Metric Entropy Duality and the Sample Complexity of Outcome Indistinguishability

Lunjia Hu LUNJIA@STANFORD.EDU
Computer Science Department, Stanford University

Charlotte Peale CPEALE@STANFORD.EDU
Computer Science Department, Stanford University

Omer Reingold REINGOLD@STANFORD.EDU
Computer Science Department, Stanford University

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

We give the first sample complexity characterizations for outcome indistinguishability, a theoretical framework of machine learning recently introduced by Dwork, Kim, Reingold, Rothblum, and Yona (STOC 2021). In outcome indistinguishability, the goal of the learner is to output a predictor that cannot be distinguished from the target predictor by a class $\mathcal{D}$ of distinguishers examining the outcomes generated according to the predictors' predictions. While outcome indistinguishability originated from the algorithmic fairness literature, it provides a flexible objective for machine learning even when fairness is not a consideration. In this work, we view outcome indistinguishability as a relaxation of PAC learning that allows us to achieve meaningful performance guarantees under data constraint.

In the distribution-specific and realizable setting where the learner is given the data distribution together with a predictor class $\mathcal{P}$ containing the target predictor, we show that the sample complexity of outcome indistinguishability is characterized by the metric entropy of $\mathcal{P}$ w.r.t. the dual Minkowski norm defined by $\mathcal{D}$, and equivalently by the metric entropy of $\mathcal{D}$ w.r.t. the dual Minkowski norm defined by $\mathcal{P}$. This equivalence makes an intriguing connection to the long-standing metric entropy duality conjecture in convex geometry. Our sample complexity characterization implies a variant of metric entropy duality, which we show is nearly tight. In the distribution-free setting, we focus on the case considered by Dwork et al. where $\mathcal{P}$ contains all possible predictors, hence the sample complexity only depends on $\mathcal{D}$. In this setting, we show that the sample complexity of outcome indistinguishability is characterized by the fat-shattering dimension of $\mathcal{D}$.

We also show a strong sample complexity separation between realizable and agnostic outcome indistinguishability in both the distribution-free and the distribution-specific settings. This is in contrast to distribution-free (resp. distribution-specific) PAC learning where the sample complexity in both the realizable and the agnostic settings can be characterized by the VC dimension (resp. metric entropy).

Keywords: outcome indistinguishability, covering number, duality conjecture, fat shattering

1. Introduction

Prediction algorithms based on machine learning are becoming increasingly influential on major decisions that affect individual's lives in settings such as medical diagnoses, loan applications, and educational admissions processes. As a result, we must be careful that the predictions of these algorithms do not discriminate against any sub-communities within the input population. Unfortunately,

standard measures of prediction accuracy don't necessarily guarantee the absence of such discrimination. Consider a binary classifier that is used to predict an individual's probability of repaying a loan. A natural way to measure the success of our classifier would be to use classification error–the fraction of instances from some representative test set that it classifies incorrectly. Ideally, we'd hope that a classifier with 5% error incorrectly classifies any particular individual in our population only 5% of the time. However, if some low-income minority group makes up 10% of the population and has a low probability (say 40%) of repaying a loan on average, a classifier that chooses to focus on getting 99% accuracy on the majority group and applies a blanket policy of classifying every individual in the minority group as unlikely to pay back the loan can still receive this 5% classification error guarantee despite incorrectly classifying minority individuals 40% of the time. If factored into loan application decisions, such a classifier could deny loans to many worthy applicants from the minority group, further cementing and potentially exacerbating the financial disparities present in the population.

This potential for discrimination was the original inspiration behind *Outcome indistinguishability* (OI), a theoretical framework for machine learning recently proposed by Dwork, Kim, Reingold, Rothblum, and Yona (2021) that aims to address disparities in the treatment of different groups by replacing the accuracy objective with a more flexible objective that can give stronger guarantees for sub-communities in a population. OI is a framework used when learning predictors, rather than classifiers. In this setting, an analogous measure to the classification error of a learned predictor p is the ℓ_1 error considered by e.g. Bartlett et al. (1996). Instead of using the ℓ_1 error as a quality measure, the main idea of OI is to measure the quality of a learned predictor p by how easy it is for a predetermined class $\mathcal{D}$ of distinguishers to distinguish between the output of p and the output of the target predictor.

More precisely, the goal of OI is to output a predictor $p : X \to [0,1]$ assigning probabilities $p(x) \in [0,1]$ to individuals $x \in X$. Given a distribution μ over X, every predictor p defines a distribution μ_p for individual-outcome pairs $(x,o) \in X \times \{0,1\}$ where the individual x is drawn from μ, and the outcome $o \sim \mathrm{Ber}(p(x))$ is drawn from the Bernoulli distribution with mean $p(x)$. The input to the OI learner consists of examples $(x_1,o_1),\ldots,(x_n,o_n)$ drawn i.i.d. from μ_{p^*} for an unknown target predictor p^*. The outputted predictor is audited by a set of distinguishers $\mathcal{D}$. Here, a distinguisher takes an individual-outcome pair $(x,o) \in X \times \{0,1\}$ as input, and either accepts or rejects the input.[1] The distinguishing advantage of a distinguisher d on two predictors p_1 and p_2 is defined as the absolute difference between the acceptance probabilities of d on (x,o) drawn from μ_{p_1} and μ_{p_2}. The quality of the learned predictor p is then measured by the maximum distinguishing advantage (MDA) on p and p^* over all distinguishers $d \in \mathcal{D}$. In the loan repayment setting described above, adding a distinguisher that only compares the rate of positive classification on the minority group could identify and prevent this disparity for a small enough allowable MDA threshold.

It can be shown that the MDA never exceeds the ℓ_1 error $\mathbb{E}_{x\sim\mu}\,|p(x) - p^*(x)|$, and when $\mathcal{D}$ contains all possible distinguishers, the MDA and the ℓ_1 error are known to be equal (see Lemma 7). However, by leveraging our ability to select only the distinguishers we care about in OI, we can

1. Dwork et al. (2021) also considered more advanced distinguishers with access to additional information about the predictors such as the predictions themselves or the predictor code, but the simplest distinguishers we describe here can already express non-trivial objectives. Also, while we describe outcome indistinguishability in the binary outcome setting, it is possible to generalize the notion to an arbitrary number of possible outcomes by considering predictors that output a description of a more general distribution.

tune the MDA to be a more suitable quality measure compared to ℓ_1 error even in settings outside of fairness. As an example, suppose we would like to learn a binary image classifier that can be used in self-driving cars to determine whether the road contains an obstruction. Ideally we would like to learn a model that gives classification error very close to zero because it means that we can expect the car to fail to detect an obstruction extremely rarely. However, what if we have insufficient data to learn a classifier with extremely low error, and must settle for larger error on the order of 1% or more? This is where we need to observe that not all errors are born equal. Failing to recognize leaves on the road is dramatically different from failing to identify a parking vehicle. Unfortunately, a 1% error may be completely concentrated on important obstructions (that may occur less than 1% of the time). An overall error rate that would guarantee minuscule errors in important cases may very well be impossible and concentrating on the errors we care about may be mandated.

This example demonstrates that classification error alone may not be enough to tell us whether this is a model we would trust on the roads. In particular, the ability to refine our measure of performance to focus on the particular types of mistakes that we care about most would give us a better understanding of performance and might potentially allow us to get good targeted accuracy guarantees even with insufficient data to achieve high accuracy against *all* types of errors. If we believe that distinguishing between instances where a large obstruction is or is not present requires a smaller number of circuit gates than distinguishing between instances that may contain more minor obstructions such as a small tree branch or plastic bag, choosing $\mathcal{D}$ to contain all distinguishers with small circuit sizes would filter out serious errors such as missing a pedestrian crossing the street even when we cannot achieve any meaningful overall accuracy due to limited data. Moreover, if $\mathcal{D}$ contains a distinguisher that accepts (x, o) if and only if x contains a large obstruction and o equals 1, a predictor that significantly underestimates the true prediction $p^*(x)$ when x contains a serious obstruction would have a high MDA. In general, when the distinguisher class $\mathcal{D}$ is properly chosen, a predictor with ℓ_1 error and MDA both being about 0.1 can be less preferable than a predictor with ℓ_1 error being 0.2 but MDA being only 0.01.

1.1. Sample Complexity Characterizations

Traditionally, outcome indistinguishability and related notions such as multicalibration (Hébert-Johnson, Kim, Reingold, and Rothblum, 2018) have been viewed as providing stronger guarantees than ℓ_1 error alone by allowing a predictor's performance to be fine-tuned at the group level rather than just looking at the entire population. However, our obstruction-identification example from the previous section demonstrates how OI can also be viewed as a useful *relaxation* of the standard ℓ_1 performance benchmark. By focusing on the important errors specified by the distinguisher class, outcome indistinguishability may allow us to achieve good performance with relatively small sample size. It is natural to ask: how much improvement in the sample size do we get from OI? This is the main focus of this paper—understanding the sample complexity of outcome indistinguishability.

It is one of the major objectives of learning theory to understand the sample complexity of various learning tasks and many influential characterizations of sample complexity have been discovered. The most notable example is the VC dimension for PAC learning (Vapnik and Chervonenkis, 1971; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Linial, Mansour, and Rivest, 1991). In PAC learning (Valiant, 1984b,a), the learner receives examples $(x_1, h^*(x_1)), \ldots, (x_n, h^*(x_n))$ where $x_1, \ldots, x_n$ are drawn i.i.d. from a distribution μ over X, and aims to output a binary classifier (a.k.a. concept or hypothesis) $h : X \to \{0, 1\}$ that is "close" to the target classifier $h^* : X \to \{0, 1\}$.

Here, the performance of h is measured by the classification error $\Pr_{x\sim\mu}[h(x) \neq h^*(x)]$. In the realizable setting, the target classifier h^* is assumed to be from a given class $\mathcal{H}$, whereas in the agnostic setting, there is no assumption on h^* but the performance of h is measured relative to the best classifier in the class $\mathcal{H}$. The main result of VC theory states that the sample complexity of PAC learning in both the realizable and the agnostic settings are characterized by the *VC dimension* (Vapnik and Chervonenkis, 1971) of the class $\mathcal{H}$, which has a simple combinatorial definition. The success of VC theory raises the possibility that the sample complexity of other learning tasks may also have natural characterizations. Below we discuss two such learning tasks that are most relevant to our work: predictor learning and distribution-specific learning.

The extension of PAC learning to predictor learning dates back to (Kearns and Schapire, 1994), in which predictors were termed probabilistic concepts. In predictor learning, binary classifiers are replaced by predictors whose predictions take values in $[0, 1]$. Kearns and Schapire (1994) introduced the notion of *fat-shattering dimension* (see Section 2.4 for a precise definition) as a generalization of VC dimension to predictor classes and showed a lower bound on the sample complexity of learning a predictor class by its fat-shattering dimension. Alon, Ben-David, Cesa-Bianchi, and Haussler (1997) and Bartlett, Long, and Williamson (1996) complemented the result with corresponding upper bounds and concluded that a predictor class is learnable from finitely many examples if and only if it has finite fat-shattering dimension. Quantitatively, these papers showed that the sample complexity of learning a predictor in a class $\mathcal{P}$ within ℓ_1 error ε is characterized by

$$(1/\varepsilon)^{O(1)}\mathsf{fat}_{\mathcal{P}}\left(\varepsilon^{\Theta(1)}\right),$$

assuming we want the success probability to be at least a constant, say $9/10$. Moreover, Bartlett et al. (1996) extended the characterization to the agnostic setting, where the objective is the mean absolute error between the learned prediction and the actual outcome. Later, the sample complexity bounds and the related uniform convergence results from Alon et al. (1997) and Bartlett et al. (1996) were improved by Bartlett and Long (1995), Bartlett and Long (1998), and Li, Long, and Srinivasan (2001).

Another natural extension of PAC learning is distribution-specific learning. In both PAC learning and predictor learning discussed above, performance of the learner is evaluated based on the worst-case distribution μ. These settings are referred to as *distribution-free* due to their lack of dependence on a particular input distribution. Since the distribution is usually not the worst-case in practice, *distribution-specific* learning focuses on the performance of the learner on a given distribution μ. In this setting, the sample complexity of learning a binary classifier in class $\mathcal{H}$ to achieve a classification error below ε is characterized by

$$(1/\varepsilon)^{O(1)} \log N_\mu(\mathcal{H}, \Theta(\varepsilon)) \tag{1}$$

using the *metric entropy*, i.e., the logarithm of the covering number N_μ of $\mathcal{H}$ w.r.t. the classification error (as a metric), which indeed depends on the specific distribution μ (Benedek and Itai, 1991).

To compare OI with these previous clasification-error/ℓ_1-error-based notions of learning in terms of sample complexity, we need a characterization of the sample complexity of OI using similar quantities such as the fat-shattering dimension or the metric entropy. While it is tempting to hope that we might directly apply such notions, OI introduces additional subtlety in that we must consider how the expressiveness of the predictor class $\mathcal{P}$ *and* the class of distinguishers $\mathcal{D}$ interact to fully understand the sample complexity requirements. This is in contrast to standard settings where characterizing the expressiveness of the concept class via VC dimension or related notions is sufficient.

We show a simple example where it is indeed important to consider the interplay between $\mathcal{P}$ and $\mathcal{D}$ rather than just considering their contributions independently: Partition the set of inputs X into two equal-sized sets, X_1 and X_2. We consider a class of predictors that are maximally expressive on X_2 and constant on X_1: Let $\mathcal{P} = \{p : X \to [0,1] | p(x) = 0, \forall x \in X_1\}$. Similarly, we can define a distinguisher class that are maximally inquisitive on one set and ignores the other set but depending on which set is ignored we will get dramatically different complexity: Define two potential distinguisher classes: $\mathcal{D}_1 = \{d : X \times \{0,1\} \to \{\text{ACCEPT}, \text{REJECT}\} | d(x,b) = \text{REJECT}, \forall x \in X_1, b \in \{0,1\}\}$, $\mathcal{D}_2 = \{d : X \times \{0,1\} \to \{\text{ACCEPT}, \text{REJECT}\} | d(x,b) = \text{REJECT}, \forall x \in X_2, b \in \{0,1\}\}$. $\mathcal{D}_1$ and $\mathcal{D}_2$ are symmetric and thus identical in any independent measure of complexity. Nevertheless, it is easy to see that while achieving ε-OI on $\mathcal{P}$ with respect to $\mathcal{D}_1$ is equivalent to learning a predictor from $\mathcal{P}$ with ε ℓ_1 error on the set X_2, learning an ε-OI predictor from $\mathcal{P}$ with respect to $\mathcal{D}_2$ is trivial as $\mathcal{P}$ is constant on all individuals that $\mathcal{D}_2$ can distinguish between, and therefore any $p \in \mathcal{P}$ will satisfy perfect OI with respect to $\mathcal{D}_2$. This example demonstrates the need for a more subtle variation of existing complexity notions in order to tightly characterize the sample complexity of OI.

Prior to our work, Dwork et al. (2021) showed that $O(\varepsilon^{-4} \log(|\mathcal{D}|/\varepsilon\delta))$ examples are sufficient for achieving an advantage below ε over a distinguisher class $\mathcal{D}$ with probability at least $1 - \delta$ in the distribution-free setting. In this work, we refine the $\log|\mathcal{D}|$ dependence on $\mathcal{D}$ to its fat-shattering dimension with a matching lower bound (Section 5). In addition, we characterize the sample complexity of OI in the distribution-specific setting with greater generality for every given predictor class $\mathcal{P}$, placing OI in the same setting as the classification-error/ℓ_1-error-based notions of learning with improved sample complexity due to a well-tuned objective based on the distinguisher class $\mathcal{D}$. The interplay between the distinguisher class $\mathcal{D}$ and the predictor class $\mathcal{P}$ connects our sample complexity characterizations to the intriguing metric entropy duality conjecture in convex geometry, which we discuss more in Section 1.2 below.

1.2. Covering for Distinguishers and Metric Entropy Duality

Before we describe our sample complexity characterizations for OI, we would like to discuss the ideas behind the metric-entropy-based sample complexity characterization (1) for distribution-specific learning by Benedek and Itai (1991). Our characterizations for distribution-specific OI are based on similar ideas, but in a more subtle and arguably surprising way, making an intriguing connection to the *metric entropy duality conjecture*, a long-standing conjecture in convex geometry.

The idea behind the sample complexity upper bound in (1) is to reduce the size of the classifier class $\mathcal{H}$ by taking an ε-covering of it. Consider the space of all binary classifiers endowed with the "classification error metric" $\eta(h_1, h_2) = \Pr_{x\sim\mu}[h_1(x) \neq h_2(x)]$. If two classifiers h_1 and h_2 are close w.r.t. this metric, choosing h_1 and h_2 would result in roughly equal classification errors w.r.t. the target predictor. This gives a natural procedure for simplifying the classifier class and consequently controlling the sample complexity: we can replace $\mathcal{H}$ by an ε-covering $\mathcal{H}'$ of it with only minor loss in its expressivity. Here, an ε-covering is a subset $\mathcal{H}' \subseteq \mathcal{H}$ such that every $h \in \mathcal{H}$ can find a close companion $h' \in \mathcal{H}'$ such that $\eta(h, h') \leq \varepsilon$. As the size of $\mathcal{H}'$ can be bounded by the covering number $N_\mu(\mathcal{H}, \varepsilon)$, we can use a relatively small amount of examples to accurately estimate the classification error of every classifier in $\mathcal{H}'$. This naturally leads to the empirical risk minimization (ERM) algorithm used in (Benedek and Itai, 1991).

To extend the ERM algorithm to outcome indistinguishability, a natural idea is to define the metric $\eta(p_1, p_2)$ between two predictors p_1 and p_2 to be the maximum distinguishing advantage for p_1 and p_2 w.r.t. the distinguisher class $\mathcal{D}$, and compute an ε-covering $\mathcal{P}'$ of the predictor class $\mathcal{P}$. However, this direct extension (Algorithm 1) of the ERM algorithm to OI does not give us the right sample complexity upper bound (we show this formally in Appendix A, especially in Lemma 33). This failure is partly because the acceptance probabilities of the distinguishers in $\mathcal{D}$ are not all preserved on a small sample. Indeed, learning a predictor p with MDA below ε w.r.t. to the target predictor p^* necessarily requires estimating the acceptance probability of every distinguisher in $\mathcal{D}$ on p^* within error ε (the estimates are simply the acceptance probabilities on p).

To meet the need of estimating the acceptance probabilities of the distinguishers in $\mathcal{D}$, we use a new algorithm (Algorithm 2) where we compute a covering of $\mathcal{D}$ w.r.t. a metric defined by the predictor class $\mathcal{P}$, i.e., we flip the roles of $\mathcal{P}$ and $\mathcal{D}$ in the covering. Using this new algorithm, we get a sample complexity upper bound as a function of the metric entropy $\log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon)$ of $\mathcal{D}$ w.r.t. $\mathcal{P}$, but the sample complexity lower bound we get by extending the arguments from (Benedek and Itai, 1991) is still a function of the metric entropy $\log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon)$ of $\mathcal{P}$ w.r.t. $\mathcal{D}$. How do we make the upper and lower bounds match?

Our idea is to first transform distinguishers into the same inner product space as the predictors, and then interpret $\mathcal{D}$ and $\mathcal{P}$ as two abstract vector sets $\mathcal{F}_1$ and $\mathcal{F}_2$ in the same inner product space that can exchange roles freely. Specifically, combining our upper and lower bounds, we know that the lower bound based on $\log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon)$ never exceeds the upper bound based on $\log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon)$. If we set $\mathcal{F}_1 = \mathcal{P}$ and $\mathcal{F}_2 = \mathcal{D}$, a more precise version of what we get is

$$\log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \varepsilon) \le O\Big(\varepsilon^{-2}\Big(1 + \log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \Omega(\varepsilon))\Big)\Big). \tag{2}$$

If $\mathcal{F}_1$ and $\mathcal{F}_2$ are two arbitrary abstract sets of vectors, we can inversely set $\mathcal{F}_1 = \mathcal{D}$ and $\mathcal{F}_2 = \mathcal{P}$ in the inequality above and get

$$\log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon) \le O\Big(\varepsilon^{-2}\Big(1 + \log N_{\mu,\mathcal{D}}(\mathcal{P}, \Omega(\varepsilon))\Big)\Big).$$

This inequality is exactly what we need to flip back the roles of $\mathcal{P}$ and $\mathcal{D}$ and make our upper bound match our lower bound.

The key inequality (2) that helps match our upper and lower bounds comes from combining the bounds themselves. Moreover, this inequality makes an intriguing connection to the long-standing *metric entropy duality conjecture* in convex geometry, which conjectures that (2) holds without the ε^{-2} factor, but with the additional assumption that $\mathcal{F}_1$ and $\mathcal{F}_2$ are convex and symmetric around the origin. Without this additional assumption, we show that the quadratic dependence on $1/\varepsilon$ in (2) is nearly tight (Lemma 14).

The metric entropy duality conjecture (formally stated in Conjecture 12) was first proposed by Pietsch (1972). While the conjecture remains open, many weaker versions of it have been proved (e.g. Bourgain, Pajor, Szarek, and Tomczak-Jaegermann, 1989). Most notably, the conjecture was proved by Artstein, Milman, and Szarek (2004b) in the special case where one of the two convex sets is an ellipsoid. This result was further strengthened by Artstein, Milman, Szarek, and Tomczak-Jaegermann (2004a). Milman (2007) proved a weaker form of the conjecture with the constant factors replaced by dimension dependent quantities.

1.3. Our Contributions

Outcome indistinguishability was originally proposed as a strong fairness and accuracy criterion, potentially requiring large sample sizes to achieve. In this work, we view OI differently as a meaningful notion when we have insufficient data for ℓ_1-error-based learning. For this reason, we focus on no-access OI, the simplest form of OI introduced by Dwork et al. (2021). For no-access OI, we show that (randomized) distinguishers can be converted to vectors in the same inner product space as the predictors (Section 2.1.2), connecting the MDA objective to the *multiaccuracy* objective used by Kim et al. (2019). This allows us to understand predictor classes $\mathcal{P}$ and distinguisher classes $\mathcal{D}$ using geometric notions such as the dual Minkowski norms $\|\cdot\|_{\mu,\mathcal{P}}, \|\cdot\|_{\mu,\mathcal{D}}$ and the covering numbers $N_{\mu,\mathcal{P}}(\cdot,\cdot), N_{\mu,\mathcal{D}}(\cdot,\cdot)$ defined by these norms (see Section 2.2).

In the distribution-specific setting, we consider realizable OI where the target predictor lies in an arbitrary given predictor class $\mathcal{P}$. Setting the failure probability bound δ in Theorem 13 to be a constant, say $1/10$, for every predictor class $\mathcal{P}$, every distinguisher class $\mathcal{D}$, every data distribution μ and every MDA bound ε, we characterize the sample complexity of distribution-specific realizable OI both as

$$(1/\varepsilon)^{O(1)} \log N_{\mu,\mathcal{D}}(\mathcal{P}, \Theta(\varepsilon)) \tag{3}$$

and as

$$(1/\varepsilon)^{O(1)} \log N_{\mu,\mathcal{P}}(\mathcal{D}, \Theta(\varepsilon)). \tag{4}$$

Our sample complexity characterizations (3) and (4) highlight an intriguing connection between learning theory and the metric entropy duality conjecture (Conjecture 12) first proposed by Pietsch (1972), which conjectures that

$$\log N_{\mu,\mathcal{F}_1}(\mathcal{F}_2, \varepsilon) \leq O(\log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \Omega(\varepsilon)))$$

whenever two function classes $\mathcal{F}_1$ and $\mathcal{F}_2$ are convex and symmetric around the origin. Our sample complexity characterizations imply a variant version of metric entropy duality (Theorem 11) where $\mathcal{F}_1$ and $\mathcal{F}_2$ are not required to be convex and symmetric, which we show is nearly tight (Lemma 14).

In the distribution-free setting, we focus on the case where $\mathcal{P}$ contains all possible predictors, which is the setting considered by Dwork et al. (2021). Setting $\delta = 1/10$ in Theorem 15, we show that the sample complexity of distribution-free OI in this setting is characterized by

$$(1/\varepsilon)^{O(1)} \mathsf{fat}_{\mathcal{D}}(\Theta(\varepsilon)).$$

This characterization extends to *multicalibration* with some modifications (see Remark 21 and Remark 24). Our result refines the $\log|\mathcal{D}|$ in the sample complexity upper bounds by Dwork et al. (2021) (for OI) and by Hébert-Johnson et al. (2018) (for multicalibration) to the fat-shattering dimension of $\mathcal{D}$ with a matching lower bound.

In addition, we show that the sample complexity of OI in the agnostic setting behaves very differently from the realizable setting. This is in contrast to many common learning settings where the sample complexities of realizable and agnostic learning usually behave similarly (a recent independent work by Hopkins et al. (2021) gives a unified explanation for this phenomenon). In both the distribution-free and the distribution-specific settings, we show that the sample complexity of agnostic OI can increase when we remove some distinguishers from $\mathcal{D}$, and it can become arbitrarily large even when the realizable sample complexity is bounded by a constant (Section 6.2). This also suggests that OI can have larger sample complexity compared to ℓ_1-error based learning in the

agnostic setting. This is because in the agnostic setting, the performance of the learned predictor is measured relative to the best predictor in the class $\mathcal{P}$, which can have a much better performance when measured using the selected objective of OI than using the ℓ_1 error. On the other hand, when the target predictor p^* belongs to the symmetric convex hull of the predictor class $\mathcal{P}$, we show that the sample complexity in the distribution-specific agnostic setting has the same characterizations as in the realizable setting (Lemma 25).

1.4. Related Work

The notion of outcome indistinguishability originated from the growing research of algorithmic fairness. Specifically, outcome indistinguishability can be treated as a generalization of multiaccuracy and multicalibration introduced by Hébert-Johnson, Kim, Reingold, and Rothblum (2018) and Kim, Ghorbani, and Zou (2019), in which the goal is to ensure that the learned predictor is accurate in expectation or calibrated conditioned on every subpopulation in a subpopulation class $\mathcal{C}$. Roughly speaking, the subpopulation class $\mathcal{C}$ in multiaccuracy and multicalibration plays a similar role as the distinguisher class $\mathcal{D}$ in outcome indistinguishability, and this connection has been discussed more extensively in (Dwork et al., 2021, Section 4). Beyond fairness, multicalibration and OI also provide strong accuracy guarantees (see e.g. Rothblum and Yona, 2021; Blum and Lykouris, 2020; Zhao, Kim, Sahoo, Ma, and Ermon, 2021; Gopalan, Kalai, Reingold, Sharan, and Wieder, 2022; Kim, Kern, Goldwasser, Kreuter, and Reingold, 2022; Burhanpurkar, Deng, Dwork, and Zhang, 2021). For a general predictor class $\mathcal{P}$ and a subpopulation class $\mathcal{C}$, Shabat, Cohen, and Mansour (2020) showed sample complexity upper bounds of uniform convergence for multicalibration based on the maximum of suitable complexity measures of $\mathcal{C}$ and $\mathcal{P}$. They complemented this result with a lower bound which does not grow with $\mathcal{C}$ and $\mathcal{P}$. In comparison, we focus on the weaker no-access OI setting where the sample complexity can be much smaller, and we provide matching upper and lower bounds in terms of the dependence on $\mathcal{D}$ and $\mathcal{P}$.

1.5. Paper Organization

The remainder of this paper is structured as follows. In Section 2, we formally define OI and related notions. We give lower and upper bounds for the sample complexity of distribution-specific realizable OI in Section 3, and turn these bounds into a characterization in Section 4 via metric entropy duality. We characterize the sample complexity of distribution-free OI in Section 5. Finally, in Section 6 we show a strong separation between the sample complexities of realizable and agnostic OI in both the distribution-free and distribution-specific settings.

2. Preliminaries

We use X to denote an arbitrary non-empty set of individuals throughout the paper. For simplicity, whenever we say μ is a distribution over X, we implicitly assume that every subset of X is measurable w.r.t. μ (this holds e.g. when μ is a discrete distribution), although all the results in this paper naturally extend to more general distributions under appropriate measurability assumptions. We use Δ_X to denote the set of all distributions over X satisfying the implicit assumption. For two sets A and B, we use B^A to denote the class of all functions $f : A \to B$. Given $r \in [0, 1]$, we use $\mathrm{Ber}(r)$ to denote the Bernoulli distribution over $\{0, 1\}$ with mean r. We use $\log$ to denote the base-2 logarithm.

2.1. Outcome Indistinguishability

Outcome indistinguishability is a theoretical framework introduced by Dwork et al. (2021) that aims to guarantee that the outcomes produced by some learned predictor $p : X \to [0,1]$ are indistinguishable to a predetermined class of distinguishers $\mathcal{D}$ from outcomes sampled from the true probabilities for each individual defined by $p^* : X \to [0,1]$.

The distinguishing task in outcome indistinguishability consists of drawing an individual $x \in X$ according to some population distribution $\mu \in \Delta_X$ and then presenting the distinguisher with an outcome/individual pair (x, o) where o is either sampled according to the true predictor p^* from the Bernoulli distribution $\mathrm{Ber}(p^*(x))$, or sampled according to the learned predictor p from $\mathrm{Ber}(p(x))$. Taking a pair $(x, o) \in X \times \{0,1\}$, a distinguisher d outputs $d(x, o) = \text{ACCEPT}$ or $d(x, o) = \text{REJECT}$. We allow distinguishers to be randomized.

Given a predictor p and a distribution μ over X, we define μ_p to be the distribution of pairs $(x, o) \in X \times \{0,1\}$ drawn using the process described above such that $x \sim \mu$ and $o \sim \mathrm{Ber}(p(x))$. With this notation in hand, we can now provide a formal definition of outcome indistinguishability:

Definition 1 (No-Access Outcome Indistinguishability, Dwork et al. (2021)) *Let $\mu \in \Delta_X$ be a distribution over a set of individuals X and $p^* : X \to [0,1]$ be some target predictor for the set of individuals. For a class of distinguishers $\mathcal{D}$ and $\varepsilon > 0$, a predictor $p : X \to [0,1]$ satisfies $(\mathcal{D}, \varepsilon)$-outcome indistinguishability (OI) if for every $d \in \mathcal{D}$,*

$$\left| \Pr_{(x,o)\sim\mu_p} [d(x, o) = \text{ACCEPT}] - \Pr_{(x,o^*)\sim\mu_{p^*}} [d(x, o^*) = \text{ACCEPT}] \right| \leq \varepsilon. \tag{5}$$

We refer to the left-hand-side of (5) as the *distinguishing advantage* (or simply *advantage*) of the distinguisher d, denoted $\mathrm{adv}_{\mu,d}(p, p^*)$. Given a class $\mathcal{D}$ of distinguishers, we use $\mathrm{adv}_{\mu,\mathcal{D}}(p, p^*)$ to denote the supremum $\sup_{d\in\mathcal{D}} \mathrm{adv}_{\mu,d}(p, p^*)$. According to Definition 1, a learned predictor p satisfies $(\mathcal{D}, \varepsilon)$-OI if and only if $\mathrm{adv}_{\mu,\mathcal{D}}(p, p^*) \leq \varepsilon$.

There are many different extensions of OI to settings in which the power of distinguishers is extended to allow access to additional information about the learned predictor p or access to more than one individual/outcome pair. We will be focusing on this most basic case in which the distinguisher only has access to a single pair (x, o) and has no additional access to the learned predictor p (hence the name *No-Access OI*).

2.1.1. OI ALGORITHMS

An OI algorithm (or learner) takes examples $(x_1, o_1), \ldots, (x_n, o_n)$ drawn i.i.d. from μ_{p^*} and outputs a predictor p that satisfies $(\mathcal{D}, \varepsilon)$-OI with probability at least $1 - \delta$:

Definition 2 (No-Access OI Algorithms) *Given a target predictor $p^* \in [0,1]^X$, a class of distinguishers $\mathcal{D}$, a distribution $\mu \in \Delta_X$, an advantage bound $\varepsilon \geq 0$, a failure probability bound $\delta \geq 0$, and a nonnegative integer n, we use $\mathsf{OI}_n(p^*, \mathcal{D}, \varepsilon, \delta, \mu)$ to denote the set of all (possibly randomized and inefficient) algorithms $\mathcal{A}$ with the following properties:*

1. *$\mathcal{A}$ takes n examples $(x_1, o_1), \ldots, (x_n, o_n) \in X \times \{0,1\}$ as input and outputs a predictor $p \in [0,1]^X$;*
2. *when the input examples are drawn i.i.d. from μ_{p^*}, with probability at least $1 - \delta$ the output predictor p satisfies $\mathrm{adv}_{\mu,\mathcal{D}}(p, p^*) \leq \varepsilon$.*

When seeking to learn an outcome indistinguishable predictor, there are two different tasks that we can consider. On one hand, in what we call the *realizable* case, we assume that the target predictor p^* lies in a known predictor class $\mathcal{P} \subseteq [0,1]^X$, and we seek to achieve low distinguishing advantages over all distinguishers in the class $\mathcal{D}$.[2] Alternatively, in the *agnostic* case, we can imagine a situation in which nothing is known about the target predictor p^*, but the performance of the learned predictor is measured relative to the best predictor in $\mathcal{P}$. In both the agnostic and realizable settings, we can also specify whether we measure performance on the worst-case distribution μ over individuals in X, or on some particular distribution μ given to the learner. We call these the *distribution-free* and *distribution-specific* settings, respectively.

Definition 3 (Algorithms in Various Settings) *Given a predictor class $\mathcal{P}$, a class of distinguishers $\mathcal{D}$, a distribution μ, an advantage bound ε, a failure probability bound δ, and a nonnegative integer n, we define the sets of algorithms that solve various OI tasks using n examples as follows:*

$$\mathsf{DS\text{-}R}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu) = \bigcap\nolimits_{p^*\in\mathcal{P}} \mathsf{OI}_n(p^*,\mathcal{D},\varepsilon,\delta,\mu), \qquad \text{(distribution-specific realizable OI)}$$

$$\mathsf{DS\text{-}A}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu) = \bigcap\nolimits_{p^*\in[0,1]^X} \mathsf{OI}_n(p^*,\mathcal{D},\inf_{p\in\mathcal{P}} \mathrm{adv}_{\mu,\mathcal{D}}(p,p^*)+\varepsilon,\delta,\mu),$$

(distribution-specific agnostic OI)

$$\mathsf{DF\text{-}R}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta) = \bigcap\nolimits_{\mu'\in\Delta_X} \mathsf{DS\text{-}R}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu'), \qquad \text{(distribution-free realizable OI)}$$

$$\mathsf{DF\text{-}A}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta) = \bigcap\nolimits_{\mu'\in\Delta_X} \mathsf{DS\text{-}A}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu'). \qquad \text{(distribution-free agnostic OI)}$$

We note that while these learning goals are all defined with respect to some predictor class $\mathcal{P}$, this class simply constrains the possible values of the target predictor p^* (or in the agnostic case, constrains the predictors used to measure the performance of the returned predictor). In particular we do not require any of the OI algorithms to be proper, i.e. always output some $p \in \mathcal{P}$, despite the fact that some of the algorithms discussed in our proofs happen to satisfy this property.

Definition 4 (Sample complexity) *Given a predictor class $\mathcal{P}$, a class of distinguishers $\mathcal{D}$, a distribution μ, an advantage bound ε, a failure probability bound δ, we define the sample complexity of various OI tasks as follows:*

$$\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu) = \inf\{n\in\mathbb{Z}_{\geq 0} : \mathsf{DS\text{-}R}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu)\neq\emptyset\},$$
$$\mathsf{SAMP\text{-}DS\text{-}A}(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu) = \inf\{n\in\mathbb{Z}_{\geq 0} : \mathsf{DS\text{-}A}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu)\neq\emptyset\},$$
$$\mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P},\mathcal{D},\varepsilon,\delta) = \inf\{n\in\mathbb{Z}_{\geq 0} : \mathsf{DF\text{-}R}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta)\neq\emptyset\},$$
$$\mathsf{SAMP\text{-}DF\text{-}A}(\mathcal{P},\mathcal{D},\varepsilon,\delta) = \inf\{n\in\mathbb{Z}_{\geq 0} : \mathsf{DF\text{-}A}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta)\neq\emptyset\}.$$

It is clear from the definition that the following monotonicity properties hold: for $\mathcal{P}' \subseteq \mathcal{P}, \mathcal{D}' \subseteq \mathcal{D}, \varepsilon' \geq \varepsilon, \delta' \geq \delta$,

$$\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}',\mathcal{D}',\varepsilon',\delta',\mu) \leq \mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu),$$
$$\mathsf{SAMP\text{-}DS\text{-}A}(\mathcal{P}',\mathcal{D},\varepsilon',\delta',\mu) \leq \mathsf{SAMP\text{-}DS\text{-}A}(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu),$$
$$\mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P}',\mathcal{D}',\varepsilon',\delta') \leq \mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P},\mathcal{D},\varepsilon,\delta),$$

2. Throughout the paper, we implicitly assume that all predictor classes and distinguisher classes are non-empty.

$$\mathsf{SAMP\text{-}DF\text{-}A}(\mathcal{P}', \mathcal{D}, \varepsilon', \delta') \leq \mathsf{SAMP\text{-}DF\text{-}A}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta). \tag{6}$$

Note that the sample complexities in the agnostic setting are not guaranteed to be monotone w.r.t. $\mathcal{D}$ (see Section 6.2). It is also clear from definition that

$$\begin{aligned} \mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) &\leq \mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta), \\ \mathsf{SAMP\text{-}DS\text{-}A}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) &\leq \mathsf{SAMP\text{-}DF\text{-}A}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta). \end{aligned} \tag{7}$$

2.1.2. NO-ACCESS DISTINGUISHERS AS FUNCTIONS OF INDIVIDUALS

In the standard definition of OI, distinguishers are thought of as randomized algorithms that take as input an individual-outcome pair $(x, o) \in X \times \{0, 1\}$ and output ACCEPT or REJECT. However, there is a natural way to transform every no-access distinguisher d into a function f_d that maps every individual $x \in X$ to a label $y \in [-1, 1]$ in such a way that the distinguishing advantage of d can be recovered from f_d.

In particular, given a randomized distinguisher d, we define $f_d : X \to [-1, 1]$ such that

$$f_d(x) = \Pr[d(x, 1) = \text{ACCEPT}] - \Pr[d(x, 0) = \text{ACCEPT}] \quad \text{for all } x \in X.$$

Given the function f_d, we show that we can always recover the distinguishing advantage of the original distinguisher d:

Lemma 5 *For any two predictors $p_1, p_2 : X \to [0, 1]$,*

$$\mathrm{adv}_{\mu,d}(p_1, p_2) = |\,\mathbb{E}_{x\sim\mu}[f_d(x)(p_1(x) - p_2(x))]|.$$

Proof For any $x \in X$ and any two possible outcomes $(o_1, o_2) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, it is easily verifiable that

$$\Pr[d(x, o_1) = \text{ACCEPT}] - \Pr[d(x, o_2) = \text{ACCEPT}] = f_d(x)(o_1 - o_2),$$

where the probabilities are over the internal randomness of the distinguisher d. The lemma is proved by the following chain of equations:

$$\begin{aligned} &\mathrm{adv}_{\mu,d}(p_1, p_2) \\ &= \left| \Pr_{(x,o)\sim\mu_{p_1}}[d(x, o) = \text{ACCEPT}] - \Pr_{(x,o)\sim\mu_{p_2}}[d(x, o) = \text{ACCEPT}] \right| \\ &= \left| \mathbb{E}_{x\sim\mu, o_1\sim\mathrm{Ber}(p_1(x)), o_2\sim\mathrm{Ber}(p_2(x))}[\Pr[d(x, o_1) = \text{ACCEPT}] - \Pr[d(x, o_2) = \text{ACCEPT}]] \right| \\ &= \left| \mathbb{E}_{x\sim\mu, o_1\sim\mathrm{Ber}(p_1(x)), o_2\sim\mathrm{Ber}(p_2(x))}[f_d(x)(o_1 - o_2)] \right| \\ &= \left| \mathbb{E}_{x\sim\mu}[f_d(x)(p_1(x) - p_2(x))] \right|. \end{aligned}$$

■

Note that the transformation from a distinguisher d to the function $f_d \in [-1, 1]^X$ is onto: given any function $f \in [-1, 1]^X$, we can construct a distinguisher d such that $f_d = f$ in the following way: if $f(x) \geq 0$, distinguisher d accepts $(x, 1)$ with probability $f(x)$, and accepts $(x, 0)$ with

probability 0; if $f(x) < 0$, distinguisher d accepts $(x, 0)$ with probability $-f(x)$, and accepts $(x, 1)$ with probability 0.

Lemma 5 shows that all distinguishers d mapped to the same function f_d have equal distinguishing advantages. It also shows that no-access OI has the same error objective as *multiaccuracy* considered by Kim et al. (2019). From now on, when we refer to a distinguisher d, it should be interpreted as the function $f_d \in [-1, 1]^X$. Similarly, we think of a distinguisher class $\mathcal{D}$ as a non-empty subset of $[-1, 1]^X$.

2.2. Inner Product, Norm, and Covering Number

The set $\mathbb{R}^X$ of all real-valued functions on X is naturally a linear space: for all $f_1, f_2 \in \mathbb{R}^X$, we define $f_1 + f_2 = g \in \mathbb{R}^X$ where $g(x) = f_1(x) + f_2(x)$ for all $x \in X$, and for $f \in \mathbb{R}^X$ and $r \in \mathbb{R}$, we define $rf = h \in \mathbb{R}^X$ where $h(x) = rf(x)$ for all $x \in X$.

For any function class $\mathcal{F} \subseteq \mathbb{R}^X$ and a real number r, we use $r\mathcal{F}$ to denote $\{rf : f \in \mathcal{F}\}$ and use $-\mathcal{F}$ to denote $(-1)\mathcal{F}$. For any two function classes $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathbb{R}^X$, we define

$$\mathcal{F}_1 + \mathcal{F}_2 = \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}, \text{and}$$
$$\mathcal{F}_1 - \mathcal{F}_2 = \{f_1 - f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}.$$

For any non-empty function class $\mathcal{F} \subseteq \mathbb{R}^X$, we say a function $f \in \mathbb{R}^X$ is in the convex hull of $\mathcal{F}$ if there exist $n \in \mathbb{Z}_{>0}$, $f_1, \ldots, f_n \in \mathcal{F}$ and $r_1, \ldots, r_n \in \mathbb{R}_{\geq 0}$ such that $r_1 + \cdots + r_n = 1$ and $f = r_1 f_1 + \cdots + r_n f_n$. We use $\bar{\mathcal{F}}$ to denote the symmetric convex hull of $\mathcal{F}$ consisting of all functions in the convex hull of $\mathcal{F} \cup (-\mathcal{F})$. When $\mathcal{F}$ is empty, we define $\bar{\mathcal{F}}$ to be the class consisting of only the all zeros function.

We say a function $f \in \mathbb{R}^X$ is bounded if there exists $M > 0$ such that $|f(x)| \leq M$ for all $x \in X$. We say a function class $\mathcal{F} \subseteq \mathbb{R}^X$ is bounded if there exists $M > 0$ such that $|f(x)| \leq M$ for all $f \in \mathcal{F}$ and $x \in X$.

For two bounded functions $f_1, f_2 \in \mathbb{R}^X$, we define their inner product w.r.t. distribution μ as

$$\langle f_1, f_2 \rangle_\mu = \mathbb{E}_{x \sim \mu}[f_1(x) f_2(x)].$$

Although we call the above quantity an inner product for simplicity, it may not be positive definite ($\langle f, f \rangle_\mu = 0$ need not imply $f(x) = 0$ for all $x \in X$). For a non-empty bounded function class $\mathcal{F}_1 \subseteq \mathbb{R}^X$ and a bounded function $f \in \mathbb{R}^X$, we define the dual Minkowski norm of f w.r.t. $\mathcal{F}_1$ to be

$$\|f\|_{\mu, \mathcal{F}_1} = \sup_{f_1 \in \mathcal{F}_1} |\langle f, f_1 \rangle_\mu|.$$

If $\mathcal{F}_1$ is empty, we define $\|f\|_{\mu, \mathcal{F}_1} = 0$. The norm $\|\cdot\|_{\mu, \mathcal{F}_1}$ is technically only a semi-norm as it may not be positive definite, but whenever it is positive definite, it is the dual norm of the Minkowski norm induced by (the closure of) $\bar{\mathcal{F}}_1$ for finite X (see e.g. Nikolov et al., 2013, Section 2.1).

For a non-empty bounded function class $\mathcal{F}_2 \subseteq \mathbb{R}^X$ and $\varepsilon \geq 0$, we say a subset $\mathcal{F}_2' \subseteq \mathcal{F}_2$ is an ε-covering of $\mathcal{F}_2$ w.r.t. the norm $\|\cdot\|_{\mu, \mathcal{F}_1}$ if for every $f_2 \in \mathcal{F}_2$, there exists $f_2' \in \mathcal{F}_2'$ such that $\|f_2 - f_2'\|_{\mu, \mathcal{F}_1} \leq \varepsilon$. We define the covering number $N_{\mu, \mathcal{F}_1}(\mathcal{F}_2, \varepsilon)$ of $\mathcal{F}_2$ w.r.t. the norm $\|\cdot\|_{\mu, \mathcal{F}_1}$ to be the minimum size of such an ε-covering of $\mathcal{F}_2'$. We refer to the logarithm of the covering number, $\log N_{\mu, \mathcal{F}_1}(\mathcal{F}_2, \varepsilon)$, as the metric entropy.

The following basic facts about the covering number are very useful:

Lemma 6 *We have the following facts:*

1. *if* $\hat{\mathcal{F}}_1 \subseteq \mathcal{F}_1 \subseteq \mathbb{R}^X$*, then* $N_{\mu,\hat{\mathcal{F}}_1}(\mathcal{F}_2, \varepsilon) \leq N_{\mu,\mathcal{F}_1}(\mathcal{F}_2, \varepsilon)$*;*
2. $N_{\mu,\mathcal{F}_1}(\mathcal{F}_2, \varepsilon) = N_{\mu,\bar{\mathcal{F}}_1}(\mathcal{F}_2, \varepsilon)$*;*
3. *for every bounded function* $f \in \mathbb{R}^X$*,* $N_{\mu,\mathcal{F}_1}(\mathcal{F}_2 + \{f\}, \varepsilon) = N_{\mu,\mathcal{F}_1}(\mathcal{F}_2, \varepsilon)$*;*
4. *for every* $a, b \in \mathbb{R}_{>0}$*,* $N_{\mu,a\mathcal{F}_1}(b\mathcal{F}_2, ab\varepsilon) = N_{\mu,\mathcal{F}_1}(\mathcal{F}_2, \varepsilon)$.

2.3. Distinguishing Advantages as Inner Products and Norms

The inner products and norms provide convenient ways for describing distinguishing advantages. Given a distinguisher $d \in [-1,1]^X$ and two predictors $p_1, p_2 \in [0,1]^X$, Lemma 5 tells us that

$$\begin{aligned}&\mathrm{adv}_{\mu,d}(p_1,p_2) = |\langle d, p_1 - p_2\rangle_\mu|, \quad \text{and thus}\\ &\mathrm{adv}_{\mu,\mathcal{D}}(p_1,p_2) = \|p_1 - p_2\|_{\mu,\mathcal{D}} \quad \text{for all distinguisher classes } \mathcal{D} \subseteq [-1,1]^X.\end{aligned}$$

Using these representations for advantages, we can prove the following lemma relating advantages to the ℓ_1 error:

Lemma 7 *It holds that* $\mathrm{adv}_{\mu,\mathcal{D}}(p_1,p_2) = \|p_1 - p_2\|_{\mu,\mathcal{D}} \leq \mathbb{E}_{x\sim\mu}[|p_1(x) - p_2(x)|]$*. Moreover, when* $\{-1,1\}^X \subseteq \mathcal{D}$*,* $\mathrm{adv}_{\mu,\mathcal{D}}(p_1,p_2) = \mathbb{E}_{x\sim\mu}[|p_1(x) - p_2(x)|]$.

Proof To prove the first statement, we recall that

$$\mathrm{adv}_{\mu,\mathcal{D}}(p_1,p_2) = \|p_1 - p_2\|_{\mu,\mathcal{D}} = \sup_{d\in\mathcal{D}} |\,\mathbb{E}_{x\sim\mu}[(p_1(x) - p_2(x))d(x)]|.$$

Because $d(x) \in [-1,1]$ for all $d \in \mathcal{D}$ and $x \in X$, we are guaranteed that $|(p_1(x) - p_2(x))d(x)| \leq |p_1(x) - p_2(x)|$, which gives

$$\sup_{d\in\mathcal{D}} |\,\mathbb{E}_{x\sim\mu}[(p_1(x) - p_2(x))d(x)]| \leq \sup_{d\in\mathcal{D}} \mathbb{E}_{x\sim\mu}[|(p_1(x) - p_2(x))d(x)|] \leq \mathbb{E}_{x\sim\mu}[|p_1(x) - p_2(x)|],$$

as desired.

For the second statement, consider the distinguisher d defined such that $d(x) = 1$ if $p_1(x) \geq p_2(x)$ and $d(x) = -1$ otherwise. For all $x \in X$, distinguisher d satisfies

$$(p_1(x) - p_2(x))d(x) = |p_1(x) - p_2(x)|.$$

Therefore,

$$\mathrm{adv}_{\mu,d}(p_1,p_2) = |\,\mathbb{E}_{x\sim\mu}[(p_1(x) - p_2(x))d(x)]| = \mathbb{E}_{x\sim\mu}[|p_1(x) - p_2(x)|].$$

Since $d \in \{-1,1\}^X \subseteq \mathcal{D}$, this proves the second statement. ■

2.4. Fat-Shattering Dimension

Given a function class $\mathcal{F} \subseteq \mathbb{R}^X$ and $\gamma \geq 0$, we say $x_1, \ldots, x_n \in X$ are γ-fat shattered by $\mathcal{F}$ w.r.t. $r_1, \ldots, r_n \in \mathbb{R}$ if for every $(b_1, \ldots, b_n) \in \{-1, 1\}^n$, there exists $f \in \mathcal{F}$ such that

$$b_i(f(x_i) - r_i) \geq \gamma \quad \text{for all } i \in \{1, \ldots, n\}.$$

We sometimes omit the mention of $r_1, \ldots, r_n$ and say $x_1, \ldots, x_n$ is γ-fat shattered by $\mathcal{F}$ if such $r_1, \ldots, r_n$ exist. The γ-fat-shattering dimension of $\mathcal{F}$ introduced first by Kearns and Schapire (1994) is defined to be

$$\mathsf{fat}_{\mathcal{F}}(\gamma) = \sup\{n \in \mathbb{Z}_{\geq 0} : \text{there exist } x_1, \ldots, x_n \in X \text{ that are } \gamma\text{-fat shattered by } \mathcal{F}\}.$$

3. Sample Complexity of Distribution-Specific Realizable OI

In this section, we give lower and upper bounds for the sample complexity of distribution-specific realizable OI for every given predictor class $\mathcal{P}$, distinguisher class $\mathcal{D}$, and distribution μ over individuals. Our lower bound is based on the metric entropy of $\mathcal{P}$ w.r.t. the norm $\|\cdot\|_{\mu,\mathcal{D}}$, whereas in our upper bound the roles of $\mathcal{P}$ and $\mathcal{D}$ are flipped. In the next section, we remove this role flip and give a complete characterization of the sample complexity using a version of metric entropy duality implied from combining our lower and upper bounds.

3.1. Lower Bound

We prove the following lemma showing that the sample complexity of distribution-specific realizable OI is lower bounded by the metric entropy of the predictor class $\mathcal{P}$ w.r.t. the dual Minkowski norm $\|\cdot\|_{\mu,\mathcal{D}}$ defined by the distinguisher class $\mathcal{D}$. This lemma generalizes (Benedek and Itai, 1991, Lemma 4.8), which considered the special case where every predictor is a binary classifier, and the distinguisher class $\mathcal{D}$ contains all possible distinguishers ($\mathcal{D} = [-1, 1]^X$).

Lemma 8 *For every predictor class $\mathcal{P} \subseteq [0, 1]^X$, every distinguisher class $\mathcal{D} \subseteq [-1, 1]^X$, every distribution $\mu \in \Delta_X$, every advantage bound $\varepsilon > 0$, and every failure probability bound $\delta \in (0, 1)$, the following sample complexity lower bound holds for distribution-specific realizable OI:*

$$\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \geq \log((1 - \delta)N_{\mu,\mathcal{D}}(\mathcal{P}, 2\varepsilon)).$$

Proof Define $M = N_{\mu,\mathcal{D}}(\mathcal{P}, 2\varepsilon)$. Let $\mathcal{P}'$ be the maximum-size subset of $\mathcal{P}$ such that

$$\|p_1 - p_2\|_{\mu,\mathcal{D}} > 2\varepsilon \quad \text{for all distinct } p_1, p_2 \in \mathcal{P}'. \tag{8}$$

It is clear that $|\mathcal{P}'| \geq M$ because otherwise $\mathcal{P}'$ is not a 2ε-covering of $\mathcal{P}$ and we can add one more predictor into $\mathcal{P}'$ without violating (8), a contradiction with the maximality of $|\mathcal{P}'|$.

Let n be a nonnegative integer such that $\mathsf{DS\text{-}R}_n(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \neq \emptyset$. Our goal is to prove

$$n \geq \log((1 - \delta)M). \tag{9}$$

Let $\mathcal{A}$ be an algorithm in $\mathsf{DS\text{-}R}_n(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu)$. We draw a predictor p^* uniformly at random from $\mathcal{P}'$, and draw examples $(x_1, o_1), \ldots, (x_n, o_n)$ i.i.d. from μ_{p^*}. We say algorithm $\mathcal{A}$ succeeds if it outputs p such that $\|p - p^*\|_{\mu,\mathcal{D}} \leq \varepsilon$. By assumption, when $(x_1, o_1), \ldots, (x_n, o_n)$ are given as

input, algorithm $\mathcal{A}$ succeeds with probability at least $1-\delta$. Now instead of drawing $x_1, \ldots, x_n$ i.i.d. from μ, we fix them so that the success probability is maximized. In other words, we can find fixed $x_1, \ldots, x_n \in X$ such that if we run algorithm $\mathcal{A}$ on examples $(x_1, o_1), \ldots, (x_n, o_n)$ where $o_i \sim \mathrm{Ber}(p^*(x_i))$ and p^* is chosen uniformly at random from $\mathcal{P}'$, the algorithm succeeds with probability at least $1-\delta$. Similarly, we can fix the internal randomness of algorithm $\mathcal{A}$ and assume that $\mathcal{A}$ is deterministic without decreasing its success probability on $(x_1, o_1), \ldots, (x_n, o_n)$. Now algorithm $\mathcal{A}$ has at most 2^n possible inputs, and thus has at most 2^n possible outputs. No output can be the success output for two different choices of p^* from $\mathcal{P}'$ because of (8). Therefore, the success probability of algorithm $\mathcal{A}$ is at most $2^n/M$, and thus

$$2^n/M \geq 1-\delta.$$

This implies (9), as desired. ■

3.2. Upper Bound

We give an algorithm for distribution-specific realizable OI to prove a sample complexity upper bound for it. Before we describe our algorithm, let us briefly discuss the empirical risk minimization algorithm (Algorithm 1). Benedek and Itai (1991, Proof of Lemma 4.6) showed that this natural algorithm works in the special case where 1) every predictor in $\mathcal{P}$ is a binary classifier, and 2) the distinguisher class $\mathcal{D}$ contains all possible distinguishers. When both 1) and 2) are satisfied, the algorithm gives a sample complexity upper bound of

$$O((1/\varepsilon)^{O(1)} \log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/2)), \tag{10}$$

which would give a satisfactory sample complexity characterization when combined with our lower bound in Lemma 8. However, in Appendix A, we show that the algorithm fails when only one of the two conditions 1) and 2) (no matter which) is true.[3] Since neither 1) nor 2) is guaranteed to hold in the distribution-specific realizable OI setting, we use a new algorithm (Algorithm 2) to prove our sample complexity upper bound (Lemma 10) where the roles of $\mathcal{P}$ and $\mathcal{D}$ flip compared to (10). In Section 4, we show how to flip them back to get a sample complexity characterization for distribution-specific realizable OI.

Our sample complexity upper bound is based on the following analysis of Algorithm 2:

Lemma 9 *For every predictor class $\mathcal{P} \subseteq [0,1]^X$, every distinguisher class $\mathcal{D} \subseteq [-1,1]^X$, every distribution $\mu \in \Delta_X$, every advantage bound $\varepsilon \in (0,1)$, and every failure probability bound $\delta \in (0,1)$, there exists a positive integer*

$$n \leq O\big(\varepsilon^{-2}\big(\log N_{\mu,\mathcal{Q}}(\mathcal{D}, \varepsilon/2) + \log(2/\delta)\big)\big) \tag{11}$$

3. There is a variant of Algorithm 1 that minimizes $\mathsf{loss}(p)$ over the entire predictor class $\mathcal{P}$ instead of the covering $\mathcal{P}'$. As discussed in (Benedek and Itai, 1991), this variant is not guaranteed to give a sample complexity upper bound close to (10) even under both conditions 1) and 2). Changing the definition of $\mathsf{loss}(p)$ to $\sup_{d\in\mathcal{D}} |\langle p, d\rangle_\mu - \frac{1}{n}\sum_{i=1}^n d(x_i)o_i|$ (mimicking Algorithm 2) also makes Algorithm 1 fail under both conditions 1) and 2). To see this, suppose μ is the uniform distribution over a finite domain X, $\mathcal{P} = \{p_0, p_1\}$ where $p_0(x) = 0$ and $p_1(x) = 1$ for every $x \in X$, and $\mathcal{D} = [-1,1]^X$. Assuming $\varepsilon \in (0,1)$ and $n < |X|/10$, when the target predictor p^* is p_1, Algorithm 1 always outputs p_0 on the new loss (note that changing the values of d on $x_1, \ldots, x_n$ can significantly change $\frac{1}{n}\sum_{i=1}^n d(x_i)o_i$, but it never changes $\langle p, d\rangle_\mu$ by more than $2n/|X|$).

Algorithm 1: Empirical Risk Minimization

Parameters: predictor class $\mathcal{P}$, distinguisher class $\mathcal{D}$, distribution μ, MDA bound ε, positive integer n.

Input : examples $(x_1, o_1), \ldots, (x_n, o_n)$.

Output : predictor $p \in \mathcal{P}$.

$\mathcal{P}' \leftarrow$ minimum-size $\varepsilon/2$-covering of $\mathcal{P}$ w.r.t. norm $\|\cdot\|_{\mu,\mathcal{D}}$;

return $p \in \mathcal{P}'$ that minimizes the empirical error

$$\mathsf{loss}(p) := \sup_{d \in \mathcal{D}} \left| \frac{1}{n} \sum_{i=1}^{n} d(x_i)(p(x_i) - o_i) \right|;$$

Algorithm 2: Distinguisher Covering

Parameters: predictor class $\mathcal{P}$, distinguisher class $\mathcal{D}$, distribution μ, MDA bound ε, positive integer n.

Input : examples $(x_1, o_1), \ldots, (x_n, o_n)$.

Output : predictor $p \in \mathcal{P}$.

$\mathcal{Q} \leftarrow \mathcal{P} - \mathcal{P}$; `/* Recall that` $\mathcal{P} - \mathcal{P} = \{p_1 - p_2 : p_1, p_2 \in \mathcal{P}\}$. `*/`

$\mathcal{D}' \leftarrow$ minimum-size $\varepsilon/2$-covering of $\mathcal{D}$ w.r.t. norm $\|\cdot\|_{\mu,\mathcal{Q}}$;

return $p \in \mathcal{P}$ that $\varepsilon/16$-minimizes

$$\mathsf{loss}(p) := \sup_{d \in \mathcal{D}'} \left| \langle p, d \rangle_\mu - \frac{1}{n} \sum_{i=1}^{n} d(x_i) o_i \right|,$$

i.e., $\mathsf{loss}(p) \leq \inf_{p' \in \mathcal{P}} \mathsf{loss}(p') + \varepsilon/16$;

$$\leq O\big(\varepsilon^{-2}\big(\log N_{\mu,\mathcal{P}}(\mathcal{D},\varepsilon/4)+\log(2/\delta)\big)\big) \tag{12}$$

such that Algorithm 2 belongs to

$$\bigcap\nolimits_{p^*\in[0,1]^X}\mathsf{OI}_n(p^*,\mathcal{D},3\inf_{p\in\mathcal{P}}\mathrm{adv}_{\mu,\mathcal{D}}(p,p^*)+\varepsilon,\delta,\mu),$$

where $\mathcal{Q}=\mathcal{P}-\mathcal{P}=\{p_1-p_2:p_1,p_2\in\mathcal{P}\}$.

Proof We first note that $\|f\|_{\mu,\mathcal{Q}}\leq 2\|f\|_{\mu,\mathcal{P}}$ for all bounded functions $f\in\mathbb{R}^X$, which implies that

$$N_{\mu,\mathcal{Q}}(\mathcal{D},\varepsilon/2)\leq N_{\mu,\mathcal{P}}(\mathcal{D},\varepsilon/4).$$

This proves inequality (12).

Define $M=N_{\mu,\mathcal{Q}}(\mathcal{D},\varepsilon/2)$ and $\varepsilon_0=\inf_{p\in\mathcal{P}}\mathrm{adv}_{\mu,\mathcal{D}}(p,p^*)$ for an arbitrary $p^*\in[0,1]^X$. It remains to prove that Algorithm 2 belongs to

$$\bigcap\nolimits_{p^*\in[0,1]^X}\mathsf{OI}_n(p^*,\mathcal{D},3\varepsilon_0+\varepsilon,\delta,\mu)$$

for some $n=O(\varepsilon^{-2}(\log M+\log(2/\delta)))$ determined below. Given n examples $(x_1,o_1),\ldots,(x_n,o_n)$, we define $K(d)=\frac{1}{n}\sum_{i=1}^n o_i d(x_i)$ for every distinguisher $d\in\mathcal{D}'$, where $\mathcal{D}'$ is the minimum-size $\varepsilon/2$ covering of $\mathcal{P}$ computed in Algorithm 2. By definition, $|\mathcal{D}'|=M$, so by the Chernoff bound and the union bound, for some $n=O(\varepsilon^{-2}(\log M+\log(2/\delta)))$, with probability at least $1-\delta$,

$$|K(d)-\langle p^*,d\rangle_\mu|\leq\varepsilon/16\quad\text{for all } d\in\mathcal{D}'. \tag{13}$$

Assuming that (13) is true, it suffices to prove that the output p of Algorithm 2 satisfies $\|p-p^*\|_{\mu,\mathcal{D}}\leq 3\varepsilon_0+\varepsilon$. Let $\tilde{p}\in\mathcal{P}$ be a predictor that satisfies $\|\tilde{p}-p^*\|_{\mu,\mathcal{D}}\leq\varepsilon_0+\varepsilon/16$. We have

$$\sup_{d\in\mathcal{D}'}|K(d)-\langle\tilde{p},d\rangle_\mu|\leq\sup_{d\in\mathcal{D}'}|\langle p^*,d\rangle_\mu-\langle\tilde{p},d\rangle_\mu|+\varepsilon/16\leq\varepsilon_0+\varepsilon/8.$$

The output predictor p of Algorithm 2 satisfies

$$\sup_{d\in\mathcal{D}'}|K(d)-\langle p,d\rangle_\mu|\leq\sup_{d\in\mathcal{D}'}|K(d)-\langle\tilde{p},d\rangle_\mu|+\varepsilon/16\leq\varepsilon_0+3\varepsilon/16.$$

Therefore,

$$\begin{aligned}\sup_{d\in\mathcal{D}'}|\langle p-\tilde{p},d\rangle_\mu|&\leq\sup_{d\in\mathcal{D}'}|\langle p,d\rangle_\mu-K(d)|+\sup_{d\in\mathcal{D}'}|\langle\tilde{p},d\rangle_\mu-K(d)|\\&\leq(\varepsilon_0+\varepsilon/8)+(\varepsilon_0+3\varepsilon/16)\\&\leq 2\varepsilon_0+5\varepsilon/16.\end{aligned}$$

Since $p-\tilde{p}\in\mathcal{Q}$ and $\mathcal{D}'$ is an $\varepsilon/2$-covering w.r.t. $\|\cdot\|_{\mu,\mathcal{Q}}$,

$$\|p-\tilde{p}\|_{\mu,\mathcal{D}}=\sup_{d\in\mathcal{D}}|\langle p-\tilde{p},d\rangle_\mu|\leq\sup_{d\in\mathcal{D}'}|\langle p-\tilde{p},d\rangle_\mu|+\varepsilon/2\leq 2\varepsilon_0+13\varepsilon/16.$$

Finally,

$$\|p-p^*\|_{\mu,\mathcal{D}}\leq\|p-\tilde{p}\|_{\mu,\mathcal{D}}+\|\tilde{p}-p^*\|_{\mu,\mathcal{D}}\leq(2\varepsilon_0+13\varepsilon/16)+(\varepsilon_0+\varepsilon/16)\leq 3\varepsilon_0+\varepsilon,$$

as desired. ■

We are now ready to state and prove our sample complexity upper bound for distribution-specific realizable OI.

Lemma 10 *For every predictor class $\mathcal{P} \subseteq [0,1]^X$, every distinguisher class $\mathcal{D} \subseteq [-1,1]^X$, every distribution $\mu \in \Delta_X$, every advantage bound $\varepsilon > 0$, and every failure probability bound $\delta \in (0,1)$, the following sample complexity upper bound holds for distribution-specific realizable OI:*

$$\begin{aligned}\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu) &\le O\big(\varepsilon^{-2}\big(\log N_{\mu,\mathcal{Q}}(\mathcal{D},\varepsilon/2)+\log(2/\delta)\big)\big)\\ &\le O\big(\varepsilon^{-2}\big(\log N_{\mu,\mathcal{P}}(\mathcal{D},\varepsilon/4)+\log(2/\delta)\big)\big),\end{aligned}$$

where $\mathcal{Q} = \mathcal{P} - \mathcal{P} = \{p_1 - p_2 : p_1, p_2 \in \mathcal{P}\}$.

Proof When $\varepsilon \ge 1$, we have $\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu) = 0$ and the lemma is trivially true. We assume $\varepsilon \in (0,1)$ henceforth.

For every $p^* \in \mathcal{P}$, we have $\inf_{p\in\mathcal{P}} \mathrm{adv}_{\mu,\mathcal{D}}(p,p^*) = 0$, so by Lemma 9, there exists a positive integer n satisfying (11) and (12) such that

$$\begin{aligned}\emptyset \neq \bigcap\nolimits_{p^*\in[0,1]^X} \mathsf{OI}_n(p^*,\mathcal{D},3\inf_{p\in\mathcal{P}}\mathrm{adv}_{\mu,\mathcal{D}}(p,p^*)+\varepsilon,\delta,\mu) &\subseteq \bigcap\nolimits_{p^*\in\mathcal{P}} \mathsf{OI}_n(p^*,\mathcal{D},\varepsilon,\delta,\mu)\\ &= \mathsf{DS\text{-}R}_n(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu).\end{aligned}$$

This completes the proof by the definition of $\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P},\mathcal{D},\varepsilon,\delta,\mu)$. ■

4. Metric Entropy Duality

In the previous section, we proved lower and upper bounds for the sample complexity of distribution-specific realizable OI, but these bounds do not yet give a satisfactory sample complexity characterization because of the exchanging roles of the predictor class $\mathcal{P}$ and the distinguisher class $\mathcal{D}$ in the lower and upper bounds. In this section, we solve the issue by proving the following theorem:

Theorem 11 *There exists an absolute constant $c \ge 1$ with the following property. For $M_1, M_2 > 0$, let $\mathcal{F}_1 \subseteq [-M_1, M_1]^X$ and $\mathcal{F}_2 \subseteq [-M_2, M_2]^X$ be two non-empty bounded function classes. For any distribution $\mu \in \Delta_X$ and any $\varepsilon > 0$, it holds that*

$$\log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1,\varepsilon) \le c(M_1M_2/\varepsilon)^2(1+\log N_{\mu,\mathcal{F}_1}(\mathcal{F}_2,\varepsilon/8)).$$

Before we prove this theorem, we note that it has a similar statement to the long-standing metric entropy duality conjecture proposed first by Pietsch (1972). The conjecture can be stated as follows using our notations (for other equivalent statements of the conjecture, see e.g. Artstein et al., 2004a):

Conjecture 12 (Pietsch (1972)) *There exist absolute constants $c_1, c_2 \ge 1$ with the following property. For any two bounded function classes $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathbb{R}^X$ over a non-empty finite set X, if $\mathcal{F}_1$ and $\mathcal{F}_2$ are convex and symmetric (i.e. $\bar{\mathcal{F}}_1 = \mathcal{F}_1, \bar{\mathcal{F}}_2 = \mathcal{F}_2$), then for any distribution $\mu \in \Delta_X$ and any $\varepsilon > 0$,*

$$\log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1,\varepsilon) \le c_1 \log N_{\mu,\mathcal{F}_1}(\mathcal{F}_2,\varepsilon/c_2).$$

Compared to Conjecture 12, our Theorem 11 gives a variant of metric entropy duality which does not require $\mathcal{F}_1$ and $\mathcal{F}_2$ to be convex and symmetric, but has the constant c_1 in Conjecture 12 replaced by a quantity dependent on the granularity ε and the scale of the functions in $\mathcal{F}_1$ and $\mathcal{F}_2$. Since the predictor class $\mathcal{P}$ and the distinguisher class $\mathcal{D}$ are not in general convex and symmetric,

our Theorem 11 is more convenient for proving sample complexity characterizations for OI (see Theorem 13). In Lemma 14, we show that the quadratic dependence on $M_1 M_2/\varepsilon$ in Theorem 11 is nearly tight.

Below we prove Theorem 11 by combining our lower and upper bounds in the previous section.

Proof By Lemma 6 Item 4, we can assume w.l.o.g. that $M_1 = M_2 = 1$. Define $\mathcal{D} = \mathcal{F}_2 \subseteq [-1,1]^X$, $\mathcal{P} = \{(1+f)/2 : f \in \mathcal{F}_1\} \subseteq [0,1]^X$, and $\mathcal{Q} = \mathcal{P} - \mathcal{P}$. Combining Lemma 8 and Lemma 10 with $\delta = 1/3$ and ε replaced by $\varepsilon/4$, there exists a constant $c \geq 1$ such that

$$\log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/2) \leq c\varepsilon^{-2}(1 + \log N_{\mu,\mathcal{Q}}(\mathcal{D}, \varepsilon/8)).$$

Using Lemma 6 Items 3 and 4,

$$\begin{aligned}
\log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \varepsilon) &= \log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/2) \\
&\leq c\varepsilon^{-2}(1 + \log N_{\mu,\mathcal{Q}}(\mathcal{D}, \varepsilon/8)) \\
&\leq c\varepsilon^{-2}(1 + \log N_{\mu,\mathcal{F}_1}(\mathcal{D}, \varepsilon/8)) \qquad (14) \\
&= c\varepsilon^{-2}(1 + \log N_{\mu,\mathcal{F}_1}(\mathcal{F}_2, \varepsilon/8)),
\end{aligned}$$

as desired. Here, inequality (14) holds because $\|f\|_{\mu,\mathcal{Q}} \leq \|f\|_{\mu,\mathcal{F}_1}$ for every bounded function $f \in \mathbb{R}^X$. ■

We are now ready to state and prove our sample complexity characterizations for distribution-specific realizable OI.

Theorem 13 *For every predictor class $\mathcal{P} \subseteq [0,1]^X$, every distinguisher class $\mathcal{D} \subseteq [-1,1]^X$, every distribution $\mu \in \Delta_X$, every advantage bound $\varepsilon > 0$, and every failure probability bound $\delta \in (0,1)$, the following sample complexity characterizations hold for distribution-specific realizable OI:*

$$\begin{aligned}
&\log N_{\mu,\mathcal{D}}(\mathcal{P}, 2\varepsilon) + \log(1-\delta) \\
\leq\ &\textsf{SAMP-DS-R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \\
\leq\ &O(\varepsilon^{-4} \log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/32) + \varepsilon^{-2}\log(2/\delta)),
\end{aligned}$$

and

$$\begin{aligned}
&\Omega(\varepsilon^2 \log N_{\mu,\mathcal{P}}(\mathcal{D}, 16\varepsilon)) - 1 + \log(1-\delta) \\
\leq\ &\textsf{SAMP-DS-R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \\
\leq\ &O(\varepsilon^{-2} \log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon/4) + \varepsilon^{-2}\log(2/\delta)).
\end{aligned}$$

Since $\log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon)$ is monotone w.r.t. $\mathcal{D}$ (Lemma 6 Item 1), compared to ℓ_1-error-based learning which corresponds to $\mathcal{D} = [-1,1]^X$ (see Lemma 7), Theorem 13 shows that a selected class $\mathcal{D}$ of distinguishers helps reduce the sample complexity, or allows us to achieve smaller ε (and potentially better performance guarantees) with the same sample size. We prove Theorem 13 below.

Proof We start with the following chain of inequalities:

$$\begin{aligned}
&\Omega(\varepsilon^2 \log N_{\mu,\mathcal{P}}(\mathcal{D}, 16\varepsilon)) - 1 + \log(1-\delta) \\
\leq\ &\log N_{\mu,\mathcal{D}}(\mathcal{P}, 2\varepsilon) + \log(1-\delta) \qquad \text{(by Theorem 11)}
\end{aligned}$$

$$\leq \mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \qquad \text{(by Lemma 8)}$$
$$\leq O(\varepsilon^{-2} \log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon/4) + \varepsilon^{-2} \log(2/\delta)), \tag{15}$$

where the last inequality is by Lemma 10. It remains to show

$$\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \leq O(\varepsilon^{-4} \log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/32) + \varepsilon^{-2} \log(2/\delta)). \tag{16}$$

If $N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/32) \leq 1$, it is clear that $\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) = 0$, so the inequality is trivial. If $N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/32) \geq 2$, we have the following inequality by Theorem 11:

$$\varepsilon^{-2} \log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon/4)) \leq O(\varepsilon^{-4}(1 + \log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/32))) \leq O(\varepsilon^{-4} \log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/32)).$$

Inequality (16) is proved by plugging the inequality above into (15). ■

We complete the section by showing near-tightness of Theorem 11.

Lemma 14 *There exists a constant $c > 0$ such that for all $M_1 > 0, M_2 > 0, \varepsilon \in (0, M_1 M_2/2)$, there exist a ground set X, a distribution μ over X, and function classes $\mathcal{F}_1 \subseteq [-M_1, M_2]^X, \mathcal{F}_2 \subseteq [-M_2, M_2]^X$ such that*

$$\log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \varepsilon) \geq c(1 + \log |\mathcal{F}_2|) \left(\frac{M_1 M_2}{\varepsilon}\right)^2 \left(\log\left(\frac{M_1 M_2}{\varepsilon}\right)\right)^{-1}.$$

Proof By Lemma 6 Item 4, we can assume w.l.o.g. that $M_1 = M_2 = 1$. Now we have $\varepsilon < M_1 M_2/2 = 1/2$, and $1/2\varepsilon^2 > 2$. Let m be the largest integer power of 2 such that $m \leq 1/2\varepsilon^2$. We choose X to be $\{1, \ldots, m\}$, and choose μ to be the uniform distribution over X. Let vec be the bijection from $\mathbb{R}^X$ to $\mathbb{R}^m$ such that $\mathsf{vec}(f) = (f(1), \ldots, f(m)) \in \mathbb{R}^m$ for all $f \in \mathbb{R}^X$. Define H_m to be the set of vectors in $\{-1, 1\}^m$ consisting of the m columns of the Hadamard matrix of size $m \times m$. We choose $\mathcal{F}_1 = [-1, 1]^X$, and $\mathcal{F}_2 = \{\mathsf{vec}^{-1}(v) : v \in H_m\}$. The intuition behind the choice of $\mathcal{F}_2$ is to keep $|\mathcal{F}_2|$ small while making $N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \varepsilon)$ large, which by Lemma 6 Items 1 and 2 roughly corresponds to making the symmetric convex hull $\bar{\mathcal{F}}_2$ large. That is why we use the Hadamard matrix to ensure that the functions f in $\mathcal{F}_2$ achieve the maximum norm (with $f(x) = \pm 1$ for every $x \in X$) and are orthogonal to each other.

Define $B = \{f \in \mathbb{R}^X : \|f\|_{\mu,\mathcal{F}_2} \leq \varepsilon\}$. By the properties of the Hadamard matrix, the functions in $\mathcal{F}_2$ form an orthonormal basis of $\mathbb{R}^X$ w.r.t. the inner product $\langle \cdot, \cdot \rangle_\mu$. This implies

$$B = \left\{ \sum\nolimits_{f \in \mathcal{F}_2} r_f f : r_f \in [-\varepsilon, \varepsilon] \text{ for all } f \in \mathcal{F}_2 \right\}. \tag{17}$$

Let $\mathcal{F}_1' \subseteq \mathcal{F}_1$ be an ε-covering of $\mathcal{F}_1$ w.r.t. norm $\|\cdot\|_{\mu,\mathcal{F}_2}$ such that $|\mathcal{F}_1'| = N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \varepsilon)$. By the definition of ε-covering,

$$\mathcal{F}_1 \subseteq \bigcup\nolimits_{f \in \mathcal{F}_1'} (\{f\} + B).$$

For a function class $\mathcal{F} \subseteq \mathbb{R}^X$, we define $\mathsf{vec}(\mathcal{F}) = \{\mathsf{vec}(f) : f \in \mathcal{F}\} \subseteq \mathbb{R}^m$. Now we have $\mathsf{vec}(\mathcal{F}_1) \subseteq \bigcup_{f \in \mathcal{F}_1'} \mathsf{vec}(\{f\} + B)$, which implies that the volume of $\mathsf{vec}(\mathcal{F}_1)$ is at most $|\mathcal{F}_1'|$ times the volume of $\mathsf{vec}(B)$.

It is clear that $\mathsf{vec}(\mathcal{F}_1) = [-1, 1]^m$ has volume 2^m. By (17),

$$\mathsf{vec}(B) = \left\{ \sum\nolimits_{v \in H_m} r_v v : r_v \in [-\varepsilon, \varepsilon] \text{ for all } v \in H_m \right\}.$$

Since the columns of H_m are orthogonal and have ℓ_2 norm $\sqrt{m}$ in $\mathbb{R}^m$, the volume of $\mathsf{vec}(B)$ is $(2\varepsilon\sqrt{m})^m$. Therefore,

$$2^m \le |\mathcal{F}_1'| \cdot (2\varepsilon\sqrt{m})^m,$$

and thus $|\mathcal{F}_1'| \ge (1/\varepsilon\sqrt{m})^m$. Now we have

$$\log N_{\mu,\mathcal{F}_2}(\mathcal{F}_1, \varepsilon) = \log |\mathcal{F}_1'| \ge m \log(1/\varepsilon\sqrt{m}) \ge \Omega(m) \ge \Omega(1/\varepsilon^2), \tag{18}$$

and

$$\log |\mathcal{F}_2| = \log m \le O(\log(1/\varepsilon)). \tag{19}$$

Combining (18) and (19) proves the lemma. ■

5. Sample Complexity of Distribution-Free OI

In this section, we consider the distribution-free setting where the OI learner has no knowledge of the distribution μ, and its performance is measured on the worst-case distribution μ. We focus on the case where $\mathcal{P} = [0,1]^X$ and characterize the sample complexity of OI in this setting using the fat-shattering dimension of $\mathcal{D}$ (Theorem 15). Without the assumption that $\mathcal{P} = [0,1]^X$, we give a sample complexity *upper bound* for realizable OI in Remark 16 and leave the intriguing question of sample complexity *characterization* to future work.

Theorem 15 *For every distinguisher class $\mathcal{D}$, every advantage bound $\varepsilon \in (0,1)$, and every failure probability bound $\delta \in (0,1)$, the following sample complexity characterization holds for distribution-free OI:*

$$\begin{aligned}
&\mathsf{fat}_{\mathcal{D}}(12\varepsilon)/8 + \log(1-\delta) \\
&\le \mathsf{SAMP\text{-}DF\text{-}R}([0,1]^X, \mathcal{D}, \varepsilon, \delta) \\
&= \mathsf{SAMP\text{-}DF\text{-}A}([0,1]^X, \mathcal{D}, \varepsilon, \delta) \\
&\le O\Big(\varepsilon^{-4}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big).
\end{aligned} \tag{20}$$

This theorem is a direct corollary of the sample complexity upper bound (Lemma 19) and lower bound (Lemma 23) we prove in the remaining of this section.[4] With some modifications, Theorem 15 also extends to *multicalibration* (see Remarks 21 and 24).

Remark 16 *If we drop the assumption that $\mathcal{P} = [0,1]^X$ and assume instead that $\mathcal{D} = [-1,1]^X$, by Lemma 7, our error objective $\mathrm{adv}_{\mu,\mathcal{D}}(\cdot,\cdot)$ becomes the ℓ_1 error. By the ideas and results in (Bartlett et al., 1996; Bartlett and Long, 1995), it holds that*

$$\mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P}, [-1,1]^X, \varepsilon, \delta) \le O\Big(\varepsilon^{-4}\Big(\mathsf{fat}_{\mathcal{P}}(\varepsilon^2/20)(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big).$$

Combining this with (20) and using the monotonicity (6) of the sample complexity of realizable OI, without assuming $\mathcal{P} = [0,1]^X$ or $\mathcal{D} = [-1,1]^X$, we have

$$\mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta) \le O\Big(\varepsilon^{-4}\Big(\min\{\mathsf{fat}_{\mathcal{D}}(\varepsilon/25), \mathsf{fat}_{\mathcal{P}}(\varepsilon^2/20)\}(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big).$$

4. If we assume $\ell := \sup_{x\in X} \sup_{d\in\mathcal{D}} |d(x)| > 3\varepsilon$ and $\delta \in (0, 1/3)$, it is rather straightforward to show a sample complexity lower bound of $\Omega(\ell^2\varepsilon^{-2}\log(1/\delta))$ by a reduction from estimating the bias of a coin from independent tosses. We omit the details as our focus is on the dependence on $\mathcal{D}$, rather than on ε and δ.

This, however, does not give a sample complexity characterization for distribution-free realizable OI because when $\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)$ *and* $\mathsf{fat}_{\mathcal{P}}(\varepsilon^2/20)$ *are both infinite, it is still possible to have* $\mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta)$ *being finite (see the example in Section 1.1).*

5.1. Upper Bound

We prove our upper bound using the *multiaccuracy boost* algorithm (Algorithm 3). Kim et al. (2019) used the algorithm to show a sample complexity upper bound for multiaccuracy based on an abstract notion of dimension of the distinguisher class. We make their upper bound concrete using the fat-shattering dimension and match it with a lower bound in Section 5.2. The following standard uniform convergence result based on the fat-shattering dimension is crucial for our upper bound:

Lemma 17 (Uniform convergence from fat shattering (Bartlett and Long, 1995)) *Let* $\mathcal{F} \subseteq [-1,1]^X$ *be a function class. For every* $\varepsilon, \delta \in (0,1)$ *there exists* $n \in \mathbb{Z}_{>0}$ *such that*

$$n = O\Big(\varepsilon^{-2}\Big(\mathsf{fat}_{\mathcal{F}}(\varepsilon/5)(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big)$$

and for every probability distribution μ *over* X*,*

$$\Pr\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - E_f\right| \geq \varepsilon\right] \leq \delta,$$

where $x_1, \ldots, x_n$ *are drawn i.i.d. from* μ *and* $E_f := \mathbb{E}_{x\sim\mu}[f(x)]$ *for all* $f \in \mathcal{F}$*.*

Lemma 18 *Let* $\mathcal{D} \subseteq [-1,1]^X$ *be a distinguisher class. For every* $\varepsilon, \delta \in (0,1)$ *there exists* $n \in \mathbb{Z}_{>0}$ *such that*

$$n = O\Big(\varepsilon^{-2}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/5)(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big)$$

and for every distribution μ *over* X *and every predictor* $p \in [0,1]^X$*,*

$$\Pr\left[\sup_{d\in\mathcal{D}}\left|\frac{1}{n}\sum_{i=1}^{n} o_i d(x_i) - \langle p, d\rangle_\mu\right| \geq \varepsilon\right] \leq \delta,$$

where $(x_1, o_1), \ldots, (x_n, o_n)$ *are drawn i.i.d. from* μ_p*.*

Proof For every $d \in \mathcal{D}$, define $\tilde{d} : X \times \{0,1\} \to [-1,1]$ to be a function that maps $(x,o) \in X \times \{0,1\}$ to $od(x)$. Define $\tilde{\mathcal{D}} = \{\tilde{d} : d \in \mathcal{D}\}$.

We show that $\mathsf{fat}_{\tilde{\mathcal{D}}}(\varepsilon) = \mathsf{fat}_{\mathcal{D}}(\varepsilon)$ for all $\varepsilon > 0$. Consider $(x_1, o_1), \ldots, (x_m, o_m)$ that are ε-fat shattered by $\tilde{\mathcal{D}}$. Since every $\tilde{d}$ in $\tilde{\mathcal{D}}$ maps $(x, 0)$ to 0 for all $x \in X$, we know that $o_1 = \cdots = o_m = 1$. This then implies that $x_1, \ldots, x_m$ are ε-fat shattered by $\mathcal{D}$. Conversely, if $x_1, \ldots, x_m$ are ε-fat shattered by $\mathcal{D}$, it is clear that $(x_1, 1), \ldots, (x_m, 1)$ are ε-fat shattered by $\tilde{\mathcal{D}}$.

The lemma then follows directly from applying Lemma 17 with $\mathcal{F} = \tilde{\mathcal{D}}$. ■

Now we state our sample complexity upper bound for distribution-free OI (Lemma 19) and prove it using Algorithm 3 and a helper lemma (Lemma 20).

Lemma 19 *For every distinguisher class $\mathcal{D}$, every advantage bound $\varepsilon \in (0,1)$, and every failure probability bound $\delta \in (0,1)$, the following sample complexity upper bound holds for distribution-free OI:*

$$\begin{aligned}\mathsf{SAMP\text{-}DF\text{-}R}([0,1]^X, \mathcal{D}, \varepsilon, \delta) &= \mathsf{SAMP\text{-}DF\text{-}A}([0,1]^X, \mathcal{D}, \varepsilon, \delta) \\ &\leq O\Big(\varepsilon^{-4}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big).\end{aligned}$$

Algorithm 3: Multiaccuracy Boost (Hébert-Johnson et al., 2018; Kim et al., 2019)

Parameters: distinguisher class $\mathcal{D}$, advantage bound $\varepsilon \in (0,1)$.
Input : examples $(x_1, o_1), \ldots, (x_n, o_n)$.
Output : predictor $p \in [0,1]^X$.

1 Initialize $p(x) = 1/2$ for all $x \in X$;
2 $T \leftarrow \lceil 16/\varepsilon^2 \rceil$;
3 $m \leftarrow \lfloor n/T \rfloor$;
4 **for** $t = 1, \ldots, T$ **do**
5 take fresh examples $(x_1^*, o_1^*), \ldots, (x_m^*, o_m^*)$ where $(x_i^*, o_i^*) = (x_{(t-1)m+i}, o_{(t-1)m+i})$;
6 draw o_i' from $\mathrm{Ber}(p(x_i^*))$ independently for all $i = 1, \ldots, n$;
7 **if** *there exists* $d \in \mathcal{D} \cup (-\mathcal{D})$ *such that* $\frac{1}{m}\sum_{i=1}^m d(x_i^*)(o_i^* - o_i') \geq 3\varepsilon/5$ **then**
8 $p(x) \leftarrow p(x) + \varepsilon d(x)/5$ for all $x \in X$;
9 $p(x) \leftarrow \max\{0, \min\{1, p(x)\}\}$ for all $x \in X$;
10 **else**
11 **return** p;
12 **end**
13 **end**
14 **return** p;

Lemma 20 *If $\langle p^* - p, d\rangle_\mu \geq \varepsilon/5$ for some predictor $p^* \in [0,1]^X$ and some distribution $\mu \in \Delta_X$ before Line 8 of Algorithm 3 is executed, then Lines 8-9 decrease $\|p - p^*\|_2^2$ by at least $\varepsilon^2/25$. Here, we use $\|f\|_2^2$ as a shorthand for $\langle f, f\rangle_\mu$.*

Proof Line 8 decreases $\|p^* - p\|_2^2$ by at least $\varepsilon^2/25$ because

$$\begin{aligned}&\|p^* - p\|_2^2 - \|p^* - (p + (\varepsilon/5)d)\|_2^2 \\ &= 2\langle p^* - p, (\varepsilon/5)d\rangle_\mu - (\varepsilon/5)^2\|d\|_2^2 \\ &\geq 2(\varepsilon/5)^2 - (\varepsilon/5)^2 \\ &= \varepsilon^2/25.\end{aligned}$$

The lemma is proved by noting that Line 9 never increases $\|p^* - p\|_2^2$. ∎

Now we finish the proof of Lemma 19.

Proof We first consider the relatively simple case where $\mathsf{fat}_{\mathcal{D}}(\varepsilon/25) = 0$. By the definition of the fat-shattering dimension, there exists $d^* \in [-1,1]^X$ such that

$$|d^*(x) - d(x)| \leq \varepsilon/25 \quad \text{for every } d \in \mathcal{D} \text{ and every } x \in X. \tag{21}$$

Therefore, for every fixed distribution $\mu \in \Delta_X$ and target predictor $p^* \in [0,1]^X$, as long as the learned predictor p satisfies $|\langle p - p^*, d^* \rangle_\mu| \le \varepsilon/2$, we get the desired error bound $\|p - p^*\|_{\mu,\mathcal{D}} \le \varepsilon$. Decomposing $d^* = d^*_+ + d^*_-$ with $d^*_+(x) := \max\{d^*(x), 0\}$ and $d^*_-(x) := \min\{d^*(x), 0\}$ for every $x \in X$, we aim for the stronger goal that $|\langle p - p^*, d^*_+ \rangle_\mu| \le \varepsilon/4$ and $|\langle p - p^*, d^*_- \rangle_\mu| \le \varepsilon/4$.

Given input examples $(x_1, o_1), \dots, (x_n, o_n)$ drawn i.i.d. from μ_{p^*}, we first compute $u_+ := \frac{1}{n}\sum_{i=1}^n d^*_+(x_i) o_i$ and $v_+ := \frac{1}{n}\sum_{i=1}^n d^*_+(x_i)$. It is clear that $0 \le u_+ \le v_+$, so there exists $r_+ \in [0,1]$ such that $u_+ = r_+ v_+$. Similarly, we compute u_-, v_- satisfying $0 \ge u_- \ge v_-$ and define $r_- \in [0,1]$ so that $u_- = r_- v_-$. We output $p \in [0,1]^X$ with $p(x) = r_+$ if $d^*(x) \ge 0$ and $p(x) = r_-$ otherwise.

By the Chernoff bound and the union bound, we can choose $n = O(\varepsilon^{-2}\log(2/\delta))$ such that with probability at least $1 - \delta/2$, both of the following inequalities hold:

$$|u_+ - \langle p^*, d^*_+ \rangle_\mu| \le \varepsilon/8, \tag{22}$$

$$|v_+ - \mathbb{E}_{x\sim\mu}[d^*_+(x)]| \le \varepsilon/8. \tag{23}$$

When multiplied by r_+, (23) implies $|u_+ - \langle p, d^*_+ \rangle_\mu| \le \varepsilon/8$. Combining this with (22), with probability at least $1 - \delta/2$, $|\langle p^* - p, d^*_+ \rangle_\mu| \le \varepsilon/4$. Similarly, with probability at least $1 - \delta/2$, $|\langle p^* - p, d^*_- \rangle_\mu| \le \varepsilon/4$. By the union bound, with probability at least $1 - \delta$, $|\langle p^* - p, d^*_+ \rangle_\mu| \le \varepsilon/4$ and $|\langle p^* - p, d^*_- \rangle_\mu| \le \varepsilon/4$ both hold as desired, so

$$\begin{aligned}\mathsf{SAMP\text{-}DF\text{-}R}([0,1]^X, \mathcal{D}, \varepsilon, \delta) &\le O(\varepsilon^{-2}\log(2/\delta)) \\ &\le O\Big(\varepsilon^{-4}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big).\end{aligned}$$

We thus assume $\mathsf{fat}_{\mathcal{D}}(\varepsilon/25) \ge 1$ from now on. We use Algorithm 3 with $n = \lceil 25/\varepsilon^2 \rceil m$ for a positive integer m chosen as follows. Given input examples $(x_1, o_1), \dots, (x_n, o_n)$ drawn i.i.d. from μ_{p^*} for some $\mu \in \Delta_X$ and $p^* \in [0,1]^X$, by Lemma 18 and the union bound, we can choose

$$m = O\Big(\varepsilon^{-2}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)(\log(2/\varepsilon))^2 + \log(2/\delta\varepsilon)\Big)\Big)$$

so that with probability at least $1 - \delta$, every time Line 7 is executed, we have

$$\sup_{d\in\mathcal{D}} \left| \frac{1}{m}\sum_{i=1}^m d(x_i^*) o_i^* - \langle d, p^* \rangle_\mu \right| \le \varepsilon/5, \text{and} \tag{24}$$

$$\sup_{d\in\mathcal{D}} \left| \frac{1}{m}\sum_{i=1}^m d(x_i^*) o_i' - \langle d, p \rangle_\mu \right| \le \varepsilon/5. \tag{25}$$

Assuming this is true, when the output predictor p is returned at Line 11, we have

$$\sup_{d\in\mathcal{D}} |\langle p - p^*, d \rangle_\mu| \le 3\varepsilon/5 + \varepsilon/5 + \varepsilon/5 \le \varepsilon,$$

as desired. Moreover, (24) and (25) imply $\langle p^* - p, d \rangle_\mu \ge 3\varepsilon/5 - \varepsilon/5 - \varepsilon/5 = \varepsilon/5$ before every time Line 8 is executed, so by Lemma 20, when the output predictor p is returned at Line 14,

$$\|p - p^*\|_2^2 \le 1 - (\varepsilon^2/25)\lceil 25/\varepsilon^2 \rceil \le 0,$$

as desired. Therefore,

$$\begin{aligned}\mathsf{SAMP\text{-}DF\text{-}R}([0,1]^X, \mathcal{D}, \varepsilon, \delta) &= \mathsf{SAMP\text{-}DF\text{-}A}([0,1]^X, \mathcal{D}, \varepsilon, \delta) \\ &\leq n \\ &= \lceil 25/\varepsilon^2 \rceil m \\ &= O\Big(\varepsilon^{-4}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)(\log(2/\varepsilon))^2 + \log(2/\delta\varepsilon)\Big)\Big) \\ &= O\Big(\varepsilon^{-4}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)(\log(2/\varepsilon))^2 + \log(2/\delta)\Big)\Big). \\ &\qquad\qquad \text{(by } \mathsf{fat}_{\mathcal{D}}(\varepsilon/25) \geq 1\text{)}\end{aligned}$$

■

Remark 21 *Hébert-Johnson et al. (2018) used a modified version of Algorithm 3 for* multicalibration*, and indeed, this gives a sample complexity upper bound for multicalibration similar to Lemma 19. In multicalibration, the only difference is in the error objective: the interval* $[0,1]$ *is partitioned into* k *subsets* $\Lambda_1, \ldots, \Lambda_k$ *for some* $k \in \mathbb{Z}_{>0}$*, and given a distribution* $\mu \in \Delta_X$ *and a distinguisher class* $\mathcal{D} \subseteq [-1,1]^X$*, we measure the error of a learned predictor* $p \in [0,1]^X$ *w.r.t. the target predictor* $p^* \in [0,1]^X$ *by*

$$\text{mc-error}_{\mu,\mathcal{D}}(p, p^*) := \sup_{d \in \mathcal{D}, 1 \leq j \leq k} |\,\mathbb{E}_{x \sim \mu}[(p(x) - p^*(x))d(x)\mathbb{1}(p(x) \in \Lambda_j)]|. \tag{26}$$

For $\varepsilon, \delta \in (0,1)$*, suppose our goal is to achieve* $\text{mc-error}_{\mu,\mathcal{D}}(p, p^*) \leq \varepsilon$ *with probability at least* $1 - \delta$ *in the distribution-free setting. Changing Line 7 of Algorithm 3 to*

$$\text{“}\textbf{if } \textit{there exists } d \in \mathcal{D} \cup (-\mathcal{D}) \textit{ and } j \in \{1, \ldots, k\} \textit{ such that } \frac{1}{m}\sum_{i=1}^{m} d(x_i^*)(o_i^* - o_i')\mathbb{1}(p(x_i^*) \in \Lambda_j) \geq 3\varepsilon/5 \textbf{ then}\text{”}$$

and changing Line 8 to

$$\text{“}p(x) \leftarrow p(x) + \varepsilon d(x)\mathbb{1}(p(x) \in \Lambda_j)/5 \text{ for all } x \in X\text{”},$$

we can prove a sample complexity upper bound of

$$O\Big(\varepsilon^{-4}\Big(\mathsf{fat}_{\mathcal{D}}(\varepsilon/25)(\log(2/\varepsilon))^2 + \log(1/\delta)\Big)\Big).$$

This bound follows from combining the proof of Lemma 19 with the observation that for every predictor $p \in [0,1]^X$*, the following distinguisher class*

$$\begin{aligned}\mathcal{D}_p := \{d' \in [-1,1]^X : &\exists d \in \mathcal{D} \textit{ and } j \in \{1, \ldots, k\} \textit{ such that} \\ &\textit{for every } x \in X, d'(x) = d(x)\mathbb{1}(p(x) \in \Lambda_j)\}\end{aligned}$$

satisfies $\mathsf{fat}_{\mathcal{D}_p}(\varepsilon) \leq \mathsf{fat}_{\mathcal{D}}(\varepsilon) + 1$ *for every* $\varepsilon > 0$*.*

5.2. Lower Bound

We first prove the following lemma showing that the fat-shattering dimension gives a lower bound for the metric entropy on a particular distribution μ.

Lemma 22 *If $x_1, \ldots, x_n \in X$ are 6ε-fat shattered by $\mathcal{D}$, then $\log N_{\mu,\mathcal{D}}([0,1]^X, \varepsilon) \geq n/8$, where μ is the uniform distribution over $x_1, \ldots, x_n$.*

Proof We can assume without loss of generality that $X = \{x_1, \ldots, x_n\}$. The assumption that $x_1, \ldots, x_n$ are 6ε-fat shattered by $\mathcal{D}$ implies the existence of a function $r \in \mathbb{R}^X$ such that for every function $b \in \{-1,1\}^X$, there exists a distinguisher $d_b \in \mathcal{D}$ satisfying

$$b(x)(d_b(x) - r(x)) \geq 6\varepsilon \quad \text{for all } x \in X. \tag{27}$$

We first show that $\{r\} + [-6\varepsilon, 6\varepsilon]^X$ is a subset of the symmetric convex hull $\bar{\mathcal{D}}$. Assume for the sake of contradiction that some $f \in \{r\} + [-6\varepsilon, 6\varepsilon]^X$ does not belong to $\bar{\mathcal{D}}$. In particular, f does not belong to the symmetric convex hull of $\{d_b : b \in \{-1,1\}^X\}$. By the hyperplane separation theorem, there exists $g \in \mathbb{R}^X$ such that

$$\langle g, d_b - f \rangle_\mu < 0 \quad \text{for all } b \in \{-1,1\}^X. \tag{28}$$

Consider the function $b \in \{-1,1\}^X$ such that $b(x) = 1$ if and only if $g(x) \geq 0$. For $x \in X$ with $b(x) = 1$, we have $d_b(x) - r(x) \geq 6\varepsilon$ by (27) and thus $d_b(x) \geq f(x)$. Similarly, for $x \in X$ with $b(x) = -1$ we have $d_b(x) \leq f(x)$. In both cases, we have $g(x)(d_b(x) - f(x)) \geq 0$, a contradiction with (28).

Now we have proved that $\{r\} + [-6\varepsilon, 6\varepsilon]^X \subseteq \bar{\mathcal{D}}$. By the symmetry of $\bar{\mathcal{D}}$, $\{-r\} + [-6\varepsilon, 6\varepsilon]^X \subseteq \bar{\mathcal{D}}$. Then by the convexity of $\bar{\mathcal{D}}$, $[-6\varepsilon, 6\varepsilon]^X \subseteq \bar{\mathcal{D}}$.

The lemma is proved by the following chain of inequalities:

$$\begin{aligned}
\log N_{\mu,\mathcal{D}}([0,1]^X, \varepsilon) &= \log N_{\mu,\bar{\mathcal{D}}}([-1,1]^X, 2\varepsilon) && (29)\\
&\geq \log N_{\mu,[-6\varepsilon,6\varepsilon]^X}([-1,1]^X, 2\varepsilon) && (30)\\
&= \log N_{\mu,[-1,1]^X}([-1,1]^X, 1/3) && (31)\\
&\geq n/8. && (32)
\end{aligned}$$

We used Lemma 6 Items 2, 3, and 4 in (29), Lemma 6 Item 1 in (30), and Lemma 6 Item 4 in (31). We used Lemma 34 to get (32). ■

Combining the lemma above with Lemma 8, we obtain the following sample complexity lower bound for distribution-free OI.

Lemma 23 *For every distinguisher class $\mathcal{D}$, every advantage bound $\varepsilon > 0$, and every failure probability bound $\delta \in (0,1)$, the following sample complexity lower bound holds for distribution-free OI:*

$$\begin{aligned}
\mathsf{SAMP\text{-}DF\text{-}R}([0,1]^X, \mathcal{D}, \varepsilon, \delta) &= \mathsf{SAMP\text{-}DF\text{-}A}([0,1]^X, \mathcal{D}, \varepsilon, \delta)\\
&\geq \mathsf{fat}_{\mathcal{D}}(12\varepsilon)/8 + \log(1-\delta).
\end{aligned}$$

Proof Define $n = \mathsf{fat}_{\mathcal{D}}(12\varepsilon)$, and suppose $x_1, \ldots, x_n \in X$ are 12ε-shattered by $\mathcal{D}$. Let μ be the uniform distribution over $x_1, \ldots, x_n$. By Lemma 22, $\log N_{\mu,\mathcal{D}}([0,1]^X, 2\varepsilon) \geq n/8$. By Lemma 8,

$$\begin{aligned}
\mathsf{SAMP\text{-}DF\text{-}A}([0,1]^X, \mathcal{D}, \varepsilon, \delta) &= \mathsf{SAMP\text{-}DF\text{-}R}([0,1]^X, \mathcal{D}, \varepsilon, \delta) \\
&\geq \mathsf{SAMP\text{-}DS\text{-}R}([0,1]^X, \mathcal{D}, \varepsilon, \delta, \mu) \qquad \text{(by (7))} \\
&\geq \log N_{\mu,\mathcal{D}}([0,1]^X, 2\varepsilon) + \log(1-\delta) \\
&\geq n/8 + \log(1-\delta).
\end{aligned}$$

■

Remark 24 *It is clear that the error objective* $\text{mc-error}_{\mu,\mathcal{D}}(p, p^*)$ *for multicalibration defined in* (26) *satisfies*

$$\text{mc-error}_{\mu,\mathcal{D}}(p, p^*) \geq \text{adv}_{\mu,\mathcal{D}}(p, p^*)/k,$$

so Lemma 23 directly implies a sample complexity lower bound for multicalibration. Specifically, assuming the predictor class $\mathcal{P}$ *is* $[0,1]^X$*, if we want to achieve* $\text{mc-error}_{\mu,\mathcal{D}}(p, p^*) \leq \varepsilon$ *with probability at least* $1-\delta$ *in the distribution-free setting, the sample complexity is at least*

$$\mathsf{fat}_{\mathcal{D}}(12k\varepsilon)/8 + \log(1-\delta).$$

6. Separation between Agnostic and Realizable OI

The sample complexities of realizable and agnostic learning have the same characterization in many settings. They are both characterized by the VC dimension in distribution-free PAC learning (Vapnik and Chervonenkis, 1971), whereas in distribution-specific PAC learning they are both characterized by the metric entropy (this characterization was proved in the realizable setting by Benedek and Itai (1991), and it extends straightforwardly to the agnostic setting (see Lemma 26)). Recently, an independent work by Hopkins et al. (2021) shows that this relationship between realizable and agnostic learning holds very broadly for a unified reason.

In this section, we study this relationship between realizable and agnostic learning in outcome indistinguishability. In contrast to many common learning settings including PAC learning, we show a strong separation between the sample complexities of realizable OI and agnostic OI in both the distribution-free and the distribution-specific settings.

6.1. No separation with additional assumptions

Before we present the sample complexity separation between realizable and agnostic OI, we first discuss several assumptions that make the separation disappear in the following two lemmas.

In the first lemma (Lemma 25), we make the assumption that the target predictor p^* belongs to the symmetric convex hull $\bar{\mathcal{P}}$ of $\mathcal{P}$.

Lemma 25 (Inside the symmetric convex hull) *For every predictor class* $\mathcal{P} \subseteq [0,1]^X$*, every distinguisher class* $\mathcal{D} \subseteq [-1,1]^X$*, every distribution* $\mu \in \Delta_X$*, every advantage bound* $\varepsilon > 0$*, and every failure probability bound* $\delta \in (0,1)$*, there exist a nonnegative integer* n *and an algorithm* $\mathcal{A}$ *such that*

$$n = O(\varepsilon^{-2}(\log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon/4) + \log(2/\delta))), \quad \text{and}$$

$$\mathcal{A} \in \bigcap\nolimits_{p^* \in \bar{\mathcal{P}}} \mathsf{OI}_n(p^*, \mathcal{D}, \varepsilon, \delta, \mu).$$

Proof By Lemma 6 Item 2,

$$\log N_{\mu,\mathcal{P}}(\mathcal{D}, \varepsilon/4)) = \log N_{\mu,\bar{\mathcal{P}}}(\mathcal{D}, \varepsilon/4)).$$

The lemma is then a consequence of applying Lemma 10 with $\mathcal{P}$ replaced by $\bar{\mathcal{P}}$. ■

In the second lemma (Lemma 26), we make the assumption that $\{-1, 1\} \subseteq \mathcal{D}$ (i.e. we consider the ℓ_1 error, see Lemma 7), and that either $\mathcal{P}$ only contains binary classifiers or the target predictor p^* is a binary classifier.

Lemma 26 (Binary classifiers with ℓ_1 error) *Assume $\{-1, 1\}^X \subseteq \mathcal{D}$. For every predictor class $\mathcal{P} \subseteq [0, 1]^X$, every distribution $\mu \in \Delta_X$, every advantage bound $\varepsilon \in (0, 1)$ and every failure probability bound $\delta \in (0, 1)$, there exists a positive integer n such that*

$$n = O((\log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/2) + \log(2/\delta))/\varepsilon^2),$$

and for every target predictor $p^* \in [0, 1]^X$, Algorithm 1 belongs to

$$\mathsf{OI}_n(p^*, \mathcal{D}, \inf_{p \in \mathcal{P}} \mathrm{adv}_{\mu,\mathcal{D}}(p, p^*) + \varepsilon, \delta, \mu)$$

whenever $\mathcal{P} \subseteq \{0, 1\}^X$ or $p^* \in \{0, 1\}^X$.

Proof We choose $n = O((\log N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/2) + \log(2/\delta))/\varepsilon^2)$ so that by the Chernoff bound and the union bound, with probability at least $1 - \delta$ over $(x_1, o_1), \ldots, (x_n, o_n)$ drawn i.i.d. from μ_{p^*}, for all $p' \in \mathcal{P}'$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} |p'(x_i) - o_i| - \mathbb{E}_{(x,o)\sim\mu_{p^*}}[|p'(x) - o|] \right| \leq \varepsilon/8. \tag{33}$$

It remains to prove that whenever the input $(x_1, o_1), \ldots, (x_n, o_n)$ to Algorithm 1 satisfies the inequality above, the output p satisfies $\mathrm{adv}_{\mu,\mathcal{D}}(p, p^*) \leq \inf_{p \in \mathcal{P}} \mathrm{adv}_{\mu,\mathcal{D}}(p, p^*) + \varepsilon$ assuming $\mathcal{P} \subseteq \{0, 1\}^X$ or $p^* \in \{0, 1\}^X$.

By definition, there exists $\hat{p} \in \mathcal{P}$ such that $\|\hat{p} - p^*\|_{\mu,\mathcal{D}} \leq \inf_{p \in \mathcal{P}} \|p - p^*\|_{\mu,\mathcal{D}} + \varepsilon/4$. Since $\mathcal{P}'$ is an $\varepsilon/2$-covering of $\mathcal{P}$, there exists $\tilde{p} \in \mathcal{P}'$ such that $\|\tilde{p} - \hat{p}\|_{\mu,\mathcal{D}} \leq \varepsilon/2$. Combining these two inequalities together,

$$\|\tilde{p} - p^*\|_{\mu,\mathcal{D}} \leq \|\tilde{p} - \hat{p}\|_{\mu,\mathcal{D}} + \|\hat{p} - p^*\|_{\mu,\mathcal{D}} \leq \inf_{p \in \mathcal{P}} \|p - p^*\|_{\mu,\mathcal{D}} + 3\varepsilon/4. \tag{34}$$

Since $\{-1, 1\}^X \subseteq \mathcal{D}$,

$$\mathsf{loss}(p') = \frac{1}{n} \sum_{i=1}^{n} |p'(x_i) - o_i| \quad \text{for all } p' \in \mathcal{P}'.$$

Therefore, the output p of Algorithm 1 satisfies

$$\frac{1}{n} \sum_{i=1}^{n} |p(x_i) - o_i| = \mathsf{loss}(p) \leq \mathsf{loss}(\tilde{p}) = \frac{1}{n} \sum_{i=1}^{n} |\tilde{p}(x_i) - o_i|. \tag{35}$$

Our assumption $\mathcal{P} \subseteq \{0,1\}^X$ or $p^* \in \{0,1\}^X$ implies that for all $p' \in \mathcal{P}'$,

$$\mathbb{E}_{(x,o)\sim\mu_{p^*}}[|p'(x) - o|] = \mathbb{E}_{x\sim\mu}[|p'(x) - p^*(x)|] = \|p' - p^*\|_{\mu,\mathcal{D}}, \tag{36}$$

where the last equation is by Lemma 7.

Combining everything together,

$$\begin{aligned}
\|p - p^*\|_{\mu,\mathcal{D}} &\le \frac{1}{n}\sum_{i=1}^{n} |p(x_i) - o_i| + \varepsilon/8 && \text{(by (33) and (36))}\\
&\le \frac{1}{n}\sum_{i=1}^{n} |\tilde{p}(x_i) - o_i| + \varepsilon/8 && \text{(by (35))}\\
&\le \|\tilde{p} - p^*\|_{\mu,\mathcal{D}} + \varepsilon/4 && \text{(by (33) and (36))}\\
&\le \inf_{p\in\mathcal{P}} \|p - p^*\|_{\mu,\mathcal{D}} + \varepsilon. && \text{(by (34))}
\end{aligned}$$

■

6.2. Separation without additional assumptions

Without the additional assumptions in Lemma 25 and Lemma 26, we give examples where the sample complexity of agnostic OI can be arbitrarily larger than that of realizable OI in both the distribution-specific and the distribution-free settings. Given a positive integer m, we choose $X = \{\perp\} \cup \{0,1\}^m$, and choose the predictor class $\mathcal{P}$ to consist of only two predictors p_1 and p_2 where $p_1(\perp) = 0, p_2(\perp) = 1$ and $p_1(x) = p_2(x) = 1/2$ for all $x \in \{0,1\}^m$.

We first show that $O(\varepsilon^{-2}\log(2/\delta))$ examples are sufficient for distribution-free agnostic OI as long as $\mathcal{D} = [-1,1]^X$.

Lemma 27 *For all $\varepsilon, \delta \in (0,1)$, there exists a positive integer $n = O(\varepsilon^{-2}\log(2/\delta))$ such that for all m and $X, \mathcal{P}$ defined as above, there exists an algorithm $\mathcal{A} \in \mathsf{DF\text{-}A}_n(\mathcal{P}, [-1,1]^X, \varepsilon, \delta)$.*

Proof We choose algorithm $\mathcal{A}$ to be the following simple algorithm: after seeing examples (x_1, o_1), $\ldots, (x_n, o_n)$, it computes

$$r := \frac{|\{i \in \{1,\ldots,n\} : (x_i, o_i) = (\perp, 1)\}|}{|\{i \in \{1,\ldots,n\} : x_i = \perp\}|}.$$

If the denominator $|\{i \in \{1,\ldots,n\} : x_i = \perp\}|$ is zero, define $r = 1/2$. Algorithm $\mathcal{A}$ outputs the predictor p such that $p(\perp) = r$, and $p(x) = 1/2$ for all $x \in \{0,1\}^m$.

It remains to show that when the input examples $(x_1, o_1), \ldots, (x_n, o_n)$ are drawn i.i.d. from μ_{p^*} for some distribution $\mu \in \Delta_X$ and some $p^* \in [0,1]^X$, the output p of algorithm $\mathcal{A}$ satisfies the following with probability at least $1 - \delta$:

$$\|p - p^*\|_{\mu,[-1,1]^X} \le \inf_{p'\in\mathcal{P}} \|p' - p^*\|_{\mu,[-1,1]^X} + \varepsilon. \tag{37}$$

Let $\mu_\perp$ denote the probability mass on $\perp$ in distribution μ. Since $p(x) = 1/2 = p'(x)$ for all $p' \in \mathcal{P}$ and $x \in \{0,1\}^X$, by Lemma 7,

$$\|p - p^*\|_{\mu,[-1,1]^X} - \mu_\perp|p(\perp) - p^*(\perp)| \le \inf_{p'\in\mathcal{P}} \|p' - p^*\|_{\mu,[-1,1]^X}. \tag{38}$$

If $\mu_\perp \le \varepsilon$, inequality (38) implies that (37) is always true. If $\mu_\perp > \varepsilon$, we choose $n = O(\varepsilon^{-2}\log(2/\delta))$ so that the following two conditions hold:

1. by the Chernoff bound, with probability at least $1-\delta/2$, $|\{i \in \{1,\dots,n\} : x_i = \perp\}| \ge \mu_\perp n/2$;
2. it holds that $\mu_\perp n/2 \ge C\mu_\perp \varepsilon^{-2}\log(2/\delta)$, where C is an absolute constant determined later.

When combined, the two conditions guarantee that with probability at least $1-\delta/2$, $|\{i \in \{1,\dots,n\} : x_i = \perp\}| \ge C\mu_\perp\varepsilon^{-2}\log(2/\delta)$. Conditioned on this being true, by the Chernoff bound, choosing C sufficiently large guarantees that with probability at least $1-\delta/2$,

$$|p^*(\perp) - p(\perp)| = \left|p^*(\perp) - \frac{|\{i : (x_i, o_i) = (\perp, 1)\}|}{|\{i : x_i = \perp\}|}\right| \le \varepsilon/\sqrt{\mu_\perp}.$$

Combining this with (38), we know (37) holds with probability at least $1-\delta$, as desired. ■

By the monotonicity properties of sample complexities (see (6) and (7)), the lemma above implies that for all $\mathcal{D} \subseteq [-1,1]^X$ and $\mu \in \Delta_X$,

$$\mathsf{SAMP\text{-}DS\text{-}A}(\mathcal{P}, [-1,1]^X, \varepsilon, \delta, \mu) \le \mathsf{SAMP\text{-}DF\text{-}A}(\mathcal{P}, [-1,1]^X, \varepsilon, \delta) \le O(\varepsilon^{-2}\log(2/\delta)), \quad (39)$$

and

$$\begin{aligned}\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \le \mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta) &\le \mathsf{SAMP\text{-}DF\text{-}R}(\mathcal{P}, [-1,1]^X, \varepsilon, \delta) \\ &\le O(\varepsilon^{-2}\log(2/\delta)). \quad (40)\end{aligned}$$

Now we show that for a specific distribution μ and a particular distinguisher class $\mathcal{D}$, it requires at least $m-20$ examples for distribution-specific agnostic OI when $\varepsilon = 1/8$ and $\delta = 1/3$ (Lemma 28). Sending m to infinity and comparing with (40), this implies that the sample complexity of agnostic OI can be arbitrarily large even when the sample complexity of realizable OI is bounded by a constant in both the distribution-specific and the distribution-free settings. Comparing this with (39), we also know that the sample complexity of agnostic OI is not monotone w.r.t. the distinguisher class $\mathcal{D}$ in both the distribution-specific and distribution-free settings.

Before we describe the μ and $\mathcal{D}$ used in Lemma 28, we need the following definitions. Let B denote the set of all functions $t : \{0,1\}^m \to \{0,1\}$. A function $f \in \mathsf{B}$ is a parity function if for a subset $S \subseteq \{1,\dots,m\}$, it holds that $f(x) = \bigoplus_{i\in S} x_i$ for all $x \in \{0,1\}^m$. A function $g \in \mathsf{B}$ is an anti-parity function if for some parity function f, it holds that $g(x) = 1 \oplus f(x)$ for all $x \in \{0,1\}^m$. Let $\mathsf{BP} \subseteq \mathsf{B}$ (resp. $\mathsf{BA} \subseteq \mathsf{B}$) denote the set of parity functions (resp. anti-parity) functions.

We choose μ so that it puts $1/3$ probability mass on $\perp$, and puts the remaining $2/3$ probability mass evenly on $\{0,1\}^m$. We choose $\mathcal{D}$ to contain 2^m hypotheses as follows: for every parity function $f \in \mathsf{BP}$, there is a distinguisher $d \in \mathcal{D}$ such that $d(\perp) = 1$ and $d(x) = (-1)^{f(x)}$ for all $x \in \{0,1\}^m$.

Lemma 28 *For $m, \mathcal{P}, \mathcal{D}, \mu$ defined as above,* $\mathsf{SAMP\text{-}DS\text{-}A}(\mathcal{P}, \mathcal{D}, 1/8, 1/3, \mu) \ge m - 20$.

To prove the lemma, we first consider the following task of list-learning parity and anti-parity functions, which we abbreviate as ListL. Given a positive integer list size k, in the task of ListL, an algorithm takes as input examples $(u_1, v_1), \dots, (u_n, v_n)$ where u_i's are drawn independently and uniformly from $\{0,1\}^m$, and $v_i = t(u_i)$ for some unknown $t \in \mathsf{B}$, and the algorithm outputs a subset (i.e. list) $L \subseteq \mathsf{B}$. The algorithm succeeds if $t \in L$ and $\min\{|L \cap \mathsf{BP}|, |L \cap \mathsf{BA}|\} \le k$.

Lemma 29 *Assuming $n \leq m$ and $k \leq 2^{m-n}$, no algorithm has failure probability smaller than $(1/2)(1-1/2^{m-n})(1-k/2^{m-n})$ for all choices of $t \in \mathsf{B}$ in the task* ListL.

Proof Assume that for some $\alpha \geq 0$, there exists an algorithm $\mathcal{A}$ that has failure probability at most α for all $t \in \mathsf{B}$. In particular, when t is drawn uniformly at random from $\mathsf{BP} \cup \mathsf{BA}$, the failure probability of $\mathcal{A}$ is at most α. For this fixed distribution of t, we can assume that algorithm $\mathcal{A}$ is deterministic without increasing its failure probability. By the law of total probability,

$$\alpha \geq \Pr[\text{failure}] = \mathbb{E}[\Pr[\text{failure}|(u_1, v_1), \ldots, (u_n, v_n)]].$$

Here, $(u_1, v_1), \ldots, (u_n, v_n)$ are the input examples to algorithm $\mathcal{A}$ where $u_1, \ldots, u_n$ are drawn independently and uniformly from $\{0,1\}^m$, and $v_i = t(u_i)$ for every $i = 1, \ldots, n$ with t drawn uniformly at random from $\mathsf{BP} \cup \mathsf{BA}$.

With probability

$$(1-2^{-m})(1-2^{-m+1})\cdots(1-2^{-m+n-1}) \geq 1 - 2^{-m} - \cdots - 2^{-m+n-1} \geq 1 - 1/2^{m-n},$$

$u_1, \ldots, u_n$ are linearly independent in $\mathbb{F}_2^m$, in which case the conditional distribution of t given $(u_1, v_1), \ldots, (u_n, v_n)$ is the uniform distribution over $L_1 \cup L_2$ for some $L_1 \subseteq \mathsf{BP}, L_2 \subseteq \mathsf{BA}$ with $|L_1| = |L_2| = 2^{m-n}$, and thus

$$\Pr[\text{failure}|(u_1, v_1), \ldots, (u_n, v_n)] \geq (1/2)(1 - k/2^{m-n}).$$

Therefore,

$$\alpha \geq \mathbb{E}[\Pr[\text{failure}|(u_1, v_1), \ldots, (u_n, v_n)]] \geq (1/2)(1-1/2^{m-n})(1-k/2^{m-n}),$$

as desired. ■

We are now ready to prove Lemma 28. Let $\mathcal{A}$ be an algorithm in $\mathsf{DS\text{-}A}_n(\mathcal{P}, \mathcal{D}, 1/8, \delta, \mu)$ for a nonnegative integer n and a positive real number δ. We use this algorithm to build an algorithm for ListL. Suppose $(u_1, v_1), \ldots, (u_n, v_n)$ are the input examples in ListL, where u_i are drawn independently and uniformly from $\{0,1\}^m$, and $v_i = t(u_i)$ for some $t \in \mathsf{B}$.

Before we construct the algorithm for ListL, we need the following definition. For every $f \in \mathsf{B}$, we define $p_f \in [0,1]^X$ such that $p_f(\perp) = 1/2$, and $p_f(x) = (1 + (-1)^{f(x)})/2$ for all $x \in \{0,1\}^m$.

Now we construct the algorithm for ListL using algorithm $\mathcal{A}$. We start by generating the input examples $(x_1, o_1), \ldots, (x_n, o_n)$ to algorithm $\mathcal{A}$. Independently for every $i = 1, \ldots, n$, with probability $1/6$ we set $(x_i, o_i) = (\perp, 0)$, with probability $1/6$ we set $(x_i, o_i) = (\perp, 1)$ and with the remaining probability $2/3$ we set $(x_i, o_i) = (u_i, (1 + (-1)^{v_i})/2)$. After running algorithm $\mathcal{A}$ and obtaining the output p, we return the list

$$L = \{f \in \mathsf{B} : \|p_f - p\|_{\mu,\mathcal{D}} \leq 1/6 + 1/8\} \cup (\mathsf{B} \backslash (\mathsf{BP} \cup \mathsf{BA})). \tag{41}$$

We prove the following two helper lemmas before we finish the proof of Lemma 28.

Lemma 30 *If $p(\perp) \leq 1/2$, then $|L \cap \mathsf{BP}| \leq 64$. Similarly, if $p(\perp) \geq 1/2$, then $|L \cap \mathsf{BA}| \leq 64$.*

Proof We prove the first half of the lemma, and the second half follows from a similar argument. Pick any function $f \in L \cap \mathsf{BP}$, and pick $d \in \mathcal{D}$ such that $d(\bot) = 1$ and $d(x) = (-1)^{f(x)}$. Define p_0 to be the predictor that maps everything to $1/2$. We have $\langle p_f - p_0, d\rangle_\mu = 1/3$. Define μ' to be the uniform distribution over $\{0,1\}^d$. We have

$$\langle p - p_0, d\rangle_\mu = (1/3)(p(\bot) - p_0(\bot)) + (2/3)\langle p - p_0, d\rangle_{\mu'} \le (2/3)\langle p - p_0, d\rangle_{\mu'}.$$

Therefore,

$$1/3 - (2/3)\langle p - p_0, d\rangle_{\mu'} \le \langle p_f - p, d\rangle_\mu \le 1/6 + 1/8.$$

This implies

$$\langle p - p_0, d\rangle_{\mu'} \ge 1/16. \tag{42}$$

However, the predictors in $\mathcal{D}$, when restricted to the sub-domain $\{0,1\}^m \subseteq X$, form an orthonormal basis for $\mathbb{R}^{\{0,1\}^m}$ w.r.t. the inner product $\langle\cdot,\cdot\rangle_{\mu'}$, so

$$\sum\nolimits_{d\in\mathcal{D}} \langle p - p_0, d\rangle_{\mu'}^2 = \langle p - p_0, p - p_0\rangle_{\mu'} \le 1/4.$$

Therefore, there can be at most 64 different $d \in \mathcal{D}$ that satisfy (42). Since d is defined differently for different $f \in L \cap \mathsf{BP}$, we get $|L \cap \mathsf{BP}| \le 64$ as desired. ■

Lemma 31 *The event $t \in L$ happens with probability at least $1 - \delta$.*

Proof By the definition of L in (41), the lemma is trivial if $t \notin \mathsf{BP} \cup \mathsf{BA}$, so we assume $t \in \mathsf{BP} \cup \mathsf{BA}$.

Define μ' to be the uniform distribution over $\{0,1\}^m$. If $t \in \mathsf{BP}$, for the predictor $d \in \mathcal{D}$ that satisfy $d(\bot) = 1$ and $d(x) = (-1)^{t(x)}$ for all $x \in \{0,1\}^m$, we have $\langle p_t - p_2, d\rangle_{\mu'} = 1/2$, so

$$\langle p_t - p_2, d\rangle_\mu = (1/3)(-1/2) + (2/3)\langle p_t - p_2, d\rangle_{\mu'} = 1/6.$$

For all other predictors $d \in \mathcal{D}$, we have $\langle p_t - p_2, d\rangle_{\mu'} = 0$, so

$$\langle p_t - p_2, d\rangle_\mu = (1/3)(-1/2) + (2/3)\langle p_t - p_2, d\rangle_{\mu'} = -1/6.$$

Therefore, $\|p_t - p_2\|_{\mu,\mathcal{D}} = 1/6$. Similarly, we can show that when $t \in \mathsf{BA}$, $\|p_t - p_1\|_{\mu,\mathcal{D}} = 1/6$.

Since the input examples $(x_1, o_1), \ldots, (x_n, o_n)$ to algorithm $\mathcal{A}$ are generated i.i.d. from μ_{p_t}, and since $\mathcal{A} \in \mathsf{DS\text{-}A}_n(\mathcal{P}, \mathcal{D}, 1/8, \delta, \mu)$, with probability at least $1 - \delta$, algorithm $\mathcal{A}$ outputs p such that $\|p - p_t\|_{\mu,\mathcal{D}} \le 1/6 + 1/8$, in which case $t \in L$, as desired. ■

Now we complete the proof of Lemma 28.

Proof Combining Lemma 29, Lemma 30, and Lemma 31, we have

$$1 - \delta \le 1 - (1/2)(1 - 1/2^{m-n})(1 - 64/2^{m-n})$$

whenever $n \le m - 8$. Setting $\delta = 1/3$, we get $n \ge m - 20$, which completes the proof. ■

Acknowledgments

LH is supported by Moses Charikar's Simons Investigator award and Omer Reingold's NSF Award IIS-1908774. CP is supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. OR is supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Sloan Foundation Grant 2020-13941 and the Simons Foundation investigators award 689988.

References

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997. ISSN 0004-5411. doi: 10.1145/263867.263927. URL https://doi.org/10.1145/263867.263927.

S. Artstein, V. Milman, S. Szarek, and N. Tomczak-Jaegermann. On convexified packing and entropy duality. *Geom. Funct. Anal.*, 14(5):1134–1141, 2004a. ISSN 1016-443X. doi: 10.1007/s00039-004-0486-3. URL https://doi.org/10.1007/s00039-004-0486-3.

S. Artstein, V. Milman, and S. J. Szarek. Duality of metric entropy. *Ann. of Math. (2)*, 159(3): 1313–1328, 2004b. ISSN 0003-486X. doi: 10.4007/annals.2004.159.1313. URL https://doi.org/10.4007/annals.2004.159.1313.

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.

Peter L. Bartlett and Philip M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. volume 56, pages 174–190. 1998. doi: 10.1006/jcss.1997.1557. URL https://doi.org/10.1006/jcss.1997.1557. Eighth Annual Workshop on Computational Learning Theory (COLT) (Santa Cruz, CA, 1995).

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

Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoret. Comput. Sci.*, 86(2):377–389, 1991. ISSN 0304-3975. doi: 10.1016/0304-3975(91)90026-X. URL https://doi.org/10.1016/0304-3975(91)90026-X.

Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151, pages 55:1–55:24, 2020.

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. ISSN 0004-5411. doi: 10.1145/76359.76371. URL https://doi.org/10.1145/76359.76371.

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. URL https://doi.org/10.1007/BFb0090048.

Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets. *arXiv preprint arXiv:2111.03135*, 2021.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

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)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:21, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-217-4. doi: 10.4230/LIPIcs.ITCS.2022.79. URL https://drops.dagstuhl.de/opus/volltexte/2022/15675.

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.

Max Hopkins, Daniel Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. *arXiv preprint arXiv:2111.04746*, 2021.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Computational learning theory and natural learning systems, Vol. I*, Bradford Book, pages 289–329. MIT Press, Cambridge, MA, 1994.

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.

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), 2022.

Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. System Sci.*, 62(3):516–527, 2001. ISSN 0022-0000. doi: 10.1006/jcss.2000.1741. URL https://doi.org/10.1006/jcss.2000.1741.

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

Emanuel Milman. A remark on two duality relations. *Integral Equations Operator Theory*, 57(2): 217–228, 2007. ISSN 0378-620X. doi: 10.1007/s00020-006-1479-4. URL https://doi.org/10.1007/s00020-006-1479-4.

Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: the sparse and approximate cases. In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 351–360. ACM, New York, 2013. doi: 10.1145/2488608.2488652. URL https://doi.org/10.1145/2488608.2488652.

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

Guy N Rothblum and Gal Yona. Multi-group agnostic PAC learnability. *arXiv preprint arXiv:2105.09989*, 2021.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13331–13340. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9a96876e2f8f3dc4f3cf45f02c61c0c1-Paper.pdf.

L. G. Valiant. Deductive learning. volume 312, pages 441–446. 1984a. doi: 10.1098/rsta.1984.0069. URL https://doi.org/10.1098/rsta.1984.0069. With discussion, Mathematical logic and programming languages.

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

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Measures of Complexity*, 16(2):11, 1971.

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, 2021.

Appendix A. Failure of Empirical Risk Minimization

The following two lemmas show that in certain cases the failure probability of Algorithm 1 approaches 1 (instead of 0) when the number of examples increases.

Lemma 32 *Assume X is finite and μ is the uniform distribution over X. Assume $\mathcal{D} = [-1,1]^X$, and $\mathcal{P}$ consists of the following three predictors: p_b that maps every $x \in X$ to b for $b = 0, 1/2, 1$. For every positive integer n, Algorithm 1 does not belong to*

$$\mathsf{OI}_n(p_{1/2}, \mathcal{D}, 1/3, 1 - 1/\sqrt{n+1}, \mu).$$

Proof Consider the n input examples to Algorithm 1: $(x_1, o_1), \ldots, (x_n, o_n)$ drawn i.i.d. from $\mu_{p_{1/2}}$. We first show that with probability above $1 - 1/\sqrt{n+1}$,

$$n' := |\{i \in \{1, \ldots, n\} : o_i = 1\}| \neq n/2.$$

This is trivially true when n is odd. When n is even, this is also true because

$$\Pr[n' = n/2] = \binom{n}{n/2}/2^n < 1/\sqrt{n+1}.$$

It remains to prove that whenever $n' \neq n/2$, the output p of Algorithm 1 satisfies $\|p - p_{1/2}\|_{\mu,\mathcal{D}} > 1/3$. By Lemma 7, $\|p_0 - p_{1/2}\|_{\mu,\mathcal{D}} = 1/2$, $\|p_{1/2} - p_1\|_{\mu,\mathcal{D}} = 1/2$, and $\|p_0 - p_1\|_{\mu,\mathcal{D}} = 1$. Therefore, the $\varepsilon/2 (= 1/6)$-covering $\mathcal{P}'$ in Algorithm 1 is equal to $\mathcal{P}$. The loss of predictor p_b is

$$\mathsf{loss}(p_b) = \frac{1}{n}\sum_{i=1}^{n} |b - o_i| = (1-b)n' + b(n-n') = n' + b(n-2n').$$

When $n' < n/2$, p_0 has the smallest loss, so Algorithm 1 returns $p = p_0$, in which case we indeed have $\|p - p_{1/2}\|_{\mu,\mathcal{D}} = 1/2 > 1/3$. Similarly, when $n' > n/2$, p_1 has the smallest loss, so $p = p_1$ and $\|p - p_{1/2}\|_{\mu,\mathcal{D}} = 1/2 > 1/3$. ■

In the lemma below, we give an example of the failure of Algorithm 1 in distribution-specific realizable OI when all the predictors in $\mathcal{P}$ are binary classifiers. In this example, $X, \mu, \mathcal{P}, \mathcal{D}$ are parametrized by two positive integers m and n as follows. We choose the individual set to be $X = \{-1, -2, -3\} \cup \{1, \ldots, m\}$, and choose the distinguisher class $\mathcal{D}$ as follows. For every size-n subset $Y \subseteq \{1, \ldots, m\}$, define $\mathcal{D}_Y \subseteq [-1,1]^X$ to be the set of all distinguishers $d \in [-1,1]^X$ satisfying $d(x) = 0$ for all $x \in \{1, \ldots, m\} \backslash Y$. The distinguisher class $\mathcal{D}$ is then defined as $\mathcal{D} = \bigcup_Y \mathcal{D}_Y$, where the union is over all size-n subsets $Y \subseteq \{1, \ldots, m\}$. The predictor class $\mathcal{P}$ consists of 4 predictors p_0, p_1, p_2, p_3 defined as follows:

$$\begin{aligned}
&p_0(-1) = 0, p_0(-2) = 0, p_0(-3) = 0, p_0(x) = 0 \text{ for all } x \in \{1, \ldots, m\},\\
&p_1(-1) = 0, p_1(-2) = 0, p_1(-3) = 0, p_1(x) = 1 \text{ for all } x \in \{1, \ldots, m\},\\
&p_2(-1) = 1, p_2(-2) = 1, p_2(-3) = 0, p_2(x) = 0 \text{ for all } x \in \{1, \ldots, m\},\\
&p_3(-1) = 0, p_3(-2) = 1, p_3(-3) = 1, p_3(x) = 1 \text{ for all } x \in \{1, \ldots, m\}.
\end{aligned}$$

The distribution μ spreads $1/2$ probability mass evenly on $\{-1, -2, -3\}$, and the spreads the remaining $1/2$ probability mass evenly on $\{1, \ldots, m\}$.

Since $N_{\mu,\mathcal{D}}(\mathcal{P}, \varepsilon/32) \leq |\mathcal{P}| = 4$, Theorem 13 tells us that

$$\mathsf{SAMP\text{-}DS\text{-}R}(\mathcal{P}, \mathcal{D}, \varepsilon, \delta, \mu) \leq O(\varepsilon^{-2}\log(2/\delta) + \varepsilon^{-4}).$$

The lemma below shows that this sample complexity upper bound cannot be achieved using Algorithm 1 when $\varepsilon \leq 1/4$ and δ is close to zero.

Lemma 33 *For every positive integer n, there exists a positive integer m such that when $X, \mathcal{P}, \mathcal{D}, \mu$ are defined as above, Algorithm 1 does not belong to*

$$\mathsf{DS\text{-}R}_n(\mathcal{P}, \mathcal{D}, 1/4, \max\{2^{-2-O(n)}, 1 - O(2^{-\Omega(n)})\}, \mu).$$

Proof By choosing m sufficiently large, we get

$$\|p_0 - p_1\|_{\mu,\mathcal{D}} = (1/2)(n/m) \leq 1/100, \tag{43}$$

$$\|p_i - p_j\|_{\mu,\mathcal{D}} \geq (1/2)(2/3) = 1/3 \text{ for all } i,j \in \{1,2,3,4\} \text{ satisfying } i < j \text{ and } (i,j) \neq (0,1). \tag{44}$$

Suppose Algorithm 1 belongs to $\mathsf{DS\text{-}R}_n(\mathcal{P}, \mathcal{D}, 1/4, \delta, \mu)$ for some $\delta \in (0,1)$. Consider the case where the input to Algorithm 1 is n examples $(x_1, o_1), \ldots, (x_n, o_n)$ drawn i.i.d. from μ_{p^*} where p^* is drawn uniformly at random from $\{p_0, p_2\}$. By assumption, the output p of Algorithm 1 satisfies $\Pr[\|p - p^*\|_{\mu,\mathcal{D}} \leq 1/4] \geq 1 - \delta$. Since Algorithm 1 only outputs $p \in \mathcal{P}$, this means that $\Pr[p \in \mathsf{Nei}(p^*)] \geq 1 - \delta$ where $\mathsf{Nei}(p_0) = \{p_0, p_1\}$ and $\mathsf{Nei}(p_2) = \{p_2\}$. However, with probability $(1/2)^n$, all the x_i's belong to $\{1, \ldots, m\}$, in which case $o_1 = \cdots = o_n = 0$, giving no information about p^*. Therefore, $\Pr[p \notin \mathsf{Nei}(p^*)] \geq (1/2)^n(1/2)$. This implies

$$\delta \geq (1/2)^n(1/2) > 2^{-2-O(n)}. \tag{45}$$

Inequalities (43) and (44) imply that the covering $\mathcal{P}'$ computed in Algorithm 1 is either $\{p_0, p_2, p_3\}$ or $\{p_1, p_2, p_3\}$. Without loss of generality, we assume $\mathcal{P}' = \{p_1, p_2, p_3\}$ because the other case can be handled similarly. Now consider the case where the input examples $(x_1, o_1), \ldots, (x_n, o_n)$ are drawn i.i.d. from μ_{p^*} with $p^* = p_0$.

By our construction of $\mathcal{D}$, the loss of every predictor in $p' \in \mathcal{P}'$ is

$$\mathsf{loss}(p') = \frac{1}{n}\sum_{i=1}^{n} |p'(x_i) - p_0(x_i)|.$$

By the Chernoff bound and the union bound, with probability above $1 - O(2^{-\Omega(n)})$, the absolute difference between $\mathsf{loss}(p')$ and $\mathbb{E}_{x\sim\mu}|p'(x) - p_0(x)|$ is below $1/100$ for all $p' \in \mathcal{P}'$. Note that

$$\mathbb{E}_{x\sim\mu}|p_1(x) - p_0(x)| = 1/2, \mathbb{E}_{x\sim\mu}|p_2(x) - p_0(x)| = 1/3, \mathbb{E}_{x\sim\mu}|p_3(x) - p_0(x)| = 5/6.$$

Therefore, with probability above $1 - O(2^{-\Omega(n)})$, Algorithm 1 returns $p = p_2$, which does not satisfy $\|p - p^*\|_{\mu,\mathcal{D}} \leq 1/4$. This implies

$$\delta > 1 - O(2^{-\Omega(n)}). \tag{46}$$

The lemma is proved by combining (45) and (46). ■

Appendix B. A Helper Lemma

Lemma 34 *Let μ be the uniform distribution over a set X of individuals with size $|X| = n \in \mathbb{Z}_{>0}$. Then for all $\varepsilon \in (0, 1/e)$,*

$$\log N_{\mu,[-1,1]^X}([-1,1]^X, \varepsilon) \geq n \log(1/e\varepsilon),$$

where e is the base of the natural logarithm.

Proof Suppose $X = \{x_1, \ldots, x_n\}$. Let vec be the bijection from $\mathbb{R}^X$ to $\mathbb{R}^n$ such that $\mathsf{vec}(f) = (f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$ for all $f \in \mathbb{R}^X$. Let $\{f_1, \ldots, f_m\} \subseteq [-1,1]^X$ be an ε-covering of $[-1,1]^X$ w.r.t. the norm $\|\cdot\|_{\mu,[-1,1]^X}$. Defining

$$B = \{f \in \mathbb{R}^X : \|f\|_{\mu,[-1,1]^X} \leq \varepsilon\},$$

we have $[-1,1]^X \subseteq \bigcup_{i=1}^m (\{f_i\} + B)$, which implies $\mathsf{vec}([-1,1]^X) \subseteq \bigcup_{i=1}^m \mathsf{vec}(\{f_i\} + B)$. Therefore, the volume of $\mathsf{vec}([-1,1]^X)$ is at most m times the volume of $\mathsf{vec}(B)$.

It is clear that $\mathsf{vec}([-1,1]^X) = [-1,1]^n$ has volume 2^n in $\mathbb{R}^n$. Moreover, by Lemma 7,

$$\mathsf{vec}(B) = \{(r_1, \ldots, r_n) \in \mathbb{R}^n : |r_1| + \cdots + |r_n| \leq \varepsilon n\},$$

which has volume $(2\varepsilon n)^n / n!$. Therefore,

$$2^n \leq m(2\varepsilon n)^n / n!,$$

and thus

$$\log m \geq \log(n!/(n\varepsilon)^n) \geq n \log(1/e\varepsilon),$$

as desired. We used the fact $\log(n!) \geq \int_1^n \log t \mathrm{d}t > n \log(n/e)$ in the last inequality. ■