# A Novel Data-Dependent Learning Paradigm for Large Hypothesis Classes

**Alireza F. Pour** ALIREZA.FATHOLLAHPOUR@UWATERLOO.CA
*University of Waterloo*

**Shai Ben-David** SHAI@UWATERLOO.CA
*University of Waterloo and the Vector Institute*

**Editors:** Matus Telgarsky and Jonathan Ullman

## Abstract

We address the general task of learning with a set of candidate models that is too large to have a uniform convergence of empirical estimates to true losses. While the common approach to such challenges is SRM (or regularization) based learning algorithms, we propose a novel learning paradigm that relies on stronger incorporation of empirical data and requires less algorithmic decisions to be based on prior assumptions. We analyze the generalization capabilities of our approach and demonstrate its merits in several common learning assumptions, including similarity of close points, clustering of the domain into highly label-homogeneous regions, Lipschitzness assumptions of the labeling rule, and contrastive learning assumptions. Our approach allows utilizing such assumptions without the need to know their true parameters a priori.

**Keywords:** Data-dependent learning, partial concept classes, structural risk minimizer.

## 1. Introduction

We propose a new learning paradigm that closely follows the objective of non-uniform learning, namely, given a set $\mathcal{C}$ of candidate hypotheses , how many samples do we need to compete with the error of some $h \in \mathcal{C}$? The canonical approach to non-uniform learning is the Structural Risk Minimization (SRM) algorithm, where the learner structures the hypotheses in a sequence of nested classes. The generalization error that SRM guarantees for any $h \in \mathcal{C}$ then scales with the index of the first class in which $h$ appears, which is implicit in some weighting function $w(.)$, and the complexity of that class, e.g., its VC dimension. Hence, the success of SRM depends on how well the structure reflects the prior belief about low approximating hypotheses, and how justified that belief is. That is, whether data matching hypotheses are given high weight and/or the class in which they appear is 'simple' to learn. For instance, one may assume that the labelling rule is Lipschitz and assign to hypotheses weights that scale with the margin around their decision boundary. The caveat is that the SRM paradigm requires fixing an a priori weighting function and selects its decision rule based on this weighting.

We introduce a different learning paradigm (for classes that are too big for enjoying uniform convergence). We introduce a novel parameter of a collection of hypothesis classes: Given a collection $\mathbb{H}$, the growth function of the collection, denoted by $\tau_{\mathbb{H}}(m)$, controls the number of equivalence classes of behaviours that the collection induces on a sample of size $m$. We show that, whenever we group $\mathcal{C}$ into a collection of hypothesis classes $\mathbb{H}$, for any $h \in \mathcal{C}$, it is possible to guarantee a generalization error that scales with $\tau_{\mathbb{H}}(m)$ and the complexity of the class to which $h$ belongs. In contrast to a weighting function $w(h)$, this parameter is uniform over any $h$ - there is no 'penalty'

to placing a data fitting $h$ in a low weight class. Thus, we shift the focus to finding a grouping of $\mathcal{C}$ that minimizes the growth function of the collection of classes and the complexity of the classes that contain low approximating hypotheses.

The paradigm we offer is a tool that can be applied on collection of partial concept classes. A partial concept is a function that can be undefined on parts of the domain with the goal of incorporating data assumptions (Alon et al., 2022). Given an assumption (or prior knowledge) about the labelling function, a total function $h$ can be transformed into a set of behaviours such that each behaviour aligns with the assumption on its $\{0,1\}$ part, namely, its support. Our paradigm leads the following approach: given a set $\mathcal{C}$ of candidate (total) concepts, find a parameter of pairs (a partial function, a finite sample) with the property that (i) it reflects the prior knowledge to enable evaluating to what extent a transformation of a concept into a partial one aligns with the assumption; and (ii) grouping the partial concepts into classes based on this parameter will result in a collection with bounded growth rate. Conceptually, this shifts the role of prior knowledge from defining an a priori hierarchy over hypotheses to inducing a data-dependent organization. For instance, under a Lipschitz-type assumption, instead of ranking hypotheses by a predefined margin parameter, we group them according to the extent to which they violate the Lipschitz condition on the training sample. In Section 3.2, this idea is formalized via forbidden behaviours equipped with penalty values, which quantify empirical violations of the prior knowledge and induce a data-dependent grouping of hypotheses with controlled growth. We will show in Section 4 that this covers several common choices such as similarity of close points, clustering of the domain, Lipschitzness assumptions on the labeling rule, and contrastive assumptions.

The proposed paradigm guarantees its error by relying on a compression-based argument. The learner goes over different subsets of the training data, on which it invokes learners $\mathcal{A}_{\mathcal{H}}, \mathcal{H} \in \mathbb{H}$ that are specifically designed to learn from the classes $\mathcal{H} \in \mathbb{H}$. It then estimates the expected error of these predictors on the left out part of the training sample. The learner guarantees the convergence of the error of $\mathcal{A}_{\mathcal{H}}$ on the left out set to its expected error with a failure probability that scales with the size of the subset that is used as input training data to $\mathcal{A}_{\mathcal{H}}$. Notably, this is one reason that the proposed learner has data-dependent guarantees. Whenever the function belongs to a class with a small data-dependent complexity, we show that the learner needs a smaller subset as training input to learn from that class and thus it will be assigned a large weight when estimating its error on the left out part. In Section 2.1, we will discuss a detailed comparison with SRM and show

## 2. Learning with Respect to a Collection of Concept Classes

We will first setup the notation and then discuss learnability results with respect to collections of concept classes. our main results.

**Notations.** For a partial concept class $\mathcal{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ and any set $U \subseteq \mathcal{X}$, we define $\mathcal{H}_{|U} = \{h_{|U} : h \in \mathcal{H}, h_{|U} \in \{0,1\}^U\}$. Note that based on this definition $\mathcal{H}_{|U}$ will always be a subset of $\{0,1\}^U$ and we never include in $\mathcal{H}_{|U}$ the behaviours that contain $\star$. For a labelled set $S \subseteq (\mathcal{X} \times \{0,1\})$, we denote by $\mathsf{dom}(S) \subseteq \mathcal{X}$ the unlabelled part of $S$. We will sometimes overload the notation and write $\mathcal{H}_{|S}$ for labeled samples $S \subseteq (\mathcal{X} \times \{0,1\})^*$ instead of $\mathcal{H}_{|\mathsf{dom}(S)}$.

For a set $S$, we will write $\mathsf{VC}(\mathcal{H}, S)$ as a shorthand notation for the VC dimension $\mathsf{VC}(\mathcal{H}_{|\mathsf{dom}(S)})$. Moreover, for any distribution $\mathcal{D}$, we denote by $\mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta)$ the smallest integer such that the probability over $S \sim \mathcal{D}^m$ that $\mathsf{VC}(\mathcal{H}, S) \leq \mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta)$ is at least $1 - \delta$. We will refer to $\mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta)$ as the effective VC dimension of $\mathcal{H}$ with respect to $\mathcal{D}$.

For any set $S$ and any hypothesis $h : \mathcal{X} \to \{0, 1, \star\}$ we will denote $\mathsf{err}(h, S) = \frac{1}{|S|} \sum_{(x,y)\in S} 1\{h(x) \neq y\}$. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ and any hypothesis $h : \mathcal{X} \to \{0, 1, \star\}$ we will define $\mathsf{err}(h, \mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}} [h(x) \neq y]$. Moreover, for any hypothesis class $\mathcal{H}$, we define $\mathsf{err}(\mathcal{H}, \mathcal{D}) = \inf_{h\in\mathcal{H}}(\mathsf{err}(h, \mathcal{D}))$.

We now give the definitions of the equivalence relation between hypothesis classes and the growth parameter of a collection.

**Definition 1** *Let $\mathcal{H}, \mathcal{H}' \subseteq \{0, 1, \star\}^{\mathcal{X}}$ be two (partial) concept classes defined on $\mathcal{X}$. Given a set $U \subseteq \mathcal{X}$, we say $\mathcal{H}$ and $\mathcal{H}'$ are equivalent on $U$ if $\mathcal{H}_{|T} = \mathcal{H}'_{|T}$ for all $T \subseteq U$ and we denote it by $\mathcal{H} \sim_U \mathcal{H}'$. For a family of partial concept classes $\mathbb{H}$ and for any hypothesis class $\mathcal{H} \in \mathbb{H}$ we denote $[\mathcal{H}_{\sim_U}]_{\mathbb{H}} = \{\mathcal{H}' \in \mathbb{H} : \mathcal{H} \sim_U \mathcal{H}'\}$. Moreover, for any set $U \subseteq \mathcal{X}$, we denote the set of all equivalence classes generated by members of $\mathbb{H}$ on $U$ by $\mathbb{H}_{|U} := \{[\mathcal{H}_{\sim_U}]_{\mathbb{H}} : \mathcal{H} \in \mathbb{H}\}$.*

**Definition 2 (Growth parameter $\tau_{\mathbb{H}}$)** *Let $\mathbb{H}$ be a family of (partial) concept classes. Given a set $U \subseteq \mathcal{X}$ we define $\tau_{\mathbb{H}}(U) = |\mathbb{H}_{|U}|$. We denote by $\tau_{\mathbb{H}}(m)$ the maximum number of the equivalence classes generated on sets of size $m$. Formally,*

$$\tau_{\mathbb{H}}(m) = \max_{U\subseteq\mathcal{X},|U|=m} \tau_{\mathbb{H}}(U) = \max_{U\subseteq\mathcal{X},|U|=m} |\mathbb{H}_{|U}|.$$

*Moreover, for a distribution $\mathcal{D}$ and $\delta \in (0, 1)$, we denote by $\tau_{\mathbb{H}}(m, \mathcal{D}, \delta)$ the smallest integer such that $\mathbb{P}_{S\sim\mathcal{D}^m} [\tau_{\mathbb{H}}(\mathsf{dom}(S)) \leq \tau_{\mathbb{H}}(m, \mathcal{D}, \delta)] \geq 1 - \delta$.*

We sometimes write $\mathbb{H}_{|S}$ instead of $\mathbb{H}_{|\mathsf{dom}(S)}$ and write $\tau_{\mathbb{H}}(S)$ instead of $\tau_{\mathbb{H}}(\mathsf{dom}(S))$. We now state our main result for learning with respect to a collection of hypothesis classes. The proof of all results in this section can be found in Appendix A.

**Theorem 3** *Let $\mathbb{H}$ be a collection of (partial) concept classes. There exists a learner $\mathcal{A}_{\mathbb{H}} : (\mathcal{X} \times \{0, 1\})^* \times \mathcal{X} \to \{0, 1\}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0, 1)$ and $m \in \mathbb{N}$ we have with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}_{\mathbb{H}}(S), \mathcal{D}) \leq \min_{\mathcal{H}\in\mathbb{H}} \left\{ \mathsf{err}(\mathcal{H}, \mathcal{D}) + O\left( \sqrt{\frac{(\mathsf{VC}(\mathcal{H}) + \log(\tau_{\mathbb{H}}(2m))) \log^2(m) + \log(\frac{1}{\delta})}{m}} \right) \right\}.$$

**Corollary 4** *Let $\mathcal{C}$ be a class of (partial) functions and $\mathbb{H}$ be a grouping of $\mathcal{C}$ such that $\bigcup_{\mathcal{H}\in\mathbb{H}} \mathcal{H} = \mathcal{C}$. There exists a learner $\mathcal{A}_{\mathbb{H}} : (\mathcal{X} \times \{0, 1\})^* \times \mathcal{X} \to \{0, 1\}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0, 1)$ and $m \in \mathbb{N}$ we have with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}_{\mathbb{H}}(S), \mathcal{D}) \leq \min_{h\in C} \left\{ \mathsf{err}(h, \mathcal{D}) + O\left( \sqrt{\frac{(\mathsf{VC}(\mathcal{H}_{(h)}) + \log(\tau_{\mathbb{H}}(2m))) \log^2(m) + \log(\frac{1}{\delta})}{m}} \right) \right\},$$

*where $\mathcal{H}_{(h)}$ is the class to which $h$ belongs.*

We give an informal overview of how the learner works and how its error bound is guaranteed. The formal definition of the learner and proofs appear in Appendix A.

The learner $\mathcal{A}_{\mathbb{H}}$ in the above works by going over subsets of the training data. On each subset, it invokes a set of learners $\{\mathcal{A}_{\mathcal{H}} : \mathcal{H} \in \mathbb{H}\}$, estimates their error on the left-out subset, and returns the one with the minimum value of its generalization bound which depends on the error on left-out subset and the size of the subset input to $\mathcal{A}_{\mathcal{H}}$. The learners $\mathcal{A}_{\mathcal{H}}$ predicts according to the majority vote of $O(\log(m))$ many One-Inclusion Graph (OIG) learners. The OIG learners that $\mathcal{A}_{\mathcal{H}}$ takes into account predict by only considering the $\{0,1\}$ behaviours of $\mathcal{H}$ on the input training set and the test point. We then observe that OIGs are weak learners for $\mathcal{H}$ with sample size that is relevant to the VC dimension *with respect to the distribution*. Using a boosting technique guarantees that for a given class $\mathcal{H}$, there exists a subset of size relevant to its VC dimension with respect to the distribution such that the majority vote of the OIGs on that subset approximately recovers the label of the rest of the sample. In other words, we have a (approximate) compression set for $\mathcal{H}$ of size proportional to the VC of $\mathcal{H}$ with respect to the distribution.

We then prove that the expected error of the learners $\{\mathcal{A}_{\mathcal{H}} : \mathcal{H} \in \mathbb{H}\}$ trained on any subset are close to their error on the left-out subset. Our guarantee is established non-uniformly and proportional to the size of the subset for training. Particularly, This is where the introduced growth parameter is used to show that on any subset, there are only a few different predictions that the learners $\{\mathcal{A}_{\mathcal{H}} : \mathcal{H} \in \mathbb{H}\}$ induce on the left out subset. Crucially, the data-dependent benefit we get is due to the fact that we prove the convergence of the error on the sample to the expected error with a probability that depends on the size of the compression set, i.e., the subset used to train the learner. The compression set size depends on the complexity of $\mathcal{H}$ with respect to the distribution, which is unknown in advance. Therefore, contrary to a fixed weighting, the classes are assigned weights based on the size of the compression set we are using to learn from them; there is no mismatch between how complex (with respect to the data distribution) it is to learn a class and how much it is preferred over other classes.

Indeed, the following shows that the same learner that guarantees the bound of Theorem 3 is also capable of guaranteeing bounds that scale with the growth parameter of the grouping and the complexity of the concept classes with respect to the data generating distribution.

**Theorem 5** *Let $\mathbb{H}$ be a collection of concept classes. There exists a learner $\mathcal{A}_{\mathbb{H}} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}_{\mathbb{H}}(S), \mathcal{D}) \leq \min_{\mathcal{H} \in \mathbb{H}} \left\{ \mathsf{err}(\mathcal{H}, \mathcal{D}) + O\left( \sqrt{\frac{\left(\mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \frac{\delta}{4}) + \log(\tau_{\mathbb{H}}(m, \mathcal{D}, \frac{\delta}{4}))\right)\log^2(m) + \log(\frac{1}{\delta})}{m}} \right) \right\}.$$

### 2.1. Comparison with Structural Risk Minimization (SRM) for Partial Concepts

When faced with a large set of candidate partial hypotheses, it may seem natural to apply the structural risk minimization (SRM) principle (see Section 4.2 in Alon et al. (2022) for a discussion on SRM over partial classes). Yet, this approach introduces several drawbacks. In what follows, we contrast our proposed paradigm with SRM over partial concept classes.

SRM requires an a priori weighting over the hypothesis classes, determined independently of the training data. The generalization bound it guarantees for any concept $h$ depends on both the weight $w(\mathcal{H}_{(h)})$ and the complexity of its class $\mathcal{H}_{(h)}$. In contrast, the guarantees in Theorems 3 and 5

depend only on the (data-dependent) complexity of $\mathcal{H}_{(h)}$, while the term $\tau_{\mathbb{H}}(m)$ (or $\tau_{\mathbb{H}}(m, \mathcal{D}, \delta)$) is uniform across all $h$. Consequently, SRM may fail to provide a reasonable bound when the well-approximating hypothesis $h^*$ is assigned a small weight. SRM selects its predictor by minimizing an upper bound on the error, derived from non-uniformly requiring uniform convergence for each hypothesis class.[1] Such worst-case guarantees scale with the (worst-case) VC dimension of each class and ignore the possibility that the effective complexity on the underlying distribution may be much smaller. For instance, a class $\mathcal{H}_{(h)}$ may have infinite VC dimension but finite $\mathsf{VC}(\mathcal{H}_{(h)}, \mathcal{D}, m, \delta)$. To address this, several works propose SRM variants based on sample-dependent complexity measures, such as Rademacher or localized complexities (Koltchinskii, 2002; Bartlett et al., 2002), which attempt to ensure non-uniform convergence relative to $\mathsf{VC}(\mathcal{H}_{(h)}, \mathcal{D}, m, \delta)$ rather than its worst-case counterpart.

The key insight is that uniform convergence must hold (even non-uniformly) across all hypothesis classes. An SRM structures hypotheses into a (possibly nested) collection and assigns each class $\mathcal{H}_{(h)}$ a weight $w(\mathcal{H}_{(h)})$ a priori, before observing data. Since it cannot anticipate the distribution-dependent complexity $\mathsf{VC}(\mathcal{H}_{(h)}, \mathcal{D}, m, \delta)$, it must rely on the worst-case $\mathsf{VC}(\mathcal{H}_{(h)})$. Indeed, a common approach is to construct a nested hierarchy of hypothesis classes based on how well they satisfy the available prior knowledge: as the index increases, the classes become richer and more complex, while the assigned weights decrease accordingly. This introduces an additive term $O(\sqrt{\log(1/w(\mathcal{H}_{(h)}))/m})$ in the generalization bound SRM considers for returning a predictor, which can undermine the benefits of small effective VC dimension. As a result, even if a high-complexity class achieves near-zero approximation error, SRM is unlikely to select it over a simpler but highly inaccurate class.

The following proposition is an abstract application of the last point above. We call an SRM learner *standard*, if the learner returns a hypothesis $\hat{h}$ that has the minimum value of the generalization upper bound

$$\mathsf{err}(\hat{h}, S) + O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{H}_{(\hat{h})}, \mathcal{D}, m, \delta) + \log(1/w(\mathcal{H}_{(\hat{h})})) + \log(1/\delta)}{m}}\right),$$

which is based on uniform convergence of the hypothesis class with failure probability proportional to the weight of class. The proof appears in Appendix A.

**Proposition 6** *There exists a collection $\mathbb{H} = \{\mathcal{H}_n : n \in \mathbb{N}\}$ with $\mathsf{VC}(\mathcal{H}_n) = n$ for all $n \in \mathbb{N}$ such that for any* standard *SRM learner $\mathcal{A}_{\mathsf{SRM}}$ the following holds. For every $\delta \in (0, 1)$ and sample size $m$, there exists a distribution $\mathcal{D}$ such that $\mathbb{P}_{S\sim\mathcal{D}^m}[\mathsf{err}(\mathcal{A}_{\mathsf{SRM}}(S), \mathcal{D}) = 1/2] \geq 1 - \delta$. Moreover, for the learner $\mathcal{A}$ in Theorem 3 we have with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}(S), \mathcal{D}) \leq O\left(\sqrt{\frac{\log^2(m) + \log(1/\delta)}{m}}\right).$$

Theorems 3 and 5 thus offer generalization bounds comparable to those of the Empirical Risk Minimizer (ERM) over any $\mathcal{H} \in \mathbb{H}$, with a uniform additive factor of $O\left(\sqrt{\frac{\log(\tau_{\mathbb{H}}(m))\log^2(m)}{m}}\right)$.

1. For comparison with SRM, we assume that the collection contains total classes mapping the domain to $0, 1$.

In contrast, the bound for SRM deviates from that of ERM by a non-uniform term depending on $w(\mathcal{H})$; see Section 7.2 in Shalev-Shwartz and Ben-David (2014).

We will instantiate our results in settings where a large candidate set $\mathcal{C}$ and an additional source of prior knowledge are available. The prior knowledge is used to group $\mathcal{C}$ into a nested family $\mathbb{H} = \{\mathcal{H}_r : r \geq 0\}$ with $\mathcal{H}_r \subseteq \mathcal{H}_s$ for $r \leq s$, where the complexity of classes grow with $r$. A standard SRM assigns larger weights to simpler classes (small $r$) to favor lower estimation error. Consequently, as $\mathcal{H}_r$ approaches the full class $\mathcal{C}$, its weight $w(\mathcal{H}_r)$ decreases, and the term $\sqrt{\frac{\log(1/w(\mathcal{H}_r))}{m}}$ in the SRM error bound increases—making it less competitive with ERM over $\mathcal{C}$. In contrast, the learner of Theorems 3 and 5 treats all classes uniformly, and thus competes with the ERM bound even when only large $r$ yield small approximation error—comparable to ignoring the additional source altogether.

Finally, when $\mathbb{H}$ is uncountable, SRM requires an a priori discretization of the family. While for nested sequences it is possible to select breakpoints $\{r_i : i \in \mathbb{N}\}$ where the VC dimension changes, such discretization still fails to guarantee competitiveness with every $\mathcal{H}_r$. Between two discretization points, one either loses approximation accuracy (when using $\mathcal{H}_{r_i}$) or incurs higher complexity (when using $\mathcal{H}_{r_{i+1}}$). Instead of fixing discretization in advance, our approach lets the data itself to partition $\mathbb{H}$ into equivalence classes, ensuring that each restricted class $\mathcal{H}_{r|S}$ is considered as a candidate. This allows the resulting bound to compete directly with $\mathsf{err}(\mathcal{H}_r, \mathcal{D})$ whenever $\tau_{\mathbb{H}}(m)$ grows polynomially, without assuming prior discretization or explicit weighting.

## 3. The Growth Parameter $\tau_{\mathbb{H}}$

We can see that while a polynomial bound on the growth of $\tau_{\mathbb{H}}(m)$ is sufficient to learn from $\mathbb{H}$, there are collection of classes with exponential growth that are still easy to learn.

For instance, assume the universe is $[m]$. For any subset $A \subset [m]$ of size $\log(m)$ let $h_A^{(1)}, \ldots, h_A^{(m)}$ be functions that shatter the set $A$ such that they all extend to $[m] \setminus A$ with a label of 1, i.e., they form a cube on $[m]$ with dimension set $A$. Let $\sigma(1), \ldots, \sigma(2^m)$ be any enumeration of the subsets of the set $[m]$. Now, for any $A \subset [m]$ and $i \in [2^m]$ define a hypothesis class $\mathcal{H}_A^{(i)} = \{h_A^{(j)} : j \in \sigma(i)\}$. Let $\mathbb{H} = \{\mathcal{H}_A^{(i)} : A \subset [m], |A| = \log(m), i \in [2^m]\}$. It is clear that on any set $S$ of $m$ distinct points, we have $\tau_{\mathbb{H}}(S) = 2^m$. On the other hand, we can simply see that for the union $\mathcal{C} = \bigcup_{\mathcal{H}_A^{(i)} \in \mathbb{H}}$, we have $\mathsf{VC}(\mathcal{C}) \leq \log(m)$ and an ERM is a good learner for the union.

We will now list some behaviours of $\tau_{\mathbb{H}}(S)$ and then discuss interesting cases in the setting of nested hypothesis classes. We denote by $\pi_{\mathcal{C}}(S)$ the number of behaviours $\mathcal{C}$ induces on $S$ and by $\pi_{\mathcal{C}}(m)$ the maximum number of $\pi_{\mathcal{C}}(S)$ over all sets $S$ of size $m$.

1. If $\mathbb{H} = \{\mathcal{H}\}$ is a singleton then $\tau_{\mathbb{H}}(S) = 1$ and $\max_{\mathcal{H} \in \mathbb{H}} \mathsf{VC}(\mathcal{H}, S) = \mathsf{VC}(\mathcal{C}, S)$
2. If $\mathbb{H} = \{\{h\} : h \in \mathcal{C}\}$ then $\tau_{\mathbb{H}}(S) = \pi_{\mathcal{C}}(S)$ so $\tau_{\mathbb{H}}(m) = \pi_C(m)$ and $\max_{\mathcal{H} \in \mathbb{H}} \mathsf{VC}(\mathcal{H}, S) = 0$
3. If $\mathbb{H} = \{\mathcal{H}_r : r \geq 0\}$ is a nested collection, i.e., $\mathcal{H}_r \subseteq \mathcal{H}_s$ for $r \leq s$ then we have $\tau_{\mathbb{H}}(S) \leq \pi_{\mathcal{C}}(S)$.
4. $\tau_{\mathbb{H}}(m)$ is non-decreasing, i.e., for any $m, m' \in \mathbb{N}$ with $m \leq m'$,$\tau_{\mathbb{H}}(m) \leq \tau_{\mathbb{H}}(m')$.

We will show some cases where $\tau_{\mathbb{H}}(S)$ is much smaller than $\pi_{\mathcal{C}}(S)$.

### 3.1. Hierarchical Clustering

Assume we have a finite domain $\mathcal{X}$ and $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. We also have, as an additional knowledge, a hierarchical clustering over the domain that we wish to incorporate in learning. We want to assume each level as a possible clustering of the domain that groups the instance into label-homogenous clusters. Therefore, we want to define a class $\mathcal{H}_i$ for each level that respects the clustering of level $i$. We now discuss formally the hierarchical clustering and the classes $\mathcal{H}_i$.

A clustering $\mathcal{C}$ is defined by a weight function $w_{\mathcal{C}} : \mathcal{X} \times \mathcal{X} \to \{0,1\}$ that is symmetric and transitive in the 1 value, i.e., $w_{\mathcal{C}}(x,x') = 1$ and $w_{\mathcal{C}}(x',x'') = 1$ implies $w_{\mathcal{C}}(x,x'') = 1$. A clustering partitions the domain into equivalence classes. A cluster $C$ in $\mathcal{C}$ is any equivalence class, i.e., for all $x \in C$, and for all $x' \in \mathcal{X}$, $w(x,x') = 1$ implies $x' \in C$.

A hierarchical clustering is a tree of clusterings such that each node $v$ in this tree is a cluster $C_v$ and all the clusters at depth $i$ create a clustering of the domain, which we denote by $\mathcal{C}_i$. In the leafs of this graph we have a clustering $\mathcal{C}_{\mathcal{X}}$ that has a cluster for every $x \in \mathcal{X}$, i.e., $w_{\mathcal{C}_{\mathcal{X}}}(x,x') = 0$ for all $x, x' \in \mathcal{X}$. For every other node $v$ in this graph, the cluster $C_v$ merges the clusters of its children, i.e., $C_v = \bigcup_{d \in d(v)} C_d$ where $d(v)$ denotes the children of a node $v$. This means that the hiearachical clustering starts by assigning every point to a cluster. Each time it merges some clusters and moves up the tree until we have at the root the clustering $\mathcal{C}_0$ that collects every point into a single cluster.

We now define the collection of hypotheses classes as those that incorporate this hierarchical clustering into $\mathcal{H}$. We want to consider the entire hierarchy in a collection because it is not clear which $\mathcal{C}_i$ is the best approximation of the data-generating distribution $\mathcal{D}$ and that we want to make a trade-off between how the clustering 'fits' the true labeling and how 'complex' it is to learn from that clustering.

The translation of $\mathcal{H}$ into a partial concept that is enforced to respect the homogeneity over the clustering $\mathcal{C}$ is defined as

$$\begin{aligned}\mathcal{H}(\mathcal{C}) :=& \{h \in \{0,1,\star\}^{\mathcal{X}} : \exists h' \in \mathcal{H}, \forall x \in \mathcal{X}, h(x) = \star \text{ or } h(x) = h'(x) \\ & \text{and } \forall x_1, x_2 \in \mathcal{X} \text{ with } h(x_1), h(x_2) \in \{0,1\}, \text{and } w_{\mathcal{C}}(x_1,x_2) = 1, h(x_1) = h(x_2)\}.\end{aligned}$$

We then define $\mathbb{H}_{\mathsf{HC}} = \{\mathcal{H}_i\}$ and note that $\mathcal{H}_i \subseteq \mathcal{H}_j$ for all $i \leq j$. The interesting observation here is that although $|\mathbb{H}_{\mathsf{HC}}|$ can be as large as $|\mathcal{X}|$ and we have many hypothesis classes, the growth of this collection is only linear in the sample size.

**Claim 7** *For the hierarchical clustering and the collection $\mathbb{H}$ defined above, we have for every $S \in \mathcal{X}^*$ that $\tau_{\mathbb{H}_{\mathsf{HC}}}(S) \leq |S|$. Hence, $\tau_{\mathbb{H}_{\mathsf{HC}}}(m) \leq m$.*

**Proof**

Let $K_i(S)$ denote the number of clusters in $\mathcal{C}_i$ such that at least a point from $S$ resides in the cluster. Observe that $K_{\mathcal{X}}(S) \leq |S|$ and that whenever $w_{\mathcal{C}_i}(x,x') = 0$ and $w_{\mathcal{C}_{i-1}}(x,x') = 1$ we have merged the two clusters that contain $x$ and $x'$. Therefore, we have $K_{i-1}(S) \leq K_i(S) - 1$ and the number of clusters that contain at least a point from $S$ decreases. Moreover, for any $i \leq j$ if $w_{\mathcal{C}_j}(x,x') = 1$ then $w_{\mathcal{C}_i}(x,x') = 1$. Also, for any $i \leq j$ such that $w_{\mathcal{C}_i}(x,x') = w_{\mathcal{C}_j}(x,x')$ for all $x, x' \in S$ we have that $\mathcal{H}_{i|T} = \mathcal{H}_{j|T}$ for all $T \subseteq S$ and therefore $\mathcal{H}_i \sim_S \mathcal{H}_j$. This means that we can have at most $K_{\mathcal{X}}(S) \leq |S|$ equivalence classes, concluding the proof. ■

### 3.2. Forbidden Behaviours on Tuples of Points

In this section, we study settings where extra information appears as forbidden behaviours on tuples of points. These constraints are graded by a real-valued penalty, indicating how strongly a behaviour is disallowed. Limiting the allowed behaviours on the sample could reduce the complexity of learning but can also increase the chance of error. The penalty value reflects this trade-off between approximation and estimation. We define partial concept classes by giving a threshold to the penalty and only allowing behaviours below a chosen value. The goal is to find a threshold of penalty, i.e., a set of allowed behaviours, that minimizes the error bound among all the values for the data-generating distribution.

Formally, let $\mathcal{Y} = \{0,1\}^k$ and $\mathcal{B} : \mathcal{X}^k \to 2^{\mathcal{Y}}$ be a function that gets as input a sample of size $k$ and outputs a set of forbidden behaviours on those $k$ instances. Moreover, let $p : \mathcal{X}^k \to \mathbb{R}$ be a penalty function. For any $r \geq 0$, the class $\mathcal{H}_{\mathcal{B},p}(r) \subseteq \{0,1,\star\}^*$ is defined as

$$\mathcal{H}_{\mathcal{B},p}(r) = \{h \in \{0,1,\star\}^* : \exists h' \in \mathcal{H}, \forall x \in \mathcal{X} \text{ if } h(x) \neq \star \text{ then } h(x) = h'(x)\ , \\ \text{and } \forall S \in \mathcal{X}^k \text{ if } p(S) \geq r \text{ then } h_{|S} \notin \mathcal{B}(S)\}$$

When the forbidden behaviours $\mathcal{B}$ are clear from context we often drop. Moreover, we will overload the notation and write $p(S)$ instead of $p(\mathsf{dom}(S))$. We sometimes drop $p$ and simply write $\mathcal{H}(r)$.

We then define the family of concept classes for the above assumption by $\mathbb{H}_{\mathcal{B},p} = \{\mathcal{H}_{\mathcal{B},p}(r) : r \geq 0\}$ and show that its growth rate is only polynomial in $m$.

**Lemma 8** *For any hypothesis class $\mathcal{H}$ and the smooth family $\mathbb{H}_{\mathcal{B},p}$ as defined above, we have $\tau_{\mathbb{H}_{\mathcal{B},p}}(m) = O(m^k)$.*

**Proof** Fix a set $S$ of size $m$ and let $\ell := \binom{m}{k}$ . Denote by $A_1, \ldots, A_\ell \subseteq S$ all the subsets of $S$ of size $k$. Let $p_1 \leq \ldots \leq p_\ell$ be the values of $\{p(A_i) : i \in [\ell]\}$ sorted in increasing order. Fix any $h \in \{0,1,\star\}^S$ such that $\exists h' \in \mathcal{H}, \forall x \in S$ if $h(x) \neq \star$ then $h(x) = h'(x)$. Then for any subset $T \subset S$ with $|T| < k$ we either have $h(x) \in \{0,1\}$ for all $x \in T$ in which case $h_{|T} \in \mathcal{H}(r)_{|T}$ for all $r \geq 0$ or $h(x) = \star$ for some $x \in T$ in which case the behaviour is not included in any $\mathcal{H}(r)_{|T}$. For any $T \subseteq S$ with $|T| \geq k$, let $A(T) \subseteq \{A_1, \ldots, A_\ell\}$ be the collection of subsets of size $k$ that are also subsets of $T$, i.e., $A(T) = \{A_i : A_i \subseteq T, i \in [\ell]\}$. Now we either have $h(x) \in \{0,1\}$ for all $x \in T$ or $h(x) = \star$ for some $x \in T$. In the former case, if $h_{|A} \notin \mathcal{B}(A)$ for all $A \in A(T)$ then $h_{|T} \in \mathcal{H}(r)_{|T}$ for all $r \geq 0$. Otherwise, $h_{|T} \in \mathcal{H}(r)_{|T}$ only for $r > \max\{p(A) : A \in A(T), h_{|A} \in \mathcal{B}(A)\}$. Assume $A_j \in A(T)$ is the subset achieving $\max\{p(A) : A \in A(T), h_{|A} \in \mathcal{B}(A)\}$. This implies that $h_{|T} \in \mathcal{H}(r)_{|T}$ for all $r > p_j$ and $h_{|T} \notin \mathcal{H}(r)_{|T}$ otherwise. In the case that $h(x) = \star$ for some $x \in T$, the behaviour is not included in any $\mathcal{H}(r)_{|T}$. This concludes that for all subsets $T \subseteq S$, the behaviours in $\mathcal{H}(r)_{|T}$ are exactly the behaviours in $\mathcal{H}(r')_{|T}$ if $p_j \leq r, r' < p_{j+1}$ for some $j \in [\ell]$ or if both $r, r' < p_1$ or both $r, r' \geq p_\ell$. Hence, $\mathcal{H}(r) \sim_S \mathcal{H}(r')$ if $p_j \leq r, r' < p_{j+1}$ for some $j \in [\ell]$, if $r, r' < p_1$, and if $r, r' \geq p_\ell$. The claim follows. ■

The concept of forbidden behaviours can be reminiscent of regularization penalties. We want to highlight that forbidden behaviours penalize violations of constraints deriven from prior knowledge, while regularization typically limits the complexity of the hypothesis class. In certain applications—such as the similarity function discussed earlier—these two notions may coincide conceptually. Yet, the results in this section focus on learning from such restriction by quantifying

how much they are violated on the training sample. Notably, it is unclear how these constraints could be incorporated into standard regularization penalties for total concepts. For instance, it is not straightforward to include the Lipschitz-type similarity restriction of a binary total concept into a regularization penalty, since almost any total function on a continuous domain would violate it. In contrast, the forbidden behaviour framework distinguishes hypotheses based on the extent to which they violate the Lipschitz criteria on the finite training set. It offers a tool to learn from such penalties by considering each candidate concept based on how well it classifies the sample and how much it violates the restrictions on it.

The bound on $\tau_{\mathbb{H}_{\mathcal{B}_k,p}}(m)$ in Lemma 8 leads the following result for cases where we group the candidate set based on the penalties over forbidden behaviours.

**Theorem 9** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class and $p$, be the penalty function on some forbidden behaviours $\mathcal{B}$. There exists a learner $\mathcal{A}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}(S),\mathcal{D}) \leq \min_{r\geq 0}\left\{\mathsf{err}(\mathcal{H}(r),\mathcal{D}) + O\left(\sqrt{\frac{(\mathsf{VC}(\mathcal{H}(r),D)+k)\log^2(m)+\log(1/\delta)}{m}}\right)\right\}.$$

**Structural risk minimization over data-dependent hierarchies.** Shawe-Taylor et al. (2002) introduce an algorithmic paradigm based on a luckiness function—a measure on data used to prefer hypotheses that appear 'luckier' on the training sample. While a complete comparison is out of the scope of this work, we highlight that their framework applies only to the realizable setting, as it relies on a technical assumption called *probably smoothness*, which bounds the number of hypotheses that are as lucky as a hypothesis on a double-sample. This condition enables an error guarantee for an ERM that selects the luckiest hypothesis. In contrast, our results address the general agnostic setting. Moreover, the framework developed here builds on a broader paradigm that allows learning from a collection of partial concepts by structuring into families of bounded growth.

## 4. Applications

In this section we discuss some application of forbidden behaviours.

### 4.1. Similarity Graph of the Domain

We assume that the prior knowledge comes as a weighted graph on the domain. Specifically, there exists a weight function $w : \mathcal{X} \times \mathcal{X} \to [0,1]$ where $w(x_1,x_2)$ reflects the certainty of the prior knowledge on the similarity between the labels of $x_1$ and $x_2$.

Based on this prior knowledge, we define the following forbidden behaviours and their penalties.

**Definition 10 ($\mathcal{B}_w, p_w$)** *The forbidden behaviours for similarity graph is defined as having opposite labels, i.e., $\mathcal{B}_w(x_1,x_2) = \{(0,1),(1,0)\}$, and its penalty is the weight of the edge between them, i.e., $p_w(x_1,x_2) = w(x_1,x_2)$.*

We then define the class of functions $\mathcal{H}_{p_w}$ as

$$\mathcal{H}_{p_w}(r) = \{h \in \{0,1,\star\}^{\mathcal{X}} : \exists h' \in \mathcal{H}, \forall x \in \mathcal{X}, h(x) = \star \text{ or } h(x) = h'(x) \\ \text{and } \forall x_1,x_2 \in \mathcal{X} \text{ with } h(x_1),h(x_2) \in \{0,1\}, \text{and } w(x_1,x_2) \geq r,\, h(x_1) = h(x_2)\}.$$

For a weight function $w$, we define shattering at weight $r$ by counting behaviours that give opposing labels only to points with weight less than $r$. Similar notions are defined for robust learning and learning with augmentation (Montasser et al., 2019; Attias et al., 2022; Shao et al., 2022).

**Definition 11** ($\mathsf{VC}_{w,r}(\mathcal{H})$) *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. For any $r \geq 0$, a set $U \in \mathcal{X}^*$ is called to be shattered at weight $r$ if (i) $U$ is shattered by $\mathcal{H}$ and (ii) for any $x_1, x_2 \in U$ we have $w(x_1, x_2) < r$. The $\mathsf{VC}_{w,r}(\mathcal{H})$ dimension of the hypothesis class $\mathcal{H}$ at distance $r$ is then defined as the size of the largest shattered set by distance $r$.*

We know from Lemma 8 that $\tau_{\mathbb{H}_{\mathcal{B}_w, p_w}}(m) = O(m^2)$. We can now apply Theorem 9 to conclude the following two bounds that combines the prior knowledge of a hypothesis class with the partial similarity knowledge.

**Theorem 12** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. There exists a learner $\mathcal{A}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}(S), \mathcal{D}) \leq \min_{r \geq 0} \left\{ \mathsf{err}(\mathcal{H}(r), \mathcal{D}) + O\left( \sqrt{\frac{(\mathsf{VC}_{w,r}(\mathcal{H}))\log^2(m) + \log(1/\delta)}{m}} \right) \right\}.$$

**Remark 13** *In Section 3.1 we defined hierarchical clustering as a way of expressing prior knowledge. We note that such a knowledge can also be given to the learner by giving the criteria of the clustering, i.e., a notion of similarity and the entire domain $\mathcal{X}$ as the unlabeled data. Each level of the hierarchy then joins clusters with a similarity above some threshold. We showed that in such a setting the growth rate of the clusterings at different levels, i.e., thresholds of the similarity, is $\tau_{\mathbb{H}_{\mathsf{HC}}}(m) \leq m$. The setting in this section gives the learner less knowledge—only the similarity measure, with no extra unlabeled data—and the resulting collection has a growth rate of $\tau_{\mathbb{H}_{\mathcal{B}_w, p_w}}(m) = O(m^2)$. As discussed earlier, the guarantee in Theorem 12 is achieved by taking multiple OIG learners trained on different subsets of the labeled sample, aggregating them by majority vote, and selecting the predictor with the tightest bound on the held-out. Changing the training subset changes the induced clustering the OIGs 'see', thereby increasing the number of equivalence classes. Contrary to Section 3.1, the knowledge of the clusterings on the entire sample is not given to them a priori. An interesting question would be understanding how we can improve over the bound in Theorem 12 by only having access to a finite unlabeled sample.*

### 4.2. Incorporating Prior Knowledge with Nearest Neighbour

In the following, we derive generalization bounds that take into account both the assumptions of a hypothesis class and also the assumption that labels of close points should be similar.

**Definition 14** ($\mathcal{B}_{\mathsf{NN}}, p_{\mathsf{NN},\phi}$) *The forbidden behaviours for nearest neighbour is defined as having opposite labels, i.e., $\mathcal{B}_{\mathsf{NN}}(x_1, x_2) = \{(0,1),(1,0)\}$, and its penalty is the inverse of the distance function $\phi : \mathcal{X}^2 \to \mathbb{R}$ between the points, i.e., $p_{\mathsf{NN},\phi}(x_1, x_2) = 1/\phi(x_1, x_2)$.*

The class of functions $\mathcal{H}_{p_{\mathsf{NN},\phi}}(1/r)$ for $r \geq 0$ based on the above forbidden behaviours will then become

$$\begin{aligned}\mathcal{H}_{p_{\mathsf{NN},\phi}}(1/r) = \{&h \in \{0,1,\star\}^{\mathcal{X}} : \exists h' \in \mathcal{H}, \forall x \in \mathcal{X}, h(x) = \star \text{ or } h(x) = h'(x) \\ &\text{and } \forall x_1, x_2 \in \mathcal{X} \text{ with } h(x_1), h(x_2) \in \{0,1\}, \text{and } \phi(x_1, x_2) \leq r, h(x_1) = h(x_2)\}.\end{aligned}$$

We define shattering at distance $r$ by only allowing behaviours that do not give opposing labels to points with distance smaller than $r$. This will then lead to the notion of VC dimension of $\mathcal{H}_{p_{\mathsf{NN},\phi}}$.

**Definition 15** ($\mathsf{VC}_{\phi,r}(\mathcal{H})$) *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. For any $r \geq 0$, a set $U \in \mathcal{X}^*$ is called to be shattered by distance $r$ if (i) $U$ is shattered by $\mathcal{H}$ and (ii) for any $x_1, x_2 \in U$ we have $\phi(x_1, x_2) > r$. The $\mathsf{VC}_{\phi,r}(\mathcal{H})$ dimension of the hypothesis class $\mathcal{H}$ at distance $r$ is then defined as the size of the largest shattered set by distance $r$.*

Noting that $\mathcal{B}_{\mathsf{NN}}$ is a forbidden behaviour on pairs of points so $\tau_{\mathbb{H}_{\mathcal{B}_{\mathsf{NN}},\phi}}(m) = O(m^2)$, and that $\mathsf{VC}(\mathcal{H}(1/r)) = \mathsf{VC}_{\phi,r}(\mathcal{H})$ we can apply Theorem 9 to conclude the following bound. This bound combines the prior knowledge of a hypothesis class with the smoothness of the labeling function, without requiring the knowledge of the optimum parameter of smoothness.

**Theorem 16** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. There exists a learner $\mathcal{A}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}(S), \mathcal{D}) \leq \min_{r \geq 0} \left\{ \mathsf{err}(\mathcal{H}(1/r), \mathcal{D}) + O\left( \sqrt{\frac{(\mathsf{VC}_{\phi,r}(\mathcal{H})) \log^2(m) + \log(1/\delta)}{m}} \right) \right\}.$$

#### 4.2.1. NEAREST NEIGHBOUR WITH NO PRIOR KNOWLEDGE ABOUT A CONCEPT CLASS.

We now show that if we set $\mathcal{H}$ to be the class of all functions, then we would recover bounds similar to some of the bounds derived for nearest neighbour in literature (Gottlieb et al., 2014b; Kontorovich et al., 2017; Hanneke and Kontorovich, 2021), etc.

**Proposition 17** *Let $\mathcal{H}$ be the class of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $\mathsf{VC}_{\phi,r}(\mathcal{H}) = |N_r|$, where $N_r$ is the $r$-net of the domain.*

**Proof** Consider any maximal set $U$ that is shattered. Any $x, x' \in U$ must have $\phi(x,x') > r$. Therefore, $U$ is a $r$-packing of $\mathcal{X}$. Assume for the sake of contradiction that $U$ is not a $r$-cover of $\mathcal{X}$. Then there exists $x_u \in \mathcal{X}$ such that $x_u \notin B(x,r)$ for all $x \in U$, where $B(x,r)$ is the ball of radius $r$ centered at $x$. But this implies that $\phi(x_u, x) > r$ for all $x \in U$. Therefore, $U \bigcup x_u$ is also shattered contradicting the maximality of $U$. This concludes that $U$ must be an $r$-net of $\mathcal{X}$. ■

**Theorem 18** *Let $\mathcal{H}$ be the class of all functions from $\mathcal{X}$ to $\{0,1\}$. Assume $\mathsf{diam}(\mathcal{X}) \leq R$ and $\mathsf{ddim}(\mathcal{X}) \leq d$, where $\dim(\mathcal{X})$ and $\mathsf{ddim}(\mathcal{X})$ are the diameter and doubling dimension of the space. There exists a learner $\mathcal{A} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\begin{aligned}
\mathsf{err}(\mathcal{A}(S), \mathcal{D}) &\leq \min_{r \geq 0} \left\{ \mathsf{err}(\mathcal{H}(1/r), \mathcal{D}) + O\left( \sqrt{\frac{|N_r| \log^2(m) + \log(1/\delta)}{m}} \right) \right\} \\
&\leq \min_{r \geq 0} \left\{ \mathsf{err}(\mathcal{H}(1/r), \mathcal{D}) + O\left( \sqrt{\frac{((R/r)^{d+1}) \log^2(m) + \log(1/\delta)}{m}} \right) \right\}.
\end{aligned}$$

### 4.3. Learning from Real-Valued Functions

In this section we show how the introduced paradigm can be used to learn from real-valued functions. We also show how the generalization bound achieved through this parameter compares with the well-known bounds on the fat-shattering dimension. We first define the necessary components to understand such bounds.

For a class $\mathcal{F}$ of real-valued functions from $\mathcal{X}$ to $\mathbb{R}$, we define by $\mathcal{F}(L) \subseteq \mathcal{F}$ the set of all functions in $\mathcal{F}$ with Lipschitz constant at most $L$. We also define by $\mathcal{F}^{0/1} = \{f^{0/1} \in \{0,1\}^{\mathcal{X}} : \exists f \in \mathcal{F}, \forall x \in \mathcal{X}, f^{0/1}(x) = \mathsf{sign}(f(x))\}$. For a real-valued function $f$ define $\mathsf{err}^{\gamma}(f,\mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[yf(x) < \gamma]$ and for any set $S$ define $\mathsf{err}^{\gamma}(f,S) = 1/|S| \sum_{(x,y)\in S} 1\{yf(x) < \gamma\}$. Moreover, define $\mathsf{err}^{\gamma}(\mathcal{F},\mathcal{D}) = \inf_{f\in\mathcal{F}} \mathsf{err}^{\gamma}(f,\mathcal{D})$ and $\mathsf{err}^{\gamma}(\mathcal{F},S) = \inf_{f\in\mathcal{F}} \mathsf{err}^{\gamma}(f,S)$.

We further define $\mathsf{err}^{*}(\mathcal{H},\mathcal{D},n) = \mathbb{E}_{S\sim\mathcal{D}^n}[\min_{h\in\mathcal{H}} \mathsf{err}(\mathcal{H},S)]$ as a related error rate for a partial concept $\mathcal{H}$. This error is defined in Alon et al. (2022) to set the benchmark of PAC learning partial concepts as competing with $\mathsf{err}^{*}(\mathcal{H},\mathcal{D}) = \lim_{n\to\infty} \mathsf{err}^{*}(\mathcal{H},\mathcal{D},n)$. All the generalization guarantees we have also hold for the benchmark of $\mathsf{err}^{*}(\mathcal{H},\mathcal{D})$. Moreover, it is shown in Alon et al. (2022) that we always have $\mathsf{err}^{*}(\mathcal{H},\mathcal{D}) \leq \mathsf{err}(\mathcal{H},\mathcal{D})$.

We now show how the $\mathsf{err}^{\gamma}$ of a class of real-valued functions is compared with the $0/1$ error of partial concepts $\mathcal{F}^{0/1}(1/r)$ that respect similarity of points at different distance levels $r$. Our bounds will be based on the error of $\mathcal{F}^{0/1}(.)$ and its VC dimension.

**Proposition 19** *Let $\mathcal{F}$ be a class of real-valued functions from $\mathcal{X}$ to $\mathbb{R}$. For any sample $S \in (\mathcal{X}\times\{0,1\})^*$, we have that $\mathsf{err}(\mathcal{F}^{0/1}(L/2\gamma),S) \leq \mathsf{err}^{\gamma}(\mathcal{F}(L),S)$. Furthermore, for any distribution $\mathcal{D}$ and any $L > 0$, we have $\mathsf{err}^{*}(\mathcal{F}^{0/1}(L/2\gamma),\mathcal{D}) \leq \mathsf{err}^{\gamma}(\mathcal{F}(L),\mathcal{D})$.*

**Proof** Let $f_L \in \mathcal{F}(L)$ and $S = \{(x_i,y_i)\}_{i=1}^n$ be any set. Define $b_i = \mathsf{sign}(f_L(x_i))$ if $|f_L(x)| \geq \gamma$ and $b_i = \star$ otherwise. We prove that the function $f_S$ with $f_S(x_i) = b_i$ for $i \in [n]$ and $f_S(x) = \star$ for $x \notin \mathsf{dom}(S)$ is in $\mathcal{F}^{0/1}(L/2\gamma)$. If $b_i = 1$ and $b_j = 0$ for some $i,j \in [n]$ then we have $f(x_i) \geq \gamma$ and $f(x_j) \leq -\gamma$. This implies that $|f(x_i) - f(x_j)| \geq 2\gamma$ and therefore, $|x_i - x_j| \geq 2\gamma/L$. This concludes that $f_S \in \mathcal{F}^{0/1}(L/2\gamma)$. On the other hand, for any $b_i = \star$ we have $y_i f_L(x_i) < \gamma$. Therefore, $f_L$ $\gamma$-errs on $x_i$ and so does $f_S$. For any $b_i \neq \star$, $f_L$ will $\gamma$-err on $x_i$ if and only if $\mathsf{sign}(f_L(x_i)) \neq y_i$, which means that $f_S$ will err on $x_i$ because $f_S(x_i) = \mathsf{sign}(f_L(x_i))$. The above concludes that for all $i \in [n]$, $f_S$ will error on $(x_i,y_i)$ if and only if $f_L$ $\gamma$-errs on $(x_i,y_i)$. We can therefore conclude that $\mathsf{err}(f_S,S) = \mathsf{err}^{\gamma}(f_L,S)$. Since we proved for any $f_L \in \mathcal{F}(L)$ such $f_S$ exists, we get that $\mathsf{err}(\mathcal{F}^{0/1}(L/2\gamma),S) \leq \mathsf{err}^{\gamma}(\mathcal{F}(L),S)$.

Moreover, the above also concludes that for any $f_L$ and $n$ we have that

$$\mathsf{err}^{\gamma}(f_L,\mathcal{D}) = \mathbb{E}_{S\sim\mathcal{D}^n}[\mathsf{err}^{\gamma}(f_L,S)] \geq \mathbb{E}_{S\sim\mathcal{D}^n}\left[\min_{f\in\mathcal{F}^{0/1}(L/2\gamma)} \mathsf{err}(f,S)\right] = \mathsf{err}^{*}(\mathcal{F}^{0/1}(L/2\gamma),\mathcal{D},n).$$

Since this holds for any $f_L \in \mathcal{F}(L)$ and any $n$, by letting $n \to \infty$ we get that $\mathsf{err}^{*}(\mathcal{F}^{0/1}(L/2\gamma),\mathcal{D}) \leq \mathsf{err}^{\gamma}(\mathcal{F}(L),\mathcal{D})$. ■

We give the definition of fat-shattering dimension and then compare it with the VC dimension of the translated partial functions $\mathcal{F}^{0/1}(.)$. This definition is introduced in Kearns and Schapire (1990) as a version of the pseudo-dimension (Pollard, 1990) that scales by a margin parameter.

**Definition 20** *We say a set $U \in \mathcal{X}^*$ is $\gamma$-shattered by a class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ if there exists a function $r : \mathcal{X} \to \mathbb{R}$ such that for every $b \in \{-1,1\}^U$, there exists a function $f \in \mathcal{F}$ with*

$b(x)(f(x) - r(x)) \geq \gamma$ *for all* $x \in U$. *The fat shattering dimension of the class* $\mathcal{F}$ *at scale* $\gamma$ *is denoted by* $\mathsf{fat}_\gamma(\mathcal{F})$ *and is the maximum size of a set that is* $\gamma$*-shattered by* $\mathcal{F}$.

**Proposition 21** *Let* $\mathcal{F}$ *be the class of* all *real-valued functions from* $\mathcal{X}$ *to* $\mathbb{R}$. *For any* $L > 0$, *we have* $\mathsf{VC}(\mathcal{F}^{0/1}(L/2\gamma)) = \mathsf{fat}_\gamma(\mathcal{F}(L))$. *Moreover, for any unlabeled set* $U$, *we have* $\mathsf{VC}(\mathcal{F}^{0/1}(L/2\gamma), U) = \mathsf{fat}_\gamma(\mathcal{F}(L), U)$, *where* $\mathsf{fat}_\gamma(\mathcal{F}(L), U) = \mathsf{fat}_\gamma(\mathcal{F}(L)_{|U})$.

**Proof** Since $\mathcal{F}(L)$ is the class of all $L$-Lipschitz functions, we know from Lemma 2 in Gottlieb et al. (2014a) that $\gamma$-fat shattering of $\mathcal{F}(L)$ does not change if we only consider $\forall x, r(x) = 0$ as a witness of shattering. In this case we know that if a set is $\gamma$-shattered, any pair of points in the set have distance at least $2\gamma/L$. Thus, it can be shattered by $\mathcal{F}^{0/1}(L/2\gamma)$. On the other hand, for any set that is shattered by $\mathcal{F}^{0/1}(L/2\gamma)$, we know that every pair is $2\gamma/L$ apart and therefore can be $\gamma$-shattered by $\mathcal{F}(L)$ too. ■

We now apply Theorem 16 by noting that $\mathcal{F}^{0/1}(L/2\gamma)$ is a translated version of $\mathcal{F}^{0/1}$ that only allows the forbidden behaviours with penalty at most $L/2\gamma$, i.e., pairs closer than $2\gamma/L$ must be labeled similarly. Hence, the growth rate of $\{\mathcal{F}^{0/1}(L/2\gamma) : \gamma > 0\}$ is bounded by $O(m^2)$.

**Theorem 22** *Let* $\mathcal{F}$ *be a class of functions from* $\mathcal{X}$ *to* $\mathbb{R}$ *with Lipschitz constant at most* $L$. *There exists a learner* $\mathcal{A} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$, *with the following property: for every distribution* $\mathcal{D}$, *every* $\delta \in (0,1)$ *and* $m \in \mathbb{N}$ *we have with probability at least* $1 - \delta$ *over samples* $S \sim \mathcal{D}^m$ *that*

$$\begin{aligned}\mathsf{err}(\mathcal{A}(S), \mathcal{D}) &\leq \min_{\gamma>0}\left\{\mathsf{err}(\mathcal{F}^{0/1}(L/2\gamma), \mathcal{D}) + O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{F}^{0/1}(L/2\gamma))\log^2(m) + \log(1/\delta)}{m}}\right)\right\} \\ &\leq \min_{\gamma>0}\left\{\mathsf{err}^\gamma(\mathcal{F}, \mathcal{D}) + O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{F}^{0/1}(L/2\gamma))\log^2(m) + \log(1/\delta)}{m}}\right)\right\}.\end{aligned}$$

*Furthermore, if* $\mathcal{F}$ *is the class of* all *functions from* $\mathcal{X}$ *to* $\mathbb{R}$ *with Lipschitz constant at most* $L$, $\mathsf{diam}(\mathcal{X}) \leq R$, *and* $\mathsf{ddim}(\mathcal{X}) \leq d$, *then we have*

$$\begin{aligned}\mathsf{err}(\mathcal{A}(S), \mathcal{D}) &\leq \min_{\gamma>0}\left\{\mathsf{err}^\gamma(\mathcal{F}, \mathcal{D}) + O\left(\sqrt{\frac{\mathsf{fat}_\gamma(\mathcal{F})\log^2(m) + \log(1/\delta)}{m}}\right)\right\} \\ &\leq \min_{\gamma>0}\left\{\mathsf{err}^\gamma(\mathcal{F}, \mathcal{D}) + O\left(\sqrt{\frac{(LR/\gamma)^{d+1}\log^2(m) + \log(1/\delta)}{m}}\right)\right\},\end{aligned}$$

*where the last line follows from the fact that* $\gamma$*-fat shattering dimension of the class of all* $L$ *Lipschitz functions from* $\mathcal{X}$ *to* $R$ *is bounded by* $(LR/\gamma)^{d+1}$.

Interestingly, since the above is proven using Theorem 16 and optimizing over $\mathcal{F}^{0/1}(1/r)$ for $r \geq 0$, we can further compete with the minimum of the bound over different Lipschitz constants. In other words, instead of requiring the concept class to only consist of Lipschitz functions with some constant, we can take any family $\mathcal{F}$ of real-valued functions. We then consider different subsets $\mathcal{F}(L/2\gamma)$ for $L > 0$ to create a collection $\mathbb{F} = \{\mathcal{F}(L/2\gamma) : \gamma, L > 0\}$ with $\tau_{\mathbb{F}}(m) = O(m^2)$, and then minimize the upper bound over $\mathcal{F}(L/2\gamma)$ for $\gamma, L > 0$. Note that these classes are embedded in each other, i.e., $\mathcal{F}(L/2\gamma) \subseteq \mathcal{F}(L'/2\gamma')$ for $L \leq L'$ and $\gamma \geq \gamma'$.

**Theorem 23** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. There exists a learner $\mathcal{A} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\begin{aligned}\mathsf{err}(\mathcal{A}(S),\mathcal{D}) &\leq \min_{\gamma, L>0}\left\{\mathsf{err}(\mathcal{F}^{0/1}(L/2\gamma),\mathcal{D}) + O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{F}^{0/1}(L/2\gamma))\log^2(m)+\log(1/\delta)}{m}}\right)\right\} \\ &\leq \min_{\gamma, L>0}\left\{\mathsf{err}^{\gamma}(\mathcal{F}(L),\mathcal{D}) + O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{F}^{0/1}(L/2\gamma))\log^2(m)+\log(1/\delta)}{m}}\right)\right\},\end{aligned}$$

*where $\mathcal{F}(L) \subseteq \mathcal{F}$ is the set of all functions in $\mathcal{F}$ with Lipschitz constant at most $L$. Furthermore, if $\mathcal{F}$ is the class of* all *functions from $\mathcal{X}$ to $\mathbb{R}$, $\mathsf{diam}(\mathcal{X}) \leq R$, and $\mathsf{ddim}(\mathcal{X}) \leq d$, we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\begin{aligned}\mathsf{err}(\mathcal{A}(S),\mathcal{D}) &\leq \min_{\gamma, L>0}\left\{\mathsf{err}^{\gamma}(\mathcal{F}(L),\mathcal{D}) + O\left(\sqrt{\frac{\mathsf{fat}_{\gamma}(\mathcal{F}(L))\log^2(m)+\log(1/\delta)}{m}}\right)\right\} \\ &\leq \min_{\gamma, L>0}\left\{\mathsf{err}^{\gamma}(\mathcal{F}(L),\mathcal{D}) + O\left(\sqrt{\frac{(LR/\gamma)^{d+1}\log^2(m)+\log(1/\delta)}{m}}\right)\right\}.\end{aligned}$$

It is useful to compare the above two theorems with the following bound in Bartlett (2002), which serves the same purpose of optimizing the value of $\gamma$ in the bound. It is derived by applying a stratification on values of $\gamma$ and taking a union bound over the guarantees that hold for learning with respect to a fixed $\gamma$. Such guarantees for fixed $\gamma$ are derived by controlling the generalization error with the covering number and then bounding the covering number based on $\gamma$-fat shattering dimension (Bartlett, 2002; Alon et al., 1997). Although it is not possible to exactly compare $\mathsf{fat}_{\gamma}(\mathcal{F})$ with $\mathsf{VC}(\mathcal{F}^{0/1}(2\gamma/L))$, we can still compare some specific cases. The following bound has an extra $O(\ln(1/\gamma)/m)$ dependence on the parameter $\gamma$, which makes the bound grow more rapidly with the decrease in $\gamma$. Particularly, the above results show that whenever the class consists of Lipschitz functions we can significantly improve the bounds. Moreover, in this setting, the following bound depends on $\mathsf{fat}_{\gamma/32}(\mathcal{F})$ which scales with an additional factor of $(32/\gamma)^d$, while the bound in the above theorem only scales with $(1/\gamma)^d$, saving an exponential factor of $32^d$.

**Theorem 24 (Corollary 9 in Bartlett (2002))** *Let $\mathcal{F}$ be a class of real-valued functions from $\mathcal{X}$ to $\mathbb{R}$. For every distribution $\mathcal{D}$ and $m \in \mathbb{N}$, we have with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ that $\forall f \in \mathcal{F}, \forall \gamma \in (0,1]$*

$$\mathsf{err}(h,\mathcal{D}) \leq \mathsf{err}^{\gamma}(h,S) + \sqrt{\frac{2}{m}\left(\mathsf{fat}_{\gamma/32}(\mathcal{F})\ln\left(\frac{34em}{\mathsf{fat}_{\gamma/32}(\mathcal{F})}\right)\log_2(578m) + \ln(\frac{8}{\gamma\delta})\right)}.$$

### 4.4. Smoothness and Sparse Margin Assumptions

In this section we discuss how smoothness and sparsness of the labeling rule can be incorporated into learning. For a data generating distribution $\mathcal{D}$, let $\eta_{\mathcal{D}}(x) = \mathcal{D}_{Y|X}[Y=1|X=x]$. We define

$$\phi_{\mathcal{D}}(\gamma) = \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[|\eta_{\mathcal{D}}(x) - 1/2| \leq \gamma\right]$$

to measure the sparseness of the probabilistic labeling rule inside a margin around $1/2$. For the metric space $(\mathcal{X}, \rho)$, we also define

$$\psi_{\mathcal{D}}(L) = \mathbb{P}_{x\sim\mathcal{D}}\left[\mathbb{P}_{z\sim\mathcal{D}_{\mathcal{X}}}\left[|\eta_{\mathcal{D}}(x) - \eta_{\mathcal{D}}(z)| > L\rho(x,z)\right] > 0\right]$$

to measure how probabilistically far is the labeling rule from being $L$-Lipschitz. It is worth noticing that whenever the labeling rule is deterministic, i.e., $\eta_{\mathcal{D}}(.) \in \{0,1\}$, the function $\psi_{\mathcal{D}}(L)$ is just the pairwise probabilistic Lipschitzness function at parameter $L$ (Urner and Ben-David, 2013; Pentina and Ben-David, 2018).

The following claims that it is possible to benefit from such conditions, even if the knowledge of the smoothness and the Lipschitz constant of the probabilistic labeling rule is not given. On the other hand, the guarantee is comparable with that of ERM up to additional factor of $O(\log(m)/\sqrt{m})$. Therefore, it offers a general paradigm with no dependence on the parameters of the underlying distribution while benefiting from its niceness, if it happens. Particularly, it offers a trade-off between the reduced complexity for restricted versions of the hypothesis class, i.e., $\mathsf{VC}(\mathcal{H}(L/2\gamma))$, and the additional error rate due to $\phi_{\mathcal{D}}(\gamma)$ and $\psi_{\mathcal{D}}(L)$. The following can be proved by noting that for $\mathbb{F} = \{\mathcal{F}(L/2\gamma) : L, \gamma > 0\}$ we have $\tau_{\mathbb{F}} = O(m^2)$ and that for any $\gamma, L > 0$, with high probability over $S \sim \mathcal{D}^m$ there exists a function in $\mathcal{F}(L/2\gamma)$ that agrees with $f^*_{\mathcal{D}}$ on almost $1 - \phi_{\mathcal{D}}(\gamma) - \psi_{\mathcal{D}}(L)$ fraction of $S$.

**Theorem 25** *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$. There exists a learner $\mathcal{A} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$, with the following property: for every distribution $\mathcal{D}$ such that $f^*_{\mathcal{D}} \in \mathcal{H}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}(S), \mathcal{D}) \leq \min_{\gamma, L>0}\left\{\mathsf{err}(f^*_{\mathcal{D}}, \mathcal{D}) + \phi_{\mathcal{D}}(\gamma) + \psi_{\mathcal{D}}(L) + O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{H}(L/2\gamma))\log^2(m) + \log(1/\delta)}{m}}\right)\right\},$$

*where $\psi_{\mathcal{D}}(L) = \mathbb{P}_{x,z\sim\mathcal{D}_{\mathcal{X}^2}}\left[|\eta_{\mathcal{D}}(x) - \eta_{\mathcal{D}}(z)| > L\rho(x,z)\right]$ is smoothness of the labeling rule $\eta_{\mathcal{D}}(.)$ and $\phi_{\mathcal{D}}(\gamma) = \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[|\eta(x;\mathcal{D}) - 1/2| \leq \gamma\right]$ is its sparseness.*

It is useful to compare Theorem 25 with the following result that is achievable by learning from the class of all real-valued $L$-Lipschitz functions using an ERM learner that minimizes the $\gamma$-error.

**Theorem 26** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ with Lipschitz constant at most $L$ and assume $\mathsf{diam}(\mathcal{X}) \leq R$, and $\mathsf{ddim}(\mathcal{X}) \leq d$. For $\gamma > 0$, we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}_{\gamma-\mathsf{ERM}}(S), \mathcal{D}) \leq \mathsf{err}(f^*_{\mathcal{D}}, \mathcal{D}) + \phi_{\mathcal{D}}(\gamma) + \psi_{\mathcal{D}}(L) + O\left(\sqrt{\frac{(LR/\gamma)^{d+1}\log^2(m) + \log(1/\delta)}{m}}\right),$$

*where $\mathcal{A}_{\gamma-\mathsf{ERM}}(S)$ chooses any hypothesis in $\mathcal{F}$ with minimum $\mathsf{err}^{\gamma}(S)$ and uses its sign for prediction.*

The proof of above again follows from the fact that the fat-shattering dimension of $\mathcal{F}$ is bounded by $(LR/\gamma)^{d+1}$ and that there exists a $L$-Lipschitz function that agrees with $f^*_{\mathcal{D}}$ except on almost a

$\phi_{\mathcal{D}}(\gamma) + \psi_{\mathcal{D}}(L)$ fraction of data. Theorem 25 improves the above in two fronts. First, it applies to any arbitrary class $\mathcal{H}$ and therefore, $\mathsf{VC}(\mathcal{H}(L/2\gamma))$ can be significantly smaller than $(LR/\gamma)^{d+1}$. Second, similar to learning real-valued functions, we get the advantage of relaxing the knowledge of $\gamma$ in advance and alternatively finding the minimum trade-off, among all values of $\gamma$, between the values of $\phi(\gamma)$, $\psi(\gamma)$, and the complexity term.

**Tsybakov Noise Condition.** The Tsybakov noise class $\mathsf{TN}(a, \alpha)$ is a common noise condition studied in the literature (Mammen and Tsybakov, 1999; Tsybakov, 2004). It is the class of all distributions $\mathcal{D}$ such that (i) the Bayes classifier $f^*_{\mathcal{D}}$ is in the hypothesis class reflecting prior knowledge and (ii) for all $\gamma > 0$ we have

$$\phi_{\mathcal{D}}(\gamma) = \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[|\eta_{\mathcal{D}}(x) - 1/2| \leq \gamma\right] \leq a'\gamma^{\alpha/(1-\alpha)},$$

where $a' = (1-\alpha)(2\alpha)^{\alpha/(1-\alpha)}a^{1/(1-\alpha)}$. This condition is a special case of the general setting discussed above. We have the following result for this class based on Theorem 25.

**Theorem 27** *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$. There exists a learner $\mathcal{A} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$, with the following property: for every $a, \alpha > 0$ and every distribution $\mathcal{D} \in \mathsf{TN}(a,\alpha)$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}(S), \mathcal{D}) \leq \min_{\gamma>0}\left\{\mathsf{err}(f^*_{\mathcal{D}}, \mathcal{D}) + a'\gamma^{\alpha/(1-\alpha)} + O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{H}(L/2\gamma))\log^2(m) + \log(1/\delta)}{m}}\right)\right\},$$

*where $L$ is the Lispchitz constant of the probabilistic labeling rule $\eta_{\mathcal{D}}(.)$.*

We want to highlight that while it is well-studied that Tsybakov noise condition is helpful in improving the dependency of the error rate on the sample size (Mammen and Tsybakov, 1999; Tsybakov, 2004; Bartlett et al., 2006), the above result is offering a trade-off that considers how this condition reduces the complexity of learning by considering subsets $\mathcal{H}(L/2\gamma) \subseteq \mathcal{H}$.

### 4.5. Incorporating Prior Knowledge with Contrastive Assumption

We consider a setting that has been studied as an approach for contrastive learning Saunshi et al. (2019); Awasthi et al. (2022); Alon et al. (2024). It is assumed that there exists a mapping under which points that are closer to each other tend to be 'more similar' in their labels. For instance, this mapping can be a result of pre-training. The prior literature mostly focuses on learning this representation and then finding a linear classifier for prediction of the mapped data. Here, we take a different approach and show that such information can be used by any hypothesis class as part of the prior knowledge.

We move beyond the strict assumption that for any $x$, if $x, x^+$ have opposing labels then any $x^-$ with $\phi(x, x^-) > \phi(x, x^-)$ must also have an opposite label to $x$. We suggest that this assumption holds only for triplets that lie inside a ball of radius at most $r$. We then consider the family of concept classes based on this assumption and apply Theorem 9 to optimize the distance $r$. We believe such a result is helpful because an algorithm that was previously designed to find the mapping as a pre-training stage is no longer required to make sure the assumption holds for every triplet in the domain. It may happen that there are mappings that only satisfy the assumption on balls where the distribution is supported and can conclude a smaller generalization bound.

**Definition 28** ($\mathcal{B}_{\mathsf{CN}}, p_{\mathsf{CN},\phi}$) *Let $x, x^+, x^- \in \mathcal{X}$ be any triplet of points such that $\phi(x, x^+) \leq \phi(x, x^-)$. The forbidden behaviours of contrastive assumption is defined as $\mathcal{B}_{\mathsf{NN}}(x, x^+, x^-) = \{(0,1,0), (1,0,1)\}$. The penalty $p_{\mathsf{CN},\phi}(x, x^+, x^-)$ is $1/\mathsf{rad}_\phi(x, x^+, x^-)$, where $\mathsf{rad}_\phi(x, x^+, x^-)$ is the minimum radius of a ball that contains $x$,$x^+$.*

The class of functions $\mathcal{H}_{p_{\mathsf{CN},\phi}}(r)$ based on the above forbidden behaviours will then become

$$\begin{aligned}\mathcal{H}_{p_{\mathsf{CN},\phi}}(1/r) = \{h \in \{0,1,\star\}^{\mathcal{X}} : \exists h' \in \mathcal{H}, \forall x \in \mathcal{X}, h(x) = \star \text{ or } h(x) = h'(x) \\ \text{and } \forall (x, x^+, x^-) \in \mathcal{X}^3 \text{ with } \mathsf{rad}_\phi(x, x^+, x^-) \leq r,\ \phi(x, x^+) < \phi(x, x^-), \\ \text{and } h(x), h(x^+), h(x^-) \in \{0,1\} \text{ if } h(x) = h(x^-) \text{ then } h(x) = h(x^+)\}.\end{aligned}$$

Motivated by the above, we define shattering at radius $r$, where at most two points can be shattered inside a ball of radius $r$.

**Definition 29** ($\mathsf{VC}^{\mathsf{rad}}_{\phi,r}(\mathcal{H})$) *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. For any $r \geq 0$, a set $S \subseteq \mathcal{X}$ is called to be shattered at radius $r$ if (i) $S$ is shattered by $\mathcal{H}$ and (ii) for any ball $B \subseteq \mathcal{X}$ of radius $r$, we have $|S \cap B| \leq 2$. The $\mathsf{VC}^{\mathsf{rad}}_{\phi,r}(\mathcal{H})$ dimension of the hypothesis class $\mathcal{H}$ at radius $r$ is then defined as the size of the largest shattered set at distance $r$.*

Noting that $\tau_{\mathbb{H}_{\mathcal{B}_{\mathsf{CN}},\phi}}(m) = O(m^3)$, we can apply Theorem 9 to the forbidden behaviours that respect the contrastive assumption. This yields an error bound that optimally, among $r \geq 0$, incorporates the contrastive assumption with the hypothesis class.

**Theorem 30** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. There exists a learner $\mathcal{A}$, with the following property: for every distribution $\mathcal{D}$, every $\delta \in (0,1)$ and $m \in \mathbb{N}$ we have with probability at least $1-\delta$ over samples $S \sim \mathcal{D}^m$ that*

$$\mathsf{err}(\mathcal{A}(S), \mathcal{D}) \leq \min_{r \geq 0} \left\{ \mathsf{err}(\mathcal{H}(1/r), \mathcal{D}) + O\left( \sqrt{\frac{\left(\mathsf{VC}^{\mathsf{rad}}_{\phi,r}(\mathcal{H})\right) \log^2(m) + \log(1/\delta)}{m}} \right) \right\}.$$

#### 4.5.1. CONTRASTIVE APPROACH WITH NO PRIOR KNOWLEDGE ABOUT A CONCEPT CLASS.

We know show that if we set the class to be the class of functions, then we would recover bounds similar to the bounds derived for nearest neighbour approach.

**Proposition 31** *Let $\mathcal{H}$ be the class of all functions from $\mathcal{X}$ to $\{0,1\}$. Then, $|N_{2r}| \leq \mathsf{VC}^{\mathsf{rad}}_{\phi,r}(\mathcal{H}) \leq 2|N_r|$, where $N_r$ denotes the $r$-net of domain $\mathcal{X}$.*

**Proof** Take a maximal set $U$ that is shattered and an $r$-net $N$ of the domain $\mathcal{X}$. Map every point $x \in U$ to an element in $a \in N$ such that $\phi(x, a) \leq r$. This must exists because $N$ is an $r$-cover. Now for any $a \in N$, the inverse of the mapping must contain at most 2 points from $U$, otherwise, three points have been mapped to $a$ which means they are all in $B(a, r)$. This concludes that $|U| \leq 2|N_r|$. Take any maximum $2r$-packing $N'$, which is indeed a $2r$-net. This set is obviously shattered because every $a, b \in N'$ has $\phi(a, b) > 2r$ and cannot lie inside a ball of radius $r$. ■

**Remark 32** *Regarding the relation between the two notions, we have that* $\mathsf{VC}(\mathcal{H}_{p_{\mathsf{NN},\phi}}(1/2r)) \leq \mathsf{VC}(\mathcal{H}_{p_{\mathsf{CN},\phi}}(1/r)) \leq 2\mathsf{VC}(\mathcal{H}_{p_{\mathsf{NN},\phi}}(1/r))$ *and* $\mathsf{err}(\mathcal{H}_{p_{\mathsf{CN},\phi}}(1/r), S) \leq \mathsf{err}(\mathcal{H}_{p_{\mathsf{NN},\phi}}(1/2r), S)$ *on any sample* $S$.

**Theorem 33** *Let* $\mathcal{H}$ *be the class of all functions from* $\mathcal{X}$ *to* $\{0,1\}$. *Assume* $\mathsf{diam}(\mathcal{X}) \leq R$ *and* $\mathsf{ddim}(\mathcal{X}) \leq d$. *There exists a learner* $\mathcal{A} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$, *with the following property: for every distribution* $\mathcal{D}$, *every* $\delta \in (0,1)$ *and* $m \in \mathbb{N}$*we have with probability at least* $1-\delta$ *over samples* $S \sim \mathcal{D}^m$ *that*

$$\begin{aligned}\mathsf{err}(\mathcal{A}(S), \mathcal{D}) &\leq \min_{r \geq 0} \left\{ \mathsf{err}(\mathcal{H}(1/r), \mathcal{D}) + O\left( \sqrt{\frac{2|N_r| \log^2(m) + \log(1/\delta)}{m}} \right) \right\} \\ &\leq \min_{r \geq 0} \left\{ \mathsf{err}(\mathcal{H}(1/r), \mathcal{D}) + O\left( \sqrt{\frac{((R/r)^{d+1}) \log^2(m) + \log(1/\delta)}{m}} \right) \right\}.\end{aligned}$$

## Acknowledgments

Shai Ben-David was supported by NSERC through a Discovery grant and by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. Alireza F. Pour was supported by Vector Institute Research Grant and Cheriton Graduate Scholarship.

## References

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

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

Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Idan Attias, Steve Hanneke, and Yishay Mansour. A characterization of semi-supervised adversarially robust pac learnability. *Advances in Neural Information Processing Systems*, 35:23646–23659, 2022.

Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. Do more negative samples necessarily hurt in contrastive learning? In *International conference on machine learning*, pages 1101–1116. PMLR, 2022.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 2002.

Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. *Advances in Neural Information Processing Systems*, 29, 2016.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014a.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. *Advances in Neural Information Processing Systems*, 27, 2014b.

Steve Hanneke and Aryeh Kontorovich. Stable sample compression schemes: New applications and an optimal svm margin bound. In *Algorithmic Learning Theory*, pages 697–721. PMLR, 2021.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

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

Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 382–391. IEEE, 1990.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2002.

Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. *Advances in Neural Information Processing Systems*, 30, 2017.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

Anastasia Pentina and Shai Ben-David. Multi-task $\{$K$\}$ ernel $\{$L$\}$ earning based on $\{$P$\}$ robabilistic $\{$L$\}$ ipschitzness. In *Algorithmic Learning Theory*, pages 682–701. PMLR, 2018.

David Pollard. Empirical processes: theory and applications. Ims, 1990.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

Robert E Schapire and Yoav Freund. Boosting: Foundations and algorithms. *Kybernetes*, 42(1): 164–166, 2013.

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

Han Shao, Omar Montasser, and Avrim Blum. A theory of pac learnability under transformation invariances. *Advances in Neural Information Processing Systems*, 35:13989–14001, 2022.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5): 1926–1940, 2002.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Ruth Urner and Shai Ben-David. Probabilistic lipschitzness a niceness assumption for deterministic labels. In *Learning Faster from Easy Data-Workshop@ NIPS*, volume 2, page 1, 2013.

## Appendix A. Missing Proofs from Section 2

We prove our generalization bounds using a compression-based argument. We will employ the one-inclusion graph predictor and a boosting algorithm to prove the existence of compression sets.

**Definition 34 (One-Inclusion Graph)** *Let $\mathcal{H}$ be a (partial) concept class and $U \subseteq \mathcal{X}$ an unlabeled set. The one-inclusion graph $\mathcal{G}(\mathcal{H}_{|S}) = (V, E)$ is a a graph with node set $V = \mathcal{H}_{|S}$ and edge set*

$$E = \{\{f, g\}, f, g \in V, |\{x \in S : f(x) \neq g(x)\}| = 1\}.$$

*In other words, each $h \in \mathcal{H}_{|S}$ is a vertex in this graph and two vertices are connected by an edge if they are different in only one element of $S$.*

**Definition 35 (Orientations of One-Inclusion Graph)** *Let $\mathcal{H}$ be a (partial) concept class and $U \subseteq \mathcal{X}$ an unlabeled set. A valid orientation of the one-inclusion graph $\mathcal{G}(\mathcal{H}_{|S})$ is a mapping $\sigma : E \to V$ such that $\sigma(e) \in e$ for all $e \in E$.*

For the one-inclusion graph $\mathcal{G}(\mathcal{H}_{|S})$, we will define the out-degree of a vertex $v \in V$ in as $\mathsf{deg}(v) = |\{e \in E : \sigma(e) \neq v\}|$ and the max out-degree is defined as $\mathsf{max} - \mathsf{deg}(\mathcal{G}(\mathcal{F}_{|S})) \max_{v \in V} \mathsf{deg}(v)$

We now define the one-inclusion predictor $\mathcal{A}_{\mathcal{H}}$ for the hypothesis class $\mathcal{H}$, introduced by Haussler et al. (1994).

**Definition 36 (One-Inclusion Graph Predictor)** *The one-inclusion predictor $\mathcal{A}_{\mathcal{H}} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$ for the hypothesis class $\mathcal{H}$, introduced by Haussler et al. (1994) predict as follows. Given the labeled set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \{0,1\})^n$ and a test point $x$ the predictor will create the one-inclusion graph $\mathcal{G}(\mathcal{H}_{|S \cup \{x\}})$. It then finds among all the valid orientations an orientation $\sigma_{S \cup \{x\}}$ that minimizes $\mathsf{max} - \mathsf{deg}(\mathcal{G}(\mathcal{F}_{|S}))$. If there exists only one hypothesis $h \in V$ that is consistent $S$, then predictor predicts $\mathcal{A}(S)(x) = h(x)$. Otherwise, it will predict according to the orientation of the edge between $e = \{f, g\}$ the two nodes $f, g$ that are consistent with $S$, i.e., $f_{|\mathsf{dom}(S)} = g_{|\mathsf{dom}(S)} = S$. Formally, we have $\mathcal{A}(S)(x) = h(x)$ if $\sigma_{S \cup \{x\}}(e) = h$. Finally, if no node is consistent with $S$, then predict $\mathcal{A}(S)(x) = 0$.*

Haussler et al. (1994) prove that for any total concept class $\mathcal{H}$ there exists an orientation with out-degree at most $\mathsf{VC}(\mathcal{H})$ and that the one-inclusion graph predictor can be used as a learner achieving $\frac{\mathsf{VC}(\mathcal{H})}{n+1}$ leave-one-out error. Alon et al. (2022) show that essentially the same result holds for any partial concept class and by requiring $\mathcal{A}_{\mathcal{H}}$ to only consider total behaviours on its nodes.

**Lemma 37 (Theorem 2.3 of Haussler et al. (1994), Lemma 35 of Alon et al. (2022))** *For any hypothesis class $\mathcal{H}$ with VC dimension $\mathsf{VC}(\mathcal{H})$ there is a function $\mathcal{A}_{\mathcal{H}} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$ such that for any $n \in \mathbb{N}$ and any sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \{0,1\})^n$ realizable by $\mathcal{H}$ we have*

$$\frac{1}{n!} \sum_{\gamma \in \Gamma} 1\left\{\mathcal{A}\left(x_{\gamma(1)}, y_{\gamma(1)}, \ldots, x_{\gamma(n-1)}, y_{\gamma(n-1)}, x_{\gamma(n)}\right) \neq y_{\gamma(n)}\right\} \leq \frac{\mathsf{VC}(\mathcal{H})}{n},$$

*where $\Gamma$ denotes the symmetric group on $[n]$.*

Finally, we define the notion of a compression scheme.

**Definition 38 (Compression Scheme)** *A compression scheme is a pair $(\rho, \kappa)$ where $\kappa : (\mathcal{X} \times \{0,1\})^* \to (\mathcal{X} \times \{0,1\})^*$ is a compression map and $\rho : (\mathcal{X} \times \{0,1\})^* \to \{0,1\}^{\mathcal{X}}$ is a decompression map. For a set $S$, the size of the compression is defined by $|\kappa(S)|$ and the reconstruction error will be defined by $\mathsf{err}(\rho(\kappa(S)), S)$.*

It is possible to show the existence of a compression scheme by a well-known technique of boosting the error of the the one-inclusion graph learners through majority voting. This boosting technique is discussed in Lemma 40 and is used in many works to prove the existence of compression schemes, David et al. (2016); Alon et al. (2022), etc. This could also be extended to find a compression scheme with small reconstruction error using the reduction to realizable technique David et al. (2016).

We first show a bound on $\tau_{\mathbb{H}}(.)$ lets us reliably estimate, for all $\mathcal{H} \in \mathbb{H}$, the generalization error of mappings that are based on aggregations of OIG learners $\mathcal{A}_{\mathcal{H}}$ that are based on learning from subsets of the sample. A similar argument is used to prove the generalization error of AdaBoost through hybrid compression schemes, that is, decompression functions that can be chosen from a class, but for the case of realizable compression schemes Schapire and Freund (2013). We prove an agnostic type guarantee and show that decompression error of schemes associated with every $\mathcal{H} \in \mathbb{H}$ is a good estimate of their expected error. Informally, we show that the behaviour of the schemes $\{\mathbb{A}_{\mathcal{H}}(.), \mathcal{H} \in \mathbb{H}\}$, as defined in the follwing, on the reconstructed part can be bounded by $\tau_{\mathbb{H}}(.)$.

For a set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ and a set of indices $I \subseteq [n]$, we denote $S_I := ((x_i, y_i))_{i \in I}$. Moreover for any two integers $a \leq b$ we denote $S_{a:b} := ((x_i, y_i))_{a \leq i \leq b}$

**Lemma 39 (Decompression schemes for $\mathbb{H}$)** *Let $\mathbb{H}$ be a collection of hypothesis classes. Let $k, T \in \mathbb{N}$ and denote $\mathbb{A}_{\mathcal{H}}(S_1, \ldots, S_T)(.) :=$ Majority$(\mathcal{A}_{\mathcal{H}}(S_1)(.), \ldots, \mathcal{A}_{\mathcal{H}}(S_T)(.))$. For any $m \in \mathbb{N}$, $\delta \in (0,1)$, and any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, denoting $B(\mathbb{H}, k, T, m) = (\log(\tau_{\mathbb{H}}(m)) + kT)\log(m)$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ for all $S_1, \ldots, S_T \in \{(x_1, y_1), \ldots, (x_m, y_m)\}^k$*

*and $\mathcal{H} \in \mathbb{H}$, that*

$$|\mathsf{err}(\mathbb{A}_{\mathcal{H}}(S_1,\ldots,S_T),\mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H}}(S_1,\ldots,S_T),S)| \leq$$
$$c\sqrt{\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),V_I)\frac{B(\mathbb{H},k,T,m)+\log(\frac{1}{\delta})}{m}} + c\frac{B(\mathbb{H},k,T,m)+\log(\frac{1}{\delta})+k}{m}$$
$$\leq c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m))+kT)\log(m)+\log(\frac{1}{\delta})}{m}}$$

*for some constants $c$ and $c'$.*

**Proof** For a fixed $S_1,\ldots,S_T$, let $V := S \setminus \bigcup S_i$. We know that the size of $[\mathcal{H} \sim_S]_{\mathbb{H}}$ is bounded by $\tau_{\mathbb{H}}(m)$. Note that for any $\mathcal{H}$ and $\mathcal{H}'$ such that $\mathcal{H} \sim_S \mathcal{H}'$, we have that $\mathcal{A}_{\mathcal{H}}(S_i)(x) = \mathcal{A}_{\mathcal{H}'}(S_i)(x)$ for all $x \in V$ and $i \in [T]$. This is because $\mathcal{H}$ and $\mathcal{H}'$ will induce the exact same (total) restriction on any subset of $S$, including $S_i \cup x$. Therefore, having a fixed mapping of graphs to valid orientations that minimize max out-degree, we can observe that we have $\mathcal{G}(\mathcal{H}_{|S\cup\{x\}} = \mathcal{G}(\mathcal{H}'_{|S\cup\{x\}}$ and the predictors $\mathcal{A}_{\mathcal{H}}$ and $\mathcal{A}_{\mathcal{H}'}$ will use the same orientation and thus have the same prediction the same on $V$. Thus, the total number of behaviours that $\{\mathbb{A}_{\mathcal{H}}(S_1,\ldots,S_T) : \mathcal{H} \in \mathbb{H}\}$ induces on $V$ is bounded by the number of equivalence classes on $S$, which is at most $\tau_{\mathbb{H}}(m)$. The above implies that if we fix the set of indices $I = (i_1,\ldots,i_{kT}) \in [m]^{kT}$ before observing the sample and consider the class $\mathcal{F}_I = \{\mathbb{A}_{\mathcal{H},I}(.) : \mathcal{H} \in \mathbb{H}\}$, that on any $S \sim \mathcal{D}^m$, runs the majority vote of OIG predictors on the indices in $I$, i.e. $\mathbb{A}_{\mathcal{H},I}(.) = \text{Majority}(\mathcal{A}_{\mathcal{H}}(S_{1:k})(.),\ldots,\mathcal{A}_{\mathcal{H}}(S_{k(T-1):kT})(.))$, then the instances in $V_I := S_{[n]\setminus I}$ will remain i.i.d. and also $\mathsf{VC}(\mathcal{F}_I,V_I) \leq \log(\tau_{\mathbb{H}}(m))$. Therefore, letting $\delta' = \frac{\delta}{m^{kT}}$, we apply the uniform convergence property of $\mathcal{F}_I$ to conclude that for some constant $c$, with probability at least $1-\delta'$ over $S \sim \mathcal{D}^m$ we have for all $\mathcal{H} \in \mathbb{H}$,

$$|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),\mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),V_I)| \leq$$
$$c\sqrt{\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),V_I)\frac{B(\mathbb{H},k,T,m)+\log(\frac{1}{\delta})}{m-kT}} + c\frac{B(\mathbb{H},k,T,m)+\log(\frac{1}{\delta})}{m-kT}.$$

Now a union bound over all $m^{kT}$ sequence of indices $I \in [m]^{kT}$ implies that with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ we have for all $I \in [m]^{kT}$ and $\mathcal{H} \in \mathbb{H}$,

$$|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),\mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),V_I)| \leq$$
$$c\sqrt{\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),V_I)\frac{B(\mathbb{H},k,T,m)+\log(\frac{1}{\delta})}{m-kT}} + c\frac{B(\mathbb{H},k,T,m)+\log(\frac{1}{\delta})}{m-kT}.$$

Note that in the above we have used a version of uniform convergence that is derived by Bernstein's inequality, see Corollary 5.2 in Boucheron et al. (2005). We will need this Bernstein version to be able to connect bound with high probability the deviation between $\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),\mathcal{D})$ and $\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),S)$. The hybrid compression argument in Schapire and Freund (2013) uses the standard bounds used to prove the uniform convergence for VC classes that does not have the multiplicative factor of $\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),V_I)$ in the bound. This is because they only discuss realizable setting and when the validation error is exactly zero.

Now we will use the above bound on the difference between $\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),\mathcal{D})$ and $\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),V_I)$ to bound the difference between $\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),\mathcal{D})$ and $\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S),S)$. This is discussed in Lemma A.1

in David et al. (2016) for a single predictor $h$, that is, a single decompression of a compressed subset. We prove this again for completeness and because we want the property to hold for the decompressions associated with every $\mathcal{H} \in \mathbb{H}$ . We know that $|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), V_I) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), S)| \leq kT/m$ and, thus, $|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), \mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), V_I)| \geq |\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), \mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), S)| - kT/m$. Since $kT \leq m/2$, for any fixed $\mathcal{H} \in \mathbb{H}$, we can conclude the event that

$$|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), \mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), V_I)| \geq$$
$$c\sqrt{\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), V_I)\frac{B(\mathbb{H},k,T,m) + \log(1/\delta)}{m - kT}} + c\frac{B(\mathbb{H},k,T,m) + \log(1/\delta)}{m - kT}. \tag{1}$$

is implied by the event

$$|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), \mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), S)| \geq$$
$$c\sqrt{\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), V_I)\frac{B(\mathbb{H},k,T,m) + \log(\frac{1}{\delta})}{m}} + c\frac{B(\mathbb{H},k,T,m) + \log(\frac{1}{\delta}) + k}{m}. \tag{2}$$

Therefore, the event $E_1$ that there exists some $\mathcal{H} \in \mathbb{H}$ satisfying 1 is implied by the event $E_2$ that there exists $\mathcal{H} \in \mathbb{H}$ satisfying 2. Since the first event $E_1$ happens with probability at most $\delta$, the event $E_2$ will happen with probability at most $\delta$ too. Hence, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$ we have for all $I \in [m]^{kT}$ and $\mathcal{H} \in \mathbb{H}$,

$$|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), \mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), S)| \leq$$
$$c\sqrt{\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), V_I)\frac{B(\mathbb{H},k,T,m) + \log(\frac{1}{\delta})}{m}} + c\frac{B(\mathbb{H},k,T,m) + \log(\frac{1}{\delta}) + k}{m},$$

as desired. ■

We now introduce the guarantee that we can get by boosting weak-learners which we will use to prove the next theorem.

**Lemma 40 (Boosting)** *Let $k, m \in \mathbb{N}$ and $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \{0,1\})^m$ be a set. Assume there exists an algorithm $\mathcal{A} : (\mathcal{X} \times \{0,1\})^k \times \mathcal{X} \to \{0,1\}$ such that for every distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$ with $\mathcal{D}[S] = 1$, there exists $S_{\mathcal{D}} \in \{(x_1, y_1), \ldots, (x_m, y_m)\}^k$ such that $\mathsf{err}(\mathcal{A}(S_{\mathcal{D}}), \mathcal{D}) \leq 1/3$. Then there exists a constant $c \geq 1$ such that there exists $S_1, \ldots, S_T \in \{(x_1, y_1), \ldots, (x_m, y_m)\}^k$ with $T = \lceil c \log m \rceil$ such that for the function $\hat{h}(.) = $ Majority$(\mathcal{A}(S_1)(.), \ldots, \mathcal{A}(S_T)(.))$ we have $\hat{h}(x_i) = y_i$ for all $i \in [m]$.*

**Proof of Theorem 3.** Fix $\mathcal{H} \in \mathbb{H}$ and let $\delta' = \delta/2$. We know with probability at least $1 - \delta'$ over $S \sim \mathcal{D}^m$ that there exists $\hat{h} \in \mathcal{H}$ with $\mathsf{err}(\hat{h}, S) \leq \mathsf{err}(\mathcal{H}, \mathcal{D}) + O(\sqrt{\frac{\log(1/\delta')}{m}})$ by e.g., a Hoeffding's inequality. Let $S'$ be the longest realizable subsequence of $S$ and note that $S' \geq (1 - \mathsf{err}(\hat{h}, S))|S|$. We now show that there exists $S_1, \ldots, S_T$ for $T = \lceil c'' \log(m) \rceil$ for some constant $c''$ such that $\mathbb{A}_{\mathcal{H}}(S_1, \ldots, S_T)(x) :=$ Majority$(\mathcal{A}_{S_1}(x), \ldots, \mathcal{A}_{S_T}(x))$ is equal to $y$ for any $(x, y) \in S'$.

Observe that for every distribution $\mathcal{P}$ over $\mathcal{X} \times \{0,1\}$ that is realizable with $\mathcal{H}$, we have for $\mathcal{A}_{\mathcal{H}}$ that

$$\begin{aligned}
&\mathbb{E}_{R\sim\mathcal{P}^n,(x,y)\sim\mathcal{P}}\left[1\{\mathcal{A}_{\mathcal{H}}(R)(x)\neq y\}\right] \\
&= \frac{1}{(n+1)!}\sum_{\gamma\in\Gamma}\mathbb{E}_{R\sim\mathcal{P}^{n+1}}\left[1\left\{\mathcal{A}\left(x_{\gamma(1)},y_{\gamma(1)},\ldots,x_{\gamma(n-1)},y_{\gamma(n-1)}\right)\left(x_{\gamma(n)}\right)\neq y_{\gamma(n)}\right\}\right] \\
&= \mathbb{E}_{R\sim\mathcal{P}^{n+1}}\left[\frac{1}{(n+1)!}\sum_{\gamma\in\Gamma}1\left\{\mathcal{A}\left(x_{\gamma(1)},y_{\gamma(1)},\ldots,x_{\gamma(n-1)},y_{\gamma(n-1)}\right)\left(x_{\gamma(n)}\right)\neq y_{\gamma(n)}\right\}\right] \\
&\leq \frac{\mathbb{E}_{R\sim\mathcal{P}^{n+1}}\left[\mathsf{VC}(\mathcal{H},R)\right]}{n+1},
\end{aligned}$$

where the last line is due to Lemma 37. Particularly, Lemma 2 in Haussler (1995) proves by a shifting argument that we can bound the density of the OIG $\mathcal{G}(\mathcal{H}_{|R})$ by $\mathsf{VC}(\mathcal{H},R)$. The proof of Lemma 37 is based on relating the error rate to the density of OIG (see Haussler et al. (1994)), which gives the above inequality. Now, for any distribution $\mathcal{P}'$ over $S'$ we know that this distribution is realizable by $\mathcal{H}$ and therefore, $\mathbb{E}_{R\sim\mathcal{P}^n,(x,y)\sim\mathcal{P}'}\left[1\{\mathcal{A}_{\mathcal{H}}(R)(x)\neq y\}\right] \leq \frac{\mathsf{VC}(\mathcal{H},R)}{n+1} \leq \frac{\mathsf{VC}(\mathcal{H})}{n+1}$. Let $n = 3\mathsf{VC}(\mathcal{H})$ so that $\mathbb{E}_{R\sim\mathcal{P}^n,(x,y)\sim\mathcal{P}'}\left[1\{\mathcal{A}_{\mathcal{H}}(R)(x)\neq y\}\right] \leq 1/3$. This proves the existence of a set $R'\in S'^k$ of size $k := 3\mathsf{VC}(\mathcal{H})$ with $\mathsf{err}(\mathcal{A}_{\mathcal{H}}(R'),\mathcal{P}') \leq 1/3$. This is sufficient to invoke Lemma 40 and conclude that the claimed $S_1,\ldots,S_T$ exists. Therefore, we have $S_1,\ldots,S_T$ such that $\mathsf{err}(\mathbb{A}_{\mathcal{H}}(S_1,\ldots,S_T),S) \leq \mathsf{err}(\hat{h},S)$. Let $I^*$ denote the sequence of the indices of the sets $S_1,\ldots,S_T$ in the same order.

We now describe the learner $\mathcal{A}_{\mathbb{H}}$. The learner will go over all sets $J\in[m]^i$ for any $i\in[m]$. For any $J$ it runs $\mathbb{A}_{\mathcal{H},J}(S)$ for all $\mathcal{H}\in\mathbb{H}$. It then evaluates $\mathsf{err}(\mathbb{A}_{\mathcal{H},J}(S),S)$ and returns a predictor $\mathcal{A}_{\mathbb{H}}(S)(.) := \mathbb{A}_{\bar{\mathcal{H}},\bar{J}}(S)(.)$ as follows:

$$\mathcal{A}_{\mathbb{H}}(S)(.) \in \arg\min_{\substack{J\in[m]^i,i\in[m]\\ \mathcal{H}\in\mathbb{H}}} \mathsf{err}(\mathbb{A}_{\mathcal{H},J}(S),S) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m))+|J|)\log(m)+\log(1/\delta)}{m}}$$

Set $\delta'' = \delta'/m$. For any fixed $k'\leq\lceil m/T\rceil$ we know from Lemma 39 that with probability at least $1-\delta''$ over $S\sim\mathcal{D}^m$ for all $\mathcal{H}\in\mathbb{H}$ and all $J\in[m]^{k'T}$ we have

$$\mathsf{err}(\mathbb{A}_{\mathcal{H},J}(S),\mathcal{D}) \leq \mathsf{err}(\mathbb{A}_{\mathcal{H},J}(S),S) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m))+k'T+1)\log(m)+\log(1/\delta')}{m}}$$

Taking a union bound over all values of $k'\leq\lceil m/T\rceil$ implies that with probability at least $1-\delta'$ over $S\sim\mathcal{D}^m$ for all $\mathcal{H}\in\mathbb{H}$, all $k'\leq\lceil m/T\rceil$ and $J\in[m]^{k'T}$ we have

$$\mathsf{err}(\mathbb{A}_{\mathcal{H},J}(S),\mathcal{D}) \leq \mathsf{err}(\mathbb{A}_{\mathcal{H},J}(S),S) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m))+k'T+1)\log(m)+\log(1/\delta')}{m}}. \quad (3)$$

We proved that there exists sets $S_1,\ldots,S_T$ such that $\mathsf{err}(\mathbb{A}_{\mathcal{H},I^*}(S),S) \leq \mathsf{err}(\hat{h},S)$ and hence for the returned predictor $\mathbb{A}_{\bar{\mathcal{H}},\bar{J}}(S)(.)$ we have

$$\mathsf{err}(\mathbb{A}_{\bar{\mathcal{H}},\bar{J}}(S),S) \leq \mathsf{err}(\hat{h},S) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m))+k'T+1)\log(m)+\log(1/\delta')}{m}}.$$

Taking a union bound over the events that $\mathsf{err}(\hat{h}, S)$ is a good estimate of $\mathsf{err}(\mathcal{H}, \mathcal{D})$, and the event in Equation 3, we get with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ that

$$\begin{aligned}
&\mathsf{err}(\mathcal{A}_{\mathbb{H}}(S), \mathcal{D}) \\
&\leq \mathsf{err}(\mathbb{A}_{\bar{\mathcal{H}},\bar{J}}(S), S) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m)) + |\bar{J}| + 1)\log(m) + \log(2/\delta)}{m}} \\
&\leq \mathsf{err}(\mathbb{A}_{\mathcal{H},I^*}(S), S) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m)) + kT + 1)\log(m) + \log(2/\delta)}{m}} \\
&\leq \mathsf{err}(\hat{h}, S) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m)) + \mathsf{VC}(\mathcal{H}))c''\log(m) + 1)\log(m) + \log(2/\delta)}{m}} \\
&\leq \mathsf{err}(\mathcal{H}, \mathcal{D}) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m)) + \mathsf{VC}(\mathcal{H}))c''\log(m) + 1)\log(m) + \log(2/\delta)}{m}} + \sqrt{\frac{\log(2/\delta)}{m}} \\
&\leq \mathsf{err}(\mathcal{H}, \mathcal{D}) + O\left(\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m)) + \mathsf{VC}(\mathcal{H}))\log^2(m) + \log(1/\delta)}{m}}\right).
\end{aligned}$$

Since the choice of $\mathcal{H} \in \mathbb{H}$ was arbitrary, we can start with setting it to be $\mathcal{H}^* \in \mathbb{H}$ that achieves the minimum value of

$$\min_{\mathcal{H}\in\mathbb{H}} \left\{ \mathsf{err}(\mathcal{H}, \mathcal{D}) + c'\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m)) + \mathsf{VC}(\mathcal{H}))c''\log(m) + 1)\log(m) + \log(2/\delta)}{m}} + \sqrt{\frac{\log(2/\delta)}{m}} \right\},$$

and argue that the predictor chosen by $\mathcal{A}_{\mathbb{H}}$ competes with the error of $\mathcal{H}^*$. This concludes the proof. ■

**Proof of Theorem 5.** The proof is similar to the proof of Theorem 3. Let $\delta' = \delta/4$. We only need to note two properties. First, argue we can find sets $S_1, \ldots, S_T$ with $|S_i| \leq 3\mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta')$ such that $\mathbb{A}_{\mathcal{H}}(S)(.) := \text{Majority}(\mathcal{A}_{\mathcal{H}}(S_1)(.), \ldots, \mathcal{A}_{\mathcal{H}}(S_T)(.))$ has $\mathsf{err}(\mathbb{A}_{\mathbb{H}}(S), S) \leq \mathsf{err}(\hat{h}, S)$ as defined in the proof of Theorem 3. Observe that we proved the existence of $S_i$ by showing that $\mathbb{E}_{R\sim\mathcal{P}^n,(x,y)\sim\mathcal{P}'}\left[1\{\mathcal{A}_{\mathcal{H}}(R)(x) \neq y\}\right] \leq \frac{\mathsf{VC}(\mathcal{H},R)}{n+1} \leq \frac{\mathsf{VC}(\mathcal{H})}{n+1}$. We can now simply argue that by definition we have with probability at least $1-\delta'$ over $S \sim \mathcal{D}^m$ that $\mathsf{VC}(\mathcal{H}, S) \leq \mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta')$ and therefore $\mathbb{E}_{R\sim\mathcal{P}^n,(x,y)\sim\mathcal{P}'}\left[1\{\mathcal{A}_{\mathcal{H}}(R)(x) \neq y\}\right] \leq \frac{\mathsf{VC}(\mathcal{H},\mathcal{D},m,\delta')}{n+1}$. This proves that we can find $S_i$'s with $|S_i| \leq 3\mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta')$ that satisfy what we desired, i.e., $\mathcal{A}_{\mathcal{H}}(S_i)$ is a weak learner for the distributions over sample that we wanted. Second, we need to note that for a set of indices $I \in [m]^{kT}$ the uniform convergence property for $\mathcal{F}_I$ as defined in Lemma 39 is satisfied with an error that depends on $\tau_{\mathbb{H}}(m, \mathcal{D}, \delta')$. The way we argued that $|\mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), \mathcal{D}) - \mathsf{err}(\mathbb{A}_{\mathcal{H},I}(S), V_I)|$ is small was by noting that the function $\mathbb{A}_{\mathcal{H},I}(S)$ is only defined using $S \setminus V_I$ and instances in $V_I$ are still i.i.d. from $\mathcal{D}$. We then invoked the uniform convergence for $\mathcal{F}_I$ on $S$. Now notice that because of the monotonicity of $\tau_{\mathbb{H}}(.)$ we have that $\mathbb{P}_{S\sim\mathcal{D}^m}\left[\tau_{\mathbb{H}}(S) \leq \tau_{\mathbb{H}}(\mathcal{D}, m, \delta')\right] \leq \mathbb{P}_{V\sim\mathcal{D}^{m-kT}}\left[\tau_{\mathbb{H}}(V) \leq \tau_{\mathbb{H}}(\mathcal{D}, m, \delta')\right]$. Therefore, we conclude that with probability at least $1-\delta'$ over $S \sim \mathcal{D}^m$ we have $\mathsf{VC}(\mathcal{F}_I, V_I) \leq \log(\tau_{\mathbb{H}}(\mathcal{D}, m, \delta'))$. Hence, we get the desired concentration of errors of $\mathbb{A}_{\mathcal{H}}$ on $V_I$ to expected error for all $\mathcal{H} \in \mathbb{H}$ with the same rate but by replacing $\tau_{\mathbb{H}}(m)$ with $\tau_{\mathbb{H}}(\mathcal{D}, m, \delta')$. Taking a union bound over the events that $\mathsf{err}(\hat{h}, S)$ is a good estimate of $\mathsf{err}(\mathcal{H}, \mathcal{D})$, that $\mathsf{VC}(\mathcal{H}, S) \leq \mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta')$, that $\tau_{\mathbb{H}}(S) \leq \tau_{\mathbb{H}}(\mathcal{D}, m, \delta')$, the event in Equation 3,

we get with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ that

$$\mathsf{err}(\mathcal{A}_{\mathbb{H}}(S), \mathcal{D}) \leq \mathsf{err}(\mathcal{H}, \mathcal{D}) + O\left(\sqrt{\frac{(\log(\tau_{\mathbb{H}}(m, \mathcal{D}, \delta/4)) + \mathsf{VC}(\mathcal{H}, \mathcal{D}, m, \delta/4))\log^2(m) + \log(1/\delta)}{m}}\right).$$

■

### A.1. Proof of Proposition 6

In this section, we prove that for every standard SRM learner there exists a distribution on which SRM has high error while the learner in Theorem 3 (or Theorem 5) achieves small generalization error on this distribution due to its data-dependent approach.

**Proof of Proposition 6.** We first define the collection $\mathbb{H}$. For any $n \in \mathbb{N}$, we define $\widetilde{\mathcal{H}}_n = \{h \in \{0,1\}^{\mathbb{N}} : \forall i \notin [n(n-1)+1, n^2],\ h(i) = 1\}$. We also define $h^*_n \in \{0,1\}^{\mathbb{N}}$ as $h^*_n(i) = 0$ for all $i \in [n^2+1, n^2+\lceil n/2 \rceil]$ and $h^*_n(i) = 1$ otherwise. In other words, the class $\widetilde{\mathcal{H}}_n$ induces every binary behaviour on $[n(n-1)+1, n^2]$ while every hypothesis in it is constant 1 on every other natural number. The hypothesis $h^*_n$ is 1 on every natural number except the interval $[n^2, n^2+\lceil n/2 \rceil]$. Define $\mathcal{H}_n = \widetilde{\mathcal{H}}_n \bigcup h^*_n$ and let $\mathbb{H} = \{\mathcal{H}_n : n \in \mathbb{N}\}$. In words, we have consecutive pairs of intervals of increasing size and reserve the $n$th pair for $\mathcal{H}_n$. Clearly, $\mathsf{VC}(\mathcal{H}_n) = n$ and $\mathbb{H}$ contains hypothesis classes of arbitrary large VC dimension. Let $\mathcal{A}_{\mathsf{SRM}}$ be any SRM learner that defines a weight function $w(n)$ to assign to each $\mathcal{H}_n$ and returns a function $h \in \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$ from the collection that minimizes

$$\mathsf{err}(h, S) + \sqrt{\frac{\mathsf{VC}(\mathcal{H}_{n(h)}) + \log(1/w(n(h))) + \log(1/\delta)}{m}},$$

where $n(h)$ is the smallest index such that $h \in \mathcal{H}_{n(h)}$.

Take the hypothesis $h_1$ with $h_1(i) = 1$ for all $i \in \mathbb{N}$. Clearly, $h_1 \in \mathcal{H}_1$. Let $w_0 \in \mathbb{N}$ be such that

$$\sqrt{w_0/m} > 1/2 + C_1\sqrt{\log(1/\delta)/m} + C_2\sqrt{\frac{\log(1/w(1)) + \log(1/\delta)}{m}},$$

where $C_1$ and $C_2$ are the constants for the Hoeffding's inequality and uniform convergence guarantee, respectively. Take the hypothesis class $\mathcal{H}_{m_0}$ such that $\log(1/w(\mathcal{H}_{m_0})) \geq w_0$. It is easy to observe that such class must exists because $\sum_{n\in\mathbb{N}} w(\mathcal{H}_n) \leq 1$ and there cannot be any positive lower bound on the weight function. Take the function $h^*_{m_0}$ and observe that $h^*_{m_0}$ is only a member of $\mathcal{H}_{m_0}$ and therefore $\mathsf{VC}(H_{n(h^*_{m_0})}) = \mathsf{VC}(\mathcal{H}_{m_0}) = m_0$. Now, let $\mathcal{D}$ be the uniform distribution on $[m_0^2+1, m_0^2+m_0+1]$ that is realized by $h^*_{m_0}$, i.e., on the second interval for the $m_0$th pair. Note that $h_1$ is constant 1 on the support of $\mathcal{D}$ and we have with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ that $\mathsf{err}(h_1, S) \leq 1/2 + C_1\sqrt{\log(1/\delta)/m}$. This means that the generalization term that $\mathcal{A}_{\mathsf{SRM}}$ considers for $h_1$ is at most $1/2 + C_1\sqrt{\log(1/\delta)/m} + C_2\sqrt{\frac{1+\log(1/w(1))+\log(1/\delta)}{m}}$. Since we chose $m_0$ such that this term is less than $\sqrt{w_0/m}$, we conclude that $\mathcal{A}_{\mathsf{SRM}}$ will never return $h^*_{m_0}$. Note that we picked $\mathcal{H}_{m_0}$ by making sure its weight is so small that the generalization bound $\mathcal{A}_{\mathsf{SRM}}$ considers is very large. Even if $\mathcal{A}_{\mathsf{SRM}}$ is careful in using $\mathsf{VC}(\mathcal{H}_{n(h)}, S)$ instead of $\mathsf{VC}(\mathcal{H}_{n(h)})$, it will still not choose the function $h^*_{m_0}$ because although $\mathsf{VC}(\mathcal{H}_{m_0}, S)$ is small, the term $\sqrt{\log(1/w(\mathcal{H}_{m_0})/m}$ is large. Now observe that for any $h \in \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$ with $h \neq h^*_{m_0}$ we have $\mathsf{err}(h, \mathcal{D}) = 1/2$. This

proves that with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ we have $\mathsf{err}(\mathcal{A}_{\mathsf{SRM}}(S), \mathcal{D}) = 1/2$. On the other hand, every $\mathcal{H}_n$ with $n \neq m_0$ induces the same behaviour on any $S \sim \mathcal{D}^m$, i.e., the constant 1 behaviour. Therefore, $\tau_{\mathbb{H}}(m, \mathcal{D}, \delta) = 2$. Moreover, $\mathsf{VC}(\mathcal{H}_{m_0}, \mathcal{D}) = 0$. This implies that for the learner $\mathcal{A}$ in Theorem 3 (or Theorem 5) we have with probability at least $1-\delta$ that $\mathsf{err}(\mathcal{A}(S), \mathcal{D}) \leq O\left(\sqrt{\frac{\log^2(m)+\log(1/\delta)}{m}}\right)$, which concludes the proof. ■