REFINING THE SAMPLE COMPLEXITY OF COMPARATIVE LEARNING

SAJAD RAHMANIAN ASHKEZARI

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

APRIL 2025

Abstract

The PAC (Probably Approximately Correct) framework is a well-established theoretical framework for analyzing the statistical (and sometimes computational) complexity of machine learning tasks. Comparative learning is a recently introduced variation of the PAC framework that interpolates between the two standard extreme settings of realizable and agnostic PAC learning. In comparative learning the labeling is assumed to be from one hypothesis class (the source) while the learner's performance is to be measured against another hypothesis class (the benchmark). This setup allows for incorporating more specific prior knowledge into PAC-type learning bounds, which are known to be otherwise overly pessimistic. In this work we study the sample complexity of a variation of this setting we call proper comparative learning where we require the learning algorithm to output a hypothesis from the benchmark class. This setting represents model distillation tasks, where a predictor with specific requirements (e.g., interpretability) is trained on the labels from another model.

Acknowledgements

I would like to dedicate this work to my family. Thank you for your constant support and for always believing in me.

I am deeply grateful to Ruth Urner for first accepting me as her student and for her invaluable guidance and encouragement. Looking back to the beginning of my master's journey, I can see how much I've grown as a researcher, and I owe a significant part of that growth to her mentorship.

I would also like to thank Aditya Potukuchi and Kelly Ramsay for taking the time to serve on my committee.

Finally, I want to thank everyone who, directly or indirectly, has supported me throughout my life.

Table of Contents

1 Introduction

The standard learning theoretic concept of Probably Approximately Correct (PAC) learnability is a well-studied framework to establish general performance guarantees for statistical learning [48, 12, 25]. In PAC learning, we assume the data is generated i.i.d. from an unknown distribution and the goal of a learning algorithm is to achieve a small loss (e.g., classification loss) on this distribution given a finite sample generated from it. However, without having any further assumptions, solving this learning problem is impossible due to no-free-lunch arguments [44]. There are two standard variations of PAC framework that avoid this issue by introducing a hypothesis class: realizable PAC learning and agnostic PAC learning. In the realizable setting, the distribution is restricted such that there exists a hypothesis in the class that achieves zero loss. In the agnostic setting, we do not make any assumptions on the data distributions, however, instead of expecting the learning algorithm to output a hypothesis that achieves arbitrarily small loss, we expect it to achieve a loss that is arbitrarily close to the smallest loss that can be achieved

by a hypothesis in the class (we formalize these definitions in Section 3.1).

The main appeal of PAC-type guarantees is that they hold uniformly over all possible data-generating distributions. A PAC learning guarantee provides a finite sample size (that depends only on the desired error and confidence parameters, and the model or hypothesis class in use) which suffices for the desired error bound, independently of any properties of the data-generating distribution. The parameters that determine these distribution-free finite sample bounds are parameters of the model or hypothesis class, and as such are controlled by the user, rather than hinging on unknown properties of the data-generation. The original framework and bounds for binary classification have thus been extended to myriad other tasks and settings such as regression, multiclass classification, active learning, adversarially robust learning to name a few [1, 9, 18, 34]. Recent years have seen new interest in these types of guarantees with novel frameworks developed and long-standing open questions resolved [22, 15, 2, 6, 28].

However, the main merit of the PAC framework, its generality, also constitutes its main drawback. Since they are required to be valid for all possible data generating distributions, PAC-type bounds are often overly pessimistic. For deep learning methods, they are mostly considered to provide vacuous bounds [50, 40, 20]: methods often perform much better than guaranteed through the PAC framework when applied to real data (rather than the worst case data-generation that PAC proofs

need to hold against). The standard PAC framework does not allow for naturally modeling prior knowledge about the data generation, nor does it, in its standard form, allow for incorporating additional requirements on the learned predictor, such as being interpretable, satisfying fairness requirements, or requiring little memory to be stored.

The above two issues, namely modeling better-than-worst-case data generation and adding requirements on the output predictor, are then typically treated separately in the PAC literature. There are various works that derive PAC-type learning rates under additional distributional assumptions. Often these are conditions on the noise [46, 45, 32], or assumptions about label-separatedness (cluster assumptions or margin conditions) for classification tasks [16, 42, 47]. These are typically assumptions about how the labeling component of the distribution relates to the marginal over the feature vectors. Requirements on the learned predictor are treated separately, and on a case by case basis. That is, there is analysis on methods to incorporate adversarial robustness requirements [34, 35], other investigations on methods to achieve an interpretable model [14] etc. One general, practical method to incorporate such additional requirements is the teacher-student framework [47, 21, 7, 5], or the closely related model distillation approach [26, 31, 41]. Here, in a first round of learning, an arbitrary (in terms of requirements) but highly accurate model is trained. The labels (or soft labels) from that model are then used

to annotate an unlabeled data set, with which a model that satisfies the specific requirements (for example being fast at prediction time, being interpretable, being robust to adversarial perturbations, or requiring little memory) is trained. In such a scenario, we have specific *prior knowledge about the type of labeling function* in the second round of this process. This, however, is not naturally captured in the existing notions of learning from beyond-worst case distributions.

Comparative learning is a recently introduced variation of the PAC learning framework that naturally models this scenario [27]. In comparative learning the labeling is assumed to be from one hypothesis class (the *source class*) while the learner's performance is measured against another hypothesis class (the *benchmark class*). Thus, this framework facilitates incorporating prior knowledge about the labeling component of the data generating process distribution (for example, being from a specific class of models in the teacher-student framework). The initial study that introduced the framework, provided first upper and lower bounds on the sample complexity of *general* comparative learning for classification with binary hypothesis classes in terms of the *mutual VC dimension* [27]. Their bounds left a gap between a linear dependence in the error parameter $\frac{1}{\epsilon}$ for the lower bound, and a quadratic dependence in the upper bound.

In this work, we shift the focus from general comparative learning (where there is no restrictions on the learner) to comparative learning that is *proper* with respect

to the benchmark class (where the learner is required to output a predictor from the benchmark class). This setting models the above discussed techniques of teacher-student learning or model distillation. We prove that in the benchmark-proper case, the sample complexity of comparative learning is not governed by the mutual VC dimension, but rather by a novel parameter we introduce, the *one-sided (mutual) graph dimension*. Our analysis incorporates the wider frameworks of multi-class predictors [18, 15] and the source and benchmark class potentially being *partial hypothesis classes* [2]. For both general and benchmark-proper comparative learning, we also identify general conditions that yield linear versus quadratic rates in the error parameter. However, obtaining a full characterization of the rates still remains open.

Overview and Summary of Contributions

Our results can be summarized as follows:

- We propose the focus on *benchmark-proper comparative learning* to model learning scenarios of the model distillation type. Our work also broadens the scope of analysis for comparative learning to multi-class settings with (potentially) partial hypothesis classes.

- In Section 4.1 we introduce a novel combinatorial parameter that measures the

relatedness between source and benchmark class, the *one-sided mutual graph dimension*. We prove that the sample complexity of benchmark-ERM, as well as general benchmark-proper learning, is upper and lower bounded in terms of this dimension. We further identify conditions for linear and quadratic dependence on $\frac{1}{\epsilon}$. Finally, our analysis shows that there are cases of total binary hypothesis classes, where benchmark-proper comparative learning is impossible, even though the pair is comparatively learnable. Such phenomena have previously been shown for multi-class learning with infinitely many labels [17] or for partial classes [2].

- In Section 4.2, we study the general case of the comparative learning framework. We introduce a broad set of assumptions under which the exact dependence of the sample complexity on $\frac{1}{\epsilon}$ is shown to be either linear or quadratic.

- The bounds in Section 4.2 are derived by relating the comparative learning framework with the previously established setting of agnostic PAC learning under deterministic labels [8]. In order to make these results applicable to our setting, we generalize some of the bounds for agnostic PAC learning under deterministic labels from total to partial hypothesis classes. These results might be of independent interest.

2 Related Work

In this work we mainly focus on PAC framework which has been extensively studied since the late 80s [44]. Here we give an informal definition of PAC learnability and refer the reader to Section 3.1 for a more formal definition. Consider any unknown distribution and any algorithm that maps a finite sample to a hypothesis, where the samples are generated i.i.d. from the unknown distribution. In PAC learning, we are interested in deriving bounds on the number of samples necessary and sufficient to ensure, with high probability over the samples, the hypothesis returned by the algorithm has small loss. Without any restrictions, PAC learnability is impossible [44, Theorem 5.1]. Thus, a hypothesis class is added to the definition of the framework. In the first version of PAC framework, which is known as realizable PAC learning, the distributions are restricted such that a hypothesis from the class can achieve zero loss on them. This condition is usually stated as the distribution is realizable by the hypothesis class. On the other hand, in agnostic PAC learning we do not make any assumptions on the data generating distribution and instead,

require the learner to achieve loss that is arbitrarily close to the smallest loss that can be achieved by a hypothesis in the class. In the above settings there are no restrictions on the output of the learning algorithm. If the output of the learning algorithm is restricted to belong to the hypothesis class, we call it a proper learner. Then a hypothesis class is proper PAC learnable if it can be learned by a proper learner [44].

Comparative learning is a recently introduced variation of PAC learning which is an interpolation between the realizable and agnostic settings [27]. In this settings, the distributions are realizable by one class referred to as the source class and the learning algorithm is compared to the smallest loss achievable by another class referred to as the benchmark class. We give a formal definition of comparative learning in Section 3.2. The initial sample complexity bounds in [27] had a gap in terms of the dependence on error parameter. Here we show both rates can be achieved and give general conditions that when satisfied by the classes one of the rates can be tight.

In the comparative learning definition, there is no constraint on the hypotheses that the learning algorithm can output. Similar to proper PAC learning we define benchmark-proper comparative learning where we restrict the learner to output a hypothesis from the benchmark class. Proper comparative learning can be viewed as a model for knowledge distillation, a practical method for training machine learning

models with specific requirements such as being small given access to more powerful models [26]. In this scenario, the source class plays the role of the more expressive model and the benchmark class can be thought of as the model satisfying our requirements. We show proper comparative learning can be strictly more difficult than comparative learning, i.e., there are classes that are comparative learnable but not benchmark-proper comparative learning. We show this even for binary hypothesis classes, which is surprising as in PAC learning, a binary hypothesis class is learnable if and only if it is proper learnable. We also show existence of different rates of sample complexity (in terms of the error parameter).

We now give a review of results and methods in the study of PAC learning. The study of PAC learning started for total binary classes where the hypotheses were defined everywhere on the domain and labeled each point by 1 or 0 [44]. For binary classes it has been shown that PAC learnability is equivalent to a property called uniform convergence. Uniform convergence is satisfied if, with high probability, the true risk of all hypotheses in the class is close to their empirical risk after observing a finite number of samples, whose size is independent of the underlying distribution and depends only on the confidence and error parameters [44]. The initial attempts to generalize the results to other classes were made by using uniform convergence. However, recent results have shown that uniform convergence is not equivalent to PAC learnability in general and since then, there has been an interest towards

learning without using uniform convergence. An integral part of these methods is the *one-inclusion graph* (OIG) algorithm introduced by [24] for binary classes and later generalized for multiclass classes by [43].

Another tool which is widely used in recent works is *sample compression schemes* introduced by [29]. Informally, sample compression scheme of constant size means the ability to reconstruct the label of samples of arbitrarily large size from a constant number of labeled sample points. It has been shown that compression implies learnability [29]. Years later, it was shown that VC-classes are compressible, which implies that for binary classes learnability is equivalent to compressibility [36]. A general learner to compression method was introduced in [19], which was improved in and generalized to real valued functions in [23]. While compression implies learnability in general, the converse does not necessarily hold. There exists a learnable multiclass hypothesis class with infinite classes that does not have a sample compression scheme of constant size [39].

Multiclass Learnability: The study of learning multiclass hypothesis classes was initiated in [38, 37]. Two natural generalizations of the VC dimension were introduced: the Natarajan dimension, whose finiteness is necessary for PAC learnability, and the Graph dimension, whose finiteness is sufficient for PAC learnability. When the label space is finite, it has been shown that these two dimensions are equivalent up to constant and logarithmic factors in the size of the label space and thus they

both characterize learnability [9]. However, for infinite label spaces finiteness of the Graph dimension is not necessary for PAC learning [37]. Whether finiteness of the Natarjan dimension is sufficient for PAC learnability remained an open problem for more than two decades until it was recently solved with a negative answer [15]. A shortcoming of the initial attempts was their use of uniform convergence. In [18] the authors showed that the equivalence between uniform convergence and PAC learnability, which holds for binary classes, no longer holds for multiclass hypothesis classes with an infinite label space. They showed there are classes that are learnable but cannot be learned by ERM algorithms. This was strengthened in [17] where it was shown there are learnable classes that cannot be learned by any proper learner, i.e., any learner that always output a hypothesis from the class. Daniely and Shalev-Shwartz [17] also showed that the OIG algorithm is an optimal multiclass learner up to logarithmic factors. They also introduced a new dimension which was later called the *Daniely-Shalev-Shwartz* (DS) dimension. They showed that finiteness of this dimension is necessary for PAC learnability but left the sufficiency as an open question. A positive answer to this problem was recently provided, showing that the DS dimension characterizes multiclass learnability [15].

Partial Hypothesis Classes: A partial hypothesis is a hypothesis that abstains from giving an output on certain points in the domain. A class containing such hypotheses is called a partial hypothesis class. Partial classes are suitable for sce-

narios where we are guaranteed that the data generating process satisfies certain conditions. An example of such a scenario is when we know that the positive and negative classes have a positive margin from the boundary that separates them. In this case, the hypotheses can abstain from prediction on data points that lie within this margin. Learnability of partial hypothesis classes was first studied in [30]. They defined the VC dimension of a partial binary classes similar to the total binary classes, with the distinction that a set is shattered if all binary patterns are produced on that set by the hypothesis class (this means there could be other patterns where some points are unlabeled by the hypotheses in the class). A formal treatment of the partial binary hypothesis classes was done in [2]. They showed that the VC dimension characterizes the learnability in both the realizable and in agnostic setting. Interestingly, they showed that the uniform convergence again fails in this setting. They showed that there is a hypothesis class that is learnable, but cannot be learned by any proper learner and in particular any ERM. They achieved their result using the OIG algorithm and sample compressions.

3 Preliminaries

We employ the standard statistical learning theoretic framework and notation. We let $\mathcal{X}$ denote the *domain*, and $\mathcal{Y}$ the *label set* of the classification task. In *binary classification* we have $\mathcal{Y} = \{0, 1\}$, for multiclass classification with finitely many labels, we can assume $\mathcal{Y} = [k] = \{1, 2, 3, \ldots, k\} \subseteq \mathbb{N}$, otherwise $\mathcal{Y}$ can in general be any set. The environment is modeled as a *data-generating distribution* P over $\mathcal{X} \times \mathcal{Y}$. We use $P_{\mathcal{X}}$ to denote the marginal distribution of P over $\mathcal{X}$ and use $\text{supp}(P)$ to denote the support of a distribution P.

A *classifier* is a function $h : \mathcal{X} \to \mathcal{Y}$. A *partial classifier* is a function $h : \mathcal{X} \to \mathcal{Y} \cup \{\star\}$, where $\star \notin \mathcal{Y}$ [2]. Assigning label $\star$ is sometimes interpreted as abstaining from prediction and is always an inaccurate label assignment in terms of the classification task. For ease of notation, we set $\tilde{\mathcal{Y}} = \mathcal{Y} \cup \{\star\}$. Equivalently, a partial classifier can be viewed as a function $h : \mathcal{Z} \to \mathcal{Y}$, where $\mathcal{Z} \subseteq \mathcal{X}$ is a subset of the domain. This subset, namely $\mathcal{Z} = h^{-1}(\mathcal{Y})$, is also referred to as *support of* h. We also refer to (partial) classifiers as *hypotheses* and we call a set of such

a *hypothesis class* $\mathcal{H}$. Following the literature, we call $\mathcal{H}$ a *partial (hypothesis) class* if its members are partial classifiers. Otherwise, we also refer to $\mathcal{H}$ as a *total (hypothesis) class*.

The *binary loss* measures the correctness of a classifier h on labeled point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ as

$$\ell(h, x, y) = \mathbb{1}\left[h(x) \neq y\right],$$

where $\mathbb{1}\left[\cdot\right]$ denotes the indicator function. Note that the loss is only defined on points with a proper label (and only such can be generated by the environment P), and thus $\ell(h, x, y) = 1$ for all points x with $h(x) = \star$. The goal of learning is to identify a classifier h with low *expected* or *true loss* (also called true risk)

$$\mathcal{L}_P(h) = \mathbb{E}_{(x,y)\sim P}[\ell(h, x, y)].$$

A *learner* takes a finite sequence of labeled data points $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, for some $n \in \mathbb{N}$ and outputs a (partial) classifier. The *empirical loss* (also called empirical risk) of h with respect to data S is defined as $\mathcal{L}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, x_i, y_i)$. A learner is called an Empirical Risk Minimizer (ERM) with respect to class $\mathcal{H}$ if it always outputs a hypothesis from the class $\mathcal{H}$ with minimal empirical loss within that class.

3.1 The standard PAC framework

In the standard PAC (Probably Approximately Correct) learning framework [48, 12, 44], the success of a learner is evaluated against the best possible performance within a fixed hypothesis class $\mathcal{H}$, the *approximation error* of the hypothesis class $\mathcal{H}$:

$$\mathrm{opt}_P(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{L}_P(h)$$

We call a distribution P *realizable by hypothesis class* $\mathcal{H}$ if $\mathrm{opt}_P(\mathcal{H}) = 0$.

Definition 1 (PAC Learner [48, 44])**.** *We say that a learner* $\mathcal{A}$ *is an* (agnostic) PAC learner *for hypothesis class* $\mathcal{H}$ *if for every* $\epsilon, \delta > 0$*, there is a sample-size* $n(\epsilon, \delta)$ *such that, for all* $n \geq n(\epsilon, \delta)$*, and all distributions* P *we have*

$$\Pr_{S \sim P^n} [\mathcal{L}_P(\mathcal{A}(S)) \leq \mathrm{opt}_P(\mathcal{H}) + \epsilon] \geq 1 - \delta.$$

We call $\mathcal{A}$ *a PAC learner for* $\mathcal{H}$ in the realizable case *if the above requirement holds for all distributions* P *that are realizable by the hypothesis class. If such a learner exists, the class* $\mathcal{H}$ *is called* PAC learnable*. The learner is called a* proper PAC learner *if it always outputs a function from* $\mathcal{H}$ *(and the class* $\mathcal{H}$ *is then called* properly PAC learnable).

It is well known that for total binary hypothesis classes, PAC learnability is equivalent to finiteness of the VC dimension of the class [49, 44].

Definition 2 (VC dimension). *A set $U = \{u_1, \ldots, u_d\} \subseteq \mathcal{X}$ is shattered by a total binary hypothesis class if $\mathcal{H}|_U = \{0,1\}^{|U|}$, where $\mathcal{H}|_U = \{\big(h(u_1), \ldots, h(u_d)\big) : h \in \mathcal{H}\}$. The VC dimension of $\mathcal{H}$, denoted by $\mathrm{vc}(\mathcal{H})$, is defined as the maximum cardinality of a set that is shattered by $\mathcal{H}$. If $\mathcal{H}$ shatters sets of arbitrary size, $\mathrm{vc}(\mathcal{H}) = \infty$.*

Before stating the main results of PAC learning for binary hypothesis classes, we first state the definition of uniform convergence which plays an important role. It is easy to see any hypothesis class that satisfies the uniform convergence property is PAC learnable by any ERM (see, for example, Section 4.1 in [44])

Definition 3 (Uniform Convergence, [44]). *We say a hypothesis class $\mathcal{H}$ satisfies the uniform convergence property, if there exists a function $n^{UC} : [0,1]^2 \to \mathbb{N}$ such that for all $\epsilon, \delta > 0$ and for all distributions P, samples $S \sim P^n$ with $n \geq n^{UC}(\epsilon, \delta)$ satisfy $|\mathcal{L}_P(h) - \mathcal{L}_S(h)| < \epsilon$ for all $h \in \mathcal{H}$ simultaneously, with probability at least $1 - \delta$ over samples.*

Theorem 1 (The Fundamental Theorem of Learning, Theorem 6.7, 6.8 in [44]). *For any binary total hypothesis class $\mathcal{H}$ the following are equivalent.*

1. $\mathrm{vc}(\mathcal{H}) < \infty$.
2. *$\mathcal{H}$ satisfies the uniform convergence property.*
3. *$\mathcal{H}$ is agnostically PAC learnable with sample complexity $\Theta(\frac{\mathrm{vc}(\mathcal{H}) + \log(\frac{1}{\delta})}{\epsilon^2})$.*

4. *$\mathcal{H}$ is realizably PAC learnable with sample complexity* $\Theta(\frac{\text{vc}(\mathcal{H})+\log(\frac{1}{\delta})}{\epsilon})$

5. *Any ERM is a PAC learner (both in the agnostic and in the realizable setting) with sample the same sample complexity up to logarithmic factors.*

The definition of VC dimension is extended to partial hypothesis classes [2]. The definition is the same as above with the modification that a set U is shattered if $\mathcal{H}|_U \supseteq \{0,1\}^{|U|}$, meaning that $\mathcal{H}$ can produce all binary patterns on U. They show that VC dimension still characterizes PAC learnability both in the agnostic and in the realizable setting. They show the sample complexity of agnostic learning is $\tilde{\Theta}(\frac{\text{vc}(\mathcal{H})+\log(\frac{1}{\delta})}{\epsilon^2})$ and the sample complexity of realizable learning is $\tilde{\Theta}(\frac{\text{vc}(\mathcal{H})+\log(\frac{1}{\delta})}{\epsilon})$ (Please refer to Appendix C in [2] for more details). It is worth noting that they achieve their results using an improper algorithm and show there are learnable partial classes that cannot be learned by a proper learner. This is in contrast to the results for total classes where any learnable class can be learned by any ERM.

Agnostic PAC learning with deterministic labeling was studied in [8]. Here we restate the following lemma from [8], which we use in the proof of Theorem 28.

Lemma 2 (Essentially Lemma 3 in [8])**.** *Let* $0 < \epsilon < \frac{1}{4}$ *and* $0 < \delta < \frac{1}{32}$*. Let* $\mathcal{H}$ *be a hypothesis class defined on* $\mathcal{X}$ *with* $|\mathcal{X}| \geq \frac{1}{\epsilon^3}$ *such that there exist two hypotheses* h_0 *and* h_1 *in* $\mathcal{H}$ *that disagree on* $\mathcal{X}$*. Then* $(\epsilon/2, \delta)$*-PAC learning this class in the agnostic setting with deterministic labels requires* $\Omega(\frac{1}{\epsilon^2})$ *samples.*

Note that Lemma 2 was originally proved for classes that contain the two constant hypotheses that label all points in the domain as 0 or 1, respectively. However, the only property that is used in this proof is that these two hypotheses disagree on every point in the domain. Thus, the lemma can be generalized to any hypothesis class that contains two hypotheses that disagree on every point in the domain [8]. This can be proven by simply adapting the current proof to a set of distributions that label $(1/2 + \epsilon)$-fraction of domain according to h_0 and the rest according to h_1, or vice versa, where h_0 and h_1 are the two hypotheses that disagree everywhere in the domain.

The VC dimension and the above results on sample complexity of PAC learning have been generalized to also work for multiclass classification. Two dimensions referred to as the Natarjan dimension (Definition 4) and the Graph dimension (Definition 5) were introduced in [37], and have been shown to give lower bound and upper bound on sample complexity of multiclass classification, respectively. It was later shown that when label space is finite, the Natarjan dimension also gives an upper bound [9]. It has also been shown that the ERM sample complexity of multiclass learning is characterized by the graph dimension [18].

Definition 4 (Natarjan dimension [37])**.** *A set $U \subseteq \mathcal{X}$ is Natarjan-shattered by a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if for any $A \subseteq U$, there exist two hypotheses f and g in $\mathcal{H}$ such $f(x) = g(x)$ for all $x \in A$ and $f(x) \neq g(x)$ for all $x \in U \backslash A$. The Natarjan*

dimension of a class $\mathcal{H}$ is the size of the largest set that is Natarjan-shattered by $\mathcal{H}$ and is denoted by $d_N(\mathcal{H})$. We say $d_N(\mathcal{H}) = \infty$ if $\mathcal{H}$ Natarjan-shatters sets with arbitrarily large cardinality.

Definition 5 (Graph dimension [37])**.** *A set $U \subseteq \mathcal{X}$ is Graph-shattered by a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a hypothesis $g \in \mathcal{Y}^{\mathcal{X}}$ such that for any $A \subseteq U$, there exist $h \in \mathcal{H}$ such that $h(x) = g(x)$ for all $x \in A$ and $h(x) \neq g(x)$ for all $x \in U \backslash A$. The Graph dimension of a class $\mathcal{H}$ is defined as the size of the largest set that is Graph-shattered by $\mathcal{H}$ and is denoted by $d_G(\mathcal{H})$. If $\mathcal{H}$ Graph-shatters sets with arbitrary size, then we defined $d_G(\mathcal{H}) = \infty$.*

3.2 The Comparative Learning Framework

In the standard PAC framework, finite sample based learning success is required to hold uniformly (in terms of the sufficient sample sizes) over all possible data-generating distributions (for reference, see Definition 1 in Section 3.1). This can lead to overly pessimistic conclusions since natural data-generating environments rarely possess the qualities of a worst-case scenario for a given method, and thus success is in practice typically achieved with much smaller sample sizes than what is predicted by PAC theory. *Comparative Learning*, a PAC-type learning framework recently introduced [27], allows for naturally modeling prior knowledge about the labeling component of the data-generating process.

The comparative learning framework presumes two hypothesis classes, the *source class* $\mathcal{S}$ and the *benchmark class* $\mathcal{B}$. The source class incorporates the prior knowledge about the data generation in that learning success is only required for distributions P realizable by $\mathcal{S}$. The source class can thus be viewed as a class containing the true labeling rule of the classification task. The learner's success is then measured against the approximation error of the benchmark class $\mathcal{B}$. The following definition is due to Hu and Peale [27], while adapted to our terminology.

Definition 6 (Comparative PAC Learner)**.** *We say that a learner $\mathcal{A}$ is a* comparative PAC learner *for source and benchmark classes $(\mathcal{S}, \mathcal{B})$ if for every $\epsilon, \delta > 0$, there is a sample-size $\mathrm{n}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta)$ such that, for all $n \geq \mathrm{n}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta)$, and all distributions P with $\mathrm{opt}_P(\mathcal{S}) = 0$ we have*

$$\Pr_{S \sim P^n} \left[\mathcal{L}_P(\mathcal{A}(S)) \leq \mathrm{opt}_P(\mathcal{B}) + \epsilon \right] \geq 1 - \delta.$$

We call $\mathcal{A}$ a benchmark-proper *comparative learner for $(\mathcal{S}, \mathcal{B})$ if it is a learner as above that always outputs a hypothesis from $\mathcal{B}$.*

If such a learner (or benchmark-proper learner) exists, we call the pair $(\mathcal{S}, \mathcal{B})$ *comparatively (benchmark-proper) PAC learnable*, and the function $\mathrm{n}_{\mathcal{S},\mathcal{B}} : (0,1)^2 \to \mathbb{N}$ an upper bound to the sample complexity of comparatively learning $(\mathcal{S}, \mathcal{B})$. The *sample complexity* is the pointwise smallest such upper bound and we will use the notation $\mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}(\cdot, \cdot)$ for the sample complexity of general comparative learning,

$\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\cdot,\cdot)$ for the benchmark-proper case and $\mathrm{n}^{\mathrm{ERM}}_{\mathcal{S},\mathcal{B}}(\cdot,\cdot)$ to denote the sample complexity of learning for learning with any empirical risk minimizing learner (ERM).

Standard PAC learning in the realizable case can be viewed as comparative leaning when source and benchmark class coincide, that is the case $\mathcal{S} = \mathcal{B}$. PAC learning under deterministic labels corresponds to comparative learning when the source class contains all functions from domain to label set, that is $\mathcal{S} = \mathcal{Y}^{\mathcal{X}}$ [8].

The study that introduced the comparative PAC learning framework provided a first upper and first lower bound for the sample complexity of this learning problem, $\mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}$, in terms of the *mutual VC dimension* $\mathrm{vc}(\mathcal{S},\mathcal{B})$ of the two classes involved. The mutual VC dimension $\mathrm{vc}(\mathcal{H},\mathcal{H}')$ of two hypothesis classes $\mathcal{H}$ and $\mathcal{H}'$, formally defined below, is the largest possible size of a set of points that is simultaneously shattered by both classes.

Definition 7 (Mutual VC Dimension [27])**.** *A set $U \subseteq \mathcal{X}$ is said to be mutually shattered by $(\mathcal{H},\mathcal{H}')$ if it is VC-shattered by both $\mathcal{H}$ and $\mathcal{H}'$. The mutual VC dimension of $(\mathcal{H},\mathcal{H}')$, denoted by $\mathrm{vc}(\mathcal{H},\mathcal{H}')$, is the maximum cardinality of a set that is mutually shattered by $(\mathcal{H},\mathcal{H}')$. If $(\mathcal{H},\mathcal{H}')$ mutually shatter sets of arbitrary size, $\mathrm{vc}(\mathcal{H},\mathcal{H}') = \infty$.*

The following citation summarizes the initial bounds by Hu and Peale [27], omitting log-factors.

Theorem 3 ([27], Theorem 3.1)**.** *Let $\mathcal{S}, \mathcal{B} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ be two hypothesis classes with $\mathrm{vc}(\mathcal{S}, \mathcal{B}) \geq 2$. Then the sample complexity of comparatively learning the pair $(\mathcal{S}, \mathcal{B})$ satisfies:*

$$\Omega\left(\frac{\mathrm{vc}(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon}\right) = \mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{vc}(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon^2}\right).$$

While both bounds are in terms of the same combinatorial complexity parameter relating the source and benchmark class, there is a gap in the $\frac{1}{\epsilon}$-dependences, namely linear versus quadratic. We will identify general conditions under which linear rates can be achieved, as well as separate conditions under which quadratic rates are unavoidable, by introducing new dimensions capturing the relatedness between the two classes.

3.3 Chernoff Bounds

Here we state some Chernoff inequalities used in some of our proofs.

Lemma 4 ([33])**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. Bernoulli random variables and let $Z = \sum_{i=1}^{n} Z_i$ denote their sum. Then:*

1. *For any $\eta \in (0, 1)$, $\Pr[Z \leq (1 - \eta)\mathbb{E}[Z]] \leq \exp(-\frac{\eta^2 \mathbb{E}[Z]}{2})$.*

2. *For any $\eta \geq 0$, $\Pr[Z \geq (1 + \eta)\mathbb{E}[Z]] \leq \exp(-\frac{\eta^2 \mathbb{E}[Z]}{2+\eta})$*

We refer to these as Chernoff's lower and upper tail bound, respectively.

4 Our Results

4.1 Proper Comparative Learning

In this section we consider general multiclass classification tasks with a potentially countably infinite label set $\mathcal{Y}$. Furthermore, unless otherwise stated, we consider general, partial hypothesis classes for source and benchmark classes so that they are a subset of $\tilde{\mathcal{Y}}^{\mathcal{X}}$.

The initial comparative learning bounds in Theorem 3 hold for any learner, in particular learners that are not proper for the benchmark class. However, the comparative learning framework is especially suitable for modeling learning settings where we are interested in outputting a classifier with specific properties, which is most suitably modeled with the benchmark-proper setting. Moreover, we note that practical methods are typically set up to minimize an empirical loss over a training data set, and are thus ERM (or approximate ERM) methods for a specific class of predictors. We start by showing that for such methods the sample complexity can be significantly higher than what is revealed in Theorem 3. For this, we define a

novel relatedness parameter between the source and benchmark class, the *one-sided graph dimension.*

Definition 8 (One-sided Mutual Graph Dimension)**.** *A set $U \subseteq \mathcal{X}$ is one-sided graph-shattered by $(\mathcal{S}, \mathcal{B})$ if there exists a function $s \in \mathcal{S}$, which is a total function when restricted to U, such that for all $V \subseteq U$, there exists $b \in \mathcal{B}$ such that*

$$\forall x \in V : b(x) = s(x) \qquad \textit{and} \qquad \forall x \in U \backslash V : b(x) \neq s(x).$$

The one-sided mutual graph dimension of $(\mathcal{S}, \mathcal{B})$ is the maximum size of a set that can be one-sided graph-shattered and is denoted by $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B})$. We define $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) = \infty$ if $(\mathcal{S}, \mathcal{B})$ one-sided graph-shatter sets of arbitrary size.

The following example illustrates the notion of shattering in Definition 8.

Example 1. *Consider a set of three points $U = \{x_1, x_2, x_3\}$. Let $s \in \mathcal{S}$ be a source hypothesis defined on U labeling it as $s|_U = (0, 0, 0)$. Also assume*

$$\begin{aligned} \mathcal{B}|_U \supseteq \{&(0,0,0), (1,0,0), (0,2,0), (0,0,3), \\ &(4,5,0), (*,0,6), (0,*,*), (7,8,9)\}. \end{aligned}$$

Then $(\mathcal{S}, \mathcal{B})$ one-sided graph-shatter U.

For two partial binary classes, it is easy to see that the one-sided mutual graph dimension can be significantly larger than their mutual VC dimension, i.e.,

$\text{vc}(\mathcal{S}, \mathcal{B}) \leq d_G^{\rightarrow}(\mathcal{S}, \mathcal{B})$. However, the next observation shows this gap can be arbitrarily large.

Observation 5. *There exist total, binary classes $\mathcal{S}$ and $\mathcal{B}$ such that* $\text{vc}(\mathcal{S}, \mathcal{B}) = 0$*, but* $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) = \infty$

Proof. Let $\mathcal{X}$ be an infinite set, e.g., $\mathcal{X} = \mathbb{N}$. Consider the classes $\mathcal{S} = \{h_0\}$, where $h_0(x) = 0$ for all $x \in \mathcal{X}$, and $\mathcal{B} = \{0,1\}^{\mathcal{X}}$ (or any binary class with $\text{vc}(\mathcal{B}) = \infty$). We get $\text{vc}(\mathcal{S}, \mathcal{B}) = 0$. To see that $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) = \infty$, consider any set $U \subseteq \mathcal{X}$ of any size that is shattered by $\mathcal{B}$. For any $V \subseteq U$, choose $s = h_0 \in \mathcal{S}$, and $b \in \mathcal{B}$ such that b is 0 on V and is 1 on $U \backslash V$. Such b exists because U is shattered by $\mathcal{B}$. Thus any such set of arbitrary size is shattered by $(\mathcal{S}, \mathcal{B})$ in the sense of Definition 8, and thus the one-sided mutual graph dimension is ∞. □

4.1.1 Benchmark-ERM comparative learning

We start by analyzing the sample complexity of benchmark-ERM learners for comparative learning. We define the *benchmark-ERM sample complexity* to be the smallest sample size function $\text{n}_{\mathcal{S},\mathcal{B}}^{\text{ERM}} : (0,1)^2 \to \mathbb{N}$ such that the conditions of Definition 6 are satisfied for all ERM learners for the benchmark class with this function.

The following Theorem summarizes the upper and lower bounds for benchmark-ERM comparative learning and establishes that this setting is governed by the one-sided mutual graph dimension:

Theorem 6. *For a pair of partial classes $(\mathcal{S}, \mathcal{B})$ with $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) \geq 2$, the sample complexity of benchmark-ERM is determined by the one-sided graph dimension and satisfies:*

$$\Omega\left(\frac{d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon}\right) \;=\; \mathrm{n}^{\mathrm{ERM}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) \;=\; \tilde{\mathcal{O}}\left(\frac{d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon^2}\right).$$

Remark 7. *It is worth noting again that in the definition of $\mathrm{n}^{\mathrm{ERM}}_{\mathcal{S},\mathcal{B}}$, we consider the worst-case ERM. Thus, $\mathrm{n}^{\mathrm{ERM}}_{\mathcal{S},\mathcal{B}} = \Omega(N)$ means that there is an ERM learner that requires at least $\Omega(N)$ samples and $\mathrm{n}^{\mathrm{ERM}}_{\mathcal{S},\mathcal{B}} = \mathcal{O}(N)$ means a sample size of $\mathcal{O}(N)$ is enough for any ERM learner. This is unlike the definition of $\mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}$ and $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}$ where the lower bound holds for any (proper) learner and the upper bound holds for at least one (proper) learner.*

As noted above, the one-sided mutual graph dimension always upper bounds the mutual VC dimension of two partial binary hypothesis classes and Observation 5 above shows that the gap between these two parameters can be arbitrarily large. This implies that, for the case of ERM learners, our lower bound is a significant strengthening of the lower bound for arbitrary learners from Theorem 3. It also implies that the sample complexity of ERM for comparative learning can differ significantly from the general sample complexity of the task (where no properness restrictions are imposed).

Examples 2 and 3 below illustrate how benchmark-ERM learning in the com-

parative setting relates to learning the benchmark class in the usual PAC setting.

Example 2. *When both source and benchmark are total, binary classes, $\mathcal{S}, \mathcal{B} \subseteq \{0,1\}^{\mathcal{X}}$, we have $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) = \mathrm{vc}(\mathcal{B})$. Thus, by Theorem 6, any pair $(\mathcal{S}, \mathcal{B})$ of total binary classes is benchmark-ERM comparative learnable if and only if $\mathcal{B}$ is agnostic PAC learnable.*

Example 3. *If $\mathcal{S} = \{h_0\}$ (the source contains only the all-0 classifier) and $\mathcal{B} = \{1,2\}^{\mathcal{X}}$, then $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) = 0$. However, the Natarjan dimension and the usual Graph dimension of the benchmark are both infinite, i.e., $d_N(\mathcal{B}) = \infty$ and $d_G(\mathcal{B}) = \infty$. Please refer to Definition 4 and Definition 5 for definitions of these two dimensions. This, again by Theorem 6, implies that as soon as we allow more than two labels, benchmark-ERM comparative learning can be easier than agnostic PAC learning of the benchmark.*

We now proceed to prove Theorem 6, which is established through the bounds provided in Lemmas 8 and 9 below.

Lemma 8. *For any pair of partial classes $(\mathcal{S}, \mathcal{B})$ with $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) \geq 2$, the sample complexity of benchmark-ERM satisfies*

$$\mathrm{n}_{\mathcal{S},\mathcal{B}}^{\mathrm{ERM}}(\epsilon, \delta) = \Omega\left(\frac{d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon}\right)$$

Proof. Let $\epsilon < \frac{1}{12}$ and $\delta < \frac{1}{100}$. We need to show that there exist benchmark-ERM learners that don't succeed in the sense of Definition 6 with less than the

stated sample sizes. Let $d = d_G^{\rightarrow}(\mathcal{S}, \mathcal{B})$ and let $U = \{x_1, \ldots, x_d\}$ be a set that is one-sided graph-shattered by $(\mathcal{S}, \mathcal{B})$ and also let $s^* \in S$ be the total function that witnesses this shattering. By definition, this means there exists $b^* \in B$ such that $s^*(x) = b^*(x)$ for all $x \in U$. Let the source function to be equal to s^* on U. Since $s^* = b^*$ on U this also means $\inf_{b \in B} \Pr[b(x) \neq s^*(x)] = 0$ for all distributions whose support is restricted to U. Consider a "bad" benchmark-ERM learner that for sample points $V \subseteq U$, returns $b \in B$ such that $b(x) = s^*(x)$ on V and $b(x) \neq s^*(x)$ on $U \backslash V$. Such b exists due to the definition of shattering. Also, note that this is a valid ERM since it has error 0 on samples. Define a distribution on U as follows:

$$\Pr[x_1] = 1 - 2\epsilon, \Pr[x_i] = \frac{2\epsilon}{d-1} \forall i \in [d] \backslash \{1\}$$

We now prove the lower bound by showing that the bad benchmark-ERM needs to see as many samples as $\max\{\frac{d-1}{6\epsilon}, \frac{1}{4\epsilon}\log(\frac{1}{\delta})\} \geq \frac{1}{2}(\frac{d-1}{6\epsilon} + \frac{1}{4\epsilon}\log(\frac{1}{\delta}))$. Let m be the size of the sample. We first show $m > \frac{d-1}{12\epsilon}$. To see this, suppose $m \leq \frac{d-1}{6\epsilon}$. Then using Chernoff's upper tail bound in Lemma 4, it is not hard to see that the sample points will be equal to x_1 except for at most $\frac{d-1}{2}$ of them with probability more than $\frac{1}{100} > \delta$. However, if this happens, since the learner has not seen at least $d - 1 - \frac{d-1}{2} = \frac{d-1}{2}$ light points (points in $U \backslash \{x_1\}$), its probability of error will be at least $\frac{d-1}{2}\frac{2\epsilon}{d-1} = \epsilon$. This means the probability of failure is more than δ and thus the algorithm cannot (ϵ, δ)-PAC learn. To see that $m > \frac{1}{4\epsilon}\log(\frac{1}{\delta})$, note that with probability $(1 - 2\epsilon)^m$, the sample will contain only x_1. This means the learner

will make an error on all light point which have mass 2ϵ, which is a failure. Thus, we need to ensure this probability is less than δ. However, if $m \leq \frac{1}{4\epsilon}\log(\frac{1}{\delta})$, then $(1-2\epsilon)^m \geq e^{-4\epsilon m} \geq \delta$. □

Lemma 9. *For any pair of partial classes $(\mathcal{S}, \mathcal{B})$ the sample complexity of benchmark-ERM satisfies*

$$\mathrm{n}^{\mathrm{ERM}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \tilde{\mathcal{O}}\left(\frac{d^{\rightarrow}_G(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon^2}\right).$$

Proof. We will show that the benchmark class $\mathcal{B}$ satisfies uniform convergence (see Definition 3) when restricted to the set of distributions that are realizable by the source class $\mathcal{S}$. This implies that any benchmark-ERM learner is successful in the sense of Definition 6. To see this, let us view both source class $\mathcal{S}$ and benchmark class $\mathcal{B}$ as collections of subsets of $\mathcal{X} \times \mathcal{Y}$. Note that realizability with respect to the source class $\mathcal{S}$ implies that we are only considering distributions P over $\mathcal{X} \times \mathcal{Y}$ whose support is included in one of the sets $s \in \mathcal{S}$, let's denote this set s^*. The benchmark-proper comparative learning task then is to choose a set $b \in \mathcal{B}$ with largest (in terms of probability weight) intersection with this set s^*.

Note that the definition of the one-sided graph dimension implies that for each $s \in \mathcal{S}$, the VC dimension of the collection of subsets $\mathcal{B}_s = \{s \cap b \mid b \in \mathcal{B}\}$ is finite. Moreover, the VC dimension of these collections is uniformly upper bounded by the one sided mutual graph dimension. That is $\mathrm{vc}(\mathcal{B}_s) \leq d^{\rightarrow}_G(\mathcal{S}, \mathcal{B})$ for all $s \in \mathcal{S}$. This implies that, restricted to the class of distributions realizable by $\mathcal{S}$, the benchmark

class satisfies uniform convergence. This yields the upper bound on the sample complexity of any benchmark-ERM learner stated in the theorem. □

4.1.2 Additional ERM Bounds in Terms of the Mutual Graph Dimension

In this section we define a new dimension which gives us upper bounds for the sample complexity of both source-ERM and benchmark-ERM. To do this, we first define an agreement loss class. Note that this is different from what is defined as the agreement class in [27].

Definition 9 (Agreement Loss Class)**.** *For classes $\mathcal{S}, \mathcal{B} \subseteq \tilde{\mathcal{Y}}^{\mathcal{X}}$ and for any $s \in \mathcal{S}$ and $b \in \mathcal{B}$, define the agreement function $a_{s,b} : \mathcal{X} \times \tilde{\mathcal{Y}} \to \{0, 1\}$ as follows:*

$$a_{s,b}(x, y) = \begin{cases} 0 & s(x) = b(x) = y \in \mathcal{Y} \\ 1 & \textit{otherwise} \end{cases} \tag{4.1}$$

Also let $A_{\mathcal{S},\mathcal{B}} = \{a_{s,b} : s \in \mathcal{S}, b \in \mathcal{B}\}$ denote the agreement class.

Next we define the mutual graph dimension of $(\mathcal{S}, \mathcal{B})$. As we will show this dimension controls VC dimension of the agreement loss class.

Definition 10 (Mutual Graph Dimension)**.** *A set $U \subseteq \mathcal{X}$ is mutually G-shattered by $(\mathcal{S}, \mathcal{B})$ if there exists a total function $g : U \to \mathcal{Y}$ such that for all $V \subseteq U$, there*

exists $s \in \mathcal{S}$ *and* $b \in \mathcal{B}$ *such that:*

$$\forall x \in V : s(x) = b(x) = g(x)$$

and

$$\forall x \in U \backslash V : s(x) \neq g(x) \vee b(x) \neq g(x)$$

The mutual graph dimension of $(\mathcal{S}, \mathcal{B})$*, denoted by* $d_G(S, B)$*, is defined as the maximum size of a set that can be mutually* G*-shattered.*

Lemma 10. *For any two hypothesis classes* $(\mathcal{S}, \mathcal{B})$*,* $d_G(\mathcal{S}, \mathcal{B}) = \mathrm{vc}(A_{\mathcal{S},\mathcal{B}})$

Proof. The proof is straightforward if you note that:

$$s(x) = b(x) = g(x) \text{ if and only if } a_{s,b}(x, g(x)) = 0,$$

and equivalently,

$$s(x) \neq g(x) \vee b(x) \neq g(x) \text{ if and only if } a_{s,b}(x, g(x)) = 1$$

Define $(U, g(U)) := \{(x, g(x)) : x \in U\}$. Then a set U is mutually G-shattered by $(\mathcal{S}, \mathcal{B})$ with g as a witness iff $(U, g(U))$ is VC-shattered by $A_{\mathcal{S},\mathcal{B}}$. □

Now we are ready to see how we can convert a proper learner for agnostic PAC learning of the agreement loss class to a benchmark-proper or source-proper comparative learner.

Theorem 11. *A pair $(\mathcal{S}, \mathcal{B})$ can be both benchmark-properly and source-properly learned in the comparative setting if $d_G(\mathcal{S}, \mathcal{B}) = \mathrm{vc}(A_{\mathcal{S},\mathcal{B}}) < \infty$. Furthermore, the sample complexity of learners achieving this task is $\tilde{\mathcal{O}}(\frac{d_G(\mathcal{S},\mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon^2})$.*

Proof. Fix any $\epsilon, \delta > 0$, and any distribution P on $\mathcal{X} \times \mathcal{Y}$ realizable with respect to $\mathcal{S}$. First, we can see that for any $a_{s,b}$ we have:

$$L(a_{s,b}) := \Pr[a_{s,b}(x, y) \neq 0] = \Pr[s(x) \neq y \vee b(x) \neq y] \tag{4.2}$$

Note that there does not necessarily exist a hypothesis $a_{s,b}$ with $L(a_{s,b}) = 0$. However, since $\mathrm{vc}(A_{\mathcal{S},\mathcal{B}}) < \infty$, by Theorem 1, we can learn $A_{\mathcal{S},\mathcal{B}}$ with $m = \tilde{\mathcal{O}}(\frac{\mathrm{vc}(A_{\mathcal{S},\mathcal{B}}) + \log(\frac{1}{\delta})}{\epsilon^2})$ by a **proper** algorithm $\mathcal{A}$. Let $S^m \in ((\mathcal{X} \times \mathcal{Y}) \times \{0\})^m$. Let $a_{s',b'} = \mathcal{A}(S^m)$. Then we have:

$$\begin{aligned}
\Pr[b'(x) \neq y] &\leq \Pr[s'(x) \neq y \vee b'(x) \neq y] \\
&= \Pr[a_{s',b'}(x, y) \neq 0] \\
&\leq \inf_{a_{s,b}} \Pr[a_{s,b}(x, y) \neq 0] + \epsilon \\
&= \inf_{s \in \mathcal{S}, b \in \mathcal{B}} \Pr[s(x) \neq y \vee b(x) \neq y] + \epsilon \\
&= \inf_{b \in \mathcal{B}} \Pr[b(x) \neq y] + \epsilon
\end{aligned}$$

The inequality is due to the PAC guarantees of $\mathcal{A}$ and the last equality follows from the fact that P is realizable with respect to $\mathcal{S}$. The same calculations are true if we replace b' with s' in the LHS. □

Corollary 12. *A pair $(\mathcal{S}, \mathcal{B})$ can be comparatively PAC learned by any source-ERM or benchmark-ERM if $d_G(\mathcal{S}, \mathcal{B}) = \mathrm{vc}(A_{\mathcal{S},\mathcal{B}}) < \infty$. Furthermore, the sample complexity of those learners is $\tilde{\mathcal{O}}(\frac{d_G(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon^2})$.*

Proof. Let b' and s' be the empirical risk minimizers over $\mathcal{B}$ and $\mathcal{S}$, respectively. That is, $b' \in \arg\min_{b \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} 1[b(x_i) \neq y_i]$ and $s' \in \arg\min_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} 1[s(x_i) \neq y_i]$. Then $a_{s',b'} \in \arg\min_{a_{s,b}} \frac{1}{n} \sum_{i=1}^{n} 1[a_{s,b}(x, y) \neq 0]$ and thus is a proper learner by Theorem 1. The results follow by the arguments in proof of Theorem 11. □

Although the bound for benchmark-ERM is not tight as we already proved a similar upper bound with the one-sided mutual graph dimension which is smaller than the mutual graph dimension, the upper bound for source-ERM is new as such result is lacking.

4.1.3 General benchmark-proper comparative learning

The lower bound in Theorem 6 is established through worst case benchmark-ERM learners. One might ask whether allowing general benchmark-proper learners yields a lower sample complexity. In Lemma 13 below we provide an example of two

hypotheses for which the sample complexity lower bound in Theorem 6 holds *for any benchmark-proper learning algorithm.* As a corollary, we show that there exist classes (and in particular binary total classes) that are comparatively learnable, but cannot be learned by any benchmark-proper learner.

Lemma 13. *For any $d \in \mathbb{N}$ such that $d \geq 2$, there exists a pair $(\mathcal{S}, \mathcal{B})$ of total, binary classes with $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) \geq \lfloor \frac{d}{2} \rfloor$ and $\mathrm{vc}(\mathcal{S}, \mathcal{B}) = 0$ such that*

$$\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \Omega\left(\frac{d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta})}{\epsilon}\right).$$

Proof. Here we assume d is even and for odd d we can replace d with $d-1$ in the rest of the proof. Let $\mathcal{X}_d = \{x_1, \ldots, x_d\}$, $\mathcal{X} = \{x_0\} \cup \mathcal{X}_d$. Let $\mathcal{S} = \{h_1\}$ consists of only the all 1 function and let $\mathcal{B} = \{1_{U \cup \{x_0\}} : U \subseteq \mathcal{X}_d, |U| = d/2\}$ consists of all functions that label half the domain and x_0 with 1 and the rest with 0. Here we have used 1_U to denote a function defined as $1_U(x) = \mathbb{1}[x \in U]$. It is not hard to see that $d_G^{\rightarrow}(\mathcal{S}, \mathcal{B}) = d/2$. Fix any benchmark-proper algorithm and let m be its sample complexity. Fix $\epsilon > 0$ and $0 < \delta \leq \epsilon$. Let $\epsilon' = \frac{2\epsilon}{1-\exp(-1/24)}$. For $U \subseteq \mathcal{X}_d$ with $|U| = d/2$, let P_U be a distribution on $\mathcal{X}$ such that $P_U(x_0) = 1 - 4\epsilon'$ and $P_U(x) = \frac{8\epsilon'}{d}$ for $x \in U$. First note that with probability $(1-4\epsilon')^m$ all samples equal x_0. In this case, for any learner that outputs a hypothesis corresponding U then there exists a distribution $P_{\mathcal{X} \setminus U}$ which makes the error of learner 1. Thus, we need to make sure the probability of this event is less than δ. Thus, we must have

$e^{-8m\epsilon'} \leq (1-4\epsilon')^m < \delta$ which means $m = \Omega(\frac{\log(\frac{1}{\delta})}{\epsilon'}) = \Omega(\frac{\log(\frac{1}{\delta})}{\epsilon})$. We now show $m \geq \frac{d}{32\epsilon'} = \Omega(\frac{d}{\epsilon})$. Assume $m < \frac{d}{32\epsilon'}$. Define $\mathcal{P}_k(A) := \{E \subseteq A : |E| = k\}$. For a sequence $S_{\mathcal{X}}$, let $\text{set}(S_{\mathcal{X}})$ denote its unique elements. We choose $U \in \mathcal{P}_{d/2}(\mathcal{X}_d)$ uniformly and get m samples from P_U. We would like to lower bound the following:

$$\mathbb{E}_{U \sim \text{Unif}(\mathcal{P}_{d/2}(\mathcal{X}_d))} \mathbb{E}_{S_{\mathcal{X}} \sim P_U^m} [\mathcal{L}_{P_U}(\mathcal{A}(S_{\mathcal{X}}))] \tag{4.3}$$

We can decompose U in the above into two parts. The part that is present in our sample and the rest of it. The number of sample points that fall into $\mathcal{X}_d$ are distributed as $N \sim \text{Bin}(m, 4\epsilon')$. Then the samples are distributed as $S_{\mathcal{X}} \sim \text{Unif}(\mathcal{X}_d^N)$ where we have ignored the samples that are x_0 for simplicity. The rest of the support is chosen uniformly from rest of the domain. However, note that the sample is a sequence and not a set. Thus, only $l = |\text{set}(S_{\mathcal{X}})| \leq N$ points have been chosen so far. Thus, the rest of the support is chosen as $A \sim \text{Unif}(\mathcal{P}_{d/2-l}(\mathcal{X}_d \backslash \text{set}(S_{\mathcal{X}})))$. Therefore, we can equivalently write 4.3 in the following form.

$$\mathbb{E}_N \mathbb{E}_{S_{\mathcal{X}}} \mathbb{E}_A [\mathcal{L}_{P_{S_{\mathcal{X}} \cup A}}(\mathcal{A}(S_{\mathcal{X}}))] \tag{4.4}$$

$$\geq \mathbb{E}_N \mathbb{E}_{S_{\mathcal{X}}} \mathbb{E}_A [\mathcal{L}_{P_{S_{\mathcal{X}} \cup A}}(\mathcal{A}(S_{\mathcal{X}})) | N \leq d/4] \Pr[N \leq d/4] \tag{4.5}$$

Using Chernoff's upper tail bound stated in Lemma 4, we can see that $\Pr[N \geq d/4] \leq \exp(-\frac{d}{24}) \leq \exp(-\frac{1}{24})$ for $m \leq \frac{d}{32\epsilon'}$. Fix any $N \leq d/4$ and $S_{\mathcal{X}} \in \mathcal{X}_d^N$. Note that $|\text{set}(S_{\mathcal{X}})| \leq N \leq d/4$. Let $1_{\{x_0\} \cup V} = \mathcal{A}(S_{\mathcal{X}})$. W.l.o.g. we assume $V = \text{set}(S_{\mathcal{X}}) \cup B$ because the learner makes a mistake on any point in the sample

that is not in its output set. Let $r = d/2 - |\text{set}(S_{\mathcal{X}})|$ be the size of the rest of the support which satisfies $r \geq d/4$. Then $\mathbb{E}_A|A \cap B| = \mathbb{E}\sum_{b\in B} 1_{b\in A} = r\Pr[b \in A] = \frac{r}{2}$ and $\mathbb{E}_A[A\backslash B] = \frac{r}{2}$. Therefore, $\mathbb{E}_A[\mathcal{L}_{P_{S_{\mathcal{X}}\cup A}}(\mathcal{A}(S_{\mathcal{X}}))] = \frac{r}{2}\frac{4\epsilon'}{d/2} \geq \epsilon'$. Putting it all together we have:

$$\mathbb{E}_{U\sim \text{Unif}(\mathcal{P}_{d/2}(\mathcal{X}_d))}\mathbb{E}_{S_{\mathcal{X}}\sim P_U^m}[\mathcal{L}_{P_U}(\mathcal{A}(S_{\mathcal{X}}))] \geq (1 - e^{-\frac{1}{24}})\epsilon' = 2\epsilon$$

Therefore, there exists U such that $\mathbb{E}_{S_{\mathcal{X}}\sim P_U^m}[\mathcal{L}_{P_U}(\mathcal{A}(S_{\mathcal{X}}))] \geq 2\epsilon$. Thus, we get that $\Pr_{S_{\mathcal{X}}\sim P_U^m}[\mathcal{L}_{P_U}(\mathcal{A}(S_{\mathcal{X}})) > \epsilon] \geq \epsilon \geq \delta$, which contradicts (ϵ, δ)-comparative learning. Thus, $m = \Omega(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$. □

Corollary 14. *There exists a pair $(\mathcal{S}, \mathcal{B})$ of total, binary classes that is comparatively learnable, but cannot be learned by any benchmark-proper learner.*

Proof. Let $(\mathcal{S}_d, \mathcal{B}_d)$ be as in Lemma 13 defined on $\mathcal{X}_d$. Let $\mathcal{X} = \{x_0\} \bigcup \cup_{d\geq 2}\mathcal{X}_d$ where $\mathcal{X}_d$'s are mutually exclusive. Let $\mathcal{B} = \cup_{d\geq 2}\mathcal{B}_d$ where $b \in \mathcal{B}_d$ labels each point in $\mathcal{X}_{d'}$ with 1 for $d' \neq d$ and let $\mathcal{S} = \{h_1\}$. Then, for each $d \geq 2$, there exist distributions as in the proof of Lemma 13 such that any proper learner needs $\Omega(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples. Considering the limit $d \to \infty$ establishes the impossibility for benchmark-proper learning. However, $\text{vc}(\mathcal{S}, \mathcal{B}) = 0$ and thus, the pair is comparatively learnable by Theorem 3. □

4.1.4 Linear versus quadratic dependence on the error parameter

Our upper and lower bound in Theorem 6 still exhibit a gap in terms of their dependence on the error parameter $\frac{1}{\epsilon}$ (linear versus quadratic). We now show that both rates can occur for proper comparative learning and provide a general condition that enforces quadratic dependence (a slow rate). When the two classes coincide (that is, $\mathcal{S} = \mathcal{B}$), the comparative learning task reduces to proper PAC learning in the realizable case (see Definition 1). Note that in case both the source and benchmark are total binary classes the one-sided mutual graph dimension corresponds to the VC dimension of the benchmark class, and thus in case $\mathcal{S} = \mathcal{B}$ the proper learning sample complexity is $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \tilde{\Theta}\left((d^{\rightarrow}_{G}(\mathcal{S}, \mathcal{B}) + \log(\frac{1}{\delta}))/(\epsilon)\right)$, corresponding to the lower bound in Theorem 6. On the other hand, if the source class contains all binary functions $\mathcal{S} = \{0,1\}^{\mathcal{X}}$, then the comparative learning task corresponds to PAC learning with deterministic labels, and it has been shown that there are classes whose ERM sample complexity and proper learning sample complexity exhibit quadratic dependence on $\frac{1}{\epsilon}$ [8, 11].

Interestingly, whether the proper learning sample complexity exhibits a $\frac{1}{\epsilon}$ or $\frac{1}{\epsilon^2}$ dependence hinges on the relatedness between the two classes rather than merely on the complexity of the source class (one might expect that a more complex source class corresponds to a more challenging learning problem).

It has been shown that proper learning can require quadratic dependence on $\frac{1}{\epsilon}$ even if the true labeling of the distribution is known to the learner [11]. This corresponds to comparative learning with a source class that is a singleton.

With the following definition and theorem we provide a sufficient condition for quadratic dependence in benchmark-proper comparative learning setting. We show that they are induced by the following relatedness parameter, which we term the *mutual star dimension* of source and benchmark class. It is closely related to the hollow star number [13]. For a hypothesis h and a set K, let $h|_K$ be the restriction of h to K. Similarly, for a hypothesis class $\mathcal{H}$, let $\mathcal{H}|_K$ be a class including hypotheses from $\mathcal{H}$ restricted to K.

Definition 11 (Mutual Star Dimension)**.** *A mutual star of size $k > 1$ for a pair of classes $\mathcal{S}, \mathcal{B} \subseteq \tilde{\mathcal{Y}}^{\mathcal{X}}$ consists of a set of k points $K = \{x_1, \ldots, x_k\} \subseteq \mathcal{X}$, $s \in \mathcal{S}$ and $b_1, \ldots, b_k \in \mathcal{B}$ with the following properties:*

- *$s|_K$ is a total functions,*
- *$s|_K$ is not realizable by $\mathcal{B}|_K$, and*
- *$s(x_i) \neq b_i(x_i)$ and $s(x_j) = b_i(x_j)$ for each $i, j \in [k]$ with $j \neq i$.*

We define the mutual star dimension of $(\mathcal{S}, \mathcal{B})$, denoted by $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B})$, to be the smallest $k > 1$ such that there exist a mutual star of size k. If no such k exists, we define $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B}) = \infty$.

Lemma 15. *For any pair of hypothesis classes $(\mathcal{S}, \mathcal{B})$, with finite mutual star dimension $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B}) = k$, $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \Omega(\frac{1}{k}\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$.*

Proof. Fix $0 < \epsilon < \frac{1}{k}$ and $0 < \delta < \frac{1}{4}$. Consider a mutual star of size k with parameters $K = \{x_1, \ldots, x_k\} \subseteq \mathcal{X}$, $s \in \mathcal{S}$ and $b_1, \ldots, b_k \in \mathcal{B}$ as defined in Definition 11. Consider a set of k distributions whose support lies in K and are defined as follows: $P_i(x_i) = \frac{1+\epsilon}{k} - \epsilon$ and $P_i(x_j) = \frac{1+\epsilon}{k}$ for all $i, j \in [k]$ with $j \neq i$. Then as s is not realizable by the benchmark class, $OPT_i = \inf_{b\in\mathcal{B}} \mathcal{L}_{P_i}(b) = \frac{1+\epsilon}{k} - \epsilon$ which is attained by b_i. On the other hand, if the learner outputs any other hypothesis in $\mathcal{B}$ that differs with s on any point other than x_i, it will occur loss of at least $\frac{1+\epsilon}{k}$ which is bigger than $OPT_i + \epsilon$ and means the learner is not a valid comparative learner. Thus, if the adversary picks the distribution uniformly random between P_i's, a successful learner must find out which point x_i has the smallest mass. We can thus reduce the *weighted dice problem* with k sides to this learning problem, and the sample complexity of this problem has been shown to be lower bounded by $\Omega(\frac{1}{k}\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$ (Theorem 2 by Ben-David and Ben-David [11]). □

The following simple example illustrates the phenomenon.

Example 4. *Consider a domain containing two points $\mathcal{X} = \{x_0, x_1\}$, a source class $\mathcal{S} = \{h_{01}\}$, and benchmark class $\mathcal{B} = \{h_{00}, h_{11}\}$, where the index of the function indicates which labels the function assigns to the two points. The set $K = \mathcal{X}$,*

$s = h_{01}$ and $b_1 = h_{00}, b_2 = h_{11}$ form a mutual star of size 1 and thus $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B}) = 1$. Now consider the class of distributions that are realizable by the source (thus label x_0 with 0 and x_1 with 1) and assign probability weights $\frac{1}{2} \pm \epsilon$ to the two points, for all $\epsilon > 0$. A benchmark proper learner then has to estimate which of the two points has heavier probability weight, in order to choose the optimal classifier among $\{h_{00}, h_{11}\}$. The proper learning problem over this set of distributions thus reduces to estimating the bias of a coin flip, which is known to require sample sizes of $\Omega(1/\epsilon^2)$ [3].

Now we show what happens when $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B}) = \infty$. Interestingly, we answer this question by formulating the mutual star dimension in another form. According to this new formulation, when $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B}) = \infty$, on each finite subset of the domain there exists a "best" benchmark hypothesis for each source hypothesis. If the learner knew the true labeling function $s^* \in \mathcal{S}$ and the domain was finite, the learner could simply output the best hypothesis for s^*. However, these assumptions do not necessarily hold. If the source class is proper realizable learnable, we can find a source hypothesis that is close the true source hypothesis in terms of classification error, which can then be used by the learner instead of the true source hypothesis. When the domain is countable, we show we can clip the ϵ-tail of the distribution and only consider a finite subset of the domain without incurring much loss. When the domain is uncountable, we show there exists a pair $(\mathcal{S}, \mathcal{B})$ proper comparative

learnability is undecidable.

For any two hypotheses h, g denote their disagreement set within K by $\mathrm{DIS}_K(h,g) := \{x \in K : h(x) \neq g(x)\}$. We say b_s is a minimal hypothesis (among all $b \in \mathcal{B}$) with respect to a hypothesis s on a set K if $\mathrm{DIS}_K(b_s, s) \subseteq \mathrm{DIS}_K(b, s)$ for all $b \in \mathcal{B}$. We also let $E_{\mathcal{B}} = \{x \in \mathcal{X} : \exists b_1, b_2 \in \mathcal{B} \text{ s.t. } b_1(x) \neq b_2(x)\}$ denote the effective domain of $\mathcal{B}$.

Lemma 16 (Equivalent form of $d_{\mathfrak{s}}$)**.** *For any pair $(\mathcal{S}, \mathcal{B})$, the mutual star dimension $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B})$ satisfies*

$$d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B}) = \inf\{k \in \mathbb{N} : \exists K = \{x_1, \ldots, x_k\} \text{ and } s \in \mathcal{S} \text{ defined on all points in } K \text{ s.t.}$$
$$\nexists b_s \in \mathcal{B} \text{ with } DIS_K(s, b_s) \subseteq DIS_K(s, b)\, \forall b \in \mathcal{B}\},$$

where we define $\inf \emptyset := \infty$

Proof. Let k be the term on the right hand side. We prove the lemma in two directions. First we show $d_{\mathfrak{s}}(\mathcal{S}, \mathcal{B}) \leq k$. If k is infinite, then the inequality holds. Thus, we assume it is finite. Then by definition, there exist a set $K = \{x_1, \ldots, x_k\}$, a source function $s \in \mathcal{S}$ such that there is no b_s. Furthermore, such k is the smallest number that satisfies these conditions. It is not hard to see that $k > 1$. The following holds by definition of k:

1. $\forall b \in \mathcal{B}, \mathrm{DIS}_K(s, b) \neq \emptyset$.

2. $\forall i \in [k]$ and $K_i := K \backslash \{x_i\}, \forall \tilde{s} \in \mathcal{S}$, there exist a minimal $b_{i,\tilde{s}}$ w.r.t. $\tilde{s}$ on K_i. In words, every source function on every subset of size $k-1$ has a minimal benchmark hypothesis.

3. Neither of the following can happen for any $x \in K$:

   (a) $\forall b \in \mathcal{B}$, $x \in \mathrm{DIS}_K(s, b)$.

   (b) $\forall b \in \mathcal{B}$, $x \notin \mathrm{DIS}_K(s, b)$.

   In words, we cannot have a point in K on which all benchmark hypotheses agree (or disagree) with s. Otherwise, we could remove such point and the conditions would still hold which contradicts the minimality of k.

We now prove the inequality using the above facts. First, for each x_i, let b_i be the minimal benchmark function w.r.t. s on $K_i = K \backslash \{x_i\}$ promised by 2. We argue that $\mathrm{DIS}_{K_i}(s, b_i) = \emptyset$, which by 1 implies $\mathrm{DIS}_K(s, b_i) = \{x_i\}$. Assume there exist $x \in \mathrm{DIS}_{K_i}(s, b_i)$, then by minimality of b_i on K_i, $x \in \mathrm{DIS}_{K_i}(s, b)$ for all $b \in \mathcal{B}$. However, this contradicts 3.(a). and thus, $\mathrm{DIS}_{K_i}(s, b_i) = \emptyset$. Fact 1 together with $\mathrm{DIS}_K(s, b_i) = \{x_i\}$ imply that s, b_is and K form a mutual hollow star and thus the inequality in this direction holds. To prove the other direction, we just need to show that a mutual hollow star also satisfies the condition of joint disagreement dimension, i.e. existence of a source hypothesis and a set K on which there is no minimal b with respect to s. This simply follows by definition of a mutual hollow

star structure as all the b_is in a hollow star satisfy $\mathrm{DIS}_K(s, b_i) = \{x_i\}$. Thus no b_i and b_j satisfy $\mathrm{DIS}_K(s, b_i) \subseteq \mathrm{DIS}_K(s, b_j)$. Moreover, if there is another $b \in \mathcal{B}$ that is minimal on K w.r.t. s, it must be that $\mathrm{DIS}_K(s, b) = \emptyset$ which contradicts the unrealizability assumption. □

Lemma 17. *If $d_{\mathfrak{s}}(\mathcal{S}|_{E_\mathcal{B}}, \mathcal{B}|_{E_\mathcal{B}}) = \infty$, $\mathcal{S}$ is proper realizable learnable on $E_\mathcal{B}$ with sample complexity $\tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$, and $\mathcal{X}$ is countable, then we can properly learn $(\mathcal{S}, \mathcal{B})$ using $\tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples.*

Proof. We first show that with $\tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples we can find $s \in \mathcal{S}$ such that with probability more than $1 - \frac{\delta}{3}$, $\Pr[s(x) \neq s^*(x), x \in E_\mathcal{B}] < \epsilon/2$ where s^* is the true labeling function. Let $\mu = \Pr[x \in E_\mathcal{B}]$. Then we have $\Pr[s(x) \neq s^*(x), x \in E_\mathcal{B}] = \mu \Pr[s(x) \neq s^*(x) | x \in E_\mathcal{B}]$. Thus, we need to find s such that $\Pr[s(x) \neq s^*(x) | x \in E_\mathcal{B}] < \frac{\epsilon}{2\mu}$. Given that $\mathcal{S}$ is proper realizable learnable on $E_\mathcal{B}$, we only need $m = \tilde{\mathcal{O}}(\mu \frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples in $E_\mathcal{B}$ and we can discard the rest of the samples. Then by using Chernoff's lower tail bound in Lemma 4, we get this many samples in $E_\mathcal{B}$ with probability at least $1 - \frac{\delta}{3}$ if we draw $\tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples from the whole distribution. To see this, consider samples $X_1, \ldots, X_N$ and define $Z_i = \mathbb{1}[X_i \in E_\mathcal{B}]$. Let $Z = \sum_{i=1}^{N} Z_i$. Then $\mathbb{E}[Z] = N\mu$. If $N \geq \frac{2m}{\mu}$ by Lemma 4 (setting $\eta = 0.5$), we have

$$\Pr[Z \leq m] \leq \Pr[Z \leq 0.5N\mu] \leq \exp(-\frac{N\mu}{8})$$

Choosing $N \geq \frac{8\log(\frac{3}{\delta})}{\mu}$ ensures that the above probability is bounded by $\frac{\delta}{3}$. Thus, getting $N = \max\{\frac{2m}{\mu}, \frac{8\log(\frac{3}{\delta})}{\mu}\} = \mathcal{O}(\frac{m+\log(\frac{1}{\delta})}{\mu}) = \tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples from the distribution suffices to guarantee our requirements. Note that for the last equality we need to assume $\mu \geq \epsilon$, however, this does not affect our results as when $\mu < \epsilon$, any proper learner is successful and proper learnability becomes trivial.

Fix an enumeration of $\mathcal{X}$. Then given $\mathcal{O}(\frac{\log(\frac{1}{\delta})}{\epsilon})$ samples, we have with probability at least $1 - \frac{\delta}{3}$, there exists x_l in samples such that $\Pr[x > x_l] < \epsilon/2$.

Thus, by a union bound, with probability at least $1 - \delta$ the above three events occur and we can condition the rest of the proof on them. Let $\mathcal{X}_l = \{x_1, \ldots, x_l\}$. Let $b_s \in \mathcal{B}$ be such that $\mathrm{DIS}_{\mathcal{X}_l}(s, b_s) \subseteq \mathrm{DIS}_{\mathcal{X}_l}(s, b)$ for all $b \in \mathcal{B}$. Such b_s exists as $d_{\mathfrak{s}}(\mathcal{S}|_{E_{\mathcal{B}}}, \mathcal{B}|_{E_{\mathcal{B}}}) = \infty$.

$$
\begin{aligned}
\Pr[b_s(x) \neq s^*(x)] &\leq \Pr[b_s(x) \neq (s^*(x) = s(x)), x \in E_{\mathcal{B}}] + \Pr[s^*(x) \neq s(x), x \in E_{\mathcal{B}}] \\
&\quad + \Pr[b_s(x) \neq s^*(x), x \notin E_{\mathcal{B}}] \\
&< \Pr[b_s(x) \neq (s^*(x) = s(x)), x \in E_{\mathcal{B}}] + \Pr[b_s(x) \neq s^*(x), x \notin E_{\mathcal{B}}] \\
&\quad + \epsilon/2 \\
&\leq \Pr[b_s(x) \neq (s^*(x) = s(x)), x \leq x_l, x \in E_{\mathcal{B}}] \\
&\quad + \Pr[b_s(x) \neq s^*(x), x \notin E_{\mathcal{B}}] + \Pr[x > x_l] + \epsilon/2 \\
&< \Pr[b_s(x) \neq (s^*(x) = s(x)), x \leq x_l, x \in E_{\mathcal{B}}] \\
&\quad + \Pr[b_s(x) \neq s^*(x), x \notin E_{\mathcal{B}}] + \epsilon \\
&\leq \Pr[b^*(x) \neq (s^*(x) = s(x)), x \leq x_l, x \in E_{\mathcal{B}}] \\
&\quad + \Pr[b^*(x) \neq s^*(x), x \notin E_{\mathcal{B}}] + \epsilon \\
&\leq OPT + \epsilon
\end{aligned}
$$

where $\mathcal{L}(b^*) = OPT$ and the second to last inequality follows from $\mathrm{DIS}_{\mathcal{X}_l}(s, b_s) \subseteq \mathrm{DIS}_{\mathcal{X}_l}(s, b^*)$. Note that for $x \notin E_{\mathcal{B}}$, we have $b_s(x) = b^*(x)$ by definition of $E_{\mathcal{B}}$ and thus $\Pr[b_s(x) \neq s^*(x), x \notin E_{\mathcal{B}}] = \Pr[b^*(x) \neq s^*(x), x \notin E_{\mathcal{B}}]$. □

Corollary 18. *If $\mathcal{S}$ is total multiclass hypothesis class, $d_{\mathfrak{s}}(\mathcal{S}|_{E_{\mathcal{B}}}, \mathcal{B}|_{E_{\mathcal{B}}}) = \infty$, $d = d_G(\mathcal{S}|_{E_{\mathcal{B}}}) < \infty$, and $\mathcal{X}$ is countable, then we can properly learn the pair $(\mathcal{S}, \mathcal{B})$ using $\tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples.*

Corollary 19. *If $\mathcal{S}$ is a total binary class, $d_{\mathfrak{s}}(\mathcal{S}|_{E_\mathcal{B}}, \mathcal{B}|_{E_\mathcal{B}}) = \infty$, $d = \mathrm{vc}(\mathcal{S}|_{E_\mathcal{B}}) < \infty$, and $\mathcal{X}$ is countable, then we can properly learn $(\mathcal{S}, \mathcal{B})$ using $\tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$ samples.*

Lemma 20. *Let $(\mathcal{S}, \mathcal{B})$ be a pair of total binary classes. If $\mathrm{vc}(\mathcal{S}|_{E_\mathcal{B}}) = \infty$ and $d_{\mathfrak{s}}(\mathcal{S}|_{E_\mathcal{B}}, \mathcal{B}|_{E_\mathcal{B}}) = \infty$ then the sample complexity of benchmark-proper comparative learning is $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \infty$.*

Proof. Consider sets $W_k = \{A_k \subseteq E_\mathcal{B} : |A_k| = k \text{ and } A_k \text{ is shattered by } \mathcal{S}\}$ for $k \in \mathbb{N}$. Such sets are nonempty as $\mathrm{vc}(\mathcal{S}|_{E_\mathcal{B}}) = \infty$. Consider the following two cases:

1. $\mathcal{B}$ shatters all $A_k \in W_k$ for all $k \in \mathbb{N}$: This would imply that $\mathrm{vc}(\mathcal{S}, \mathcal{B}) = \infty$ which in turn means $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \infty$ by Theorem 3.

2. There exists a set $A_k \in W_k$ that is not shattered by $\mathcal{B}$. Consider the smallest A_k that is shattered by $\mathcal{S}$ and not shattered by $\mathcal{B}$. We first note that $k = |A_k| > 1$. By definition, all $A_1 \in W_1$ satisfy $A_1 \subseteq E_\mathcal{B}$. As $\mathcal{B}$ produces both 0 and 1 labels on points in $E_\mathcal{B}$, we have that all single points in $E_\mathcal{B}$ and thus all $A_1 \in W_k$ are shattered by $\mathcal{B}$, which implies $|A_k| > 1$. Given that A_k is not shattered by $\mathcal{B}$, each benchmark hypothesis b makes a mistake on at least one point and since there is another hypothesis that does not make a mistake on that point (as it is in $E_\mathcal{B}$), b cannot be a minimal hypothesis on A_k. On the other hand, each proper subset of A_k is shattered by $\mathcal{B}$ and hence on each of those subsets there exists a benchmark hypothesis that makes zero

mistakes. These imply existence of a mutual star of size k defined on A_k which contradicts the assumption that $d_{\mathfrak{s}}(\mathcal{S}|_{E_{\mathcal{B}}}, \mathcal{B}|_{E_{\mathcal{B}}}) = \infty$, and thus this case cannot happen.

□

We are now ready to state one of our main results, which shows how the mutual star dimension governs the dependence of $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}$ on $\frac{1}{\epsilon}$ for total binary classes. Namely, when the mutual star dimension is finite, the slow rate of $\frac{1}{\epsilon^2}$ is unavoidable for any proper learner. Conversely, when the mutual star dimension is infinite, if the domain is countable and the source class has finite VC dimension on the effective domain of the benchmark class, then a proper learner can achieve the fast rate of $\frac{1}{\epsilon}$. Finally, if both the mutual star dimension and the VC dimension of the source on the effective domain are infinite, then no proper learner can succeed with a finite amount of data.

Theorem 21. *For a pair $(\mathcal{S}, \mathcal{B})$ of total binary hypothesis classes one the following cases happen:*

- *If $d_{\mathfrak{s}}(\mathcal{S}|_{E_{\mathcal{B}}}, \mathcal{B}|_{E_{\mathcal{B}}}) = k < \infty$, $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \Omega(\frac{1}{k\epsilon^2})$*
- *If $d_{\mathfrak{s}}(\mathcal{S}|_{E_{\mathcal{B}}}, \mathcal{B}|_{E_{\mathcal{B}}}) = \infty$, $\mathrm{vc}(\mathcal{S}|_{E_{\mathcal{B}}}) = d < \infty$, and $\mathcal{X}$ is countable, then $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \tilde{\mathcal{O}}(\frac{d+\log(\frac{1}{\delta})}{\epsilon})$.*

- *If* $d_{\mathfrak{s}}(\mathcal{S}|_{E_{\mathcal{B}}}, \mathcal{B}|_{E_{\mathcal{B}}}) = \infty$ *and* $\mathrm{vc}(\mathcal{S}|_{E_{\mathcal{B}}}) = \infty$*, then* $\mathrm{n}^{\mathrm{prop}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \infty$.

Proof. The first case is proven in Lemma 15, the second case is a result of Corollary 19, and the third case is the result of Lemma 20. □

It now remains open to see what happens in the case of uncountable domain. It was recently shown in [4] that proper PAC learnability can be undecidable (within the axioms of ZFC). The class whose proper PAC learnability is undecidable is a multiclass class with infinite classes defined on an uncountable set. This raised a question that whether proper comparative learnability on an uncountable domain can be undecidable as well. Interestingly, we find that even for total binary classes, the sample complexity of benchmark-proper comparative learning can be undecidable (this is in contrast to binary PAC learnability where proper learnability and learnability are both characterized by VC dimension). Similar to [4] our proof follows by reducing the problem to EMX problem introduced in [10] defined as follows:

Definition 12 (EMX problem [10])**.** *Let* $\mathcal{X}$ *be any set and let* P *be any unknown distribution supported on countably many points in* $\mathcal{X}$*. A class* $\mathcal{H} \subseteq 2^{\mathcal{X}}$ *is EMX learnable if there exists an algorithm* $\mathcal{A}$ *that given* $n = n_{EMX}(\epsilon, \delta)$ *i.i.d. samples* $S \sim P^n$*, with probability more than* $1 - \delta$ *returns* $h_S = \mathcal{A}(S) \in \mathcal{H}$ *such that* $P[h_s] > \sup_{h \in \mathcal{H}} P[h] - \epsilon$.

Theorem 22 ([10]). *Let $\mathcal{X} = \mathbb{R}$ and let $\mathcal{H} = \{h \subseteq \mathbb{R} : |h| < \infty\}$. Then the EMX learnability of $\mathcal{H}$ is undecidable.*

Theorem 23. *There exist a pair $(\mathcal{S}, \mathcal{B})$ of total binary classes whose benchmark-proper comparative learnability is undecidable.*

Proof. Let $\mathcal{X} = \mathbb{R}$. Define $\mathcal{S}, \mathcal{B} \subseteq \{0,1\}^{\mathcal{X}}$ as $\mathcal{S} = \{h_1\}$ and $\mathcal{B} = \{h_A : A \subseteq \mathcal{X}, |A| < \infty\}$ where h_1 is a hypothesis that labels the entire domain with 1 and h_A only labels the points in A with 1 and is 0 on other points. Consider the set of distributions with countable support on $\mathcal{X}$. For a distribution P, the error of a hypothesis $h_A \in \mathcal{B}$ is then equal to $\mathcal{L}_P(h_A) = P[h_A(X) = 0] = 1 - P[A]$. Thus, the goal of proper comparative learning can be reformulated as finding A such $P[A] > sup_{\tilde{A} \subseteq \mathcal{X}} P[\tilde{A}] - \epsilon$ with high probability, which is the definition of EMX-learnability of $\mathcal{B}$. However, by Theorem 22 the learnability of this problem is undecidable, which means proper comparative learnability of $(\mathcal{S}, \mathcal{B})$ is undecidable. □

Remark 24. *The distributions in Theorem 23 are restricted to have countable support as required by the definition of EMX problem. However, the reason for this assumption as stated in [10], is that we need the loss function for all the hypotheses in our class to be measurable. Thus, this result holds for all distributions under which, the loss function for all the hypotheses in the class is measurable. We note that this assumption is often implicitly made in PAC-type settings (for example, see*

Remark 3.1 in [44]).

Similar to [4], we show that restriction of hypotheses to finite subsets of the domain does not determine proper learnability, i.e., there are two pairs of hypotheses that are equal when restricted to finite subsets, but one pair is proper comparative learnable while the learnability of the other is undecidable. They also show that proper learnability is not a monotone property, i.e., if $\mathcal{H}_1 \subseteq \mathcal{H}_2$ and $\mathcal{H}_2$ is proper learnable, we cannot conclude $\mathcal{H}_1$ is learnable. We show similar results hold in the comparative setting.

Theorem 25 (Proper learnability is not a local property)**.** *There exist pairs* $(\mathcal{S}, \mathcal{B}_1)$ *and* $(\mathcal{S}, \mathcal{B}_2)$ *such that* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *agree on every finite subset of domain, i.e.,* $\mathcal{B}_1|_S = \mathcal{B}_2|_S$ *for all* $S \subseteq \mathcal{X}$ *with* $|S| < \infty$*, but the proper comparative learnability of one pair is decidable and the other not.*

Proof. Let $(\mathcal{S}, \mathcal{B}_1) = (\mathcal{S}, \mathcal{B})$ from Theorem 23 and let $\mathcal{B}_2 = \mathcal{B}_1 \cup \{h_1\}$. It is easy to see that $\mathcal{B}_1$ and $\mathcal{B}_2$ are equal on every finite subset of the domain. However, proper learnability of $(\mathcal{S}, \mathcal{B}_1)$ is undecidable by Theorem 23, but $(\mathcal{S}, \mathcal{B}_2)$ is properly learnable by a learner that outputs h_1 without seeing any samples. □

Theorem 26 (Proper learnability is not a monotone property)**.** *There exist classes* $\mathcal{S}$*,* $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2$ *satisfying* $\mathcal{B}_0 \subset \mathcal{B}_1 \subset \mathcal{B}_2$ *such that* $(\mathcal{S}, \mathcal{B}_0)$ *and* $(\mathcal{S}, \mathcal{B}_2)$ *are proper comparative learnable but proper comparative learnability of* $(\mathcal{S}, \mathcal{B}_1)$ *is undecidable.*

Proof. Let $\mathcal{S}, \mathcal{B}_1, \mathcal{B}_2$ be as in the proof of Theorem 25. Let $\mathcal{B}_0 = \{h_A\}$ for some finite $A \subseteq \mathcal{X}$. The results for $(\mathcal{S}, \mathcal{B}_1)$ and $(\mathcal{S}, \mathcal{B}_2)$ follow by Theorem 25. Clearly, $(\mathcal{S}, \mathcal{B}_0)$ is proper comparative learnable as it only has one hypothesis that the learner can always output. □

4.2 General Comparative Learning

We now consider the general comparative learning setting, where no properness requirement is imposed. We start by again noting that both linear and quadratic dependence in the error parameter $\frac{1}{\epsilon}$ can occur even in the case of total binary classes: as noted above, the case where source and benchmark coincide, $\mathcal{S} = \mathcal{B}$, yields sample complexity $\mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \tilde{\Theta}(\frac{\mathrm{vc}(\mathcal{B})+\log(\frac{1}{\delta})}{\epsilon})$ in the realizable case (see Theorem 1 in Section 3.1). The case where the source consists of all binary functions, $\mathcal{S} = \{0,1\}^{\mathcal{X}}$ corresponds to agnostic PAC learning with deterministic labels. For binary total classes it has been shown that this setting has sample complexity with quadratic dependence $\Omega(1/\epsilon^2)$ if and only if the hypothesis class has infinite *diameter* [8]. The diameter $\mathrm{diam}(\mathcal{H})$ of a total hypothesis class $\mathcal{H}$ is defined to be the largest set on which two functions from the class disagree:

$$\mathrm{diam}(\mathcal{H}) = \sup\{k \in \mathbb{N} \mid \exists U \subseteq \mathcal{X}, |U| = k, \exists h_1, h_2 \in \mathcal{H} \text{ s.t. } h_1(x) \neq h_2(x) \forall x \in U\}$$

Thus, for the benchmark class $\mathcal{B} = \{h_0, h_1\}$ containing only the constant 0 and constant 1 functions (or any benchmark class including $\{h_0, h_1\}$), the comparative learning sample complexity with source $\mathcal{S} = \{0,1\}^{\mathcal{X}}$ (or any source class that shatters arbitrarily large sets) is $\mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \Omega(1/\epsilon^2)$.

We start by generalizing some of the results for agnostic learning under deterministic labels to partial classes to derive more fine grained bounds for comparative learning of partial classes.

4.2.1 Agnostic learning with deterministic labels for partial classes

We start by extending the notion of diameter to partial hypothesis classes. Recall that for a partial hypothesis $h \in \tilde{\mathcal{Y}}^{\mathcal{X}}$, we define the support of h to be the subset of the domain, where h assigns labels, that is $\mathrm{supp}(h) = \{x \in \mathcal{X} \mid h(x) \in \mathcal{Y}\}$.

Definition 13 (Diameter and joint diameter for partial classes)**.** *Let $\mathcal{H} \subseteq \tilde{\mathcal{Y}}^{\mathcal{X}}$ be a partial binary hypothesis class. We define the* diameter *of the partial class as*

$$\mathrm{diam}(\mathcal{H}) = \sup_{h,h' \in \mathcal{H}} |\{x \in \mathrm{supp}(h) \mid h(x) \neq h'(x)\}|.$$

We further define the joint diameter *of the partial class as*

$$\mathrm{diam}'(\mathcal{H}) = \sup_{h,h' \in \mathcal{H}} |\{x \in \mathrm{supp}(h) \cap \mathrm{supp}(h') \mid h(x) \neq h'(x)\}|.$$

Note that for any binary partial class $\mathcal{H}$ we have $\mathrm{vc}(\mathcal{H}) \leq \mathrm{diam}'(\mathcal{H}) \leq \mathrm{diam}(\mathcal{H})$ where VC dimesion of a partial class, denoted by $\mathrm{vc}(\mathcal{H})$, is defined in Section 3.1.

In case $\mathcal{H}$ is a total class, the two latter notions of diameter coincide and equal the diameter for total classes as defined above.

We will now show that bounded diameter implies a fast rate, $\tilde{O}(\frac{1}{\epsilon})$ while infinite joint diameter implies a slow rate of $\Omega(\frac{1}{\epsilon^2})$

Theorem 27. *Let $\mathcal{H} \subseteq \tilde{\mathcal{Y}}^{\mathcal{X}}$ be a partial hypothesis class with finite diameter* $\operatorname{diam}(\mathcal{H}) = d < \infty$. *Then the sample complexity of agnostically PAC learning $\mathcal{H}$ under deterministic labels is upper bounded by*

$$\mathcal{O}(\frac{d}{\epsilon}[\log(\frac{d}{\epsilon}) + \log(\frac{1}{\delta})])$$

Proof. Let $h_0 \in \mathcal{H}$ be an arbitrary function in the hypothesis class. Similar to the proof of Theorem 9 by Ben-David and Urner [8], we consider the following learner $\mathcal{A}$: given a sample $S = ((x_1, y_1), (x_2, y_2), \dots (x_n, y_n))$ the learner outputs a classifier $f = \mathcal{A}(S)$ with $f(x) = y_i$ in case $x = x_i$ for some $i \in [n]$ and $f(x) = h_0(x)$ otherwise. Note that this is not a proper learner for the class $\mathcal{H}$ (even if $\mathcal{H}$ is a total class).

Now consider any deterministic distribution P over $\mathcal{X} \times \mathcal{Y}$ and let $h^* \in \mathcal{H}$ be the hypothesis that achieves the approximation error of the class, that is $\mathcal{L}_P(h^*) = \operatorname{opt}_P(\mathcal{H})$. A sample of size $\mathcal{O}(\frac{d}{\epsilon}[\log(\frac{d}{\epsilon}) + \log(\frac{1}{\delta})])$ contains every point with mass at least $\frac{\epsilon}{d}$ with probability more than $1 - \delta$. To see this, note that the probability of a point with mass at least $\frac{\epsilon}{d}$ not appearing in a sample of size m is upper bounded by

$(1 - \frac{\epsilon}{d})^m \leq \exp(-m\frac{\epsilon}{d})$. There are at most $\frac{d}{\epsilon}$ such points, thus, probability that at least one of them does not appear in the sample is upper bounded by $\frac{d}{\epsilon}\exp(-m\frac{\epsilon}{d})$ by a union bound. Setting this to be less than δ gives the desired results. We now condition the rest of the proof on this event and show that the learner achieves error at most $\mathcal{L}_P(f) \leq \text{opt}_P(\mathcal{H}) + \epsilon$ (the failure probability is thus bounded by δ as required).

Since the distribution P is deterministic, the classifier $f = \mathcal{A}(S)$ will not make any mistake on points from S. Further, due to the finite diameter, outside of the sample the classifier f disagrees with h^* on at most $d = \text{diam}(\mathcal{H})$ points from the support $\text{supp}(h^*)$ of the optimal classifier (since f classifies according to the default function $h_0 \in \mathcal{H}$ on these points). As we assumed that the sample contained all points with probability mass at least ϵ/d, the d points on which f and h^* disagree have joint mass at most $d \cdot \epsilon/d = \epsilon$. This implies $\mathcal{L}_P(f) \leq \mathcal{L}_P(h^*) + \epsilon$, which completes the proof. □

We now show that infinite joint diameter implies a slow rate for learning under deterministic labels:

Theorem 28. *Let $\mathcal{H} \subseteq \tilde{\mathcal{Y}}^{\mathcal{X}}$ be a partial hypothesis class with infinite joint diameter* $\text{diam}'(\mathcal{H}) = \infty$. *Then the sample complexity of agnostically PAC learning $\mathcal{H}$ under deterministic labels is lower bounded by* $\Omega\left(\frac{1}{\epsilon^2}\right)$.

Proof. Since $\text{diam}'(\mathcal{H}) = \infty$, for any $\epsilon > 0$, there exist two hypotheses $h_0, h_1 \in \mathcal{H}$ and a set U with cardinality $\Omega(\frac{1}{\epsilon^3})$ such that $U \subseteq \text{supp}(h_0) \cap \text{supp}(h_1)$ and $h_0(x) \neq h_1(x)$ for all $x \in U$. We now consider the set of all label-deterministic distributions P over $\mathcal{X} \times \mathcal{Y}$ with marginal support included in U, that is $\text{supp}(P_{\mathcal{X}}) \subseteq U$. Learning the class $\mathcal{H}$ with respect to this set distributions requires at least as many samples as learning the smaller class $\{h_0, h_1\}$. Thus, by Lemma 3 and Remark 5 in [8], restated in Lemma 2, and noting that restrictions of h_0 and h_1 to U are total hypotheses, the sample complexity is lower bounded by $\Omega\left(\frac{1}{\epsilon^2}\right)$. □

4.2.2 Fast and slow rates in general comparative learning

We now use the results from the previous section to derive novel bounds for the comparative learning scenario. We state out main results in this section for the case where the source and benchmark are both partial binary classes, but in Remark 31, we discuss how they can be generalized to general label spaces. As in the original work on the comparative setting, we will employ the *agreement class* of source and benchmark.

Definition 14 (Agreement Class, due to [27])**.** *For two (partial) hypotheses s and*

b, we define their agreement hypothesis *as follows:*

$$a_{s,b}(x) = \begin{cases} y & s(x) = b(x) = y \in \mathcal{Y} \\ \star & \textit{otherwise} \end{cases}$$

Furthermore, for two (partial) hypothesis classes $\mathcal{S}$ *and* $\mathcal{B}$*, their* agreement hypothesis class *is defined as* $A_{\mathcal{S},\mathcal{B}} = \{a_{s,b} \mid s \in \mathcal{S}, b \in \mathcal{B}\}$.

Note that the agreement class is a partial class (except for degenerate cases) even if both $\mathcal{S}$ and $\mathcal{B}$ are total classes. It has been shown that the VC dimension of the agreement class coincides with the mutual VC dimension of the two involved classes, $\mathrm{vc}(A_{\mathcal{S},\mathcal{B}}) = \mathrm{vc}(\mathcal{S},\mathcal{B})$, and that any agnostic learner for the agreement class comparatively learns source and benchmark in the sense of Definition 6 [27]. Combining this with our upper bound in Theorem 27, and known sample complexity bounds for learning partial classes in the realizable case, yields:

Theorem 29. *Let* $(\mathcal{S},\mathcal{B})$ *be a pair of partial binary classes, and define* d *as* $d = \min\{\mathrm{vc}(\mathcal{S}), \mathrm{diam}(\mathcal{B}), \mathrm{diam}(A_{\mathcal{S},\mathcal{B}})\}$. *Then the sample complexity of comparatively learning the pair* $(\mathcal{S},\mathcal{B})$ *is upper bounded by*

$$\mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}(\epsilon,\delta) = \tilde{O}\left(\frac{d + \log(\frac{1}{\delta})}{\epsilon}\right).$$

Proof. If the source has finite VC dimension, then the sample complexity of learning this partial class in the realizable case is upper bounded by $\tilde{O}(\frac{\mathrm{vc}(\mathcal{S})+\log(\frac{1}{\delta})}{\epsilon})$ [2],

and realizably learning the source class satisfies the success criterion of comparative learning. Similarly, if the agreement class or the benchmark class have finite diameter, we can learn these classes with sample complexity $\tilde{O}(\frac{\text{diam}(A_{\mathcal{S},\mathcal{B}})+\log(\frac{1}{\delta})}{\epsilon})$ and $\tilde{O}(\frac{\text{diam}(\mathcal{B})+\log(\frac{1}{\delta})}{\epsilon})$ respectively, and in those cases the resulting learner satisfies the success criterion for comparative learning. Thus, the sample complexity of comparatively learning $(\mathcal{S},\mathcal{B})$ will be determined by the sample complexity of learner that achieves smallest sample complexity among these three. □

One might ask how the three dimensions that play a role in Theorem 29 relate to each other. In Example 5 below, we show that each of them can be finite while the other two are infinite. Finally, we identify a general condition which yields sample complexity growing quadratically in $\frac{1}{\epsilon}$.

Example 5. *For $y \in \{0,1,\star\}$, let h_y denote a function that labels every point in the domain with y.*

- *Let $\mathcal{S} = \{h_0\}$ and $\mathcal{B} = \{h_0, h_1\}$ so that $A_{\mathcal{S},\mathcal{B}} = \{h_0, h_\star\}$. Then $\text{vc}(\mathcal{S}) = 0 < \infty$, but $\text{diam}(\mathcal{B}), \text{diam}(A_{\mathcal{S},\mathcal{B}}) = \infty$.*

- *Let $\mathcal{S} = \{0,1\}^{\mathcal{X}}$ and $\mathcal{B} = \{h_0\}$ so that $A_{\mathcal{S},\mathcal{B}} = \{0,\star\}^{\mathcal{X}}$. Then $\text{diam}(\mathcal{B}) = 0 < \infty$, but $\text{vc}(\mathcal{S}), \text{diam}(A_{\mathcal{S},\mathcal{B}}) = \infty$.*

- *Let $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ where $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint and both have an infinite cardinality. Let $\mathcal{S}$ be such that $\mathcal{S}|_{\mathcal{X}_1} = \{0,1\}^{\mathcal{X}_1}$ and $\mathcal{S}|_{\mathcal{X}_2} = \{h_1\}$. Moreover,*

let $\mathcal{B}$ *be such that* $\mathcal{B}|_{\mathcal{X}_1} = \{h_\star\}$ *and* $\mathcal{B}|_{\mathcal{X}_2} = \{0, \star\}^{\mathcal{X}_2}$*. In this case* $A_{\mathcal{S},\mathcal{B}} = \{h_\star\}$ *and thus* $\operatorname{diam}(A_{\mathcal{S},\mathcal{B}}) = 0 < \infty$*. However,* $\operatorname{vc}(\mathcal{S}), \operatorname{diam}(\mathcal{B}) = \infty$*.*

Definition 15 (Mutual VC-Diameter)**.** *The mutual VC-Diameter of a pair of partial classes* $(\mathcal{S}, \mathcal{B})$ *is defined as follows:*

$$\operatorname{vcdiam}(\mathcal{S}, \mathcal{B}) = \sup\{n \in \mathbb{N} \mid \exists U \subseteq \mathcal{X}, |U| = n, \textit{such that}$$
$$\mathcal{S} \textit{ VC-shatters } U, \textit{ and } \operatorname{diam}'(\mathcal{B}\,|_U) = n\}$$

Theorem 30. *For a pair of partial binary classes, if* $\operatorname{vcdiam}(S, B) = \infty$*, then the comparative sample complexity satisfies* $\mathrm{n}^{\mathrm{gen}}_{\mathcal{S},\mathcal{B}}(\epsilon, \delta) = \Omega(\frac{1}{\epsilon^2})$*.*

Proof. This bound follows directly from the quadratic lower bound of Theorem 28. □

Remark 31. *The result of Theorem 29 can be generalized to any label space if we replace* $\operatorname{vc}(\mathcal{S})$ *by a dimension that characterizes PAC learning of* $\mathcal{S}$ *in the realizable setting for that label space. Theorem 30 can be generalized to any label space if we replace VC-shattering in Definition 15 with the following: for all* $A \subseteq U$*, there exist* $s \in \mathcal{S}$ *such that* $s(x) = b_0(x)$ *for* $x \in A$ *and* $s(x) = b_1(x)$ *for* $x \in U \backslash A$ *where* $b_0, b_1 \in \mathcal{B}$ *and* $b_0(x) \neq b_1(x)$ *for* $x \in U$*.*

5 Conclusion and Future Work

In this work, we mainly focus on proper comparative learning, a variation of comparative learning [27]. While a total binary class is PAC learnable if and only if it is proper PAC learnable, we surprisingly show that this does not hold for comparative learning. In other words, there are total binary classes that are comparative learnable but not proper comparative learnable. We study ERMs in this setting as the most popular type of proper learners. We introduce a new dimension we call the one-sided graph dimension and show it controls ERM learnability of a pair of classes. We also introduce a new dimension, mutual star dimension, and show this dimension along with VC dimension of the source class control the sample complexity rates that can be achieved by a proper learner. We further show that proper comparative learnability can be undecidable within the axioms of ZFC. We also introduce general conditions for the exact rate of convergence in the general comparative learning framework.

While we achieved a full landscape for the exact rates for total binary classes

in the proper settings, it would be interesting to obtain similar results for the multiclass and/or partial classes. Moreover, our results for ERM hold for the worst-case scenario. It might be possible to design ERMs that achieve better rates. A full landscape of the exact rates in the general comparative learning problem is also left open for future work.

Similar to the PAC framework, one can derive many interesting extensions of (proper) comparative learning. Here we mainly focused on the classification task. A natural question then is whether similar results can be achieved for the regression task.

In the original work of [27], comparative learning was also studied in the online setting where the data is adversarially selected rather than statistically from a distribution. Another possible direction for future work would be to generalize proper comparative learning to online proper comparative learning.

Bibliography

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *34th Annual Symposium on Foundations of Computer Science*, pages 292–301. IEEE Computer Society, 1993.

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

[3] Martin Anthony and Peter L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

[4] Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Proper learnability and the role of unlabeled data. In *36th International Conference on Algorithmic Learning Theory*, 2025.

[5] Idan Attias, Steve Hanneke, and Yishay Mansour. A characterization of semi-supervised adversarially robust PAC learnability. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2022.

[6] Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: PAC learning and online learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*, 2023.

[7] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. *CoRR*, abs/1706.09773, 2017.

[8] Shai Ben-David and Ruth Urner. The sample complexity of agnostic learning under deterministic labels. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 527–542, 2014.

[9] Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler, and Philip M. Long. Characterizations of learnability for classes of {0, ..., n}-valued functions. *J. Comput. Syst. Sci.*, 50(1):74–86, 1995.

[10] Shai Ben-David, Pavel Hrubes, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Learnability can be undecidable. *Nat. Mach. Intell.*, 1(1):44–48, 2019.

[11] Shalev Ben-David and Shai Ben-David. Learning a classifier when the labeling is known. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory - 22nd International Conference, ALT*, volume 6925 of *Lecture Notes in Computer Science*, pages 440–451. Springer, 2011.

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

[13] Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal SVM bound. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 582–609. PMLR, 2020.

[14] Marco Bressan, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A theory of interpretable approximations. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, COLT*, volume 247 of *Proceedings of Machine Learning Research*, pages 648–668. PMLR, 2024.

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

[16] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In Robert G. Cowell and Zoubin Ghahramani, editors,

Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS, pages 57–64. Society for Artificial Intelligence and Statistics, 2005.

[17] Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 287–316. JMLR.org, 2014.

[18] Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *J. Mach. Learn. Res.*, 16:2377–2404, 2015.

[19] Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 2784–2792, 2016.

[20] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Gal Elidan, Kristian Kersting, and Alexander Ihler, editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017*. AUAI Press, 2017.

[21] Nicholas Frosst and Geoffrey E. Hinton. Distilling a neural network into a soft decision tree. In Tarek R. Besold and Oliver Kutz, editors, *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*, volume 2071 of *CEUR Workshop Proceedings*, 2017.

[22] Steve Hanneke. The optimal sample complexity of PAC learning. *J. Mach. Learn. Res.*, 17:38:1–38:15, 2016.

[23] Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Sample compression for real-valued learners. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory, ALT 2019*, volume 98 of *Proceedings of Machine Learning Research*, pages 466–488. PMLR, 2019.

[24] D. Haussler, N. Littlestone, and M.K. Warmuth. Predicting 0, 1-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994. ISSN 0890-5401.

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

[26] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[27] Lunjia Hu and Charlotte Peale. Comparative learning: A sample complexity theory for two hypothesis classes. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference, ITCS*, volume 251 of *LIPIcs*, pages 72:1–72:30. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

[28] Tosca Lechner and Shai Ben-David. Inherent limitations of dimensions for characterizing learnability of distribution classes. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, COLT*, volume 247 of *Proceedings of Machine Learning Research*, pages 3353–3374. PMLR, 2024.

[29] Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. *Unpublished*, 1986.

[30] Philip M. Long. On agnostic learning with {0, *, 1}-valued and real-valued hypotheses. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 289–302. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-44581-4.

[31] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR*, 2016.

[32] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326 – 2366, 2006.

[33] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[34] Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 2019.

[35] Omar Montasser, Steve Hanneke, and Nati Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems NeurIPS*, 2022.

[36] Shay Moran and Amir Yehudayoff. Sample compression schemes for vc classes. In *2016 Information Theory and Applications Workshop (ITA)*, pages 1–14, 2016.

[37] B. K. Natarajan. On learning sets and functions. *Mach. Learn.*, 4:67–97, 1989.

[38] B.K. Natarajan and P. Tadepalli. Two new frameworks for learning. In John Laird, editor, *Machine Learning Proceedings 1988*, pages 402–415. Morgan Kaufmann, 1988.

[39] Chirag Pabbaraju. Multiclass learnability does not imply sample compression. In *International Conference on Algorithmic Learning Theory*, pages 930–944. PMLR, 2024.

[40] María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *J. Mach. Learn. Res.*, 22:227:1–227:40, 2021.

[41] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5142–5151. PMLR, 2019.

[42] Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, 8:1369–1392, 2007.

[43] Benjamin Rubinstein, Peter Bartlett, and J. Rubinstein. Shifting, one-inclusion mistake bounds and tight multiclass expected risk bounds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[44] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[45] Ingo Steinwart and Clint Scovel. Fast rates for support vector machines. In Peter Auer and Ron Meir, editors, *Learning Theory, 18th Annual Conference*

on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings, volume 3559 of *Lecture Notes in Computer Science*, pages 279–294. Springer, 2005.

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

[47] Ruth Urner, Shai Shalev-Shwartz, and Shai Ben-David. Access to unlabeled data can speed up prediction time. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML*, pages 641–648. Omnipress, 2011.

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

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

[50] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021.