Comparative Learning: A Sample Complexity Theory for Two Hypothesis Classes

Lunjia Hu
Computer Science Department, Stanford University, CA, USA

Charlotte Peale
Computer Science Department, Stanford University, CA, USA

Abstract

In many learning theory problems, a central role is played by a hypothesis class: we might assume that the data is labeled according to a hypothesis in the class (usually referred to as the *realizable setting*), or we might evaluate the learned model by comparing it with the best hypothesis in the class (the *agnostic setting*). Taking a step beyond these classic setups that involve only a single hypothesis class, we study a variety of problems that involve two hypothesis classes simultaneously.

We introduce *comparative learning* as a combination of the realizable and agnostic settings in PAC learning: given *two* binary hypothesis classes S and B, we assume that the data is labeled according to a hypothesis in the *source class* S and require the learned model to achieve an accuracy comparable to the best hypothesis in the *benchmark class* B. Even when both S and B have infinite VC dimensions, comparative learning can still have a small sample complexity. We show that the sample complexity of comparative learning is characterized by the *mutual VC dimension* $\mathsf{VC}(S, B)$ which we define to be the maximum size of a subset shattered by both S and B. We also show a similar result in the online setting, where we give a regret characterization in terms of the analogous *mutual Littlestone dimension* $\mathsf{Ldim}(S, B)$. These results also hold for partial hypotheses.

We additionally show that the insights necessary to characterize the sample complexity of comparative learning can be applied to other tasks involving two hypothesis classes. In particular, we characterize the sample complexity of realizable *multiaccuracy* and *multicalibration* using the *mutual fat-shattering dimension*, an analogue of the mutual VC dimension for real-valued hypotheses. This not only solves an open problem proposed by Hu, Peale, Reingold (2022), but also leads to independently interesting results extending classic ones about regression, boosting, and covering number to our two-hypothesis-class setting.

2012 ACM Subject Classification Theory of computation → Machine learning theory; Theory of computation → Sample complexity and generalization bounds; Computing methodologies → Learning settings

Keywords and phrases Comparative learning, mutual VC dimension, realizable multiaccuracy and multicalibration, sample complexity

Digital Object Identifier 10.4230/LIPIcs.ITCS.2023.72

Related Version *Full Version*: `https://arxiv.org/abs/2211.09101` [47]

Funding *Lunjia Hu*: Supported by Moses Charikar's and Omer Reingold's Simons Investigators awards and Omer Reingold's NSF Award IIS-1908774.
Charlotte Peale: Supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.

1 Introduction

The seminal theoretical framework of *PAC learning* [75] provides a formalization of machine learning that allows for rigorous theoretical analysis. In PAC learning, a learning algorithm (learner) receives individual/label pairs $(x, y) \in X \times \{-1, 1\}$ as input data, drawn i.i.d. from an unknown distribution μ. The learner's goal is to output a model $f : X \to \{-1, 1\}$ that assigns each individual in X a binary label. The performance of the model f is measured by its classification error,

$\mathsf{error}(f) := \Pr_{(x,y)\sim\mu}[f(x) \neq y]$.

Because the classification error is evaluated over the entire distribution μ, a good learner must go beyond simply memorizing the individuals and labels seen in the input data and be able to correctly predict the labels of unseen individuals as well. This can be a difficult task, and to make it possible to achieve a meaningfully small error given a limited amount of input data, additional assumptions or relaxations are needed. This leads to two standard settings of PAC learning: *realizable* and *agnostic* learning. In realizable learning, we assume that all data points are labeled according to an unknown hypothesis $h : X \to \{-1, 1\}$, i.e., $y = h(x)$ for every data point (x, y) drawn from μ, and we assume that h belongs to a hypothesis class H known to the learner. Under this assumption, realizable learning requires the output model f to achieve a low classification error ($\mathsf{error}(f) \leq \varepsilon$) with large probability. In agnostic learning there is also a hypothesis class H known to the learner, but it does not impose any assumption on the data. Instead, we aim for a relaxed goal specified by H: achieving $\mathsf{error}(f) \leq \inf_{h\in H} \mathsf{error}(h) + \varepsilon$ with large probability.

At a high level, both realizable and agnostic learning involve the introduction of a hypothesis class H, but H plays a very different role in each setting. In realizable learning, H constrains the potential *source hypotheses* that might determine the ground-truth labeling of the data. In contrast, agnostic learning places no assumptions on the ground-truth labeling, but instead uses H as a *benchmark class* and only requires the learner to perform well compared to the best benchmark hypothesis in H. Thus, realizable and agnostic learning highlight two natural ways to simplify a learning task: constrain the potential hypotheses that the ground-truth labeling is generated from, or constrain the set of hypotheses that the output model is compared against.

Our work originates from the observation that these two ways of simplifying a learning task need not be mutually exclusive. Instead, they can be treated as two "knobs" that can be simultaneously adjusted to create new hybrid learning tasks. For any *two* hypothesis classes S and B, we can define a learning task by letting them play the two roles of H in the realizable and agnostic settings, respectively. That is, we assume that there exists a *source hypothesis* $s \in S$ such that $y = s(x)$ for every data point (x, y) drawn from μ, and we aim for achieving, with large probability, an error comparable to the best *benchmark hypothesis* $b \in B$: $\mathsf{error}(f) \leq \min_{b\in B} \mathsf{error}(b) + \varepsilon$. We term this hybrid notion *comparative learning*.

Our research reveals that the notion of comparative learning is far more insightful than just a thought experiment: it serves as an unexplored playground for the study of *sample complexity*, and the new connections we establish to characterize the sample complexity of comparative learning can be fruitfully applied to open questions about existing learning tasks. Here, "sample complexity" refers to one of the key characteristics of every learning task: the minimum number of data points needed by a learner to solve the task. VC theory provides a thorough understanding of the sample complexity of classic PAC learning in both the realizable and agnostic settings: in both cases it is characterized by the *VC dimension* of the hypothesis class H, defined as the maximum size of a subset of X on which all possible labelings of the individuals can be realized by some hypothesis in H (we say a set is *shattered* by H when this condition holds; see Section 2 for the exact definition) [76, 17, 63]. Since then, understanding the sample complexity of a wide variety of new and existing learning tasks has remained an exciting area of research. These tasks include online learning [64, 13, 1, 26], reliable and useful learning [69, 60, 59, 61], statistical query learning [53, 15], learning real-valued hypotheses [56, 2, 11], multiclass learning [12, 19], learning partial hypotheses [66, 3],

active learning [7, 8, 52, 43, 44, 42], property testing [30, 55, 14], differentially private learning [4, 20, 27, 73, 50, 31], bounded-memory learning [32], and online learning in the smoothed analysis model [38, 39]. A commonality of these learning tasks is that each of them only explicitly involves a single hypothesis class, and thus the sample complexity is studied in terms of complexity measures of single hypothesis classes, such as the VC dimension, the Littlestone dimension, the statistical query dimension, the fat-shattering dimension, and the DS dimension. To tightly characterize the sample complexity of comparative learning where a *pair* of hypothesis classes S and B are involved, it is not sufficient to apply existing complexity measures to S and B separately (see Section 1.1 for a more detailed discussion). Instead, we must create new notions that measure the complexity of the *interaction* between the two classes. We show that the correct way to measure the complexity of this interaction in comparative learning is to look at the subsets of X that S and B *both* shatter, and we define the *mutual VC dimension*, $\mathsf{VC}(S, B)$, to be the maximum size of such subsets. We show that the mutual VC dimension gives both upper and lower bounds on the sample complexity of comparative learning. Similarly, in an online analogue of comparative learning, we define the *mutual Littlestone dimension* and prove upper and lower *regret* bounds.

Our sample complexity characterization for comparative learning turns out to be a powerful tool for studying the sample complexity of other tasks involving two hypothesis classes. In fact, our interest in comparative learning is derived in part from open questions related to the sample complexity of realizable *multiaccuracy* (MA) and *multicalibration* (MC) [41, 57, 48]. In these tasks, the hypothesis class H plays the same role as in realizable learning, while the classification error $\mathsf{error}(f)$ is replaced with an alternative error measure $\mathsf{MA\text{-}error}_D(f)$ or $\mathsf{MC\text{-}error}_D(f)$ specified by an additional hypothesis class D that is sometimes called the *distinguisher class*.[1] For example, the multiaccuracy error $\mathsf{MA\text{-}error}_D(f)$ is defined as follows:

$$\mathsf{MA\text{-}error}_D(f) := \sup_{d \in D} |\mathbb{E}_{(x,y)\sim\mu}[(f(x) - y)d(x)]|,$$

where the supremum is over all the *distingushers* $d : X \to [-1, 1]$ in the distinguisher class D. As demonstrated by Hu, Peale and Reingold [48], the freedom in choosing the class D allows the error to adapt to different goals that may arise in practice.

The introduction of the distinguisher class D makes sample complexity characterization challenging because the characterization needs to depend on both the class H in realizable learning and the additional distinguisher class D. Hu, Peale and Reingold [48] give a sample complexity characterization for realizable multiaccuracy using a particular *metric entropy* defined for every pair (H, D) (see Section 1.2 for more details), but their characterization is in the *distribution-specific* setting where the marginal distribution $\mu|_X$ of x in a pair (x, y) generated from the data distribution μ is fixed and known to the learner. In contrast, the VC dimension characterization for PAC learning is in the *distribution-free* setting where the learner has no explicit knowledge about $\mu|_X$ and must perform well for every $\mu|_X$. The sample complexity characterization for realizable multiaccuracy in the distribution-free setting is left as an open question by Hu et al. [48].

[1] The name "distinguisher class" comes from the observation that the *no-access outcome indistinguishability* task studied in [48] can be equivalently framed as multiaccuracy (see Section 2.1.2 in [48]). In addition to the difference in the error from realizable learning, realizable multiaccuracy and multicalibration also allow the hypothesis class H and the model f to be real-valued (see Section 1.2 and Section 5). It is also possible to replace the error in *agnostic* learning with MA-error and MC-error to get agnostic multiaccuracy and agnostic multicalibration, but Hu et al. [48] show that the sample complexity of agnostic multiaccuracy exhibits a non-monotone dependence on the complexity of the distinguisher class D. We focus on defining multiaccuracy and multicalibration in the realizable setting throughout the paper.

In this work, we answer this open question by characterizing the sample complexity of realizable multiaccuracy and multicalibration in the distribution-free setting using the *mutual fat-shattering dimension*, which we define similarly to the mutual VC dimension but for two *real-valued* hypothesis classes. Our results on comparative learning turn out to be especially useful for obtaining this characterization because there is an intimate relationship between achieving the comparative learning goal $\mathsf{error}(f) \leq \min_{b \in B} \mathsf{error}(b) + \varepsilon$ and achieving a low multiaccuracy (or multicalibration) error: $\mathsf{MA\text{-}error}_B(f) \leq \varepsilon$. Here, the benchmark class B in comparative learning plays the role of the distinguisher class D in multiaccuracy and multicalibration. This relationship has been observed in [41] and [34] in a single-hypothesis-class setting, i.e., without the assumption that the labels from the data distribution are generated according to a hypothesis in a pre-specified source class. We generalize this relationship to our two-hypothesis-class setting by showing a reduction from realizable multiaccuracy and multicalibration to comparative learning while preserving the interaction between the source and distinguisher/benchmark classes. This reduction leads to a number of new learning tasks that also involve a pair of hypothesis classes. Specifically, the reduction is accomplished via an intermediate task which we call *correlation maximization*, and we show that with some adaptation the reduction also allows us to efficiently *boost* a *weak* comparative learner to a strong one. Once we achieve multiaccuracy and multicalibration, we apply the *omnipredictor* result of Gopalan et al. [34] to solve *comparative regression*, an analogue of comparative learning but with real-valued hypotheses and general convex and Lipschitz loss functions. We believe that there is a rich collection of learning tasks where two or more hypothesis classes may interact in interesting ways, and our work is just a small step towards a better understanding of a tiny fraction of these tasks.

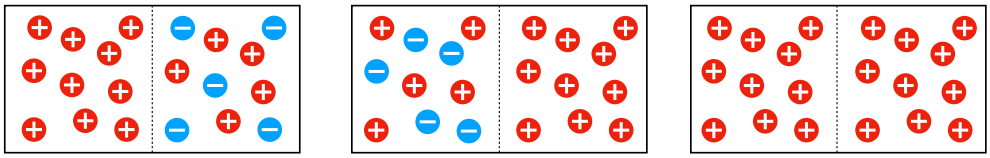
1.1 Sample Complexity of Comparative Learning

As mentioned earlier, VC theory has provided a thorough understanding of the sample complexities of both realizable and agnostic learning.

For any binary hypothesis class H, VC theory characterizes the sample complexity of realizable and agnostic learning using the VC dimension $\mathsf{VC}(H)$ of the hypothesis class H, a combinatorial quantity with a simple definition: the maximum size of a subset of X *shattered* by H (see Section 2 for exact definition) [76, 17, 63]. Moreover, the optimal sample complexity in both the realizable and agnostic settings can be achieved by a simple algorithm: the empirical risk minimization algorithm (ERM), which outputs the hypothesis in H with the minimum empirical error on the input data points.

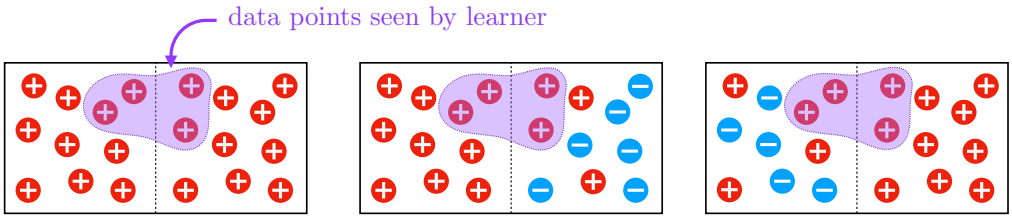
Because our notion of comparative learning combines these two settings, it would seem natural to use techniques from VC theory to understand its sample complexity as well. Compared to realizable learning for S, comparative learning for (S, B) has a relaxed goal (specified by the benchmark class B), and thus any learner solving realizable learning for S also solves comparative learning for (S, B). This gives us a sample complexity upper bound in terms of $\mathsf{VC}(S)$ for comparative learning. Similarly, any learner solving agnostic learning for B also solves comparative learning for (S, B) because comparative learning only makes additional assumptions on data (specified by the source class S), so we get another sample complexity upper bound in terms of $\mathsf{VC}(B)$.

However, perhaps surprisingly, these sample complexity upper bounds provided by the classic VC theory are not optimal. Even when $\mathsf{VC}(S)$ and $\mathsf{VC}(B)$ are both infinite, comparative learning may still have a finite sample complexity. Imagine that the domain X of individuals is partitioned into two large subsets X_1 and X_2. Suppose the source class S consists of all binary hypotheses $s : X \to \{-1, 1\}$ satisfying $s(x) = 1$ for every $x \in X_1$, and the benchmark



$s \in S$ $b \in B$ f solving CompL

Figure 1 An example where comparative learning requires no data points when $\mathsf{VC}(S)$ and $\mathsf{VC}(B)$ are both infinite. The left two images show examples of hypotheses in S and B, both of which are very complex, but on disjoint portions of the domain. In this case, a learner that always outputs the model f in the rightmost image solves comparative learning because f always achieves smaller or equal error compared to any benchmark hypothesis $b \in B$ when the ground-truth labelling is generated by a source hypothesis $s \in S$. See in-text description for more details.



$s \in S$ ERM for S ERM for B

Figure 2 Empirical risk minimization (ERM) may fail to give us optimal sample complexity in the same setting as Figure 1, where S and B are both very complex, but on disjoint domains. When the source hypothesis $s \in S$ is the constant function shown in the left image, the right two images show examples of output models of ERM when run on S and B. Neither model is guaranteed to achieve the low error required by comparative learning. See in-text description for more details.

class B consists of all binary hypotheses $b : X \to \{-1, 1\}$ satisfying $b(x) = 1$ for every $x \in X_2$. Both $\mathsf{VC}(S)$ and $\mathsf{VC}(B)$ can be large and even infinite, but comparative learning in this case requires no data points: the learner can simply output the model f that maps every $x \in X$ to 1 because no benchmark hypothesis in B can achieve a smaller error than f when the data points $(x, y) \sim \mu$ satisfy $y = s(x)$ for a source hypothesis $s \in S$ (see Figure 1). Beyond demonstrating that comparative learning may require far fewer samples than what our initial naïve upper bound might suggest, this example also shows that the standard empirical risk minimization (ERM) algorithm used for PAC learning does not give us the optimal sample complexity for comparative learning. Assume that the source hypothesis $s \in S$ maps every $x \in X$ to 1 and μ is the uniform distribution over $X \times \{1\}$. In this case $\min_{b \in B} \mathsf{error}(b) = 0$ and thus comparative learning requires a low classification error $\mathsf{error}(f) \leq \varepsilon$ with large probability. We have shown that this requirement can be achieved without any input data points, but the ERM algorithm cannot achieve this requirement in general unless there are many input data points: there can be many hypotheses in S and B that achieve zero empirical error on the input data points, but when the data points are few, most of such hypotheses do not achieve low classification error over the entire distribution μ (see Figure 2).

The example above shows that the VC dimensions $\mathsf{VC}(S)$ and $\mathsf{VC}(B)$ alone are not informative enough to characterize the sample complexity of comparative learning. These VC dimensions only tell us the complexity of S and B *separately*, but we also need to know the complexity of their *interaction*. We measure the complexity of this interaction by defining

the *mutual VC dimension* $\mathsf{VC}(S, B)$ to be the maximum size of a subset of X shattered by *both* S and B, and we give a tight characterization for the sample complexity of comparative learning in terms of $\mathsf{VC}(S, B)$.

As discussed earlier, new ideas are needed to prove this sample complexity characterization. In particular, we need to design a learner that is different from the ERM algorithm. Our technique is based on an interesting connection to *learning partial binary hypotheses*, a learning task considered first in [9, 66] and studied more systematically in a recent work by Alon, Hanneke, Holzman and Moran [3]. A partial binary hypothesis is a function $h : X \to \{-1, 1, *\}$ that may assign some individuals $x \in X$ the *undefined label* $h(x) = *$. The notion of partial hypotheses is motivated in previous work either as an intermediate step towards understanding *real-valued* hypotheses or as a way to describe data-dependent assumptions that could not be captured by the standard PAC learning model. In this work, we show that partial hypotheses have yet another application and can be used to express the interaction between a source hypothesis $s \in S$ and a benchmark hypothesis $b \in B$ in comparative learning: we construct an *agreement hypothesis* $\mathbf{a}_{s,b}$ which is a partial hypothesis assigning the undefined label $*$ to an individual x whenever $s(x)$ is different from $b(x)$, and giving the same label to x as s and b if $s(x)$ equals $b(x)$. We show that comparative learning for (S, B) can be reduced to *agnostically* learning the class $\mathbf{A}_{S,B}$ which consists of all the partial hypotheses $\mathbf{a}_{s,b}$ for $s \in S$ and $b \in B$, and conversely, we show that *realizable* learning for $\mathbf{A}_{S,B}$ reduces to comparative learning for (S, B). Our sample complexity characterization for comparative learning then follows immediately from the results by Alon et al. [3] for learning partial hypotheses. Moreover, our characterization holds even when the source hypotheses and benchmark hypotheses themselves are partial. We also show that this connection between comparative learning and learning the agreement hypotheses $\mathbf{a}_{s,b}$ extends to the online setting, allowing us to show a *regret* characterization for *comparative online learning*.

Our definition of the mutual VC dimension is clearly symmetric: $\mathsf{VC}(S, B) = \mathsf{VC}(B, S)$, and thus our sample complexity characterization for comparative learning reveals an intriguing phenomenon which we call *sample complexity duality*: comparative learning for (S, B) and comparative learning for (B, S) always have similar sample complexities. In other words, swapping the roles of the source class and the benchmark class does not change the sample complexity by much. Previously in [48], Hu et al. showed that this phenomenon holds for realizable multiaccuracy in the distribution-specific setting, drawing an insightful connection to a long-standing open question in convex geometry: the *metric entropy duality conjecture* [68, 18, 6, 5, 67]. Our sample complexity characterizations imply that sample complexity duality also holds in the distribution-*free* setting for realizable multiaccuracy as well as multicalibration. We also show that sample complexity duality does *not* hold for many learning tasks that we consider, including distribution-specific comparative learning, distribution-specific realizable multicalibration, correlation maximization, and comparative regression. In Table 1 we list whether sample complexity duality holds in general for every two-class learning task we consider in this paper in both the distribution-specific and distribution-free settings.

1.2 Multiaccuracy and Multicalibration

A direct motivation of our work is a recent paper by Hu, Peale, and Reingold [48] that studies a learning task called *multiaccuracy*, which was introduced in [41] and [57] originally as a notion of multi-group fairness. In multiaccuracy, the learned model (presumably making predictions about people) is required to be accurate in expectation when conditioned on

Table 1 Duality (yes) VS non-duality (no). *The sample complexity duality result in [48] for distribution-specific multiaccuracy assumes that all hypotheses are total.

	Distribution-specific	Distribution-free
Comparative learning	no	yes
Correlation maximization	no	no
Realizable multiaccuracy	yes* [48]	yes
Realizable multicalibration	no	yes
Comparative regression	no	no

each sub-community in a rich class (possibly defined based on demographic groups and their intersections). This ensures that the predictions made by the model are not systematically biased in any of the sub-communities.

Taking a broad perspective beyond fairness, Hu et al. [48] view multiaccuracy as providing a general, meaningful, and flexible performance measure for prediction models, and study PAC learning with the usual classification error replaced by this new performance measure from multiaccuracy. To be specific, let us consider a *real-valued* source hypothesis class S consisting of source hypotheses $s : X \to [-1, 1]$. We use S to replace the binary hypothesis class H in realizable learning and assume that every input data point $(x, y) \in X \times [-1, 1]$ is generated i.i.d. from a distribution μ satisfying $\mathbb{E}_{(x,y)\sim\mu}[y|x] = s(x)$ for an unknown $s \in S$. Suppose a learner which tries to learn s given the input data points produces an output model $f : X \to [-1, 1]$. This is a more general setting than binary classification because we allow $f(x)$ and $s(x)$ to take any value in the interval $[-1, 1]$, and accordingly, let us use the ℓ_1 error $\ell_1\text{-error}(f) := \mathbb{E}_{x\sim\mu|_X}[|f(x) - s(x)|]$ as a generalization of the classification error (as in, e.g., [11]). The *multiaccuracy error* of f is defined to be

$$\text{MA-error}_{\mu,B}(f) := \sup_{b\in B}|\mathbb{E}_{(x,y)\sim\mu}[(f(x) - y)b(x)]|, \tag{1}$$

where B is a *distinguisher class* consisting of *distinguishers* $b : X \to [-1, 1]$. Here we use B (rather than D) to denote the distinguisher class because a key idea we use in our work is to relate the distinguisher class to the benchmark class in comparative learning. Due to our assumption $\mathbb{E}_{(x,y)\sim\mu}[y|x] = s(x)$, the multiaccuracy error can be written equivalently as

$$\text{MA-error}_{\mu,B}(f) = \sup_{b\in B}|\mathbb{E}_{x\sim\mu|_X}[(f(x) - s(x))b(x)]|.$$

The multiaccuracy error is a generalization and relaxation of the ℓ_1 error in that if we choose the distinguisher class B to contain all distinguishers $b : X \to [-1, 1]$, then the two errors are equal: $\text{MA-error}_B(f) = \ell_1\text{-error}(f)$. The multiaccuracy error can become a more suitable performance measure than the ℓ_1 error if we customize B to reflect the goal we want to achieve: we can choose B to consist of indicator functions of demographic groups to achieve a fairness goal, and we can also choose B to specifically catch serious errors that we want to avoid (see [48] for more discussions).

The sample complexity of achieving a small $\ell_1\text{-error}(f)$ has been studied in [56], [2] and [11], who give a characterization in the distribution-free setting using the *fat-shattering dimension* of the source class S, defined as the maximum size of a subset of X *fat-shattered* by S (see Section 2.1 for a precise definition). Their results are further improved in [9, 10] and [62]. For a general distinguisher class B, the sample complexity of achieving a small $\text{MA-error}_B(f)$ depends on both classes S and B, and thus it becomes more challenging to characterize. In the distribution-specific setting where $\mu|_X$ is fixed and known to the

learner, Hu et al. [48] characterize the sample complexity of achieving $\mathsf{MA\text{-}error}_B(f) \le \varepsilon$ using $\log N_{\mu|_X,B}(S, \Theta(\varepsilon))$: the metric entropy of S w.r.t. the *dual Minkowski norm* defined based on B and $\mu|_X$. They also give an equivalent characterization using $\log N_{\mu|_X,S}(B, \Theta(\varepsilon))$ with the roles of S and B swapped. In the distribution-free setting where $\mu|_X$ is not known to the learner, they only give a sample complexity characterization when S contains all functions $s : X \to [-1, 1]$ using the fat-shattering dimension of B, and they leave the case of a general source class S as an open question. In this work, we answer this open question by giving a sample complexity characterization for arbitrary S and B in the distribution-free setting using the *mutual fat-shattering dimension* of (S, B), which we define to be the largest size of a subset of X fat-shattered by both S and B.

To prove this sample complexity characterization for distribution-free realizable multiaccuracy, we need a lower and an upper bound on the sample complexity. While we prove the lower bound using relatively standard techniques, the upper bound is much more challenging to prove. We prove the upper bound by reducing multiaccuracy for (S, B) to comparative learning for multiple pairs of *binary* hypothesis classes (S', B') with $\mathsf{VC}(S', B')$ bounded in terms of the mutual fat-shattering dimension of (S, B). We implement this reduction via an intermediate task which we call *correlation maximization*, and the main challenge here is that our learner L solving comparative learning for (S', B') is limited by the source class S' and can only handle data points realizable by a binary hypothesis in S'. Therefore, we must carefully transform the data points from multiaccuracy to ones acceptable by the comparative learner L. We implement this transformation by combining a rejection sampling technique with a *non-uniform covering* type of technique used in a recent work by Hopkins, Kane, Lovett, and Mahajan [45]. The difference between the real-valued class B and the binary class B' also poses a challenge, which we solve by taking multiple choices of B' and show that, roughly speaking, the convex hull of the chosen B' approximately includes B.

Our characterization using the mutual fat-shattering dimension holds not only for multiaccuracy, but also for a related task called *multicalibration* [41]. Here, we replace $\mathsf{MA\text{-}error}$ by the multicalibration error:

$$\begin{aligned}\mathsf{MC\text{-}error}_{\mu,B}(f) &:= \sup_{b\in B}\sum_{v\in V}|\mathbb{E}_{(x,y)\sim\mu}[(f(x)-y)b(x)\mathbf{1}(f(x)=v)]| \\ &= \sup_{b\in B}\sum_{v\in V}|\mathbb{E}_{x\sim\mu|_X}[(f(x)-s(x))b(x)\mathbf{1}(f(x)=v)]|, \qquad (2)\end{aligned}$$

where V is the range of f which we require to be countable. Multicalibration provides a strong guarantee: Gopalan et al. [34] show that it implies a notion called *omnipredictors*, allowing us to use our multicalibration results to show a sample complexity upper bound for *comparative regression*.

Our results imply that multiaccuracy and multicalibration share the same sample complexity characterization in the distribution-free realizable setting. In comparison, we show that this is *not* the case in the distribution-specific setting where there is a strong sample complexity separation between them (see details in full paper). This strong separation only appears in our two-hypothesis-class setting: if the source class S contains all hypotheses $s : X \to [-1, 1]$, then realizable multiaccuracy and multicalibration share the same sample complexity characterization (the metric entropy of B in the distribution-specific setting, and the fat-shattering dimension of B in the distribution-free setting).

1.3 Our Contributions

Below we summarize the main contributions of our paper.

1.3.1 Comparative Learning

We introduce the task of comparative learning (Definition 3) by combining realizable learning and agnostic learning. Specifically, we define comparative learning for any pair of hypothesis classes S and B each consisting of partial binary hypotheses $h : X \to \{-1, 1, *\}$ (denoted by $S, B \subseteq \{-1, 1, *\}^X$). As in realizable learning, we assume the learner receives data points $(x, y) \in X \times \{-1, 1\}$ generated i.i.d. from a distribution μ satisfying $\Pr_{(x,y)\sim\mu}[s(x) = y] = 1$ for a source hypothesis $s \in S$ (in particular, $\Pr_{(x,y)\sim\mu}[s(x) = *] = 0$). As in agnostic learning, we require the learner to output a model $f : X \to \{-1, 1\}$ satisfying

$$\Pr_{(x,y)\sim\mu}[f(x) \neq y] \leq \inf_{b\in B} \Pr_{(x,y)\sim\mu}[b(x) \neq y] + \varepsilon \tag{3}$$

with probability at least $1 - \delta$.

We characterize the sample complexity of comparative learning, which we denote by $\#\mathsf{CompL}(S, B, \varepsilon, \delta)$, using the mutual VC dimension $\mathsf{VC}(S, B)$ which we define as the largest size of a subset $X' \subseteq X$ shattered by both S and B (see Section 2.1 for the formal definition of shattering). In Theorem 4, assuming $\varepsilon, \delta \in (0, 1/4)$ and $\mathsf{VC}(S, B) \geq 2$, we show a sample complexity upper bound of

$$\#\mathsf{CompL}(S, B, \varepsilon, \delta) \leq O\left(\frac{\mathsf{VC}(S, B)}{\varepsilon^2} \log^2\left(\frac{\mathsf{VC}(S, B)}{\varepsilon}\right) + \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right), \tag{4}$$

and a lower bound of

$$\#\mathsf{CompL}(S, B, \varepsilon, \delta) \geq \Omega\left(\frac{\mathsf{VC}(S, B)}{\varepsilon} + \frac{1}{\varepsilon} \log\left(\frac{1}{\delta}\right)\right). \tag{5}$$

These bounds imply that the sample complexity of comparative learning is finite if and only if the mutual VC dimension $\mathsf{VC}(S, B)$ is finite. We show a similar sample complexity characterization for a learning task involving an arbitrary number of hypothesis classes in the full version of the paper.

1.3.2 Correlation Maximization

As an intermediate step towards characterizing the sample complexity of realizable multiaccuracy and multicalbration, we extend comparative learning to real-valued hypothesis classes by introducing correlation maximization (Definition 10). Here, the hypothesis classes S and B can contain any partial real-valued hypotheses $h : X \to [-1, 1] \cup \{*\}$ (denoted by $S, B \subseteq ([-1, 1] \cup \{*\})^X$), and every data point $(x, y) \in X \times [-1, 1]$ can have a label y taking any value in $[-1, 1]$. We assume that the data points are drawn i.i.d. from a distribution μ over $X \times [-1, 1]$ satisfying $\mathbb{E}_{(x,y)\sim\mu}[y|x] = s(x)$ for a source hypothesis $s \in S$, and we require the output model $f : X \to \{-1, 1\}$ to satisfy

$$\mathbb{E}_{(x,y)\sim\mu}[yf(x)] \geq \sup_{b\in B} \mathbb{E}_{(x,y)\sim\mu}[y \diamondsuit b(x)] - \varepsilon$$

with probability at least $1 - \delta$. Here, we define the *generalized product* $u_1 \diamondsuit u_2$ for $u_1 \in \mathbb{R}$ and $u_2 \in [-1, 1] \cup \{*\}$ such that $u_1 \diamondsuit u_2 = u_1 u_2$ if $u_2 \in [-1, 1]$, and $u_1 \diamondsuit u_2 = -|u_1|$ if $u_2 = *$. The requirement that the output model $f : X \to \{-1, 1\}$ produces binary values $f(x) \in \{-1, 1\}$ rather than real values $f(x) \in [-1, 1]$ is naturally satisfied by our learners for correlation maximization, but it is not essential to any of our results related to correlation maximization. In the special case where S and B are both binary, correlation maximization and comparative learning become equivalent for values of ε differing by exactly a factor of 2, i.e., the goal (3) of comparative learning can be equivalently written as

$$\mathbb{E}_{(x,y)\sim\mu}[yf(x)] \geq \sup_{b\in B} \Pr_{(x,y)\sim\mu}[y \diamondsuit b(x)] - 2\varepsilon.$$

We give an upper bound on the sample complexity of correlation maximization using the mutual fat-shattering dimension $\mathsf{fat}_\eta(S,B)$ which we define as the largest size of a subset $X' \subseteq X$ that is η-fat shattered by both S and B (see Section 2.1 for the definition of fat shattering). In Theorem 12, assuming $\varepsilon, \delta \in (0, 1/2)$, we show that the sample complexity of correlation maximization is upper bounded by

$$O\left(\frac{\mathsf{fat}_{\varepsilon/5}(S,B)}{\varepsilon^4}\log^2\left(\frac{\mathsf{fat}_{\varepsilon/5}(S,B)}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)+\frac{1}{\varepsilon^4}\log\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)\right).$$

We also consider a deterministic-label setting in the full version of the paper, which is a special case of correlation maximization where the data distribution μ satisfies $\Pr_{(x,y)\sim\mu}[y = s(x)] = 1$ for a source class $s \in S$. In this case, we prove the following improved sample complexity upper bound:

$$O\left(\frac{\mathsf{fat}_{\varepsilon/5}(S,B)}{\varepsilon^2}\log^2\left(\frac{\mathsf{fat}_{\varepsilon/5}(S,B)}{\varepsilon}\right)+\frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon\delta}\right)\right).$$

We also show that the mutual fat-shattering dimension does not in general give a lower bound for the sample complexity of correlation maximization. This is because sample complexity duality does not hold for correlation maximization (see full paper for examples where duality does not hold). In Theorem 12 we state a refined sample complexity upper bound for correlation maximization and we leave it as an open question to determine whether there is a matching lower bound.

1.3.3 Realizable Multiaccuracy and Multicalibration

We study multiaccuracy and multicalibration in the same setting as in [48] with a focus on the distribution-free realizable setting (Definitions 13 and 14). As in correlation maximization, the classes $S, B \subseteq ([-1,1] \cup \{*\})^X$ can contain any partial real-valued hypotheses $h : X \to [-1,1] \cup \{*\}$, and we assume that the data distribution μ satisfies $\mathbb{E}_{(x,y)\sim\mu}[y|x] = s(x)$ for a source hypothesis $s \in S$. The goal is to output a model $f : X \to [-1,1]$ such that $\mathsf{MA\text{-}error}_{\mu,B}(f) \le \varepsilon$ (in multiaccuracy) or $\mathsf{MC\text{-}error}_{\mu,B}(f) \le \varepsilon$ (in multicalibration) with probability at least $1-\delta$, where we generalize the definitions of $\mathsf{MA\text{-}error}_{\mu,B}$ and $\mathsf{MC\text{-}error}_{\mu,B}$ to partial hypothesis classes B. In Theorems 15 and 17, assuming $\varepsilon, \delta \in (0, 1/2)$, we show the following lower and upper bounds on the sample complexity of realizable multiaccuracy and multicalibration (denoted by $\#\mathsf{MA}(S,B,\varepsilon,\delta)$ and $\#\mathsf{MC}(S,B,\varepsilon,\delta)$, respectively):

$$\begin{aligned}
&\Omega\left(\mathsf{fat}_{\sqrt{3\varepsilon}}(S,B)\right)-1\\
&\le \#\mathsf{MA}(S,B,\varepsilon,\delta)\\
&\le \#\mathsf{MC}(S,B,\varepsilon,\delta)\\
&\le O\left(\frac{\mathsf{fat}_{\varepsilon/7}(S,B)}{\varepsilon^6}\log^2\left(\frac{\mathsf{fat}_{\varepsilon/7}(S,B)}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)+\frac{1}{\varepsilon^6}\log\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon\delta}\right)\right). \qquad (6)
\end{aligned}$$

This implies that the sample complexity of realizable multiaccuracy and multicalibration is finite for every $\varepsilon > 0$ if and only if $\mathsf{fat}_\eta(S,B)$ is finite for every $\eta > 0$. Also, the sample complexity is polynomial in $1/\varepsilon$ if and only if $\mathsf{fat}_\eta(S,B)$ is polynomial in $1/\eta$. This answers an open question in [48]. We also show an improved sample complexity upper bound in Theorem 16 for the special case where S is binary.

Our sample complexity upper and lower bounds stated in Theorems 15 and 17 are actually stronger, and they use a finer definition of the mutual fat-shattering dimension. Specifically, if we define $\mathsf{fat}_{\eta_1,\eta_2}(S,B)$ to be the largest size of a subset $X' \subseteq X$ that is η_1-fat shattered by S and η_2-fat shattered by B, then $\#\mathsf{MA}(S,B,\varepsilon,\delta)$ and $\#\mathsf{MC}(S,B,\varepsilon,\delta)$ are both finite

if $\mathsf{fat}_{\eta_1,\eta_2}(S,B)$ is finite for some η_1,η_2 satisfying $2\eta_1+4\eta_2<\varepsilon$, and $\#\mathsf{MA}(S,B,\varepsilon,\delta)$ and $\#\mathsf{MC}(S,B,\varepsilon,\delta)$ are both infinite if $\mathsf{fat}_{\eta_1,\eta_2}(S,B)$ is infinite for some η_1,η_2 satisfying $\eta_1\eta_2>2\varepsilon$. An open question is whether this gap can be closed to provide an exact characterization of the finiteness of $\#\mathsf{MA}(S,B,\varepsilon,\delta)$ and $\#\mathsf{MC}(S,B,\varepsilon,\delta)$ for every choice of (S,B,ε,δ).

1.3.4 Covering Number Bound

The sample complexity characterization for distribution-specific realizable multiaccuracy in [48] is in terms of a covering number defined for every pair of total hypothesis classes (S,B). A consequence of our sample complexity characterization for distribution-free realizable multiaccuracy and multicalibration is an upper bound on this covering number in terms of the mutual fat-shattering dimension of (S,B). This can be viewed as a generalization of a classic upper bound on the covering number of a binary hypothesis class H in terms of its VC dimension. Interestingly, our covering number upper bounds in the two-hypothesis-class setting hold despite the fact that a corresponding *uniform convergence bound* does not hold. See the full paper for more details.

1.3.5 Boosting

Analogous to the weak agnostic learning task considered in [51] and [25], we introduce *weak* comparative learning, where the goal (3) of comparative learning is relaxed to

$$\Pr_{(x,y)\sim\mu}[f(x)\neq y]\leq 1/2-\gamma,$$

under the additional assumption that

$$\inf_{b\in B}\Pr_{(x,y)\sim\mu}[b(x)\neq y]\leq 1/2-\alpha.$$

Here, $\alpha,\gamma\in(0,1/2)$ are parameters of the weak comparative learning task. Extending results in [25], we show an efficient boosting algorithm that solves (strong) comparative learning given oracle access to a learner solving weak comparative learning. This result also applies to correlation maximization for real-valued S and B in the deterministic-label setting. Due to space constraints, formal results and discussion of boosting appear only in the full version of this paper.

1.3.6 Comparative Regression

We define *comparative regression* by allowing the classes S and B in comparative learning to be real-valued and replacing the classification error $\Pr_{(x,y)\sim\mu}[f(x)\neq y]$ with the expected loss $\mathbb{E}_{(x,y)\sim\mu}[\ell(y,f(x))]$ for a general loss function ℓ. Specifically, we take a partial hypothesis class $S\subseteq([-1,1]\cup\{*\})^X$ as the source class, and for simplicity, we take a *total* hypothesis class $B\subseteq[-1,1]^X$ as the benchmark class. Given a loss function $\ell:[-1,1]\times[-1,1]\to\mathbb{R}$, we define the comparative regression task as follows. We assume that the data distribution μ over $X\times[-1,1]$ satisfies $\mathbb{E}_{(x,y)\sim\mu}[y|x]=s(x)$ for a source hypothesis $s\in S$, and the goal is to output a model $f:X\to[-1,1]$ such that the following holds with probability at least $1-\delta$:

$$\mathbb{E}_{(x,y)\sim\mu}[\ell(y,f(x))]\leq\inf_{b\in B}\mathbb{E}_{(x,y)\sim\mu}[\ell(y,b(x))]+\varepsilon.$$

As an application of our sample complexity characterization for realizable multicalibration and the omnipredictors result by Gopalan et al. [34], in Theorem 19, we give a sample complexity upper bound in terms of $\mathsf{fat}_\eta(S,B)$ for a special case of comparative regression

(Definition 18) where we assume that the label y in each data point is binary and the loss function ℓ is convex and Lipschitz. We leave the study of other interesting settings of comparative regression to future work.

1.3.7 Comparative Online Learning

We extend our notion of comparative learning to the online setting, where we assume that the data points (x, y) are given sequentially, and the learner is required to predict the label of the individual x in each data point before its true label y is shown. For binary hypothesis classes $S, B \subseteq \{-1, 1, *\}^X$, we introduce *comparative online learning* (Definition 23) where we assume that every data point (x, y) satisfies $y = s(x)$ for some source hypothesis $s \in S$ and we measure the performance of the learner by its *regret*, defined as the number of mistakes it makes minus the minimum number of mistakes made by a benchmark hypothesis $b \in B$. The goal of comparative online learning is to ensure that the expected regret does not exceed εn, where n is the total number of data points given to the learner. In Section 7, we introduce the *mutual Littlestone dimension* $m := \mathsf{Ldim}(S, B)$ and show that it characterizes the smallest ε achievable in comparative online learning, denoted by ε^* (Theorem 25):

$$\min\left\{\frac{1}{2}, \frac{m}{2n}\right\} \leq \varepsilon^* \leq O\left(\sqrt{\frac{m}{n}\log\frac{2m+n}{m}}\right).$$

To match the form of our other sample complexity bounds, we can fix $\varepsilon \in (0, 1/2)$ and bound the smallest n (denoted by n^*) for which we can ensure that the expected regret does not exceed εn:

$$\frac{m}{2\varepsilon} \leq n^* \leq O\left(\frac{m}{\varepsilon^2}\log\frac{1}{\varepsilon}\right).$$

1.3.8 Sample Complexity Duality

Learning tasks involving two hypothesis classes can potentially satisfy *sample complexity duality*, meaning that the sample complexity of the task changes minimally when we swap the roles of the two hypothesis classes. Hu et al. [48] show that sample complexity duality holds for distribution-specific realizable multiaccuracy, assuming that the source class S and the distinguisher class B are both total. Specifically, for $S, B \subseteq [-1, 1]^X, \varepsilon, \delta \in (0, 1/2)$ and a distribution μ_X over X, defining $m := \#\mathsf{MA}^{(\mu_X)}(S, B, \varepsilon/8, \delta)$ to be the sample complexity of realizable multiaccuracy with source class S and distinguisher class B in the distribution-specific setting where the data distribution μ satisfies $\mu|_X = \mu_X$, Hu et al. [48] show that

$$\#\mathsf{MA}^{(\mu_X)}(B, S, \varepsilon, \delta) \leq O\Big(\varepsilon^{-2}(m + \log(1/\delta))\Big).$$

Results in our work imply that sample complexity duality also holds for comparative learning and realizable multiaccuracy/multicalibration in the distribution-free setting. Specifically, if we define $m := \#\mathsf{CompL}(S, B, \varepsilon, \delta) + 1$ for $\varepsilon, \delta \in (0, 1/4)$ and any partial binary hypothesis classes $S, B \subseteq \{-1, 1, *\}^X$, then the following holds by (4) and (5):

$$\#\mathsf{CompL}(B, S, \varepsilon, \delta) \leq O\left(\frac{m\log^2 m}{\varepsilon} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right).$$

Similarly, if we define $m := \#\mathsf{MA}(S, B, \varepsilon^2/147, \delta) + 1$ for $\varepsilon, \delta \in (0, 1/2)$ and any partial real-valued hypothesis classes $S, B \subseteq ([-1, 1] \cup \{*\})^X$, then the following holds because of (6):

$$\#\mathsf{MA}(B,S,\varepsilon,\delta) \leq O\left(\frac{m}{\varepsilon^6}\log^2\left(\frac{m}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right)+\frac{1}{\varepsilon^6}\log\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)\right),$$

and the same inequality holds after replacing #MA with #MC. In the full version of this paper, we show that sample complexity duality does *not* hold for other learning tasks we consider in this paper, completing Table 1.

1.4 Related Work

Motivated by multi-group/sub-group fairness, many recent papers also study learning tasks involving two (or more) hypothesis classes. Multi-group agnostic learning, introduced in [16] and [71], involves a subgroup class G and a benchmark class B, where each subgroup $g \in G$ is a subset of the individual set X. The goal in multi-group agnostic learning is to learn a model such that the loss experienced by each subgroup $g \in G$ is not much larger than the minimum loss for that group achievable by a benchmark $b \in B$. Tosh and Hsu [74] show sample complexity upper bounds for multi-group agnostic learning in terms of the individual complexities of G and B. Thus, the upper bound does not depend on the interaction of the two classes. In contrast, the individual complexities of the source and benchmark (resp. distinguisher) classes are not sufficient for our sample complexity characterizations for comparative learning (resp. realizable multiaccuracy and multicalibration). In [29], the authors propose algorithms that can improve the loss on subgroups in the spirit of multi-group agnostic learning based on suggestions from auditors. A constrained loss minimization task introduced in [54] also involves a subgroup class and a benchmark class, but the subgroup class is used to impose (fairness) constraints on the learned model and the loss/error is evaluated over the entire population (not on each subgroup). The results in [54] assume that the complexities of both classes are bounded, whereas our sample complexity upper bounds (for different tasks) in this paper can be finite even when the complexities of both classes are infinite. Motivated by the goal of learning *proxies* for sensitive features that can be used to achieve fairness in downstream learning tasks, Diana et al. [22] consider a learning task involving *three* hypothesis classes: a source class, a *proxy class*, and a *downstream class*. Again, the sample complexity upper bounds in [22] are in terms of the individual complexities of these classes. Two other recent works [72, 70] show uniform convergence bounds for multicalibration in a two-hypothesis-class setting, but their bounds are yet again in terms of the individual complexities of the two classes and are finite only when the complexities of both classes are finite.

The notions of multiaccuracy and multicalibration can be viewed in the framework of outcome indistinguishability [23, 24]. Multicalibrated predictors have been applied to solve loss minimization for rich families of loss functions and/or under a variety of constraints, leading to the notion of omnipredictors [34, 46, 28]. Recently, Gopalan et al. [33] show that certain omnipredictors can be obtained from the weaker condition of *calibrated multiaccuracy*. Multicalibrated predictors can also be used for statistical inference on rich families of target distributions [58]. The notion of multicalibration has been extended to various settings in [49, 77, 36, 35].

Many of our results in this paper are based on sample complexity characterizations of learning partial hypotheses by Alon et al. [3]. Some of the key techniques used in [3] include the 1-inclusion graph algorithm [40], sample compression schemes [65], sample compression generalization bounds [37], and a reduction from agnostic learning to realizable learning [21].

▶ Note. Due to space constraints, this is an abridged version of our work. Readers are directed to the full version of the paper for detailed proofs of all results and some extensions of the main theorems.

2 Preliminaries

Throughout the paper, we use X to denote a non-empty set and we refer to the elements in X as *individuals*. We use the term *hypothesis* to refer to an arbitrary function $h : X \to \mathbb{R} \cup \{*\}$ assigning a label $h(x)$ to each individual $x \in X$.

The label $h(x)$ can be a real number or the *undefined label* $*$. We say a hypothesis h is *total* if $h(x) \neq *$ for every $x \in X$. When we do not require a hypothesis h to be total, we often say h is *partial* to emphasize that h may or may not be total. We say a hypothesis h is *binary* if $h(x) \in \{-1, 1, *\}$ for every $x \in X$, and we say h is *real-valued* if h may or may not be binary.

A *hypothesis class* H is a set consisting of hypotheses $h : X \to \mathbb{R} \cup \{*\}$, i.e., $H \subseteq (\mathbb{R} \cup \{*\})^X$ where we use B^A to denote the set of all functions $f : A \to B$ for any two sets A and B. A *total* hypothesis class is a set $H \subseteq \mathbb{R}^X$, and a *binary* hypothesis class is a set $H \subseteq \{-1, 1, *\}^X$. We say a hypothesis class H is *partial* if it may or may not be total, and we say H is *real-valued* if it may or may not be binary.

To avoid measurability issues, all probability distributions in this paper are assumed to be discrete, i.e., to have a countable support. For any distribution μ over $X \times \mathbb{R}$, we use $\mu|_X$ to denote the marginal distribution of x with (x, y) drawn from μ.

2.1 VC and Fat-shattering Dimensions for Partial Hypothesis Classes

The *VC dimension* was introduced in [76] for any total binary hypothesis class. As in [9] and [3], we consider a natural generalization of the VC dimension to all partial binary hypothesis classes $H \subseteq \{-1, 1, *\}^X$ as follows. We say a subset $X' \subseteq X$ is *shattered* by H if for every total binary function $\xi : X' \to \{-1, 1\}$ there exists $h \in H$ such that $h(x) = \xi(x)$ for every $x \in X'$. The VC dimension of H is defined to be

$$\mathsf{VC}(H) := \sup\{|X'| : X' \subseteq X, X' \text{ is shattered by } H\}.$$

An analogous notion of the VC dimension for real-valued hypothesis classes is the *fat-shattering dimension* introduced in [56]. The fat-shattering dimension was originally defined for total hypothesis classes, but it is natural to generalize it to all partial hypothesis classes in a similar fashion to the generalization of the VC dimension to partial binary classes: given a hypothesis class $H \subseteq (\mathbb{R} \cup \{*\})^X$ and a margin $\eta \geq 0$, we say a subset $X' \subseteq X$ is η-fat shattered by H w.r.t. a reference function $r : X' \to \mathbb{R}$ if for every total binary function $\xi : X' \to \{-1, 1\}$, there exists $h \in H$ such that for every $x \in X'$,

$$h(x) \neq * \text{ and } \xi(x)(h(x) - r(x)) > \eta.$$

We sometimes omit the mention of r and say X' is η-fat shattered by H if such a function r exists. The η-fat-shattering dimension of H is defined to be

$$\mathsf{fat}_\eta(H) := \sup\{|X'| : X' \subseteq X, X' \text{ is } \eta\text{-fat shattered by } H\}.$$

2.2 An Abstract Learning Task

We study a variety of learning tasks throughout the paper, and to help define each task concisely, we first define an abstract learning task Learn, of which each specific task we consider is a special case.

Let Z and F be two non-empty sets. In the abstract learning task Learn, an algorithm (*learner*) takes data points in Z as input and it outputs a model in F. We choose a *distribution class* P consisting of distributions μ over Z, and for each distribution $\mu \in P$, we choose a subset $F_\mu \subseteq F$ to be the *admissible set*. When the input data points are drawn i.i.d. from a distribution $\mu \in P$, we require the learner to output a model f in the admissible set F_μ with large probability. Formally, for $n \in \mathbb{Z}_{\geq 0}$ and $\delta \in \mathbb{R}_{\geq 0}$, we say a (possibly inefficient and randomized) learner L solves the learning task $\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$ if

1. L takes n data points $z_1, \ldots, z_n \in Z$ as input;
2. L outputs a model $f \in F$;
3. For any distribution $\mu \in P$, if the data points $z_1, \ldots, z_n$ are drawn i.i.d. from μ, then with probability at least $1 - \delta$, the output model f belongs to F_μ. The probability is over the randomness in the data points $z_1, \ldots, z_n$ and the internal randomness in learner L.

By a slight abuse of notation, we also use $\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$ to denote the set of all learners L that solve the learning task. Clearly, the learner set $\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$ is monotone w.r.t. n: for any nonnegative integers n and n' satisfying $n \leq n'$, we have

$$\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta) \subseteq \mathsf{Learn}_{n'}(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$$

because when given n' data points, a learner can choose to ignore $n' - n$ data points and only use the remaining n data points. We define the *sample complexity* $\#\mathsf{Learn}(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$ to be the smallest n for which there exists a learner in $\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$:

$$\#\mathsf{Learn}(Z, F, P, (F_\mu)_{\mu\in P}, \delta) := \inf\{n \in \mathbb{Z}_{\geq 0} : \mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta) \neq \emptyset\}. \tag{7}$$

2.3 Learning Partial Binary Hypotheses

We define realizable learning and agnostic learning for any partial binary hypothesis class H as special cases of the abstract learning task Learn in Section 2.2. These learning tasks have been studied in [9, 66, 3], and the results in these previous works are important for many of our results throughout the paper.

▶ **Definition 1** (Realizable learning (ReaL)). *Given a partial binary hypothesis class $H \subseteq \{-1, 1, *\}^X$, an error bound $\varepsilon \geq 0$, a failure probability bound $\delta \geq 0$, and a nonnegative integer n, we define $\mathsf{ReaL}_n(H, \varepsilon, \delta)$ to be $\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$ where $Z = X \times \{-1, 1\}$, $F = \{-1, 1\}^X$, P consists of all distributions μ over $X \times \{-1, 1\}$ satisfying $\Pr_{(x,y)\sim\mu}[h(x) = y] = 1$ for some $h \in H$, and F_μ consists of all models $f : X \to \{-1, 1\}$ satisfying*

$$\Pr_{(x,y)\sim\mu}[f(x) \neq y] \leq \varepsilon.$$

A key assumption in realizable learning is that any data distribution $\mu \in P$ is consistent with some hypothesis $h \in H$, i.e., $\Pr_{(x,y)\sim\mu}[h(x) = y] = 1$. In particular, this implies that $\Pr_{(x,y)\sim\mu}[h(x) = *] = 0$ because $y \in \{-1, 1\}$ cannot be the undefined label $*$. In agnostic learning, we remove such assumptions on the data distribution:

▶ **Definition 2** (Agnostic learning (AgnL)). *Given a partial binary hypothesis class $H \subseteq \{-1, 1, *\}^X$, an error bound $\varepsilon \geq 0$, a failure probability bound $\delta \geq 0$, and a nonnegative integer n, we define $\mathsf{AgnL}_n(H, \varepsilon, \delta)$ to be $\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu\in P}, \delta)$ where $Z = X \times \{-1, 1\}$, $F = \{-1, 1\}^X$, P consists of all distributions μ over $X \times \{-1, 1\}$, and F_μ consists of all models $f : X \to \{-1, 1\}$ satisfying*

$$\Pr_{(x,y)\sim\mu}[f(x) \neq y] \leq \inf_{h\in H} \Pr[h(x) \neq y] + \varepsilon. \tag{8}$$

There is no assumption on the data distributions $\mu \in P$ in agnostic learning: μ can be any distribution over $X \times \{-1,1\}$. The hypothesis class H is used to relax the objective in agnostic learning: instead of requiring the error $\Pr_{(x,y)\sim\mu}[f(x) \neq y]$ of the model f to be at most ε, we compare the error of f with the smallest error of a hypothesis $h \in H$ as in (8). Note that for $(x,y) \in X \times \{-1,1\}$ and $h : X \to \{-1,1,*\}$, we have $h(x) \neq y$ whenever $h(x) = *$.

For every learning task we define throughout the paper, we also implicitly define the corresponding sample complexity as in (7). For example, the sample complexity of realizable learning is

$$\#\mathsf{ReaL}(H,\varepsilon,\delta) := \inf\{n \in \mathbb{Z}_{\geq 0} : \mathsf{ReaL}_n(H,\varepsilon,\delta) \neq \emptyset\}.$$

We omit the sample complexity definitions for all other learning tasks.

2.4 Other Notation

For a statement P, we define its indicator $\mathbf{1}(P)$ such that $\mathbf{1}(P) = 1$ if P is true, and $\mathbf{1}(P) = 0$ if P is false. We define $\mathsf{sign} : \mathbb{R} \to \{-1,1\}$ such that for every $u \in \mathbb{R}$, $\mathsf{sign}(u) = 1$ if $u \geq 0$, and $\mathsf{sign}(u) = -1$ if $u < 0$. For functions $f_1 : U_1 \to U_2$ and $f_2 : U_2 \to U_3$, we use $f_2 \circ f_1 : U_1 \to U_3$ to denote their composition, i.e., $(f_2 \circ f_1)(u) = f_2(f_1(u))$ for every $u \in U_1$. We use $\log(\cdot)$ to denote the base-2 logarithm. For $u \in \mathbb{R}$, we define $\log_+(u) := \log(\max\{2,u\})$. For $u \in [-1,1]$, we use $\mathsf{Ber}^*(u)$ to denote the distribution over $\{-1,1\}$ with mean u (by analogy with the Bernoulli distribution over $\{0,1\}$).

3 Sample Complexity of Comparative Learning

Given a source class $S \subseteq \{-1,1,*\}^X$ and a benchmark class $B \subseteq \{-1,1,*\}^X$, we formally define the task of comparative learning below by combining the distribution assumption in realizable learning and the relaxed objective in agnostic learning:

▶ **Definition 3** (Comparative learning (CompL)). *Given two binary hypothesis classes $S, B \subseteq \{-1,1,*\}^X$, an error bound $\varepsilon \geq 0$, a failure probability bound $\delta \geq 0$, and a nonnegative integer n, we define $\mathsf{CompL}_n(S,B,\varepsilon,\delta)$ to be $\mathsf{Learn}_n(Z,F,P,(F_\mu)_{\mu\in P},\delta)$ where $Z = X \times \{-1,1\}$, $F = \{-1,1\}^X$, P consists of all distributions μ over $X \times \{-1,1\}$ such that $\Pr_{(x,y)\sim\mu}[s(x) = y] = 1$ for some $s \in S$, and F_μ consists of all models $f : X \to \{-1,1\}$ such that*

$$\Pr_{(x,y)\sim\mu}[f(x) \neq y] \leq \inf_{b\in B}\Pr_{(x,y)\sim\mu}[b(x) \neq y] + \varepsilon. \tag{9}$$

The data distribution $\mu \in P$ in comparative learning is constrained to be consistent with a *source hypothesis* $s \in S$, i.e., $\Pr_{(x,y)\sim\mu}[s(x) = y] = 1$, and the error of the output model f is compared with the smallest error of a *benchmark hypothesis* $b \in B$ as in (9).

In this section, we characterize the sample complexity of comparative learning for every source class $S \subseteq \{-1,1,*\}^X$ and every benchmark class $B \subseteq \{-1,1,*\}^X$ according to Theorem 4 below. Our characterization is based on the *mutual VC dimension* $\mathsf{VC}(S,B)$, which we define as follows:

$$\mathsf{VC}(S,B) := \{|X'| : X' \subseteq X, X' \text{ is shattered by both } S \text{ and } B\}. \tag{10}$$

▶ **Theorem 4.** *Let $S, B \subseteq \{-1, 1, *\}^X$ be binary hypothesis classes. For any $\varepsilon, \delta \in (0, 1/4)$, the sample complexity of comparative learning satisfies the following upper bound:*

$$\#\mathsf{CompL}(S, B, \varepsilon, \delta) = O\left(\frac{\mathsf{VC}(S,B)}{\varepsilon^2}\log_+^2\left(\frac{\mathsf{VC}(S,B)}{\varepsilon}\right) + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right). \tag{11}$$

When $\mathsf{VC}(S, B) \geq 2$, we have the following lower bound:

$$\#\mathsf{CompL}(S, B, \varepsilon, \delta) = \Omega\left(\frac{\mathsf{VC}(S,B)}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right). \tag{12}$$

Our proof of Theorem 4 is based on results by Alon et al. [3] that characterize the sample complexity of realizable and agnostic learning for a partial hypothesis class $H \subseteq \{-1, 1, *\}^X$:

▶ **Theorem 5** ([3])**.** *Let $H \subseteq \{-1, 1, *\}^X$ be a binary hypothesis class. For any $\varepsilon, \delta \in (0, 1/4)$, the sample complexity of agnostic learning satisfies*

$$\#\mathsf{AgnL}(H, \varepsilon, \delta) = O\left(\frac{\mathsf{VC}(H)}{\varepsilon^2}\log_+^2\left(\frac{\mathsf{VC}(H)}{\varepsilon}\right) + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right). \tag{13}$$

When $\mathsf{VC}(H) \geq 2$, the sample complexity of realizable learning satisfies

$$\#\mathsf{ReaL}(H, \varepsilon, \delta) = \Omega\left(\frac{\mathsf{VC}(H)}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right). \tag{14}$$

We prove the sample complexity upper bound (11) by reducing comparative learning for a pair of binary hypothesis classes (S, B) to agnostic learning for a single partial hypothesis class $\mathbf{A}_{S,B}$ we define below.

For every pair of hypotheses $s, b : X \to \{-1, 1, *\}$, we define an *agreement hypothesis* $\mathbf{a}_{s,b} : X \to \{-1, 1, *\}$ by

$$\mathbf{a}_{s,b}(x) = \begin{cases} 0, & \text{if } s(x) = b(x) = 0; \\ 1, & \text{if } s(x) = b(x) = 1; \\ *, & \text{otherwise.} \end{cases}$$

For every pair of hypothesis classes $S, B \subseteq \{-1, 1, *\}^X$, we define the *agreement hypothesis class* $\mathbf{A}_{S,B}$ to be $\{\mathbf{a}_{s,b} : s \in S, b \in B\} \subseteq \{-1, 1, *\}^X$.

The following claim follows immediately from the definition of $\mathbf{a}_{s,b}$:

▷ Claim 6. For every $(x, y) \in X \times \{-1, 1\}$ and every pair of hypotheses $s, b : X \to \{-1, 1, *\}$, we have $\mathbf{a}_{s,b}(x) = y$ if and only if $s(x) = b(x) = y$.

The following claim shows that the mutual VC dimension of (S, B) is equal to the VC dimension of $\mathbf{A}_{S,B}$:

▷ Claim 7. Let $S, B \subseteq \{-1, 1, *\}^X$ be binary hypothesis classes. Then $\mathsf{VC}(\mathbf{A}_{S,B}) = \mathsf{VC}(S, B)$.

Proof. A subset $X' \subseteq X$ is shattered by both S and B if and only if for every $\xi : X' \to \{-1, 1\}$, there exists $s \in S$ and $b \in B$ such that

$$s(x) = b(x) = \xi(x) \text{ for every } x \in X'. \tag{15}$$

Similarly, by the definition of $\mathbf{A}_{S,B}$, a subset $X' \subseteq X$ is shattered by $\mathbf{A}_{S,B}$ if and only if for every $\xi : X' \to \{-1, 1\}$, there exists $s \in S$ and $b \in B$ such that

$$\mathbf{a}_{s,b}(x) = \xi(x) \text{ for every } x \in X'. \tag{16}$$

By Claim 6, the conditions (15) and (16) are equivalent. ◁

We are now ready to state and prove the reduction that allows us to prove (11):

▶ **Lemma 8.** *Let $S, B \subseteq \{-1, 1, *\}^X$ be binary hypothesis classes. For any $\varepsilon, \delta \in \mathbb{R}_{\geq 0}$ and $n \in \mathbb{Z}_{\geq 0}$, we have $\mathsf{AgnL}_n(\mathbf{A}_{S,B}, \varepsilon, \delta) \subseteq \mathsf{CompL}_n(S, B, \varepsilon, \delta)$. In other words, any learner solving agnostic learning for $\mathbf{A}_{S,B}$ also solves comparative learning for (S, B) with the same parameters ε and δ.*

Proof. Let L be a learner in $\mathsf{AgnL}_n(\mathbf{A}_{S,B}, \varepsilon, \delta)$. For $s \in S$, let μ be a distribution over $X \times \{-1, 1\}$ satisfying $\Pr_{(x,y)\sim\mu}[s(x) = y] = 1$. By the guarantee of $L \in \mathsf{AgnL}_n(\mathbf{A}_{S,B}, \varepsilon, \delta)$, given n data points $(x_1, y_1), \ldots, (x_n, y_n)$ drawn i.i.d. from μ, with probability at least $1 - \delta$, L outputs a model f satisfying

$$\begin{aligned}
\Pr_{(x,y)\sim\mu}[f(x) \neq y] &\leq \inf_{h \in \mathbf{A}_{S,B}} \Pr_{(x,y)\sim\mu}[h(x) \neq y] + \varepsilon \\
&= \inf_{s' \in S, b \in B} \Pr_{(x,y)\sim\mu}[\mathbf{a}_{s',b}(x) \neq y] + \varepsilon && \text{(by definition of } \mathbf{A}_{S,B}) \\
&\leq \inf_{b \in B} \Pr_{(x,y)\sim\mu}[\mathbf{a}_{s,b}(x) \neq y] + \varepsilon \\
&= \inf_{b \in B} \Pr_{(x,y)\sim\mu}[s(x) \neq y \text{ or } b(x) \neq y] + \varepsilon && \text{(by Claim 6)} \\
&= \inf_{b \in B} \Pr_{(x,y)\sim\mu}[b(x) \neq y] + \varepsilon. && \text{(by } \Pr_{(x,y)\sim\mu}[s(x) = y] = 1)
\end{aligned}$$

This proves that $L \in \mathsf{CompL}_n(S, B, \varepsilon, \delta)$, as desired. ◀

Our upper bound (11) follows immediately from Claim 7, Lemma 8, and (13). We defer the detailed proof to the end of the section. To prove the lower bound (12), we reduce realizable learning for $\mathbf{A}_{S,B}$ to comparative learning for (S, B):

▶ **Lemma 9.** *Let $S, B \subseteq \{-1, 1, *\}^X$ be binary hypothesis classes. For any $\varepsilon, \delta \in \mathbb{R}_{\geq 0}$ and $n \in \mathbb{Z}_{\geq 0}$, we have $\mathsf{CompL}_n(S, B, \varepsilon, \delta) \subseteq \mathsf{RealL}_n(\mathbf{A}_{S,B}, \varepsilon, \delta)$. In other words, any learner solving comparative learning for (S, B) also solves realizable learning for $\mathbf{A}_{S,B}$ with the same parameters ε and δ.*

Proof. Let L be a learner in $\mathsf{CompL}_n(S, B, \varepsilon, \delta)$. Let μ be a distribution over $X \times \{-1, 1\}$ satisfying

$$\Pr_{(x,y)\sim\mu}[h(x) = y] = 1 \text{ for some } h \in \mathbf{A}_{S,B}. \tag{17}$$

By the definition of $\mathbf{A}_{S,B}$, our assumption (17) implies that $\Pr_{(x,y)\sim\mu}[\mathbf{a}_{s,b}(x) = y] = 1$ for some $s \in S$ and $b \in B$. By Claim 6, we have $\Pr_{(x,y)\sim\mu}[s(x) = y] = \Pr_{(x,y)\sim\mu}[b(x) = y] = 1$.

By the guarantee of $L \in \mathsf{CompL}_n(S, B, \varepsilon, \delta)$, given n data points $(x_1, y_1), \ldots, (x_n, y_n)$ drawn i.i.d. from μ, with probability at least $1 - \delta$, L outputs a model f satisfying

$$\Pr_{(x,y)\sim\mu}[f(x) \neq y] \leq \inf_{b' \in B} \Pr_{(x,y)\sim\mu}[b'(x) \neq y] + \varepsilon = \varepsilon,$$

where the last equation holds because $\Pr_{(x,y)\sim\mu}[b(x) = y] = 1$. The inequality above implies $L \in \mathsf{RealL}_n(\mathbf{A}_{S,B}, \varepsilon, \delta)$, as desired. ◀

Proof of Theorem 4. Define $m := \mathsf{VC}(S, B)$. By Claim 7, we have $m = \mathsf{VC}(\mathbf{A}_{S,B})$. Our upper bound (11) holds because

$$\begin{aligned}
\#\mathsf{CompL}(S, B, \varepsilon, \delta) &\leq \#\mathsf{AgnL}(\mathbf{A}_{S,B}, \varepsilon, \delta) && \text{(by Lemma 8)} \\
&\leq O\left(\frac{m}{\varepsilon^2} \log_+^2\left(\frac{m}{\varepsilon}\right) + \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right). && \text{(by (13))}
\end{aligned}$$

Our lower bound (12) holds because

$$\begin{aligned}\#\mathsf{CompL}(S,B,\varepsilon,\delta) &\geq \#\mathsf{ReaL}(\mathbf{A}_{S,B},\varepsilon,\delta) && \text{(by Lemma 9)}\\ &\geq \Omega\left(\frac{m}{\varepsilon}+\frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right). && \text{(by (14))}\end{aligned}$$

◀

4 Sample Complexity of Correlation Maximization

As we define in Section 3, the comparative learning task CompL requires the hypothesis classes S and B to be binary. Here we introduce a natural generalization of CompL to real-valued hypothesis classes $S, B \subseteq ([-1,1]\cup\{*\})^X$ which we call *correlation maximization*.

We first generalize the product u_1u_2 of two real numbers $u_1, u_2 \in \mathbb{R}$ to the case where u_2 may be the undefined label $*$. Specifically, for $u_1 \in \mathbb{R}$ and $u_2 \in [-1,1]\cup\{*\}$, we define their *generalized product* $u_1 \diamondsuit u_2$ to be

$$u_1 \diamondsuit u_2 := \begin{cases} u_1u_2, & \text{if } u_2 \in [-1,1],\\ -|u_1|, & \text{if } u_2 = *. \end{cases}$$

The idea behind the definition is that when $u_2 = *$, we treat u_2 as being an unknown number u' in $[-1,1]$ and define the product $u_1 \diamondsuit u_2$ to be the smallest possible value of u_1u', i.e., $u_1 \diamondsuit u_2 = \inf_{u'\in[-1,1]} u_1u' = -|u_1|$.

This generalized product allows us to rewrite the goal (9) of comparative learning. For any $y \in \{-1,1\}$ and $u \in \{-1,1,*\}$, it is easy to verify that

$$\mathbf{1}(y \neq u) = \frac{1}{2}(1 - y \diamondsuit u). \tag{18}$$

Therefore, the goal (9) of comparative learning can be equivalently written as

$$\mathbb{E}_{(x,y)\sim\mu}[yf(x)] \geq \sup_{b\in B}\mathbb{E}_{(x,y)\sim\mu}[y \diamondsuit b(x)] - 2\varepsilon.$$

This reformulation is meaningful even when we relax B to be a *real-valued* hypothesis class $B \subseteq ([-1,1]\cup\{*\})^X$. If we also relax the source class S, we obtain the definition of correlation maximization:

▶ **Definition 10** (Correlation maximization (CorM)). *Given two real-valued hypothesis classes $S, B \subseteq ([-1,1]\cup\{*\})^X$, an error bound $\varepsilon \geq 0$, a failure probability bound $\delta \geq 0$, and a nonnegative integer n, we define $\mathsf{CorM}_n(S,B,\varepsilon,\delta)$ to be $\mathsf{Learn}_n(Z,F,P,(F_\mu)_{\mu\in P},\delta)$ with Z, F, P, F_μ chosen as follows. We choose $Z = X\times[-1,1]$ and $F = \{-1,1\}^X$. The distribution class P consists of all distributions μ over $X\times[-1,1]$ satisfying the following property:*

$$\text{there exists } s\in S \text{ such that } \Pr_{x\sim\mu|_X}[s(x)\neq *] = 1 \text{ and } \mathbb{E}_{(x,y)\sim\mu}[y|x] = s(x). \tag{19}$$

The admissible set F_μ consists of all models $f : X \to \{-1,1\}$ satisfying

$$\mathbb{E}_{(x,y)\sim\mu}[yf(x)] \geq \sup_{b\in B}\mathbb{E}_{(x,y)\sim\mu}[y \diamondsuit b(x)] - \varepsilon.$$

The name "correlation maximization" comes from viewing $\mathbb{E}_{(x,y)\sim\mu}[yf(x)]$ as the (uncentered) correlation between random variables y and $f(x)$. In correlation maximization, any data distribution $\mu \in P$ needs to satisfy $\mathbb{E}_{(x,y)\sim\mu}[y|x] = s(x)$ for a source hypothesis $s \in S$. This restricts the conditional *expectation* of y given x, but we allow the conditional *distribution* of

y given x to be otherwise unrestricted. That is, when conditioned on $x \in X$ being fixed, the label $y \in [-1,1]$ could be deterministically equal to $s(x)$, but y could be also be random as long as it has conditional expectation $s(x)$.

In this section, we show a sample complexity upper bound for correlation maximization for any source class $S \subseteq ([-1,1] \cup \{*\})^X$ and benchmark class $B \subseteq ([-1,1] \cup \{*\})^X$. Since S and B may no longer be binary, we cannot apply the mutual VC dimension as a way to characterize their complexity. Instead, we turn to a classic generalization of the VC dimension for real-valued hypotheses, the fat-shattering dimension (see Section 2.1 for definition). Our upper bound is in terms of the *mutual fat-shattering dimension* defined as follows, which generalizes the mutual VC dimension to real-valued hypothesis classes.

Given a pair of real-valued hypothesis classes $S, B \subseteq ([-1,1] \cup \{*\})^X$ and a margin $\eta \in \mathbb{R}_{\geq 0}$, we define the *mutual fat-shattering dimension* $\mathsf{fat}_\eta(S,B)$ as follows:

$$\mathsf{fat}_\eta(S,B) := \sup\{|X'| : X' \subseteq X, X' \text{ is } \eta\text{-fat shattered by both } S \text{ and } B\}. \tag{20}$$

In other words, $\mathsf{fat}_\eta(S,B)$ is the largest size of a subset $X' \subseteq X$ such that X' is η-fat shattered by S w.r.t. a function $r_1 : X' \to \mathbb{R}$ and X' is η-fat shattered by B w.r.t. a function $r_2 : X' \to \mathbb{R}$ (recall the definition of fat shattering in Section 2.1).

Another equivalent way to define the mutual fat-shattering dimension for real-valued hypothesis classes $S, B \subseteq ([-1,1] \cup \{*\})^X$ is by transforming them into binary classes and using the mutual VC dimension after the transformation. These transformations are also crucial in our proof of the sample complexity upper bound for correlation maximization in this section. Given a real-valued hypothesis $h : X \to [-1,1] \cup \{*\}$, a reference function $r : X \to \mathbb{R}$, and a margin $\eta \in \mathbb{R}_{\geq 0}$, we define a binary hypothesis $h_\eta^{(r)} : X \to \{-1,1,*\}$ such that

$$h_\eta^{(r)}(x) = \begin{cases} 1, & \text{if } h(x) \neq * \text{ and } h(x) > r(x) + \eta; \\ -1, & \text{if } h(x) \neq * \text{ and } h(x) < r(x) - \eta; \\ *, & \text{otherwise.} \end{cases}$$

Given a real-valued hypothesis class $H \subseteq ([-1,1] \cup \{*\})^X$, we define the binary hypothesis class $H_\eta^{(r)} \subseteq \{-1,1,*\}^X$ as

$$H_\eta^{(r)} = \{h_\eta^{(r)} : h \in H\}.$$

We can now transform any real-valued hypothesis class $H \subseteq ([-1,1] \cup \{*\})^X$ into a binary hypothesis class $H_\eta^{(r)}$ for every choice of $\eta \in \mathbb{R}_{\geq 0}$ and $r : X \to \mathbb{R}$. This allows us to measure the complexity of a pair of real-valued hypothesis classes $S, B \subseteq ([-1,1] \cup \{*\})^X$ using the mutual VC dimensions $\mathsf{VC}(S_{\eta_1}^{(r_1)}, B_{\eta_2}^{(r_2)})$ of the binary hypothesis classes $S_{\eta_1}^{(r_1)}, B_{\eta_2}^{(r_2)}$ for various choices of η_1, η_2, r_1, r_2. The following claim shows that the mutual fat-shattering dimension $\mathsf{fat}_\eta(S,B)$ can be defined equivalently in this way.

▷ Claim 11. Let $S, B \subseteq ([-1,1] \cup \{*\})^X$ be real-valued hypothesis classes. For every $\eta \in \mathbb{R}_{\geq 0}$, $\mathsf{fat}_\eta(S,B) = \sup_{r_1,r_2} \mathsf{VC}(S_\eta^{(r_1)}, B_\eta^{(r_2)})$, where the supremum is over all function pairs $r_1, r_2 : X \to \mathbb{R}$.

The claim follows from the fact that a subset $X' \subseteq X$ is η-fat shattered by S if and only if X' is shattered by the binary hypothesis class $S_\eta^{(r)}$ for some $r : X \to \mathbb{R}$, and the same holds with S replaced by B.

Before we state our sample complexity upper bound for correlation maximization in Theorem 12, we make some additional definitions to simplify the statement. Let $h : X \to [-1,1] \cup \{*\}$ be a real-valued hypothesis and $H \subseteq ([-1,1] \cup \{*\})^X$ be a real-valued hypothesis class. For every real number $\theta \in \mathbb{R}$, we use $h_\eta^{(\theta)}$ and $H_\eta^{(\theta)}$ to denote $h_\eta^{(r)}$ and $H_\eta^{(r)}$ with the reference function $r : X \to \mathbb{R}$ being the constant function satisfying $r(x) = \theta$ for every $x \in X$.

▶ **Theorem 12.** *Let $S, B \subseteq ([-1,1] \cup \{*\})^X$ be real-valued hypothesis classes. For $\eta_1, \eta_2, \beta, \delta \in (0, 1/2)$, defining $m := \sup_{\theta \in \mathbb{R}} \mathsf{VC}(S_{\eta_1}^{(0)}, B_{\eta_2}^{(\theta)})$, we have*

$$\begin{aligned} &\#\mathsf{CorM}(S, B, \beta + 2\eta_1 + 2\eta_2, \delta) \\ &\leq O\left(\frac{m}{\beta^4} \log_+^2\left(\frac{m}{\beta}\right) \log\left(\frac{1}{\eta_1}\right) + \frac{1}{\beta^4} \log\left(\frac{1}{\eta_1}\right) \log\left(\frac{1}{\delta}\right) + \frac{1}{\beta^2} \log\left(\frac{1}{\eta_2}\right)\right). \end{aligned}$$

Moreover, for every $\varepsilon \in (0, 1/2)$, choosing $\beta = \eta_1 = \eta_2 = \varepsilon/5$, we have $m \leq \mathsf{fat}_{\varepsilon/5}(S, B)$ and

$$\#\mathsf{CorM}(S, B, \varepsilon, \delta) \leq O\left(\frac{m}{\varepsilon^4} \log_+^2\left(\frac{m}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon}\right) + \frac{1}{\varepsilon^4} \log\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\delta}\right)\right).$$

It remains an open question whether there is a sample complexity lower bound that matches Theorem 12, although our sample complexity characterization for multiaccuracy and multicalibration in Section 5 does not rely on such a lower bound. Qualitatively, Theorem 12 implies that $\#\mathsf{CorM}(S, B, \beta + 2\eta_1 + 2\eta_2, \delta)$ is finite if $\sup_{\theta \in \mathbb{R}} \mathsf{VC}(S_{\eta_1}^{(0)}, B_{\eta_2}^{(\theta)})$ is finite. We thus propose the following question about a qualitative lower bound: let $S, B \subseteq ([-1,1] \cup \{*\})^X$ be real-valued hypothesis classes. Suppose $\mathsf{VC}(S_{\eta_1}^{(0)}, B_{\eta_2}^{(\theta)})$ is infinite for some $\eta_1, \eta_2 > 0$ and $\theta \in \mathbb{R}$. Does this imply that $\#\mathsf{CorM}(S, B, \varepsilon, \delta)$ is infinite for some $\varepsilon, \delta > 0$?

5 Sample Complexity of Realizable Multiaccuracy and Multicalibration

In this section, we give a sample complexity characterization for realizable multiaccuracy and multicalibration in the distribution-free setting. These tasks have been studied in [48] for total hypothesis classes. Here we generalize their definitions to partial hypothesis classes.

Given a distribution μ over $X \times [-1,1]$ and a model $f : X \to [-1,1]$, we first generalize the definition of $\mathsf{MA\text{-}error}_{\mu,B}(f)$ and $\mathsf{MC\text{-}error}_{\mu,B}(f)$ in (1) and (2) to partial hypothesis classes $B \subseteq [-1,1] \cup \{*\}$. It is not enough to directly use the generalized product $\diamondsuit$ as in Section 4. For example, suppose we define $\mathsf{MA\text{-}error}_{\mu,B}(f)$ to be

$$\sup_{b \in B} |\mathbb{E}_{(x,y)\sim\mu}[(f(x) - y) \diamondsuit b(x)]|.$$

Then $\mathsf{MA\text{-}error}_{\mu,B}(f)$ is equal to the ℓ_1 error $\mathbb{E}_{(x,y)\sim\mu}[|f(x) - y|]$ even when B only contains a single hypothesis b which assigns every individual $x \in X$ the undefined label $b(x) = *$, making it challenging to achieve a low $\mathsf{MA\text{-}error}$ even when B has fat-shattering dimension zero. To avoid this issue, we note that for any $u \in \mathbb{R}$, the absolute value $|u|$ can be equivalently written as $\sup_{\sigma \in \{-1,1\}} u\sigma$, leading us to the following definitions:

$$\mathsf{MA\text{-}error}_{\mu,B}(f) := \sup_{b \in B} \sup_{\sigma \in \{-1,1\}} \mathbb{E}\Big[\Big((f(x) - y)\sigma\Big) \diamondsuit b(x)\Big], \quad \text{and} \tag{21}$$

$$\mathsf{MC\text{-}error}_{\mu,B}(f) := \sup_{b \in B} \sum_{v \in V} \sup_{\sigma \in \{-1,1\}} \mathbb{E}\Big[\Big((f(x) - y)\mathbf{1}(f(x) = v)\sigma\Big) \diamondsuit b(x)\Big]. \tag{22}$$

In the definition of $\mathsf{MC\text{-}error}$, we use V to denote the range of f which we assume to be countable. The supremum over $\sigma \in \{-1,1\}$ is inside the sum over $v \in V$, so σ is allowed to depend on v.

We can now define realizable multiaccuracy and multicalibration for partial hypothesis classes:

▶ **Definition 13** (Realizable Multiaccuracy (MA)). *Given two hypothesis classes $S, B \subseteq ([-1,1] \cup \{*\})^X$, an error bound $\varepsilon \geq 0$, a failure probability bound $\delta \geq 0$, and a nonnegative integer n, we define $\mathsf{MA}_n(S, B, \varepsilon, \delta)$ to be $\mathsf{Learn}_n(Z, F, P, (F_\mu)_{\mu \in P}, \delta)$ where $Z = X \times [-1,1]$, $F = [-1,1]^X$, P consists of all distributions μ over $X \times [-1,1]$ satisfying (19), and F_μ consists of all models $f : X \to [-1,1]$ such that*

$$\mathsf{MA\text{-}error}_{\mu,B}(f) \leq \varepsilon. \tag{23}$$

▶ **Definition 14** (Realizable Multicalibration (MC)). *We define $\mathsf{MC}_n(S, B, \varepsilon, \delta)$ in the same way as we define $\mathsf{MA}_n(S, B, \varepsilon, \delta)$ in Definition 13 except that we replace (23) with*

$$\mathsf{MC\text{-}error}_{\mu,B}(f) \leq \varepsilon.$$

We prove the following upper bound (Theorems 15 and 16) and lower bound (Theorem 17) on the sample complexity of realizable multiaccuracy and multicalibration in the full version of this paper.

▶ **Theorem 15.** *Let $S, B \subseteq ([-1,1] \cup \{*\})^X$ be real-valued hypothesis classes. For $\beta, \eta_1, \eta_2, \delta \in (0, 1/2)$, defining $m := \sup_{r:X\to\mathbb{R}} \sup_{\theta\in\mathbb{R}} \mathsf{VC}(S_{\eta_1}^{(r)}, B_{\eta_2}^{(\theta)})$, we have*

$$\begin{aligned}
&\#\mathsf{MA}(S, B, \beta + 2\eta_1 + 4\eta_2, \delta) \\
&\leq \#\mathsf{MC}(S, B, \beta + 2\eta_1 + 4\eta_2, \delta) \\
&\leq O\left(\frac{m}{\beta^6} \log_+^2\left(\frac{m}{\beta}\right) \log\left(\frac{1}{\eta_1}\right) + \frac{1}{\beta^6} \log\left(\frac{1}{\eta_1}\right) \log\left(\frac{1}{\beta\delta}\right) + \frac{1}{\beta^4} \log\left(\frac{1}{\eta_2}\right)\right).
\end{aligned}$$

For $\varepsilon \in (0, 1/2)$, choosing $\beta = \eta_1 = \eta_2 = \varepsilon/7$, we have $m \leq \mathsf{fat}_{\varepsilon/7}(S, B)$ and

$$\#\mathsf{MA}(S, B, \varepsilon, \delta) \leq \#\mathsf{MC}(S, B, \varepsilon, \delta) \leq O\left(\frac{m}{\varepsilon^6} \log_+^2\left(\frac{m}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon}\right) + \frac{1}{\varepsilon^6} \log\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon\delta}\right)\right).$$

▶ **Theorem 16.** *In the setting of Theorem 15, assume in addition that S is binary, i.e., $S \subseteq \{-1, 1, *\}^X$ and define $m := \sup_{\theta\in\mathbb{R}} \mathsf{VC}(S, B_{\eta_2}^{(\theta)})$. Then,*

$$\begin{aligned}
\#\mathsf{MA}(S, B, \beta + 4\eta_2, \delta) &\leq O\left(\frac{m}{\beta^4} \log_+^2\left(\frac{m}{\beta}\right) + \frac{1}{\beta^4} \log\left(\frac{1}{\eta_2\beta\delta}\right)\right), \\
\#\mathsf{MC}(S, B, \beta + 4\eta_2, \delta) &\leq O\left(\frac{m}{\beta^4} \log_+^2\left(\frac{m}{\beta}\right) + \frac{1}{\beta^4} \log\left(\frac{1}{\eta_2\beta\delta}\right) + \frac{1}{\beta^5}\right).
\end{aligned}$$

▶ **Theorem 17.** *Let $S, B \subseteq ([-1,1] \cup \{*\})^X$ be real-valued hypothesis classes. For $\eta_1, \eta_2 \in \mathbb{R}_{>0}$ and $\delta \in (0,1)$, defining $m := \sup_{r_1,r_2:X\to\mathbb{R}} \mathsf{VC}(S_{\eta_1}^{(r_1)}, B_{\eta_2}^{(r_2)})$, we have*

$$\#\mathsf{MC}(S, B, \eta_1\eta_2/3, \delta) \geq \#\mathsf{MA}(S, B, \eta_1\eta_2/3, \delta) \geq \log(1-\delta) + \Omega(m).$$

For any $\varepsilon \in (0, 1/2)$, choosing $\eta_1 = \eta_2 = \sqrt{3\varepsilon}$, we have $m = \mathsf{fat}_{\sqrt{3\varepsilon}}(S, B)$ and

$$\#\mathsf{MC}(S, B, \varepsilon, \delta) \geq \#\mathsf{MA}(S, B, \varepsilon, \delta) \geq \log(1-\delta) + \Omega(m).$$

Moreover, the constant 3 in the theorem can be replaced by any absolute constant $c > 2$.

6 Comparative Regression via Omnipredictors

We formally define the *comparative regression* task where the learning objective is to minimize a general loss function.

▶ **Definition 18** (Comparative Regression (CompR)). *Given a partial hypothesis class* $S \subseteq ([-1,1]\cup\{*\})^X$*, a total hypothesis class* $B \subseteq [-1,1]^X$*, a loss function* $\ell : \{-1,1\}\times[-1,1] \to \mathbb{R}$*, an error bound* $\varepsilon \geq 0$*, a failure probability bound* $\delta \geq 0$*, and a nonnegative integer* n*, we define* $\mathsf{CompR}_n(S,B,\ell,\varepsilon,\delta)$ *to be* $\mathsf{Learn}_n(Z,F,P,(F_\mu)_{\mu\in P},\delta)$ *with* Z,F,P,F_μ *chosen as follows. We choose* $Z = X\times\{-1,1\}$ *and* $F = [-1,1]^X$*. The distribution class* P *consists of all distributions* μ *over* $X\times\{-1,1\}$ *satisfying the following property:*

$$\text{there exists } s\in S \text{ such that } \Pr_{x\sim\mu|_X}[s(x)\neq *] = 1 \text{ and } \mathbb{E}_{(x,y)\sim\mu}[y|x] = s(x). \tag{24}$$

The admissible set F_μ *consists of all models* $f : X \to [-1,1]$ *such that*

$$\mathbb{E}_{(x,y)\sim\mu}[\ell(y,f(x))] \leq \inf_{b\in B}\mathbb{E}_{(x,y)\sim\mu}\mathbb{E}[\ell(y,b(x))] + \varepsilon. \tag{25}$$

In the definition above, we assume that the benchmark class B is total so that we do not need to define $\ell(y,b(x))$ when $b(x) = *$. We assume that the ranges of the model f and any benchmark $b \in B$ are bounded between -1 and 1, but any bounded range can be reduced to this setting by a scaling. We also assume that the label y in a data point $(x,y) \sim \mu$ is binary: $y \in \{-1,1\}$, and thus (24) implies that the conditional distribution of y given x is $\mathsf{Ber}^*(s(x))$. We focus on this binary-label setting because it is the main setting of the *omnipredictor* result in [34], which our results are based on. There are certainly other natural and interesting settings of comparative regression (e.g. the deterministic-label setting where $y \in [-1,1]$ and $\Pr_{(x,y)\sim\mu}[s(x) = y] = 1$ for some $s \in S$). We leave further study of these settings for future work.

We prove the following sample complexity upper bound for comparative regression in terms of the mutual fat-shattering dimension:

▶ **Theorem 19.** *Let* $S \subseteq ([-1,1]\cup\{*\})^X$ *be a partial hypothesis class and* $B \subseteq [-1,1]^X$ *be a total hypothesis class. Let* $\ell : \{-1,1\}\times[-1,1] \to \mathbb{R}$ *be a loss function such that* $\ell(y,\cdot)$ *is convex and* κ*-Lipschitz for any* $y \in \{-1,1\}$*. For* $\beta,\eta_1,\eta_2,\delta \in (0,1/2)$*, defining* $m := \sup_{r:X\to\mathbb{R}}\sup_{\theta\in\mathbb{R}}\mathsf{VC}(S^{(r)}_{\eta_1}, B^{(\theta)}_{\eta_2})$*,*

$$\begin{aligned}&\#\mathsf{CompR}(S,B,\ell,\kappa(\beta+2\eta_1+4\eta_2),\delta)\\ &\leq O\left(\frac{m}{\beta^6}\log_+^2\left(\frac{m}{\beta}\right)\log\left(\frac{1}{\eta_1}\right) + \frac{1}{\beta^6}\log\left(\frac{1}{\eta_1}\right)\log\left(\frac{1}{\beta\delta}\right) + \frac{1}{\beta^4}\log\left(\frac{1}{\eta_2}\right)\right).\end{aligned}$$

We prove Theorem 19 using the *omnipredictor* result of Gopalan et al. [34], which shows that any model with a low MC-error w.r.t. B and a low *overall calibration* error can be easily transformed to a model that achieves a low loss compared to the best benchmark in B. Here, the overall calibration error of a model f is defined as follows:

$$\mathsf{C\text{-}error}_\mu(f) := \sum_{v\in V}|\mathbb{E}_{(x,y)\sim\mu}[(y-f(x))\mathbf{1}(f(x)=v)]|, \tag{26}$$

where V is the range of f which we require to be countable. The name "overall calibration error" comes from the fact that $\mathsf{C\text{-}error}_\mu(f) = \mathsf{MC\text{-}error}_{\mu,B}(f)$ when B only contains a single hypothesis b such that $b(x) = 1$ for every $x \in X$.

▶ **Theorem 20** (Omnipredictor [34]). *Let $\ell : \{-1,1\} \times [-1,1] \to \mathbb{R}$ be a loss function such that $\ell(y,\cdot)$ is convex and κ-Lipschitz for any $y \in \{-1,1\}$. Define $\tau : [-1,1] \to [-1,1]$ such that*

$$\tau(u) \in \arg\min_{q\in[-1,1]} \mathbb{E}_{y\sim \mathsf{Ber}^*(u)}[\ell(y,q)].$$

Let μ be a distribution over $X \times \{-1,1\}$ and $B \subseteq [-1,1]^X$ be a total hypothesis class. Let $f : X \to [-1,1]$ be a model satisfying $\mathsf{MC\text{-}error}_{\mu,B}(f) \le \alpha$ and $\mathsf{C\text{-}error}_\mu(f) \le \varepsilon$. Then,

$$\mathbb{E}_{(x,y)\sim\mu}[\ell(y,\tau(f(x)))] \le \inf_{b\in B} \mathbb{E}_{(x,y)\sim\mu}[\ell(y,b(x))] + (\alpha + 3\varepsilon)\kappa.$$

The authors of [34] only proved Theorem 20 in the special case where $\varepsilon = 0$. Since achieving $\mathsf{C\text{-}error}_\mu(f) = 0$ is in general impossible with finitely many data points, it is important to prove Theorem 20 for a general $\varepsilon > 0$. We include a proof of Theorem 20 in the full version of this paper.

We prove Theorem 19 by combining Theorem 20 with our sample complexity upper bound for realizable multicalibration in Section 5. A challenge here is that in addition to achieving $\mathsf{MC\text{-}error}_{\mu,B}(f) \le \alpha$, Theorem 20 also requires us to achieve $\mathsf{C\text{-}error}_\mu(f) \le \varepsilon$. This is similar to the situation in boosting (Section 1.3) where we need to simultaneously achieve a low MA-error and a low sign-C-error. We include a more detailed proof of Theorem 19 in the full paper.

7 Comparative Online Learning

In this section, we study comparative learning in the online setting. We show that the connections we make in Section 3 between comparative learning and learning partial hypotheses can be extended to the online setting, allowing us to show regret bounds for *comparative online learning* in Theorem 25.

Specifically, in the online setting, the data points $(x,y) \in X \times \{-1,1\}$ are not given to the learner all at once. Instead, they come one-by-one and the learner sequentially makes predictions about the label of every individual x before the true label y is revealed. Additionally, the data points are *not* assumed to be drawn i.i.d. from some distribution. Formally, a (possibly inefficient and randomized) online learner L does the following on a stream of data points $(x_1,y_1),\dots,(x_n,y_n)$: for every $i = 1,\dots,n$, given $i-1$ labeled data points $(x_1,y_1),\dots,(x_{i-1},y_{i-1}) \in X \times \{-1,1\}$ and an extra unlabeled data point $x_i \in X$, the learner outputs a prediction $\hat{y}_i \in \{-1,1\}$. The performance of the learner L is measured by

$$\mathsf{mistake}(L;(x_i,y_i)_{i=1}^n) := \frac{1}{n}\sum_{i=1}^n \Pr[\hat{y}_i \ne y_i],$$

where the probability is over the internal randomness in A. We also measure the performance of a hypothesis $h : X \to \{-1,1,*\}$ by

$$\mathsf{mistake}(h;(x_i,y_i)_{i=1}^n) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(h(x_i) \ne y_i).$$

Researchers have studied online learning for partial binary hypothesis classes $H \subseteq \{-1,1,*\}^X$ in the realizable and agnostic settings:

▶ **Definition 21** (Realizable online learning). *Given a hypothesis class $H \subseteq \{-1,1,*\}^X$, a regret bound $\varepsilon \geq 0$, and a positive integer n, we use $\mathsf{ReaOL}_n(H,\varepsilon)$ to denote the set of all online learners L such that for every $h \in H$ and every sequence of data points $(x_1,y_1),\ldots,(x_n,y_n) \in X \times \{-1,1\}$ satisfying $y_i = h(x_i)$ for every $i = 1,\ldots,n$, it holds that*

$$\mathsf{mistake}(L;(x_i,y_i)_{i=1}^n) \leq \varepsilon.$$

▶ **Definition 22** (Agnostic online learning). *Given a hypothesis class $H \subseteq \{-1,1,*\}^X$, a regret bound $\varepsilon \geq 0$, and a positive integer n, we use $\mathsf{AgnOL}_n(H,\varepsilon)$ to denote the set of all online learners L such that for any sequence of data points $(x_1,y_1),\ldots,(x_n,y_n) \in X \times \{-1,1\}$, it holds that*

$$\mathsf{mistake}(L;(x_i,y_i)_{i=1}^n) \leq \inf_{h\in H} \mathsf{mistake}(h;(x_i,y_i)_{i=1}^n) + \varepsilon.$$

We combine the realizable and agnostic settings to define comparative online learning:

▶ **Definition 23** (Comparative online learning). *Given hypothesis classes $S,B \subseteq \{-1,1,*\}^X$, a regret bound $\varepsilon \geq 0$, and a positive integer n, we use $\mathsf{CompOL}_n(S,B,\varepsilon)$ to denote the set of all online learners L such that for every $s \in S$ and every sequence of data points $(x_1,y_1),\ldots,(x_n,y_n) \in X \times \{-1,1\}$ satisfying $y_i = s(x_i)$ for every $i = 1,\ldots,n$, it holds that*

$$\mathsf{mistake}(L;(x_i,y_i)_{i=1}^n) \leq \inf_{b\in B} \mathsf{mistake}(b;(x_i,y_i)_{i=1}^n) + \varepsilon.$$

Analogous to the question of sample complexity, a basic question in online learning is to understand the optimal *regret*, i.e., the minimum ε for which there exists a learner that solves the tasks above given a sequence of n data points. Given a total binary hypothesis class H, the optimal regret in realizable and agnostic online learning has been characterized in [64] and [13] using the *Littlestone dimension*, and this characterization has been extended to partial hypothesis classes in [3]. The Littlestone dimension of a partial hypothesis class $H \subseteq \{-1,1,*\}^X$ is defined as follows. Given $m \in \mathbb{Z}_{\geq 0}$, suppose we associate an individual $x_\zeta \in X$ to every binary string $\zeta \in \cup_{i=0}^{m-1}\{-1,1\}^i$. There are $2^m - 1$ such strings ζ in total, so $(x_\zeta)_{\zeta\in\cup_{i=0}^{m-1}\{-1,1\}^i} \in X^{2^m-1}$. We say $(x_\zeta)_{\zeta\in\cup_{i=0}^{m-1}\{-1,1\}^i}$ is *shattered* by H if for every $\xi = (\xi_1,\ldots,\xi_m) \in \{-1,1\}^m$, there exists $h \in H$ such that $h(x_{\xi_{<i}}) = \xi_i$ for every $i = 1,\ldots,m$, where $\xi_{<i} \in \{-1,1\}^{i-1}$ is the prefix of ξ of length $i-1$. The Littlestone dimension of H is defined to be

$$\mathsf{Ldim}(H) := \sup\{m \in \mathbb{Z}_{\geq 0} : \text{there exists } (x_\zeta)_{\zeta\in\cup_{i=0}^{m-1}\{-1,1\}^i} \in X^{2^m-1} \text{ shattered by } H\}.$$

To give a regret characterization for comparative online learning, we define the *mutual Littlestone dimension* for a pair of hypothesis classes S and B to be

$$\mathsf{Ldim}(S,B) :=$$
$$\sup\{m \in \mathbb{Z}_{\geq 0} : \text{there exists } (x_\zeta)_{\zeta\in\cup_{i=0}^{m-1}\{-1,1\}^i} \in X^{2^m-1} \text{ shattered by both } S \text{ and } B\}.$$

Similarly to Claim 7 for the mutual VC dimension, the mutual Littlestone dimension of (S,B) is equal to the Littlestone dimension of the agreement hypothesis class $\mathbf{A}_{S,B}$ (defined in Section 3):

▷ Claim 24. Let $S,B \subseteq \{-1,1,*\}^X$ be partial binary hypothesis classes. We have $\mathsf{Ldim}(S,B) = \mathsf{Ldim}(\mathbf{A}_{S,B})$.

We omit the proof of the claim because the proof of Claim 7 can be applied here with only minor changes. Our main result in this section is the following regret characterization for comparative online learning.

▶ **Theorem 25.** *For every* $S, B \subseteq \{-1, 1, *\}^X$ *and* $n \in \mathbb{Z}_{>0}$*, define* $m := \mathsf{Ldim}(S, B)$ *and*

$$\varepsilon^* := \inf\{\varepsilon \in \mathbb{R}_{\geq 0} : \mathsf{CompOL}_n(S, B, \varepsilon) \neq \emptyset\}.$$

Then

$$\min\left\{\frac{1}{2}, \frac{m}{2n}\right\} \leq \varepsilon^* \leq O\left(\sqrt{\frac{m}{n}\log\frac{2m+n}{m}}\right).$$

Our proof of Theorem 25 uses the same strategy as in our proof of Theorem 4. Relying on the sample complexity characterizations proved in [3] for AgnOL and ReaOL, we reduce CompOL for (S, B) to AgnOL for $\mathbf{A}_{S,B}$, and conversely reduce ReaOL for $\mathbf{A}_{S,B}$ to CompOL for (S, B).

References

1. Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 447–455. ACM, New York, 2021. doi:10.1145/3406325.3451041.
2. Noga Alon, Nicolò Cesa-Bianchi, Shai Ben-David, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *34th Annual Symposium on Foundations of Computer Science (Palo Alto, CA, 1993)*, pages 292–301. IEEE Comput. Soc. Press, Los Alamitos, CA, 1993. doi:10.1109/SFCS.1993.366858.
3. Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021*, pages 658–671. IEEE Computer Soc., Los Alamitos, CA, 2022. doi:10.1109/FOCS52979.2021.00070.
4. Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860. ACM, New York, 2019. doi:10.1145/3313276.3316312.
5. S. Artstein, V. Milman, S. Szarek, and N. Tomczak-Jaegermann. On convexified packing and entropy duality. *Geom. Funct. Anal.*, 14(5):1134–1141, 2004. doi:10.1007/s00039-004-0486-3.
6. S. Artstein, V. Milman, and S. J. Szarek. Duality of metric entropy. *Ann. of Math. (2)*, 159(3):1313–1328, 2004. doi:10.4007/annals.2004.159.1313.
7. Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *J. Comput. System Sci.*, 75(1):78–89, 2009. doi:10.1016/j.jcss.2008.07.003.
8. Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Mach. Learn.*, 80(2-3):111–139, 2010. doi:10.1007/s10994-010-5174-y.
9. Peter L Bartlett and Philip M Long. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 392–401, 1995.
10. Peter L. Bartlett and Philip M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. In *Eighth Annual Workshop on Computational Learning Theory (COLT) (Santa Cruz, CA, 1995)*, volume 56, pages 174–190. Journal of Computer and System Sciences, 1998. doi:10.1006/jcss.1997.1557.

11 Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Seventh Annual Workshop on Computational Learning Theory (COLT) (New Brunswick, NJ, 1994)*, volume 52, pages 434–452. Journal of Computer and System Sciences, 1996. doi:10.1006/jcss.1996.0033.

12 Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler, and Philip M. Long. Characterizations of learnability for classes of $\{0, \cdots, n\}$-valued functions. *J. Comput. System Sci.*, 50(1):74–86, 1995. doi:10.1006/jcss.1995.1008.

13 Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

14 Eric Blais, Renato Ferreira Pinto, Jr., and Nathaniel Harms. VC dimension and distribution-free sample-based testing. In *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 504–517. ACM, New York, 2021. doi:10.1145/3406325.3451104.

15 Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262, 1994.

16 Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In Thomas Vidick, editor, *11th Innovations in Theoretical Computer Science Conference, ITCS 2020, January 12-14, 2020, Seattle, Washington, USA*, volume 151 of *LIPIcs*, pages 55:1–55:24. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.ITCS.2020.55.

17 Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. doi:10.1145/76359.76371.

18 J. Bourgain, A. Pajor, S. J. Szarek, and N. Tomczak-Jaegermann. On the duality problem for entropy numbers of operators. In *Geometric aspects of functional analysis (1987–88)*, volume 1376 of *Lecture Notes in Math.*, pages 50–63. Springer, Berlin, 1989. doi:10.1007/BFb0090048.

19 Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. *arXiv preprint*, 2022. arXiv:2203.01550.

20 Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 389–402. IEEE Computer Soc., Los Alamitos, CA, 2020. doi:10.1109/FOCS46700.2020.00044.

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

22 Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. Multiaccurate proxies for downstream fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1207–1239, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3531146.3533180.

23 Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108. ACM, New York, 2021. doi:10.1145/3406325.3451064.

24 Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature. In *International Conference on Algorithmic Learning Theory*, pages 342–380. PMLR, 2022.

25 Vitaly Feldman. Distribution-specific agnostic boosting. In Andrew Chi-Chih Yao, editor, *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 241–250. Tsinghua University Press, 2010. URL: http://conference.iiis.tsinghua.edu.cn/ICS2010/content/papers/20.html.

26 Yuval Filmus, Steve Hanneke, Idan Mehalel, and Shay Moran. Optimal prediction using expert advice and randomized Littlestone dimension, 2022.

27 Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper PAC learning with approximate differential privacy. In *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 183–196. ACM, New York, 2021. doi:10.1145/3406325.3451028.

28 Ira Globus-Harris, Varun Gupta, Christopher Jung, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Multicalibrated regression for downstream fairness. *arXiv preprint*, 2022. arXiv:2209.07312.

29 Ira Globus-Harris, Michael Kearns, and Aaron Roth. An algorithmic framework for bias bounties. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1106–1124, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3531146.3533172.

30 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 339–348. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996. doi:10.1109/SFCS.1996.548493.

31 Noah Golowich. Differentially private nonparametric regression under a growth condition. In *Conference on Learning Theory*, pages 2149–2192. PMLR, 2021.

32 Alon Gonen, Shachar Lovett, and Michal Moshkovitz. Towards a combinatorial characterization of bounded-memory learning. *Advances in Neural Information Processing Systems*, 33:9028–9038, 2020.

33 Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint*, 2022. arXiv:2210.08649.

34 Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference, ITCS 2022, January 31 - February 3, 2022, Berkeley, CA, USA*, volume 215 of *LIPIcs*, pages 79:1–79:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPIcs.ITCS.2022.79.

35 Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022.

36 Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. Multicalibrated partitions for importance weights. In *International Conference on Algorithmic Learning Theory*, pages 408–435. PMLR, 2022.

37 Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1):55–76, 2005.

38 Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. *Advances in Neural Information Processing Systems*, 33:9203–9215, 2020.

39 Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021*, pages 942–953. IEEE Computer Soc., Los Alamitos, CA, 2022. doi:10.1109/FOCS52979.2021.00095.

40 D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points (extended abstracts). In *Proceedings of the 1988 Workshop on Computational Learning Theory (Cambridge, MA, 1988)*, pages 280–296. Morgan Kaufmann, San Mateo, CA, 1989. doi:10.1016/0315-0860(89)90031-1.

41 Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

42 Max Hopkins, Daniel Kane, and Shachar Lovett. The power of comparisons for actively learning linear classifiers. *Advances in Neural Information Processing Systems*, 33:6342–6353, 2020.

43 Max Hopkins, Daniel Kane, Shachar Lovett, and Gaurav Mahajan. Noise-tolerant, reliable active classification with comparison queries. In *Conference on Learning Theory*, pages 1957–2006. PMLR, 2020.

44 Max Hopkins, Daniel Kane, Shachar Lovett, and Gaurav Mahajan. Point location and active learning: learning halfspaces almost optimally. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 1034–1044. IEEE Computer Soc., Los Alamitos, CA, 2020. doi:10.1109/FOCS46700.2020.00100.

45 Max Hopkins, Daniel M Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Conference on Learning Theory*, pages 3015–3069. PMLR, 2022.

46 Lunjia Hu, Inbal Livni-Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. *arXiv preprint*, 2022. arXiv:2209.07463.

47 Lunjia Hu and Charlotte Peale. Comparative learning: A sample complexity theory for two hypothesis classes. *arXiv preprint*, 2022. arXiv:2211.09101.

48 Lunjia Hu, Charlotte Peale, and Omer Reingold. Metric entropy duality and the sample complexity of outcome indistinguishability. In *International Conference on Algorithmic Learning Theory*, pages 515–552. PMLR, 2022.

49 Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

50 Young Jung, Baekjin Kim, and Ambuj Tewari. On the equivalence between online and private learnability beyond binary classification. *Advances in Neural Information Processing Systems*, 33:16701–16710, 2020.

51 Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *STOC'08*, pages 629–638. ACM, New York, 2008. doi:10.1145/1374376.1374466.

52 Daniel M. Kane, Shachar Lovett, Shay Moran, and Jiapeng Zhang. Active classification with comparison queries. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 355–366. IEEE Computer Soc., Los Alamitos, CA, 2017. doi:10.1109/FOCS.2017.40.

53 Michael Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of Computing*, pages 392–401, 1993.

54 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

55 Michael Kearns and Dana Ron. Testing problems with sublearning sample complexity. *J. Comput. System Sci.*, 61(3):428–456, 2000. doi:10.1006/jcss.1999.1656.

56 Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *31st Annual Symposium on Foundations of Computer Science (FOCS) (St. Louis, MO, 1990)*, volume 48, pages 464–497. Journal of Computer and System Sciences, 1994. doi:10.1016/S0022-0000(05)80062-5.

57 Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

58 Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

59 J. Kivinen. Learning reliably and with one-sided error. *Math. Systems Theory*, 28(2):141–172, 1995. doi:10.1007/BF01191474.

60 Jyrki Kivinen. Reliable and useful learning. In *Proceedings of the Second Annual Workshop on Computational Learning Theory (Santa Cruz, CA, 1989)*, pages 365–380. Morgan Kaufmann, San Mateo, CA, 1989.

61 Jyrki Kivinen. Reliable and useful learning with uniform probability distributions. In *ALT*, pages 209–222, 1990.

62 Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 2000)*, pages 309–318. ACM, New York, 2000.

63 Nathan Linial, Yishay Mansour, and Ronald L. Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Inform. and Comput.*, 90(1):33–49, 1991. doi:10.1016/0890-5401(91)90058-A.

64 Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

65 Nick Littlestone and Manfred Warmuth. Relating data compression and learnability, 1986.

66 Philip M. Long. On agnostic learning with $\{0, *, 1\}$-valued and real-valued hypotheses. In *Computational learning theory (Amsterdam, 2001)*, volume 2111 of *Lecture Notes in Comput. Sci.*, pages 289–302. Springer, Berlin, 2001. doi:10.1007/3-540-44581-1_19.

67 Emanuel Milman. A remark on two duality relations. *Integral Equations Operator Theory*, 57(2):217–228, 2007. doi:10.1007/s00020-006-1479-4.

68 Albrecht Pietsch. *Theorie der Operatorenideale (Zusammenfassung)*. Wissenschaftliche Beiträge der Friedrich-Schiller-Universität Jena. Friedrich-Schiller-Universität, Jena, 1972.

69 Ronald L. Rivest and Robert Sloan. Learning complicated concepts reliably and usefully (extended abstract). In *Proceedings of the 1988 Workshop on Computational Learning Theory (Cambridge, MA, 1988)*, pages 69–79. Morgan Kaufmann, San Mateo, CA, 1989.

70 Harrison Rosenberg, Robi Bhattacharjee, Kassem Fawaz, and Somesh Jha. An exploration of multicalibration uniform convergence bounds. *arXiv preprint*, 2022. arXiv:2202.04530.

71 Guy N Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In *International Conference on Machine Learning*, pages 9107–9115. PMLR, 2021.

72 Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *Advances in Neural Information Processing Systems*, 33:13331–13340, 2020.

73 Satchit Sivakumar, Mark Bun, and Marco Gaboardi. Multiclass versus binary differentially private pac learning. *Advances in Neural Information Processing Systems*, 34:22943–22954, 2021.

74 Christopher J Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In *International Conference on Machine Learning*, pages 21633–21657. PMLR, 2022.

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

76 V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

77 Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.