbb{N}} p_i \leq 1, \quad \sum_{j=i+1}^{\infty} \frac{p_j}{\sqrt{n_j}} \leq \frac{p_i}{\sqrt{n_i}}, \quad p_i = CR(n_i), \quad \forall i \in \mathbb{N},$$

which is proved in Lemma 57. In other words, this implies that $\mathcal{E}(n, P) \gtrsim R(n) \cdot n^{-1/2}$ holds for infinitely many $n \in \mathbb{N}$. ∎

Lemma 39 (Lemma 20 restated) *Any VC class $\mathcal{H}$ is agnostically universally learnable at rate $o(n^{-1/2})$ by ERM.*

We first prove a refined version of a classic *uniform Bernstein* inequality, which refines a result of Vapnik and Chervonenkis (1974) by a logarithmic factor. The result follows from known localized entropy integral arguments.

Proposition 40 *Let $\mathcal{H}$ be any measurable class of functions from $\mathcal{X}$ to $\{0,1\}$ with $VC(\mathcal{H}) < \infty$, and P be any probability measure on $\mathcal{X} \times \{0,1\}$. Fix any $n \in \mathbb{N}$ and $\delta \in (0,1)$. For any $h, g \in \mathcal{H}$, define $\sigma^2(h,g) = P_{\mathcal{X}}\{x : h(x) \neq g(x)\}$. Let $S_n := \{(X_i, Y_i)\}_{i=1}^n \sim P^n$, then with probability at least $1 - \delta$, every $h, g \in \mathcal{H}$ satisfy the following*

$$\left| (er_P(h) - er_P(g)) - (\hat{er}_{S_n}(h) - \hat{er}_{S_n}(g)) \right|$$

$$\leq \sqrt{\sigma^2(h,g)\frac{c_0}{n}\left(VC(\mathcal{H})\log\left(\frac{1}{\sigma^2(h,g)} \wedge \frac{n}{VC(\mathcal{H})}\right) + \log\left(\frac{1}{\delta}\right)\right)}$$

$$+ \frac{c_0}{n}\left(VC(\mathcal{H})\log\left(\frac{1}{\sigma^2(h,g)} \wedge \frac{n}{VC(\mathcal{H})}\right) + \log\left(\frac{1}{\delta}\right)\right),$$

where c_0 is a universal constant.

Proof [Proof of Proposition 40] The proof applies a combination of localization (Bartlett et al., 2004, 2005; Koltchinskii, 2006), a concentration inequality accounting for variances of loss differences (Bousquet, 2002), and an entropy integral bound on the rate of uniform convergence which also accounts for variances of loss differences (van der Vaart and Wellner, 1996; Giné and Koltchinskii, 2006; van der Vaart and Wellner, 2011), together with well-known bounds on the covering numbers of VC classes (Haussler, 1995).

Since the inequality in the proposition is vacuous when $n < 2VC(\mathcal{H})$ (for $c_0 \geq 2$), to focus on non-trivial case we assume that $n \geq 2VC(\mathcal{H})$. For any $h, g : \mathcal{X} \to \{0,1\}$, let $\ell_{h,g}(x,y) := \mathbb{1}\{h(x) \neq y\} - \mathbb{1}\{g(x) \neq y\}$, and define $\mathcal{F} := \{\ell_{h,g} : h, g \in \mathcal{H}\}$. Note that $\mathcal{F}$ is a measurable class of functions from $\mathcal{X} \times \{0,1\}$ to $\{-1,0,1\}$. Let $(X,Y) \sim P$ be independent of S_n. For each $\Delta > 0$, we will apply Lemma 60 on the following class of zero-mean functions:

$$\mathcal{F}_\Delta := \left\{(x,y) \mapsto (f(x,y) - \mathbb{E}[f(X,Y)])/2 : f \in \mathcal{F}, \; \mathbb{E}[(f(X,Y))^2] \leq 4\Delta\right\},$$

where the $1/2$ constant factor is included to ensure that $\sup_{f \in \mathcal{F}_\Delta} \|f\|_\infty \leq 1$. We remark that, while Lemma 60 requires $\mathcal{F}_\Delta$ to be countable, the purpose of this assumption in its original proof as stated in Bousquet (2002) is merely to ensure the measurability of the suprema $\sup_{f \in \mathcal{F}_\Delta} |\sum_{i=1}^n f(X_i, Y_i)|$ and $\sup_{f \in \mathcal{F}_\Delta} |\sum_{i \neq k} f(X_i, Y_i)|$, for all integers $k \leq n$. In our case, measurability of these suprema follows directly from the fact that $\mathcal{F}$ is a measurable class (and hence $\mathcal{F}_\Delta$ is as well), and thus the conclusions of Lemma 60 hold for this class $\mathcal{F}_\Delta$ as well.

Note that the variance bound is straightforward from the definition of $\mathcal{F}_\Delta$, that is

$$\sup_{f \in \mathcal{F}_\Delta} \text{Var}[f(X,Y)] = \sup_{f \in \mathcal{F}_\Delta} \mathbb{E}\left[(f(X,Y))^2\right]$$

$$= \sup_{f \in \mathcal{F} : \mathbb{E}[(f(X,Y))^2] \leq 4\Delta} \frac{1}{4}\left(\mathbb{E}[(f(X,Y))^2] - (\mathbb{E}[f(X,Y)])^2\right)$$

$$\leq \sup_{f \in \mathcal{F} : \mathbb{E}[(f(X,Y))^2] \leq 4\Delta} \frac{1}{4}\mathbb{E}[(f(X,Y))^2] \leq \Delta. \tag{4}$$

Furthermore, we have to calculate the following quantity

$$\nu := n\Delta + 2\mathbb{E}\left[\sup_{f \in \mathcal{F}_\Delta} \left|\sum_{i=1}^n f(X_i, Y_i)\right|\right].$$

By the second inequality in Lemma 60, we know that for any $\delta \in (0, 1)$, on an event $E_2(\Delta, n, \delta)$ of probability at least $1 - \delta/2$, the following holds

$$\sup_{f \in \mathcal{F}_\Delta} \left| \sum_{i=1}^n f(X_i, Y_i) \right| \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}_\Delta} \left| \sum_{i=1}^n f(X_i, Y_i) \right| \right] + \sqrt{2\nu \ln\left(\frac{2}{\delta}\right)} + \frac{1}{3}\ln\left(\frac{2}{\delta}\right). \quad (5)$$

Next, we bound the expectation on the RHS of the above (5), based on a uniform entropy integral. Letting $d = \mathrm{VC}(\mathcal{H})$, we will begin with an argument that the uniform ϵ-covering numbers of $\mathcal{F}_\Delta$ are of order $(C/\epsilon)^{C'd}$ for some universal constants C and C'. Indeed, the classic result of Haussler (1995) implies that, for any $\epsilon > 0$ and distribution Q over $\mathcal{X} \times \{0, 1\}$, the ϵ-covering number of $\mathcal{H}$ under the $L_1(Q)$ pseudo-metric, denoted by $N(\epsilon, \mathcal{H}, L_1(Q))$, is at most $(C_0/\epsilon)^d$, for some universal constant C_0 (see also van der Vaart and Wellner, 1996). Moreover, the functions in the ϵ-net of $\mathcal{H}$ (under $L_1(Q)$) of size $N(\epsilon, \mathcal{H}, L_1(Q))$ can be taken to be measurable $\{0, 1\}$-valued functions. Specifically, for any $h, g \in \mathcal{H}$, let h_ϵ and g_ϵ denote the functions within $L_1(Q)$ distance ϵ to h and g, respectively, from the above ϵ-net of $\mathcal{H}$. Then for any $(x, y) \in \mathcal{X} \times \{0, 1\}$, the following holds:

$$\begin{aligned} \left| \ell_{h_\epsilon, g_\epsilon}(x, y) - \ell_{h,g}(x, y) \right| \leq & \left| \mathbb{1}\{h_\epsilon(x) \neq y\} - \mathbb{1}\{h(x) \neq y\} \right| + \left| \mathbb{1}\{g_\epsilon(x) \neq y\} - \mathbb{1}\{g(x) \neq y\} \right| \\ = & \left| h_\epsilon(x) - h(x) \right| + \left| g_\epsilon(x) - g(x) \right|. \end{aligned} \quad (6)$$

For any distribution Q' on $\mathcal{X} \times \{0, 1\}$, we consider $Q := (1/2)Q' + (1/2)P$ for the above analysis. We know that both the $L_1(Q')$ and $L_1(P)$ distances between the functions $\ell_{h_\epsilon, g_\epsilon}$ and $\ell_{h,g}$ are at most 4ϵ, since otherwise it will contradict (6). Moreover, by Jensen's inequality, we have

$$\left| \mathbb{E}\left[\ell_{h_\epsilon, g_\epsilon}(X, Y) \right] - \mathbb{E}\left[\ell_{h,g}(X, Y) \right] \right| \leq \mathbb{E}\left[\left| \ell_{h_\epsilon, g_\epsilon}(X, Y) - \ell_{h,g}(X, Y) \right| \right] \leq 4\epsilon.$$

It follows that the functions $(\ell_{h_\epsilon, g_\epsilon} - \mathbb{E}[\ell_{h_\epsilon, g_\epsilon}(X, Y)])/2$ and $(\ell_{h,g} - \mathbb{E}[\ell_{h,g}(X, Y)])/2$ have an $L_1(Q')$ distance at most 4ϵ. Thus, the set of functions $(\ell_{h',g'} - \mathbb{E}[\ell_{h',g'}(X, Y)])/2$ with h', g' in the ϵ-net of $\mathcal{H}$ (under $L_1(Q)$) is a 4ϵ-net of $\mathcal{F}_\Delta$ under $L_1(Q')$, and hence $N(4\epsilon, \mathcal{F}_\Delta, L_1(Q')) \leq N(\epsilon, \mathcal{H}, L_1(Q))^2 \leq (C_0/\epsilon)^{2d}$. Since this argument holds for any Q' and $\epsilon > 0$, we conclude that the uniform ϵ-covering number of $\mathcal{F}_\Delta$, that is $\sup_{Q'} N(\epsilon, \mathcal{F}_\Delta, L_1(Q'))$, is at most $(4C_0/\epsilon)^{2d}$. Moreover, since the above functions in the corresponding ϵ-net of $\mathcal{F}_\Delta$ are $[-1, 1]$-valued, so are the functions in $\mathcal{F}_\Delta$, and it follows that the uniform ϵ-covering number of $\mathcal{F}_\Delta$ under the L_2 pseudo-metric satisfies

$$N(\epsilon, \mathcal{F}_\Delta, L_2) := \sup_Q N(\epsilon, \mathcal{F}_\Delta, L_2(Q)) \leq \sup_Q N(\epsilon^2/2, \mathcal{F}_\Delta, L_1(Q)) \leq \left(\frac{8C_0}{\epsilon^2}\right)^{2d} = \left(\frac{C_1}{\epsilon}\right)^{4d}, \quad (7)$$

with $C_1 := \sqrt{8C_0}$. Next, we combine this with an entropy integral bound on the expectation on the RHS of (5). Specifically, we let

$$J\left(\sqrt{\Delta}, \mathcal{F}_\Delta, L_2\right) = \int_0^{\sqrt{\Delta}} \sqrt{1 + \log N(\epsilon, \mathcal{F}_\Delta, L_2)} d\epsilon. \quad (8)$$

Applying Lemma 61 with $\mathcal{F} = \mathcal{F}_\Delta$, $\delta = \sqrt{\Delta}$ and envelope function $F = 1$, we get

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_\Delta} \left| \sum_{i=1}^n f(X_i, Y_i) \right| \right] \leq C_2 \sqrt{n} J\left(\sqrt{\Delta}, \mathcal{F}_\Delta, L_2\right) + \frac{C_2}{\Delta} J\left(\sqrt{\Delta}, \mathcal{F}_\Delta, L_2\right)^2 \quad (9)$$

for some universal constant C_2. Plugging in the bound of uniform covering numbers (7) into the uniform entropy integral (8), we have

$$J\left(\sqrt{\Delta}, \mathcal{F}_\Delta, L_2\right) \le \int_0^{\sqrt{\Delta}} \sqrt{1 + 4d \log\left(\frac{C_1}{\epsilon}\right)} d\epsilon \le C_3 \sqrt{\Delta d \log\left(\frac{1}{\Delta}\right)}$$

for some universal constant C_3. Plugging back into (9) yields

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_\Delta} \left|\sum_{i=1}^n f(X_i, Y_i)\right|\right] \le C_4 \sqrt{n\Delta d \log\left(\frac{1}{\Delta}\right)} + C_4 d \log\left(\frac{1}{\Delta}\right)$$

with $C_4 = C_2 C_3^2$. Now together with (5), we have that on the event $E_2(\Delta, n, \delta)$,

$$\sup_{f \in \mathcal{F}_\Delta} \left|\sum_{i=1}^n f(X_i, Y_i)\right| \le C_4 \sqrt{n\Delta d \log\left(\frac{1}{\Delta}\right)} + C_4 d \log\left(\frac{1}{\Delta}\right) + \sqrt{2\nu \ln\left(\frac{2}{\delta}\right)} + \frac{1}{3} \ln\left(\frac{2}{\delta}\right), \quad (10)$$

where

$$\nu = n\Delta + 2\mathbb{E}\left[\sup_{f \in \mathcal{F}_\Delta} \left|\sum_{i=1}^n f(X_i, Y_i)\right|\right] \le n\Delta + 2C_4 \sqrt{n\Delta d \log\left(\frac{1}{\Delta}\right)} + 2C_4 d \log\left(\frac{1}{\Delta}\right). \quad (11)$$

Plugging (11) into (10) and relaxing to simplify the expression yields that, on the event $E_2(\Delta, n, \delta)$,

$$\sup_{f \in \mathcal{F}_\Delta} \left|\sum_{i=1}^n f(X_i, Y_i)\right| \le C_5 \sqrt{n\Delta \left(d \log\left(\frac{1}{\Delta}\right) + \log\left(\frac{1}{\delta}\right)\right)} + C_5 \left(d \log\left(\frac{1}{\Delta}\right) + \log\left(\frac{1}{\delta}\right)\right)$$

$$(12)$$

for some universal constant C_5. Note that for any $h, g \in \mathcal{H}$, the following hold:

$$\mathbb{E}\left[(\ell_{h,g}(X, Y))\right] = \text{er}_P(h) - \text{er}_P(h),$$

$$\mathbb{E}\left[(\ell_{h,g}(X, Y))^2\right] = P_{\mathcal{X}}(x : h(x) \ne g(x)) = \sigma^2(h, g).$$

Hence, the definition of $\mathcal{F}_\Delta$ implies that for any $h, g \in \mathcal{H}$, $(\ell_{h,g} - \mathbb{E}[\ell_{h,g}(X, Y)])/2 \in \mathcal{F}_\Delta$ if and only if $\sigma^2(h, g) \le 4\Delta$, and thus

$$\sup_{h,g \in \mathcal{H}:\sigma^2(h,g) \le 4\Delta} \left|\left(\text{er}_P(h) - \text{er}_P(g)\right) - \left(\hat{\text{er}}_{S_n}(h) - \hat{\text{er}}_{S_n}(g)\right)\right|$$

$$= \sup_{h,g \in \mathcal{H}:\sigma^2(h,g) \le 4\Delta} \left|\frac{1}{n}\sum_{i=1}^n \left(\ell_{h,g}(X_i, Y_i) - \mathbb{E}\left[\ell_{h,g}(X, Y)\right]\right)\right| = \frac{2}{n} \sup_{f \in \mathcal{F}_\Delta} \left|\sum_{i=1}^n f(X_i, Y_i)\right|.$$

Together with (12), this has the immediate implication that, on the event $E_2(\Delta, n, \delta)$,

$$\sup_{h,g \in \mathcal{H}:\sigma^2(h,g) \le 4\Delta} \left|\left(\text{er}_P(h) - \text{er}_P(g)\right) - \left(\hat{\text{er}}_{S_n}(h) - \hat{\text{er}}_{S_n}(g)\right)\right|$$

$$\le 2C_5 \sqrt{\frac{\Delta}{n}\left(d \log\left(\frac{1}{\Delta}\right) + \log\left(\frac{1}{\delta}\right)\right)} + \frac{2C_5}{n}\left(d \log\left(\frac{1}{\Delta}\right) + \log\left(\frac{1}{\delta}\right)\right). \quad (13)$$

23

To complete the proof, we apply the above inequality (13) with various levels Δ, so that it holds for each $h, g \in \mathcal{H}$ with a value of $\Delta \approx \sigma^2(h,g)/4$. In more details, fixing any $\delta \in (0,1)$, for any $i \in \{1, \ldots, \lceil \log_2(n/d) \rceil - 1\}$, we have that on the event $E_2(2^{-i-1}, n, 2^{-i}\delta)$, every pair of $h, g \in \mathcal{H}$ with $2^{-i} < \sigma^2(h,g) \le 2^{1-i}$ satisfy

$$
\big| (\mathrm{er}_P(h) - \mathrm{er}_P(g)) - (\hat{\mathrm{er}}_{S_n}(h) - \hat{\mathrm{er}}_{S_n}(g)) \big|
$$

$$
\overset{(13)}{\le} 2C_5 \sqrt{\frac{2^{-i-1}}{n} \left(d \log \left(2^{i+1} \right) + \log \left(\frac{2^i}{\delta} \right) \right)} + \frac{2C_5}{n} \left(d \log \left(2^{i+1} \right) + \log \left(\frac{2^i}{\delta} \right) \right)
$$

$$
\le 2C_5 \sqrt{\frac{\sigma^2(h,g)}{2n} \left(d \log \left(\frac{4}{\sigma^2(h,g)} \right) + \log \left(\frac{2}{\sigma^2(h,g)\delta} \right) \right)}
$$

$$
+ \frac{2C_5}{n} \left(d \log \left(\frac{4}{\sigma^2(h,g)} \right) + \log \left(\frac{2}{\sigma^2(h,g)\delta} \right) \right)
$$

$$
\le C_6 \sqrt{\frac{\sigma^2(h,g)}{n} \left(d \log \left(\frac{1}{\sigma^2(h,g)} \right) + \log \left(\frac{1}{\delta} \right) \right)} + \frac{C_6}{n} \left(d \log \left(\frac{1}{\sigma^2(h,g)} \right) + \log \left(\frac{1}{\delta} \right) \right)
$$

$$
\tag{14}
$$

for some universal constant C_6. Moreover, when $i = \lceil \log_2(n/d) \rceil$, on the event $E_2(2^{-i-1}, n, 2^{-i}\delta)$, (13) implies that every pair of $h, g \in \mathcal{H}$ with $\sigma^2(h,g) \le 2^{1-i}$ satisfy

$$
\big| (\mathrm{er}_P(h) - \mathrm{er}_P(g)) - (\hat{\mathrm{er}}_{S_n}(h) - \hat{\mathrm{er}}_{S_n}(g)) \big|
$$

$$
\le 2C_5 \sqrt{\frac{2^{-i-1}}{n} \left(d \log \left(2^{i+1} \right) + \log \left(\frac{2^i}{\delta} \right) \right)} + \frac{2C_5}{n} \left(d \log \left(2^{i+1} \right) + \log \left(\frac{2^i}{\delta} \right) \right)
$$

$$
\le 2C_5 \sqrt{\frac{d}{2n^2} \left(d \log \left(\frac{4n}{d} \right) + \log \left(\frac{2n}{d\delta} \right) \right)} + \frac{2C_5}{n} \left(d \log \left(\frac{4n}{d} \right) + \log \left(\frac{2n}{d\delta} \right) \right)
$$

$$
\le \left(1 + \frac{1}{\sqrt{2}} \right) \frac{2C_5}{n} \left(d \log \left(\frac{4n}{d} \right) + \log \left(\frac{2n}{d\delta} \right) \right) \le \frac{C_7}{n} \left(d \log \left(\frac{n}{d} \right) + \log \left(\frac{1}{\delta} \right) \right) \tag{15}
$$

for some universal constant C_7. Note that every pair of $h, g \in \mathcal{H}$ either satisfy $2^{-i} < \sigma^2(h,g) \le 2^{1-i}$ for some $i \in \{1, \ldots, \lceil \log_2(n/d) \rceil - 1\}$, and hence also $\sigma^2(h,g) > d/n$, or else satisfy $\sigma^2(h,g) \le 2^{1-i}$ for $i = \lceil \log_2(n/d) \rceil$. Thus, altogether from (14) and (15) we have that, on the following event

$$
E_2(n, \delta) := \bigcap_{i=1}^{\lceil \log_2(n/d) \rceil} E_2(2^{-i-1}, n, 2^{-i}\delta),
$$

every pair of $h, g \in \mathcal{H}$ satisfy

$$
\big| (\mathrm{er}_P(h) - \mathrm{er}_P(g)) - (\hat{\mathrm{er}}_{S_n}(h) - \hat{\mathrm{er}}_{S_n}(g)) \big|
$$

$$
\le C_8 \sqrt{\frac{\sigma^2(h,g)}{n} \left(d \log \left(\frac{1}{\sigma^2(h,g)} \wedge \frac{n}{d} \right) + \log \left(\frac{1}{\delta} \right) \right)} + \frac{C_8}{n} \left(d \log \left(\frac{1}{\sigma^2(h,g)} \wedge \frac{n}{d} \right) + \log \left(\frac{1}{\delta} \right) \right)
$$

for some universal constant C_8. Finally, by union bound, the event $E_2(n, \delta)$ holds with probability at least $1 - \sum_{i=1}^{\lceil \log_2(n/d) \rceil} 2^{-i}\delta \ge 1 - \delta$, which completes the proof. ∎

Before proceeding to the proof of the super-root rate upper bound for VC classes (Lemma 20), we introduce certain useful notations and give a refined version of the lemma. For any concept class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) < \infty$, any $\epsilon \geq 0$, and any distribution P on $\mathcal{X} \times \{0, 1\}$, we define

$$\mathcal{H}(\epsilon; P) = \left\{ h \in \mathcal{H} : \mathrm{er}_P(h) - \inf_{h' \in \mathcal{H}} \mathrm{er}_P(h') \leq \epsilon \right\} \tag{16}$$

and also

$$\sigma_\epsilon^2(\mathcal{H}, P) = \sup_{h \in \mathcal{H}(\epsilon; P)} \lim_{\tau \to 0} \inf_{h_\tau \in \mathcal{H}(\tau; P)} P_\mathcal{X}(x : h(x) \neq h_\tau(x)). \tag{17}$$

The following lemma states that $\sigma_\epsilon^2(\mathcal{H}, P)$ defined in (17) converges to 0 as $\epsilon \to 0$ at some rate which is probably dependent on $\mathcal{H}$, P and the rate of decay of ϵ. A specific form of such a convergence rate might be of independent interest, while showing it converges is enough for our purposes.

Lemma 41 *For any concept class $\mathcal{H}$ and distribution P on $\mathcal{X} \times \{0, 1\}$, if $\mathcal{H}$ is totally bounded in $L_1(P_\mathcal{X})$, then we have*

$$\lim_{\epsilon \to 0} \sigma_\epsilon^2(\mathcal{H}, P) = 0.$$

Proof [Proof of Lemma 41] The proof uses a similar argument as in the proof of Lemma 24. Since $\sigma_\epsilon^2(\mathcal{H}, P)$ is non-decreasing in ϵ, it suffices to show that there exists a sequence $\epsilon_i \to 0$ with $\sigma_{\epsilon_i}^2(\mathcal{H}, P) \to 0$. For each $j \in \mathbb{N}$, let $h_j \in \mathcal{H}(2^{-j}; P)$ satisfy

$$\lim_{\tau \to 0} \inf_{h_\tau \in \mathcal{H}(\tau; P)} P_\mathcal{X}(x : h_j(x) \neq h_\tau(x)) \geq \sigma_{2^{-j}}^2(\mathcal{H}, P)/2. \tag{18}$$

Since $\mathcal{H}$ is totally bounded in $L_1(P_\mathcal{X})$, there exists an increasing sequence $\{j_i\}_{i \in \mathbb{N}}$ such that $\{h_{j_i}\}_{i \in \mathbb{N}}$ is a Cauchy sequence in the $L_1(P_\mathcal{X})$ pseudo-metric, that is,

$$\lim_{i \to \infty} \sup_{i' > i} P_\mathcal{X}(x : h_{j_i}(x) \neq h_{j_{i'}}(x)) = 0. \tag{19}$$

Then we have

$$\lim_{i \to \infty} \sigma_{2^{-j_i}}^2(\mathcal{H}, P) \overset{(18)}{\leq} \lim_{i \to \infty} 2 \lim_{\tau \to 0} \inf_{h_\tau \in \mathcal{H}(\tau; P)} P_\mathcal{X}(x : h_{j_i}(x) \neq h_\tau(x))$$

$$\leq 2 \lim_{i \to \infty} \sup_{i' > i} P_\mathcal{X}(x : h_{j_i}(x) \neq h_{j_{i'}}(x)) \overset{(19)}{=} 0,$$

where the second inequality follows from the fact that, for any $\tau > 0$, all sufficiently large i' satisfy $h_{j_{i'}} \in \mathcal{H}(\tau; P)$. Together with the fact that $\sigma_\epsilon^2(\mathcal{H}, P) \geq 0$ for all $\epsilon > 0$, the lemma follows. ∎

Next, for any $n \in \mathbb{N}$, we define

$$\mathcal{B}_\epsilon(n; \mathcal{H}, P) = \tilde{c} \sqrt{\sigma_\epsilon^2(\mathcal{H}, P) \frac{\mathrm{VC}(\mathcal{H})}{n} \log \left(\frac{1}{\sigma_\epsilon^2(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right)} + \tilde{c} \frac{\mathrm{VC}(\mathcal{H})}{n} \log \left(\frac{1}{\sigma_\epsilon^2(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right),$$

where $\tilde{c}$ is a universal constant. Following classic localization arguments (Bartlett et al., 2004, 2005; Koltchinskii, 2006), we define

$$\epsilon_n(\mathcal{H}, P) = \inf \left\{ \epsilon := \epsilon(n) > 0 : \mathcal{B}_\epsilon(n; \mathcal{H}, P) \leq 2\epsilon \right\}. \tag{20}$$

To prove Lemma 39, it suffices to prove the following lemma, which describes a detailed form of the $o(n^{-1/2})$ rate of decay.

Lemma 42 (Refined Lemma 39) *Let $\mathcal{H}$ be any measurable concept class with $VC(\mathcal{H}) < \infty$. For any distribution P, we define*

$$\phi_{\text{total}}(n; \mathcal{H}, P) = \min \left\{ c \sqrt{\frac{VC(\mathcal{H})}{n}}, \epsilon_n(\mathcal{H}, P) \right\}$$

for some universal constant c (described in the proof). Letting $S_n := \{(x_i, y_i)\}_{i=1}^n \sim P^n$, we have

$$\mathcal{E}(n, P) \leq \phi_{\text{total}}(n; \mathcal{H}, P).$$

Moreover, as a function of n, $\phi_{\text{total}}(n; \mathcal{H}, P) = o(n^{-1/2})$.

Remark 43 *Consider $\phi_{\text{total}}(n; \mathcal{H}, P)$ as a function of n, we note that $\epsilon_n(\mathcal{H}, P)$ decays roughly at the same rate as $\mathcal{B}_{\epsilon_n}(n; \mathcal{H}, P)$, where the dominant term is $\sqrt{\sigma_{\epsilon_n}^2(\mathcal{H}, P)/n}$. According to Lemma 41, we know that $\sigma_{\epsilon_n}^2(\mathcal{H}, P)$ decays as $n \to \infty$, which makes $\epsilon_n(\mathcal{H}, P) = o(n^{-1/2})$.*

Proof [Proof of Lemma 42] The proof combines Proposition 40 with a well-known localization argument. However, while traditional localization arguments would aim to bound the excess risk by an approximate fixed point, this will not be necessary for our purposes, and hence the argument here is somewhat simpler than one aiming for sharp localized excess risk bounds (see e.g., Bartlett et al., 2004, 2005; Koltchinskii, 2006). We will again suppose that $n \geq 2VC(\mathcal{H})$ to focus on the non-vacuous case. By the classic result of Talagrand (1994), for any $\delta \in (0, 1)$, on an event $E_0(n, \delta)$ of probability at least $1 - \delta$, we have

$$\sup_{h \in \mathcal{H}} \left| \hat{\text{er}}_{S_n}(h) - \text{er}_P(h) \right| \leq C_0 \sqrt{\frac{1}{n} \left(VC(\mathcal{H}) + \log \left(\frac{1}{\delta} \right) \right)}, \tag{21}$$

where C_0 is some universal constant. In particular, on the event $E_0(n, \delta)$, by applying (21) for both $\hat{h}_n$ and $\inf_{h \in \mathcal{H}} \text{er}_P(h)$, we further have

$$\text{er}_P(\hat{h}_n) - \inf_{h \in \mathcal{H}} \text{er}_P(h) \leq 2C_0 \sqrt{\frac{1}{n} \left(VC(\mathcal{H}) + \log \left(\frac{1}{\delta} \right) \right)}. \tag{22}$$

We will use this fact to derive two implications. The first one is classical, namely, by setting the RHS of (22) to ϵ and solving for δ, we can rewrite this same inequality as follow:

$$\mathbb{P} \left(\text{er}_P(\hat{h}_n) - \inf_{h \in \mathcal{H}} \text{er}_P(h) > \epsilon \right) \leq \min \left\{ 1, e^{VC(\mathcal{H}) - n\epsilon^2/4C_0^2} \right\}.$$

It follows immediately that

$$\mathcal{E}(n, P) = \int_0^\infty \mathbb{P} \left(\text{er}_P(\hat{h}_n) - \inf_{h \in \mathcal{H}} \text{er}_P(h) > \epsilon \right) \mathrm{d}\epsilon$$

$$\leq \int_0^\infty \min \left\{ 1, e^{VC(\mathcal{H}) - n\epsilon^2/4C_0^2} \right\} \mathrm{d}\epsilon \lesssim \sqrt{\frac{VC(\mathcal{H})}{n}}. \tag{23}$$

For a second implication, we recall the fact that every VC class satisfies the UGC property (Definition 8) and thus is totally bounded in $L_1(P_\mathcal{X})$ (e.g. Haussler, 1995). Let $d \in \mathbb{N}$ be such that

$VC(\mathcal{H}) \le d$, and let $h^* : \mathcal{X} \to \{0, 1\}$ be any target function, i.e. $er_P(h^*) = \inf_{h \in \mathcal{H}} er_P(h)$. For any measurable $f : \mathcal{X} \times \{0, 1\} \to \mathbb{R}$, we denote by $\hat{P}_n(f) := \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i)$ and $P(f) := \mathbb{E}[f(X, Y)]$. For any $h \in \mathcal{H}$ and $(x, y) \in \mathcal{X} \times \{0, 1\}$, we define

$$\ell_h(x, y) := (\mathbb{1}(h(x) \ne y) - \mathbb{1}(h^*(x) \ne y)),$$

and $\mathcal{F} := \{\ell_h : h \in \mathcal{H}\}$, which is a measurable class of functions. Note that the class $\{(x, y) \mapsto \mathbb{1}(h(x) \ne y) : h \in \mathcal{H}\}$ is a measurable class of functions with VC dimension at most d. Fix any integer $n \ge d$, by applying the classic uniform convergence result of Talagrand (1994) to this class, we obtain that there is a universal constant c such that, for any $\delta \in (0, 1)$, there is an event $E_1(n, \delta)$ of probability at least $1 - \delta/2$, on which the following holds:

$$\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \le c \sqrt{\frac{1}{n} \left(d + \ln \left(\frac{1}{\delta} \right) \right)}. \tag{24}$$

Similarly, we define $\epsilon_n(\delta) := 2c\sqrt{(d + \ln(1/\delta))/n}$. Note that for any $h \in \mathcal{H}$,

$$P(\ell_h) = \mathbb{E}\left[\ell_h(X, Y)\right] = \mathbb{E}\left[\mathbb{1}(h(X) \ne Y) - \mathbb{1}(h^*(X) \ne Y)\right] = er_P(h) - \inf_{h' \in \mathcal{H}} er_P(h'),$$

and hence we have from (16) that $\mathcal{H}(\epsilon_n(\delta); P) = \{h \in \mathcal{H} : P(\ell_h) \le \epsilon_n(\delta)\}$. The remaining part of the proof aims to establish the following two claims:

(1) that $\hat{h}_n \in \mathcal{H}(\epsilon_n(\delta); P)$ on the event $E_1(n, \delta)$.

(2) that the differences $|\hat{P}_n(\ell_h) - P(\ell_h)|$ are uniformly bounded for all $h \in \mathcal{H}(\epsilon_n(\delta); P)$ by a function of $\epsilon_n(\delta)$ and P that is $o(n^{-1/2})$, on a further event.

Then together with a tail integration bound, these two claims will establish the second implication. The first claim follows from the classic argument of Vapnik and Chervonenkis (1974), namely, on $E_1(n, \delta)$,

$$P(\ell_{\hat{h}_n}) = er_P(\hat{h}_n) - er_P(h^*) = er_P(h) - \inf_{h \in \mathcal{H}} er_P(h)$$

$$\overset{(24)}{\le} \hat{er}_{S_n}(\hat{h}_n) - \min_{h \in \mathcal{H}} \hat{er}_{S_n}(h) + \epsilon_n(\delta) = \epsilon_n(\delta).$$

To show the second claim, we will apply Proposition 40. Note that for any $h \in \mathcal{H}(\epsilon_n(\delta); P)$, any $\tau > 0$ and $h_\tau \in \mathcal{H}(\tau; P)$, we have that on an event $E_2(n, \delta)$ of probability at least $1 - \delta/2$,

$$\left| \hat{P}_n(\ell_h) - P(\ell_h) \right| = \left| (er_P(h) - er_P(h^*)) - (\hat{er}_{S_n}(h) - \hat{er}_{S_n}(h^*)) \right|$$

$$\le \left| (er_P(h) - er_P(h_\tau)) - (\hat{er}_{S_n}(h) - \hat{er}_{S_n}(h_\tau)) \right| + \left| er_P(h_\tau) - er_P(h^*) + \hat{er}_{S_n}(h^*) - \hat{er}_{S_n}(h_\tau) \right|$$

$$\overset{\text{Proposition 40}}{\le} \sqrt{\sigma^2(h, h_\tau) \frac{c_0}{n} \left(VC(\mathcal{H}) \log \left(\frac{1}{\sigma^2(h, h_\tau)} \wedge \frac{n}{VC(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right)}$$

$$+ \frac{c_0}{n} \left(VC(\mathcal{H}) \log \left(\frac{1}{\sigma^2(h, h_\tau)} \wedge \frac{n}{VC(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right)$$

$$+ \left| er_P(h_\tau) - er_P(h^*) + \hat{er}_{S_n}(h^*) - \hat{er}_{S_n}(h_\tau) \right|.$$

Since the above holds for any $\tau > 0$ and any $h_\tau \in \mathcal{H}(\tau; P)$,

$$\sup_{h \in \mathcal{H}(\epsilon_n(\delta); P)} \left| \hat{P}_n(\ell_h) - P(\ell_h) \right|$$

$$\leq \sup_{h \in \mathcal{H}(\epsilon_n(\delta); P)} \lim_{\tau \to 0} \inf_{h_\tau \in \mathcal{H}(\tau; P)} \left\{ \sqrt{\sigma^2(h, h_\tau) \frac{c_0}{n} \left(\mathrm{VC}(\mathcal{H}) \log \left(\frac{1}{\sigma^2(h, h_\tau)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right)} \right.$$

$$+ \frac{c_0}{n} \left(\mathrm{VC}(\mathcal{H}) \log \left(\frac{1}{\sigma^2(h, h_\tau)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right)$$

$$\left. + \left| \mathrm{er}_P(h_\tau) - \mathrm{er}_P(h^*) + \hat{\mathrm{er}}_{S_n}(h^*) - \hat{\mathrm{er}}_{S_n}(h_\tau) \right| \right\}$$

$$\overset{(17)}{=} \sqrt{\sigma^2_{\epsilon_n(\delta)}(\mathcal{H}, P) \frac{c_0}{n} \left(\mathrm{VC}(\mathcal{H}) \log \left(\frac{1}{\sigma^2_{\epsilon_n(\delta)}(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right)}$$

$$+ \frac{c_0}{n} \left(\mathrm{VC}(\mathcal{H}) \log \left(\frac{1}{\sigma^2_{\epsilon_n(\delta)}(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right). \tag{25}$$

To see that the RHS of (25) is $o(n^{-1/2})$, we note that $\epsilon_n(\delta)$ decays as $n \to \infty$. Furthermore, Lemma 41 states that $\sigma^2_\epsilon(\mathcal{H}, P) \to 0$ as $\epsilon \to 0$. Therefore, $\sigma^2_{\epsilon_n(\delta)}(\mathcal{H}, P)$ decays at some rate as a function of n, which makes the RHS of (25) faster than $n^{-1/2}$. Now we have

$$\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) = P(\ell_{\hat{h}_n})$$

$$\leq \left| \hat{P}_n(\ell_{\hat{h}_n}) - P(\ell_{\hat{h}_n}) \right| + \hat{P}_n(\ell_{\hat{h}_n})$$

$$= \left| \hat{P}_n(\ell_{\hat{h}_n}) - P(\ell_{\hat{h}_n}) \right| + \hat{\mathrm{er}}_{S_n}(\hat{h}_n) - \hat{\mathrm{er}}_{S_n}(h^*)$$

$$= \left| \hat{P}_n(\ell_{\hat{h}_n}) - P(\ell_{\hat{h}_n}) \right| + \min_{h \in \mathcal{H}(\epsilon_n(\delta); P)} \hat{\mathrm{er}}_{S_n}(h) - \hat{\mathrm{er}}_{S_n}(h^*)$$

$$= \left| \hat{P}_n(\ell_{\hat{h}_n}) - P(\ell_{\hat{h}_n}) \right| + \min_{h \in \mathcal{H}(\epsilon_n(\delta); P)} \hat{P}_n(\ell_h)$$

$$\leq \left| \hat{P}_n(\ell_{\hat{h}_n}) - P(\ell_{\hat{h}_n}) \right| + \min_{h \in \mathcal{H}(\epsilon_n(\delta); P)} \left\{ \left| \hat{P}_n(\ell_h) - P(\ell_h) \right| + P(\ell_h) \right\}$$

$$\leq 2 \sqrt{\sigma^2_{\epsilon_n(\delta)}(\mathcal{H}, P) \frac{c_0}{n} \left(\mathrm{VC}(\mathcal{H}) \log \left(\frac{1}{\sigma^2_{\epsilon_n(\delta)}(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right)}$$

$$+ \frac{2c_0}{n} \left(\mathrm{VC}(\mathcal{H}) \log \left(\frac{1}{\sigma^2_{\epsilon_n(\delta)}(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right) + \log \left(\frac{2}{\delta} \right) \right),$$

on the event $E_1(n, \delta) \cap E_2(n, \delta)$ of probability at least $1 - \delta$. Then, a tail integration bound yields

$$\mathcal{E}(n, P) = \int_0^\infty \mathbb{P} \left(\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) > \epsilon \right) \mathrm{d}\epsilon$$

$$\leq \frac{\tilde{c}}{2} \sqrt{\sigma^2_{\epsilon_n}(\mathcal{H}, P) \frac{\mathrm{VC}(\mathcal{H})}{n} \log \left(\frac{1}{\sigma^2_{\epsilon_n}(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right)}$$

$$+ \frac{\tilde{c}}{2} \cdot \frac{\mathrm{VC}(\mathcal{H})}{n} \log \left(\frac{1}{\sigma^2_{\epsilon_n}(\mathcal{H}, P)} \wedge \frac{n}{\mathrm{VC}(\mathcal{H})} \right) =: \frac{\mathcal{B}_{\epsilon_n}(n; \mathcal{H}, P)}{2} \leq \epsilon_n(\mathcal{H}, P), \tag{26}$$

where $\epsilon_n = \theta(\sqrt{d/n})$ and $\tilde{c}$ is a universal constant. Altogether, (23) and (26) complete the proof. ∎

Lemma 44 (Lemma 21 restated) *If $\mathcal{H}$ has an infinite VC-eluder sequence centered at h^*, then h^* requires at least arbitrarily slow rates to be agnostically universally learned by ERM.*

Proof [Proof of Lemma 44] Let $\{(x_1, y_1), (x_2, y_2), \ldots\}$ be an infinite VC-eluder sequence in $\mathcal{H}$ centered at h^*. For notation simplicity, we denote $\mathcal{X}_k := \{x_{n_k+1}, \ldots, x_{n_k+k}\}$ for every integer $k \geq 1$, where $\{n_k\}_{k \geq 1}$ is defined in Definition 12. According to the definition of the VC-eluder sequence, we know that for any $k \geq 1$, $\mathcal{X}_k$ is a shattered set of $\{h \in \mathcal{H} : h(x_i) = h^*(x_i), \forall i \leq n_k\}$. In other words, for any $k \geq 1$ and any label vector $\boldsymbol{y}_k \in \{0, 1\}^k$, there exists $h_{k, \boldsymbol{y}_k} \in \mathcal{H}$ such that $h_{k, \boldsymbol{y}_k}(x_i) = h^*(x_i)$ for all $i \leq n_k$ and $h_{k, \boldsymbol{y}_k}((x_{n_k+1}, \ldots, x_{n_k+k})) = \boldsymbol{y}_k$. We then consider the following distribution P defined on the VC-eluder sequence:

$$P_{\mathcal{X}}(x \in \mathcal{X}_k) = p_k, \quad P_{\mathcal{X}}(x) = p_k/k, \quad \forall x \in \mathcal{X}_k;$$
$$P(y = h^*(x)|x) = 1/2 + \epsilon_k, \quad P(y = 1 - h^*(x)|x) = 1/2 - \epsilon_k, \quad \forall x \in \mathcal{X}_k, \forall k \geq 1,$$

where $\{p_k\}_{k \geq 1}$ is a sequence of probabilities satisfying $\sum_{k \geq 1} p_k \leq 1$ and $\{\epsilon_k\}_{k \geq 1}$ is a sequence of positive reals. We would like the constructed distribution P to be centered at h^*. To this end, we first note that for any $h \in \mathcal{H}$, it holds $\text{er}_P(h) \leq \text{er}_P(h^*)$. Moreover, for any $k \geq 1$, we consider a label vector $\boldsymbol{y}_k \in \{0, 1\}^k$ with every coordinate does not match h^*. Following the same argument used in the proof of Lemma 17, suppose that $\sum_{j \geq t+1} p_{k_j} \epsilon_{k_j} \leq p_{k_t} \epsilon_{k_t}$ for all $t \geq 1$ on some increasing subsequence $\{k_t\}_{t \geq 1}$, then we have $\text{er}_P(h_{k_{t+1}, \boldsymbol{y}_{k_{t+1}}}) \leq \text{er}_P(h_{k_t, \boldsymbol{y}_{k_t}})$, and thus $\inf_{h \in \mathcal{H}} \text{er}_P(h) = \lim_{t \to \infty} \text{er}_P(h_{k_t, \boldsymbol{y}_{k_t}}) = \lim_{k \to \infty} \text{er}_P(h_{k, \boldsymbol{y}_k}) = \text{er}_P(h^*)$. Furthermore, to have $\inf_{h \in \mathcal{H}} P_{\mathcal{X}}(x : h(x) \neq h^*(x)) = \lim_{k \to \infty} p_k/k = 0$, it suffices to assume further $\{p_k\}_{k \geq 1}$ is decreasing.

Now let $R(n) \to 0$ be any rate function, and we aim to show that $\mathcal{H}$ cannot be agnostically universally learned at rate faster than $R(n)$ under the distribution P by ERM. To this end, we let $S_n := \{(x_i, y_i)\}_{i=1}^n \sim P^n$. For any $k \geq 1$ and any $l \in [k]$, we consider the following event:

$$\Lambda_{k,l}(S_n) := \left\{ \sum_{i=1}^n \mathbb{1}(x_i \in \mathcal{X}_{>k}) = 0 \text{ and } \sum_{i=1}^n \mathbb{1}(x_i = x_{n_k+l}) = 0 \right\},$$

that is, the dataset S_n contains no copy of the points in $\mathcal{X}_{>k} \cup \{x_{n_k+l}\}$. Note that if $\Lambda_{k,l}(S_n)$ holds, then a bad ERM algorithm will have an excess risk $2p_k \epsilon_k/k$. Moreover, the probability of such an event is not hard to bound:

$$\mathbb{P}(\Lambda_{k,l}(S_n)) = \mathbb{P}\left(\left\{x_i \in \bigcup_{j \leq k} \mathcal{X}_j \setminus \{x_{n_k+l}\}, \forall i \in [n]\right\}\right) = \left(1 - \sum_{j > k} p_j - \frac{p_k}{k}\right)^n.$$

We simply choose $\epsilon_k = 1/4$ for all $k \geq 1$, and then the next key step is to show there is a sequence of integers $\{k_t\}_{t \geq 1}$ such that all the following hold with some universal constant $1/2 \leq C \leq 1$:

(1) $\sum_{k \geq 1} p_k \leq 1$, $\{p_k\}_{k \geq 1} \downarrow$, $\sum_{j \geq t+1} p_{k_j} \epsilon_{k_j} \leq p_{k_t} \epsilon_{k_t}$ for all $t \geq 1$;

(2) $\sum_{j > k_t} p_j \leq 1/n_t$ for all $t \geq 1$;

(3) $n_t p_{k_t} \leq k_t$ for all $t \geq 1$;

(4) $p_{k_t} = CR(n_t)$ for all $t \geq 1$.

By Lemma 58, we know that there exists a sequence of probabilities $\{p_k\}_{k \geq 1}$ satisfying $\sum_{k \geq 1} p_k = 1$ and two sequences of integers $\{k_t\}_{t \geq 1}$ and $\{n_t\}_{t \geq 1}$ such that the above (2-4) hold. We consider any rate function $R(n)$ that decays slower than $1/n$ (note that if this holds, then it also holds for even faster $R(n)$ since we are showing a lower bound), and immediately get (1), i.e.

$$\sum_{j \geq t+1} p_{k_j} \leq \sum_{j \geq k_{t+1}} p_j \leq \sum_{j > k_t} p_j \overset{(2)}{\leq} \frac{1}{n_t} \leq CR(n_t) = p_{k_t}.$$

Finally, we have for all $t \geq 1$ (that is, for infinitely many $n \in \mathbb{N}$),

$$\mathcal{E}(n_t, P) \geq \sum_{l \in [k_t]} \frac{2 p_{k_t} \epsilon_{k_t}}{k_t} \left(1 - \sum_{j > k_t} p_j - \frac{p_{k_t}}{k_t} \right)^{n_t} \overset{(2-4)}{\geq} \frac{C}{2} \left(1 - \frac{2}{n_t} \right)^{n_t} R(n_t) \geq \frac{R(n_t)}{36}.$$

∎

Appendix C. Missing proofs from Section 4

Lemma 45 (Lemma 22 restated) *Let $\mathcal{H}$ be a concept class satisfying the UGC property (see Definition 8). If the Condition 1 holds for h^*, then h^* is agnostically universally learnable at rate e^{-n} by ERM.*

Proof [Proof of Lemma 45] Let P be any distribution centered at h^*. Since Condition 1 holds, we can define

$$\epsilon_0 := \inf_{h \in \mathcal{H}: er_P(h) > er_P(h^*)} \{ er_P(h) - er_P(h^*) \} > 0,$$

which is a distribution-dependent constant. Furthermore, since $\mathcal{H}$ is a UGC class, we know that there exists a distribution-dependent integer $n_0 := n_0(P) < \infty$ such that

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \right] \leq \epsilon_0/4, \quad \forall n \geq n_0. \tag{27}$$

Assume that $n \geq n_0$, then Lemma 59 implies that with probability of at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \right] + \sqrt{\frac{1}{2n} \ln \left(\frac{4}{\delta} \right)} \overset{(27)}{\leq} \frac{\epsilon_0}{4} + \sqrt{\frac{1}{2n} \ln \left(\frac{4}{\delta} \right)}.$$

By choosing $\delta := 4 e^{-n \epsilon_0^2/8}$, we have for any $n \geq n_0$, with probability of at least $1 - 4 e^{-n \epsilon_0^2/8}$,

$$\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \leq \epsilon_0/2. \tag{28}$$

Since $\mathcal{H}$ does not have an infinite eluder sequence centered at h^* (based on Condition 1), we know by Lemma 23 that there exists $h \in \mathcal{H}$ such that $P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0$. Now note that if the inequality in (28) holds, we will have

$$\mathrm{er}_P(\hat{h}_n) \leq \hat{\mathrm{er}}_{S_n}(\hat{h}_n) + \epsilon_0/2 \leq \hat{\mathrm{er}}_{S_n}(h) + \epsilon_0/2 \leq \mathrm{er}_P(h) + \epsilon_0 = \mathrm{er}_P(h^*) + \epsilon_0.$$

Therefore, it follows that for all $n \geq n_0$,

$$\mathcal{E}(n, P) \leq \mathbb{P}\left(\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) \geq \epsilon_0\right) \leq \mathbb{P}\left(\sup_{h \in \mathcal{H}} \left|\hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h)\right| > \epsilon_0/2\right) \leq 4e^{-n\epsilon_0^2/8}.$$

Finally, we obtain $\mathcal{E}(n, P) \leq \max\{4, e^{n_0\epsilon_0^2/8}\} \cdot e^{-n\epsilon_0^2/8}$ for all $n \in \mathbb{N}$. ∎

Lemma 46 (Lemma 23 restated) *If $\mathcal{H}$ does not have an infinite eluder sequence centered at h^*, then for any distribution P centered at h^*, let $P_{\mathcal{X}}$ be the associated marginal distribution, the following hold:*
  (1) There exists $h \in \mathcal{H}$ such that $P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0$.
  (2) $\inf_{h \in \mathcal{H}: P_{\mathcal{X}}\{x:h(x) \neq h^*(x)\} > 0} P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} > 0$.

Proof [Proof of Lemma 46] Since P is centered at h^*, $\inf_{h \in \mathcal{H}} P_{\mathcal{X}}(x : h(x) \neq h^*(x)) = 0$. Then for every $k \in \mathbb{N}$, let $r_k := 2^{-k}$ and we know that there exists a concept $h_k \in \mathcal{H}$ such that $P_{\mathcal{X}}(x : h_k(x) \neq h^*(x)) \leq r_k$. Let $\{x_1, x_2, \ldots\} \sim P_{\mathcal{X}}^\infty$, then for every integer $t \geq 1$, a union bound gives

$$\sum_{k=1}^\infty \mathbb{P}\left(h_k(x_s) \neq h^*(x_s) \text{ for some } s \leq t\right) \leq t \sum_{k=1}^\infty P_{\mathcal{X}}(x : h_k(x) \neq h^*(x)) \leq t \sum_{k=1}^\infty r_k < \infty.$$

By the Borel-Cantelli's lemma, we further have

$$\mathbb{P}\left(\lim_{k \to \infty} h_k(x_s) = h^*(x_s), \forall s \leq t\right) = 1, \tag{29}$$

that is, with probability one, there exists for every $t \geq 1$ a concept $h \in \mathcal{H}$ satisfying $h(x_s) = h^*(x_s)$ for all $s \leq t$. We assume first that for any $h \in \mathcal{H}$, $P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} > 0$. Since $\mathcal{H}$ does not have an infinite eluder sequence centered at h^*, then there exists a distribution-dependent integer $\tau := \tau(P_{\mathcal{X}}, h^*) < \infty$ such that for $x_1, \ldots, x_\tau \overset{i.i.d.}{\sim} P_{\mathcal{X}}$,

$$\mathbb{P}\left(\{h \in \mathcal{H} : h(x_i) = h^*(x_i), \forall 1 \leq i \leq \tau\} = \emptyset\right) \geq 1/2,$$

which contradicts (29). Hence, there must exist some $h \in \mathcal{H}$ such that $P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0$.
  To prove the second statement, we assume to the contrary that

$$\inf_{h \in \mathcal{H}: P_{\mathcal{X}}\{x:h(x) \neq h^*(x)\} > 0} P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0,$$

and then directly construct an infinite eluder sequence centered at h^* via the following procedure:
  1. Let $\epsilon_1 > 0$, then there exists $h_1 \in \mathcal{H}$ such that $0 < P_{\mathcal{X}}\{x : h_1(x) \neq h^*(x)\} < \epsilon_1$. We pick any point $x_1 \in \mathrm{DIS}(\{h_1, h^*\})$ such that $P_{\mathcal{X}}(x_1) > 0$.

2. Now we choose $\epsilon_2 = P_{\mathcal{X}}(x_1) > 0$, and know there exists $h_2 \in \mathcal{H}$ such that $0 < P_{\mathcal{X}}\{x : h_2(x) \neq h^*(x)\} < \epsilon_2$. Note that such h_2 must satisfy $h_2(x_1) = h^*(x_1)$, since otherwise we will have $P_{\mathcal{X}}\{x : h_2(x) \neq h^*(x)\} \geq P_{\mathcal{X}}(x_1) = \epsilon_2$. We then pick any point $x_2 \in$ DIS$(\{h_2, h^*\})$ such that $P_{\mathcal{X}}(x_2) > 0$.

3. Similarly, we set $\epsilon_3 = P_{\mathcal{X}}(x_2) > 0$ and find $h_3 \in \mathcal{H}$ such that $0 < P_{\mathcal{X}}\{x : h_3(x) \neq h^*(x)\} < \epsilon_3$. Such h_3 must satisfy $h_3(x_1) = h^*(x_1)$ and $h_3(x_2) = h^*(x_2)$, since otherwise we will reach a contradiction that $P_{\mathcal{X}}\{x : h_3(x) \neq h^*(x)\} \geq \min\{P_{\mathcal{X}}(x_1), P_{\mathcal{X}}(x_2)\} = P_{\mathcal{X}}(x_2) = \epsilon_3$. We then pick any point $x_3 \in$ DIS$(\{h_3, h^*\})$ such that $P_{\mathcal{X}}(x_3) > 0$.

...

Finally, it is clear that $\{(x_1, h^*(x_1)), (x_2, h^*(x_2)), \ldots\}$ is an infinite eluder sequence centered at h^*, witnessed by the concepts $\{h_1, h_2, \ldots\} \subseteq \mathcal{H}$. ∎

Lemma 47 (Lemma 24 restated) *Let $\mathcal{H}$ be any concept class and P be any distribution such that $\mathcal{H}$ is totally bounded in $L_1(P_{\mathcal{X}})$ pseudo-metric. If the Condition 1 fails for P, i.e. the following holds:*

$$\inf_{h\in\mathcal{H}:er_P(h)>\inf_{h'\in\mathcal{H}} er_P(h')} \left\{ er_P(h) - \inf_{h'\in\mathcal{H}} er_P(h') \right\} = 0,$$

then there exists h^* such that $er_P(h^*) = \inf_{h'\in\mathcal{H}} er_P(h')$ and

$$\inf_{h\in\mathcal{H}:P_{\mathcal{X}}\{x:h(x)\neq h^*(x)\}>0} P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0.$$

Proof [Proof of Lemma 47] Based on the assumption as in the lemma statement, we know that there exists a sequence of concepts $\{h_i\}_{i\geq 1} \subseteq \mathcal{H}$ such that

$$er_P(h_i) > \inf_{h'\in\mathcal{H}} er_P(h') \quad \text{and} \quad er_P(h_i) \to \inf_{h'\in\mathcal{H}} er_P(h') \text{ as } i \to \infty.$$

Since $\mathcal{H}$ is totally bounded in $L_1(P_{\mathcal{X}})$ pseudo-metric, every infinite sequence therein admits a Cauchy subsequence. Hence, there exists an increasing sequence of integers $\{i_t\}_{t\geq 1}$ such that $\{h_{i_t}\}_{t\geq 1}$ is a Cauchy sequence, i.e.

$$\lim_{t\to\infty} \sup_{t'>t} P_{\mathcal{X}}\{x : h_{i_t}(x) \neq h_{i_{t'}}(x)\} = 0.$$

We simply choose h^* such that $\lim_{t\to\infty} P_{\mathcal{X}}\{x : h_{i_t}(x) \neq h^*(x)\} = 0$, which also satisfies

$$er_P(h^*) = \lim_{t\to\infty} er_P(h_{i_t}) = \lim_{i\to\infty} er_P(h_i) = \inf_{h'\in\mathcal{H}} er_P(h').$$

∎

Lemma 48 (Lemma 26 restated) *Let $\mathcal{H}$ be any UGC concept class and h^* be any classifier. Then h^* is agnostically universally learnable at rate $o(n^{-1/2})$ by ERM if and only if the Condition 2 holds for h^*.*

Proof [Proof of Lemma 48] We first prove the sufficiency. Suppose that the Condition 2 holds for h^*. For any distribution P centered at h^*, there exists $\epsilon_0 := \epsilon_0(P) > 0$ such that $\mathrm{VC}(\mathcal{H}(\epsilon_0; P, h^*)) < \infty$. Furthermore, since $\mathcal{H}$ is a UGC class, following the same argument in the proof of Lemma 22, we know that there exists a distribution-dependent integer $n_0 := n_0(P) < \infty$ such that for any $n \geq n_0$, with probability of at least $1 - 4e^{-n\epsilon_0^2/8}$, $\sup_{h \in \mathcal{H}} |\hat{\mathrm{er}}_{S_n}(h) - \mathrm{er}_P(h)| \leq \epsilon_0/2$. It follows that

$$\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) \leq \hat{\mathrm{er}}_{S_n}(\hat{h}_n) - \min_{h \in \mathcal{H}} \hat{\mathrm{er}}_{S_n}(h) + \epsilon_0 = \epsilon_0$$

holds on an event $E_1(n)$ with probability of at least $1 - 4e^{-n\epsilon_0^2/8}$ for any $n \geq n_0$. In other words, $\hat{h}_n \in \mathcal{H}(\epsilon_0; P, h^*)$, namely, any ERM algorithm will then output a hypothesis from a VC class. Applying the inequality $\mathbb{P}(E) \leq \mathbb{P}(E|F) + \mathbb{P}(\neg F)$, we have for any $n \geq n_0$,

$$\begin{aligned}
\mathcal{E}(n, P) &= \int_0^\infty \mathbb{P}\left(\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) > \epsilon\right) d\epsilon \\
&\leq \int_0^\infty \left[\mathbb{P}\left(\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) > \epsilon \big| E_1(n)\right) + \mathbb{P}\left(\neg E_1(n)\right)\right] d\epsilon \\
&= \int_0^{\epsilon_0} \mathbb{P}\left(\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) > \epsilon \big| E_1(n)\right) d\epsilon + \int_0^1 \mathbb{P}\left(\neg E_1(n)\right) d\epsilon \\
&\leq \int_0^{\epsilon_0} \mathbb{P}\left(\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) > \epsilon \big| E_1(n)\right) d\epsilon + 4e^{-n\epsilon_0^2/8} \\
&\leq \mathbb{E}\left[\mathrm{er}_P(\hat{h}_n) - \mathrm{er}_P(h^*) \big| E_1(n)\right] + 4e^{-n\epsilon_0^2/8} \overset{\text{Lemma 20}}{=} o(n^{-1/2}),
\end{aligned}$$

where the rate function is dependent on the distribution-dependent constants ϵ_0, $\mathrm{VC}(\mathcal{H}(\epsilon_0; P, h^*))$ and n_0.

To show the necessity, we assume to the contrary that the Condition 2 fails for h^*, i.e. there exists a distribution P centered at h^* such that for any $\epsilon > 0$, $\mathrm{VC}(\mathcal{H}(\epsilon; P, h^*)) = \infty$. Following a similar idea of the proof for Lemma 23, we aim to show there exists an infinite VC-eluder sequence centered at h^*, which will give a contradiction in conjunction with Lemma 21. Such an infinite VC-eluder sequence can be constructed via the following procedure:

1. Let $\epsilon_1 > 0$. We pick a point $x_1 \in \mathrm{DIS}(\mathcal{H}(\epsilon_1; P, h^*))$ shattered by $\mathcal{H}(\epsilon_1; P, h^*) \subseteq \mathcal{H}$ such that

$$\inf_{h \in \mathcal{H}(\epsilon_1; P, h^*): h(x_1) \neq h^*(x_1)} \{\mathrm{er}_P(h) - \mathrm{er}_P(h^*)\} > 0. \tag{30}$$

Note that such an x_1 can actually be any point satisfying $\mathbb{P}(h^*(x) \neq y, x = x_1) < 1/2$.

2. The next step is to find two points $x_2, x_3 \in \mathcal{X}$ shattered by $V_{(x_1, h^*(x_1))}(\mathcal{H}(\epsilon_1; P, h^*))$, and it suffices to show that the version space also has infinite VC dimension. To this end, we show that there exists $0 < \epsilon_2 \leq \epsilon_1$ such that $\mathcal{H}(\epsilon_2; P, h^*) \subseteq V_{(x_1, h^*(x_1))}(\mathcal{H}(\epsilon_1; P, h^*))$, i.e. any $h \in \mathcal{H}(\epsilon_2; P, h^*)$ satisfies $h(x_1) = h^*(x_1)$. This can be realized by setting ϵ_2 to be the LHS of (30) since any $h \in \mathcal{H}(\epsilon_1; P, h^*)$ satisfying $h(x_1) \neq h^*(x_1)$ will have an excess risk larger than ϵ_2. Therefore, it follows $\infty = \mathrm{VC}(\mathcal{H}(\epsilon_2; P, h^*)) \leq \mathrm{VC}(V_{(x_1, h^*(x_1))}(\mathcal{H}(\epsilon_1; P, h^*)))$, and thus such x_2, x_3 must exist. Similarly, we can make the following happen:

$$\inf_{h \in \mathcal{H}(\epsilon_2; P, h^*): \mathrm{DIS}(\{h, h^*\}) \cap \{x_2, x_3\} \neq \emptyset} \{\mathrm{er}_P(h) - \mathrm{er}_P(h^*)\} > 0.$$

3. Find three points $x_4, x_5, x_6 \in \mathcal{X}$ shattered by $V_{\{(x_1, h^*(x_1)), (x_2, h^*(x_2)), (x_3, h^*(x_3))\}}(\mathcal{H}(\epsilon_1; P, h^*))$.

...

According to the definition, $\{x_1, x_2, \ldots\}$ is an infinite VC-eluder sequence centered at h^*. ∎

Appendix D. Missing proofs from Section 5

Theorem 49 (Theorem 29 restated) *Let $\mathcal{H}$ be any concept class and h be any classifier. Then for any distribution P such that h is a Bayes-optimal classifier with respect to P (see Definition 10), $\mathcal{H}$ is agnostically universally learnable at rate e^{-n} by ERM under P if and only if $\mathcal{H}$ does not have an infinite eluder sequence centered at h.*

Proof [Proof of Theorem 49] We first prove the sufficiency. Suppose to the contrary that there is a distribution P such that $h = h^*_{\text{Bayes}}$, $\mathcal{H}$ does not have an infinite eluder sequence centered at h, but is not agnostically universally learnable at exponential rate (under P). Lemma 22 tells us that the Condition 1 fails (under P), and then Lemma 24 guarantees that there exists an optimal function h^*, i.e. $\mathrm{er}_P(h^*) = \inf_{h \in \mathcal{H}} \mathrm{er}_P(h)$ such that $\inf_{h \in \mathcal{H}: P_{\mathcal{X}}\{x:h(x) \neq h^*(x)\} > 0} P_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0$, and these two limits can be achieved by the same sequence of concepts $\{h_i\}_{i \geq 1} \subseteq \mathcal{H}$, that is,

$$\lim_{i \to \infty} \mathrm{er}_P(h_i) = \mathrm{er}_P(h^*), \quad \lim_{i \to \infty} P_{\mathcal{X}}\{x : h_i(x) \neq h^*(x)\} = 0 \tag{31}$$

$$\mathrm{er}_P(h_i) > \mathrm{er}_P(h^*), \quad P_{\mathcal{X}}\{x : h_i(x) \neq h^*(x)\} > 0, \quad \forall i \in \mathbb{N} \tag{32}$$

Furthermore, according to the second part proof of Lemma 23, we know there exists an infinite eluder sequence $\{(x_t, h^*(x_t))\}_{t \geq 1}$ centered at h^* witnessed by some subsequence $\{h_{i_t}\}_{t \geq 1} \subseteq \{h_i\}_{i \geq 1}$.

We first assume that P is supported on the aforementioned infinite eluder sequence. Note that for these eluder points, $h(x_t) = h^*(x_t)$ cannot happen for infinitely many $t \in \mathbb{N}$, since otherwise h will also admit an infinite eluder sequence. We assume there are $k < \infty$ many of such eluder points, denoted by $\{x_{t_1}, \ldots, x_{t_k}\}$, and consider a new infinite eluder sequence $\{(x_{t_j}, h^*(x_{t_j}))\}_{j \geq 1}$ centered at h^* witnessed by $\{h_{i_{t_j}}\}_{j \geq 1} \subseteq \{h_{i_t}\}_{t \geq 1}$. Now since $h = h^*_{\text{Bayes}}$, any concept $h_{i_{t_j}}$ with $j > k$ satisfies $\mathrm{er}_P(h_{i_{t_j}}) < \mathrm{er}_P(h^*)$ because h^* disagrees with the Bayes-optimal classifier on every eluder point with index larger than k. This contradicts our assumption that $\mathrm{er}_P(h^*) = \inf_{h \in \mathcal{H}} \mathrm{er}_P(h)$. Finally, if P is not simply supported on the eluder sequence, we consider the following conditional distribution $P' := P(\cdot|\{x_t\}_{t \geq 1})$. Note that h is still the Bayes-optimal classifier since we have the same $P(\cdot|\mathcal{X})$. Moreover, since $\{h_{i_t}\}_{t \geq 1} \subseteq \{h_i\}_{i \geq 1}$, we still have $\inf_{h \in \mathcal{H}} \mathrm{er}_{P'}(h) = \mathrm{er}_{P'}(h^*)$ and $\inf_{h \in \mathcal{H}: P'_{\mathcal{X}}\{x:h(x) \neq h^*(x)\} > 0} P'_{\mathcal{X}}\{x : h(x) \neq h^*(x)\} = 0$ based on (31) and (32). Therefore, the previous argument can be applied on P'.

To show the necessity, we suppose to the contrary that $\mathcal{H}$ has an infinite eluder sequence centered at $h = h^*_{\text{Bayes}}$. Consider the distribution P as constructed in Lemma 17, for which h is not only the Bayes-optimal classifier, but also satisfies $\inf_{h' \in \mathcal{H}} \mathrm{er}_P(h') = \mathrm{er}_P(h)$. Therefore, it is not agnostically universally learnable at rate faster than $o(n^{-1/2})$ by ERM, which gives a contradiction! ∎

Lemma 50 *Let $\mathcal{H}$ be any concept class and h be any classifier. Then for any distribution P such that h is a Bayes-optimal classifier with respect to P, $\mathcal{H}$ is agnostically universally learnable at rate $o(n^{-1/2})$ by ERM under P if and only if $\mathcal{H}$ does not have an infinite VC-eluder sequence centered at h.*

Proof [Proof of Lemma 50] The necessity is straightforward: if $\mathcal{H}$ has an infinite VC-eluder sequence centered at h, then the distribution P constructed in the proof of Lemma 21 is centered at h and has h be a Bayes-optimal classifier. Therefore, we know that $\mathcal{H}$ requires at least arbitrarily slow rates to be agnostically universally learned by ERM under distribution P.

To show the sufficiency, we suppose to the contrary that $\mathcal{H}$ is not agnostically universally learnable at rate $o(n^{-1/2})$ under some distribution P having h as a Bayes-optimal classifier, then Lemma 26 implies that the Condition 2 fails for P, and furthermore, there exists an infinite VC-eluder sequence $\{x_1, x_2, \ldots\}$ centered at some target function h^* satisfying $\mathrm{er}_P(h^*) = \inf_{h \in \mathcal{H}} \mathrm{er}_P(h)$. We note that an infinite VC-eluder sequence is an infinite eluder sequence, and it is impossible for a distribution P having a target function realize an infinite eluder sequence but a Bayes-optimal classifier does not, according to Theorem 29. Therefore, a contradiction follows. ∎

Theorem 51 (Theorem 30 restated) *Let $\mathcal{H}$ be any concept class and h be any classifier. For any distribution P such that h is a Bayes-optimal classifier with respect to P, $\mathcal{H}$ is agnostically universally learnable at exact rate $o(n^{-1/2})$ by ERM under P if and only if $\mathcal{H}$ does not have an infinite VC-eluder sequence, but has an infinite eluder sequence centered at h.*

Theorem 52 (Theorem 31 restated) *Let $\mathcal{H}$ be any concept class and h be any classifier. For any distribution P such that h is a Bayes-optimal classifier with respect to P, $\mathcal{H}$ requires at least arbitrarily slow rates to be agnostically universally learned by ERM under P if and only if $\mathcal{H}$ has an infinite VC-eluder sequence centered at h.*

Proof [Proof of Theorems 51 and 52] We will prove Theorem 51 and then the proof of Theorem 52 is straightforward. To show the necessity, we note that Theorem 49 implies that "an infinite eluder sequence centered at h" is necessary, since otherwise an exponential rate is attainable. Moreover, we consider the distribution P as constructed in Lemma 21, it is not hard to show that h is not only the Bayes-optimal classifier, but also a target with respect to P. Therefore, if $\mathcal{H}$ has as infinite VC-eluder sequence centered at h, then it requires at least arbitrarily slow universal rates according to Lemma 21. This implies that the necessary condition also has to include "no infinite VC-eluder sequence centered at h". To show the sufficiency, we have to prove an upper bound as well as a lower bound for proving an $o(n^{-1/2})$ exact rate. Note that Lemma 50 implies the upper bound. The lower bound follows from the proof of Theorem 49, where we show an infinite eluder sequence centered at h yields a no faster than $o(n^{-1/2})$ universal rate. ∎

Appendix E. Technical Lemmas

Lemma 53 (Hoeffding's Inequality) *Let $Z_1, \ldots, Z_n$ be independent random variables in $[a, b]$ with some constants $a < b$, let $\bar{Z} := \frac{1}{n} \sum_{i=1}^n Z_i$. Then for any $t > 0$,*

$$\mathbb{P}\left(\left|\bar{Z} - \mathbb{E}[\bar{Z}]\right| > t\right) \leq 2e^{-\frac{2nt^2}{(b-a)^2}}.$$

Lemma 54 (McDiarmid's Inequality) *Let V be some set and let $f : V^n \to \mathbb{R}$ be a function of n variables such that for some $c > 0$, for all $i \in [n]$ and for all $x_1, \ldots, x_n, x_i' \in V$, we have the following "bounded differences property":*

$$\left|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)\right| \leq c.$$

Then let $X_1, \ldots, X_n$ be independent random variables taking values in V, we have

$$\left| f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right] \right| \leq c\sqrt{\frac{n}{2}\ln\left(\frac{2}{\delta}\right)},$$

with probability at least $1 - \delta$.

Lemma 55 *For any $a, x > 0$,*

$$\int_x^\infty e^{-a\epsilon^2}\,d\epsilon \leq \int_x^\infty \frac{\epsilon}{x}e^{-a\epsilon^2}\,d\epsilon = \frac{1}{2ax}e^{-ax^2}.$$

Lemma 56 (Slud, 1977) *Let $X \sim Binomial(n, p)$ and assume that $p = (1 - \epsilon)/2$, then*

$$\mathbb{P}\left(X \geq \frac{n}{2}\right) \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp\left(-n\epsilon^2/(1 - \epsilon^2)\right)}\right).$$

Lemma 57 *For any function $R(n) \to 0$, there exist probabilities $\{p_t\}_{t\in\mathbb{N}}$ satisfying $\sum_{t\geq 1} p_t = 1$, an increasing sequence of integers $\{n_t\}_{t\in\mathbb{N}}$, and a constant $1/2 \leq C \leq 1$ such that the following hold for all $t \in \mathbb{N}$:*

(1) $\sum_{k\geq t+1} \frac{p_k}{\sqrt{n_k}} \leq \frac{p_t}{\sqrt{n_t}}$.

(2) $p_t = CR(n_t)$.

Proof [Proof of Lemma 57] We may assume without loss of generality that $R(1) = 1$, and start by a recursive definition of the sequence $\{n_t\}_{t\in\mathbb{N}}$. Let $n_1 = 1$, and for $t > 1$, we define

$$n_t := \inf\left\{n > n_{t-1} : R(n) \leq \min_{j<t}\left\{R(n_j) \cdot 2^{j-t}\right\}\right\}.$$

Moreover, we define $p_t = CR(n_t)$ for all $t \in \mathbb{N}$, where the constant $C = 1/(\sum_{t\geq 1} R(n_t))$. Note that $R(n_t) \leq 2^{-t+1}$ for all $t > 1$, it holds that $1/2 \leq C \leq 1$. With such a construction, we have

$$\sum_{k\geq t+1} \frac{p_k}{\sqrt{n_k}} = \sum_{k>t} \frac{CR(n_k)}{\sqrt{n_k}} \leq \sum_{k>t} \frac{CR(n_k)}{\sqrt{n_t}} \leq \sum_{k>t} \frac{CR(n_t)2^{t-k}}{\sqrt{n_t}} \leq \frac{p_t}{\sqrt{n_t}}.$$

$\blacksquare$

Lemma 58 (Bousquet et al., 2021, Lemma 5.12) *For any function $R(n) \to 0$, there exist probabilities $\{p_t\}_{t\in\mathbb{N}}$ satisfying $\sum_{t\geq 1} p_t = 1$, two increasing sequences of integers $\{n_t\}_{t\in\mathbb{N}}$ and $\{k_t\}_{t\in\mathbb{N}}$, and a constant $1/2 \leq C \leq 1$ such that the following hold for all $t \in \mathbb{N}$:*

(1) $\sum_{k>k_t} p_k \leq \frac{1}{n_t}$.

(2) $n_t p_{k_t} \leq k_t$.

(3) $p_{k_t} = CR(n_t)$.

Lemma 59 *Let $\mathcal{H}$ be a concept class and P be any distribution, let $S_n \sim P^n$ for any $n \in \mathbb{N}$. For any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ we have*

$$\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \right] + \sqrt{\frac{1}{2n} \ln \left(\frac{4}{\delta} \right)}.$$

Proof [Proof of Lemma 59] Note that

$$\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| = \max \left\{ \sup_{h \in \mathcal{H}} \hat{er}_{S_n}(h) - er_P(h), \sup_{h \in \mathcal{H}} er_P(h) - \hat{er}_{S_n}(h) \right\}.$$

We will consider the first term $\sup_{h \in \mathcal{H}} \hat{er}_{S_n}(h) - er_P(h)$ below, and the other can be analyzed similarly. Let $S_n := \{(x_i, y_i)\}_{i=1}^n$ and $S'_{n,i} := \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x'_i, y'_i), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$ for any $(x'_i, y'_i) \sim P$, we have the following "bounded differences property":

$$\left| \sup_{h \in \mathcal{H}} \hat{er}_{S_n}(h) - er_P(h) - \sup_{h \in \mathcal{H}} \hat{er}_{S'_{n,i}}(h) - er_P(h) \right|$$

$$\leq \sup_{h \in \mathcal{H}} \left| (\hat{er}_{S_n}(h) - er_P(h)) - (\hat{er}_{S'_{n,i}}(h) - er_P(h)) \right|$$

$$= \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \left(\mathbb{1}\{h(x_i) \neq y_i\} - \mathbb{1}\{h(x'_i) \neq y'_i\} \right) \right| \leq \frac{1}{n}.$$

By applying the McDiarmid's inequality (Lemma 54), we have

$$\sup_{h \in \mathcal{H}} \hat{er}_{S_n}(h) - er_P(h) \leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} \hat{er}_{S_n}(h) - er_P(h) \right] + \sqrt{\frac{1}{2n} \ln \left(\frac{4}{\delta} \right)},$$

with probability at least $1 - \delta/2$. The conclusion follows immediately from a union bound and the following fact:

$$\max \left\{ \mathbb{E} \left[\sup_{h \in \mathcal{H}} \hat{er}_{S_n}(h) - er_P(h) \right], \mathbb{E} \left[\sup_{h \in \mathcal{H}} er_P(h) - \hat{er}_{S_n}(h) \right] \right\} \leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} \left| \hat{er}_{S_n}(h) - er_P(h) \right| \right].$$

∎

Lemma 60 (Bousquet, 2002, Theorem 2.3) *Let P be a distribution and $\{X_i\}_{1 \leq i \leq n} \subseteq \mathcal{X}$ be a sequence of i.i.d. P-distributed random variables. Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ are P-measurable, square-integrable and satisfy $\mathbb{E}[f] = 0$. We assume that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq 1$ and denote*

$$Z := \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n f(X_i) \right|.$$

Let σ be a positive real number such that $\sup_{f \in \mathcal{F}} \text{Var}[f(X_1)] \leq \sigma^2$ almost surely, let $\nu = n\sigma^2 + 2\mathbb{E}[Z]$. Then for all $x \geq 0$, we have

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + x \right) \leq \exp\left(-\nu h(x/\nu) \right),$$

where $h(x) := (1 + x) \log(1 + x) - x$, and also

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + \sqrt{2\nu x} + \frac{x}{3} \right) \leq e^{-x}.$$

Lemma 61 (van der Vaart and Wellner, 2011, Theorem 2.1) *Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope function $F \leq 1$ and such that $\mathcal{F}^2$ is P-measurable. If $Pf^2 < \delta^2 PF^2$ for every f and some $\delta \in (0, 1)$, then*

$$\mathbb{E}_P^* \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - Pf) \right| \right] \lesssim J(\delta, \mathcal{F}, L_2) \left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\delta^2 \sqrt{n} \|F\|_{P,2}} \right) \|F\|_{P,2},$$

where $\|f\|_{P,2}$ denotes the norm of a function f in $L_2(P)$ and

$$J(\delta, \mathcal{F}, L_2) := \sup_P \int_0^\delta \sqrt{1 + \log N(\epsilon \|F\|_{P,2}, \mathcal{F}, L_2(P))} d\epsilon.$$