Towards Fundamental Limits for Active Multi-distribution Learning

Chicheng Zhang CHICHENGZ@CS.ARIZONA.EDU
The University of Arizona

Yihan Zhou JOEYZHOU@CS.UTEXAS.EDU
The University of Texas at Austin

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Multi-distribution learning extends agnostic Probably Approximately Correct (PAC) learning to the setting in which a family of k distributions, $\{D_i\}_{i\in[k]}$, is considered and a classifier's performance is measured by its error under the worst distribution. This problem has attracted a lot of recent interests due to its applications in collaborative learning, fairness, and robustness. Despite a rather complete picture of sample complexity of passive multi-distribution learning, research on active multi-distribution learning remains scarce, with algorithms whose optimality remaining unknown.

In this paper, we develop new algorithms for active multi-distribution learning and establish improved label complexity upper and lower bounds, in distribution-dependent and distribution-free settings. Specifically, in the near-realizable setting we prove an upper bound of $\widetilde{O}\Big(\theta_{\max}(d+k)\ln\frac{1}{\varepsilon}\Big)$ and $\widetilde{O}\Big(\theta_{\max}(d+k)\Big(\ln\frac{1}{\varepsilon}+\frac{\nu^2}{\varepsilon^2}\Big)+\frac{k\nu}{\varepsilon^2}\Big)$ in the realizable and agnostic settings respectively, where $\theta_{\max}$ is the maximum disagreement coefficient among the k distributions, d is the VC dimension of the hypothesis class, ν is the multi-distribution error of the best hypothesis, and ε is the target excess error. Moreover, we show that the bound in the realizable setting is information-theoretically optimal and that the $k\nu/\varepsilon^2$ term in the agnostic setting is fundamental for proper learners. We also establish instance-dependent sample complexity bound for passive multidistribution learning that smoothly interpolates between realizable and agnostic regimes (Blum et al., 2017; Zhang et al., 2024), which may be of independent interest.

Keywords: Active Learning, Multi-distribution Learning, Statistical Learning Theory, Sample Complexity

1. Introduction

Multi-Distribution Learning (MDL) (Blum et al., 2017; Haghtalab et al., 2022) is an emerging machine learning paradigm that has gained popularity in recent years. It naturally generalizes the classical PAC (Valiant, 1984; Kearns et al., 1992) learning framework. In traditional PAC learning, the objective is to approximately identify the optimal hypothesis h^* from a hypothesis class $\mathcal{H}$ within error tolerance ε, with probability at least $1-\delta$, under a single unknown distribution D. MDL extends this framework by considering k unknown distributions $\{D_i\}_{i\in[k]}$ and evaluating the performance of a hypothesis h based on its worst-case error across these k distributions, $\max_{i\in[k]} L(h, D_i)$. The goal is to output a classifier $\hat{h}$, such that $\max_{i\in[k]} L(\hat{h}, D_i) \le \nu + \varepsilon$, where $\nu = \min_{h\in\mathcal{H}} \max_{i\in[k]} L(h, D_i)$ is the optimal worst-case error in class $\mathcal{H}$. This setting has found diverse applications, including collaborative and federated learning (Blum et al., 2017; Mohri et al., 2019), fairness (Rothblum and Yona, 2021; Du et al., 2021), and robustness (Wang et al., 2023; Deng et al., 2020), highlighting its significance in addressing complex learning scenarios.

In certain real-world applications, such as cancer detection (Gal et al., 2017), unlabeled data is significantly more abundant and less costly to obtain than labeled data. Consequently, a more practical objective in these scenarios is to minimize the number of labels required to achieve the PAC multi-distribution learning goal, a concept known as *label complexity*. This learning paradigm, termed active learning, has been extensively studied over the past three decades (Cohn et al., 1994; Dasgupta, 2005; Hanneke, 2014). An active learner can access an unlimited number of unlabeled data points and selectively query labels for certain instances, whereas a passive learner relies solely on randomly sampled feature-label pairs from the underlying distribution. Cohn et al. (1994); Freund et al. (1997); Dasgupta (2004) demonstrated that, in the realizable setting, active learning algorithms can achieve exponentially lower label complexity than passive learners when learning geometric concepts such as 1-dimensional threshold functions and d-dimensional linear separators. Subsequently, many follow-up works (e.g. Balcan et al., 2006; Hanneke, 2007; Dasgupta et al., 2007) showed that such exponential improvements in label complexity are also attainable in the agnostic setting. Notably, recent work of Rittler and Chaudhuri (2023) studied active learning in the MDL setting and provided algorithms and upper bounds on label complexity, demonstrating that an improvement over its passive counterpart is possible (see below for more details).

	Passive	Active
PAC ($k = 1$)	$\widetilde{O}\left(\frac{d(\nu+\varepsilon)}{\varepsilon^2}\right)$	$\widetilde{O}\left(d\theta\left(\ln\frac{1}{\varepsilon}+\frac{\nu^2}{\varepsilon^2}\right)\right)$
MDL	$\widetilde{O}\left(\frac{k+d}{\varepsilon^2}\right)$	$\widetilde{O}\left(\frac{\nu^2}{\varepsilon^2}kd\theta_{\max}^2+\frac{k}{\varepsilon^2}\right)$

Table 1: Label complexity upper bounds for different settings prior to our work

In the classical PAC ($k = 1$) setting, it is well known that the sample complexity upper bound in the agnostic passive setting is $\widetilde{O}\left(\frac{d(\nu+\varepsilon)}{\varepsilon^2}\right)$ (where $\widetilde{O}$ hides logarithmic dependencies on problem parameters) (Vapnik, 1982), where d is the Vapnik-Chervonenkis (VC) dimension of $\mathcal{H}$. A straightforward approach for the MDL setting is to sample S_i, a set of $\widetilde{O}\left(\frac{d(\nu+\varepsilon)}{\varepsilon^2}\right)$ iid examples from each D_i and find $\hat{h}$ that minimizes the worst-case empirical error $\hat{h} = \arg\min_{h\in\mathcal{H}} \max_i L(h, S_i)$. This approach yields a sample complexity upper bound of $\widetilde{O}\left(\frac{kd(\nu+\varepsilon)}{\varepsilon^2}\right)$. Remarkably, a series of works have developed more sample-efficient algorithms and analyses, improving the sample requirement to $\widetilde{O}\left(\frac{k+d}{\varepsilon}\right)$ or $\widetilde{O}\left(\frac{k+d}{\varepsilon^2}\right)$ in realizable and agnostic settings, respectively (Blum et al., 2017; Nguyen and Zakynthinou, 2018; Chen et al., 2018; Zhang et al., 2024; Peng, 2024). This shift from a multiplicative to an additive dependence on k and d is a significant improvement, requiring substantial technical innovation. On the other hand, one of the state-of-the-art label complexity upper bounds for agnostic single-distribution active learning is $\widetilde{O}\left(d\theta\left(\ln\frac{1}{\varepsilon}+\frac{\nu^2}{\varepsilon^2}\right)\right)$ (Dasgupta et al., 2007; Hanneke, 2007; Ailon et al., 2012), where θ is a distribution-dependent parameter known as the *disagreement coefficient* (Hanneke, 2007). The disagreement coefficient can be shown to be small under many favorable assumptions of the data distributions, leading to improved label efficiency of active learning than passive learning; see e.g. Hanneke (2014); Friedman (2009); Wang (2011) for example distributions for which the disagreement coefficients are small.

For active MDL problems, Rittler and Chaudhuri (2023) established label complexity upper bounds of $\widetilde{O}\left(kd\theta_{\max}\ln\frac{1}{\varepsilon}\right)$ and $\widetilde{O}\left(\frac{\nu^2}{\varepsilon^2}kd\theta_{\max}^2+\frac{k}{\varepsilon^2}\right)$ in near-realizable and nonrealizable settings

respectively, where $\theta_{\max}$ is the maximum disagreement coefficient among all k distributions. While this result provides a novel upper bound for active MDL, the dependence on kd and the quadratic factor $\theta_{\max}^2$ may be undesirable. It is worth noticing that in some cases where the disagreement coefficient is large, the label complexity of the active MDL algorithm could be worse than its passive counterpart. In addition, distinct from single-distribution active learning label complexity results, an extra $O\left(\frac{k}{\varepsilon^2}\right)$ additive factor appears in the agnostic label complexity bound, and it is unclear if is information-theoretically necessary. A comparison of the best known label complexity upper bounds across different settings is presented in Table 1. Given these, we ask the following questions:

In active MDL, is it possible to design algorithms with label complexity bounds that are never worse than passive learners? Is it possible to improve the multiplicative kd factor in the label complexity bounds to an additive $k + d$?

1.1. Our Contributions

In this paper, we establish novel and tighter label complexity bounds for active MDL, providing an affirmative answer to those open questions. Our main contributions are as follows:

1. For the large ε regime ($\varepsilon \geq 100\nu$), we develop an algorithm with a distribution-dependent sample complexity bound of $\widetilde{O}\left(\theta_{\max}\left(d+k\right)\ln\frac{1}{\varepsilon}\right)$, matching the bound shown in Table 1 for the PAC setting (Section 4). We prove that this bound is information-theoretically optimal by establishing a matching lower bound (Section 6).

2. For the small ε regime ($\varepsilon < 100\nu$), we propose a two-stage algorithm that achieves a distribution-dependent label complexity of $\widetilde{O}\left(\theta_{\max}(d+k)\left(\ln\frac{1}{\varepsilon}+\frac{\nu^2}{\varepsilon^2}\right)+\frac{k\nu}{\varepsilon^2}\right)$ (Section 5), accompanied by a lower bound of $\Omega(k\frac{\nu}{\varepsilon^2})$ for proper learners (Section 6). This lower bound reveals an interesting "phase transition" behavior in label complexity unique to active multi-distribution learning, since it does not appear in the large ε regime. As part of our analysis, we strengthen the existing passive MDL sample complexity bound to $\widetilde{O}\left(\frac{(k+d)(\nu+\varepsilon)}{\varepsilon^2}\right)$, which smoothly interpolates between realizable and agnostic regimes (Blum et al., 2017; Zhang et al., 2024) —a result that may be of independent interest beyond active learning.

3. In the large ε regime ($\varepsilon \geq 100\nu$), we establish a distribution-free upper bound of $\widetilde{O}\left(\mathfrak{s}(d+k)\ln\frac{1}{\varepsilon}\right)$, where $\mathfrak{s}$ denotes the star number of the hypothesis class $\mathcal{H}$ (Hanneke and Yang, 2015). We further show that when $\varepsilon \geq 100\left(d+k\right)\nu$, this bound tightens to $\widetilde{O}\left(\mathfrak{s}\ln\frac{1}{\varepsilon}\right)$—a novel result even in the context of single-distribution PAC active learning (Section 7).

Table 2 summarizes our new label complexity upper bounds.

	$\varepsilon \geq 100\nu$	$\varepsilon < 100\nu$
Distribution-dependent	$\widetilde{O}\left(\theta_{\max}\left(d+k\right)\ln\frac{1}{\varepsilon}\right)$	$\widetilde{O}\left(\theta_{\max}(d+k)(\ln\frac{1}{\varepsilon}+\frac{\nu^2}{\varepsilon^2})+\frac{k\nu}{\varepsilon^2}\right)$
	$\varepsilon \geq 100\nu$	$\varepsilon \geq 100(d+k)\nu$
Distribution-free	$\widetilde{O}\left(\mathfrak{s}(d+k)\right)$	$\widetilde{O}\left(\mathfrak{s}\ln\frac{1}{\varepsilon}\right)$

Table 2: Summary of our improved label complexity upper bounds.

2. Related Work

Multi-Distribution Learning Blum et al. (2017) first established the sample complexity upper bound of $\widetilde{O}\left(\frac{d+k}{\varepsilon}\right)$ for MDL in the realizable and large ε regime. Their analysis was later refined to remove extra logarithmic factors by Nguyen and Zakynthinou (2018); Chen et al. (2018). Their method iteratively learns the best hypothesis under the average of a carefully-maintained subset of distributions. If a hypothesis performs well on the average, it must also perform well on at least a constant fraction of the distributions, allowing the elimination of those where good performance has already been achieved. By the end of the process, a well-performing hypothesis is obtained for each distribution. For the nonrealizable setting, Haghtalab et al. (2022) introduced a new approach that reduces the problem to solving a two-player zero-sum game, where one player learns the best hypothesis while an adversary simultaneously selects the hardest distribution. However, their algorithm is only applicable when the hypothesis class is finite. Awasthi et al. (2023) modified the game dynamics algorithm of Haghtalab et al. (2022) to extend it to the infinite hypothesis class setting; however, their algorithm suffers from an undesirable $O\left(\frac{1}{\varepsilon^4}\right)$ multiplicative factor. Hanashiro and Jaillet (2023) designed algorithms inspired by best-arm identification in bandit literature that achieve instance-dependent rates; such rates depend on suboptimality gaps of hypotheses which may be arbitrarily small. The optimal sample complexity bound in the infinite hypothesis class case was later established independently by Zhang et al. (2024) and Peng (2024). While Zhang et al. (2024) employed the same game dynamics algorithm with a more careful control of the Hedge algorithm's trajectory and a delicate analysis, Peng (2024) solved the problem using recursive width reduction. Our algorithms for the large ε regime incorporates Blum et al. (2017)'s algorithm as a subroutine, while our distribution-dependent algorithm for the small ε regime builds on Zhang et al. (2024) as a subroutine.

Active Learning Cohn et al. (1994) designed the first active learning algorithm in the realizable setting, and was first analyzed by Hanneke (2007). In a phased variant of the algorithm (Hsu, 2010, Chapter 2), the algorithm begins with a constant error tolerance and learns a good hypothesis from a passive learner. It then eliminates all apparently bad hypotheses and reduces the error tolerance by half in each iteration. Since the algorithm only queries labels from the disagreement region—where hypotheses in the class disagree on labels—the disagreement region shrinks as more hypotheses are eliminated, leading to significant label savings. Balcan et al. (2006) extended this idea to the agnostic setting, and Hanneke (2007); Dasgupta et al. (2007) further refined the analysis to obtain a tighter sample complexity bound $\widetilde{O}\left(d\theta\left(\ln\frac{1}{\varepsilon}+\frac{\nu^2}{\varepsilon^2}\right)\right)$. Our work builds upon this algorithmic framework but replaces the passive learner with a passive MDL learner in each round. Beyond disagreement region-based methods, alternative approaches for active learning exist. For instance, Freund et al. (1997); Dasgupta (2004, 2005); Castro and Nowak (2008); Nowak (2011); Tosh and Dasgupta (2017); Zhou and Price (2024) proposes methods that generalize binary search to query most informative data point at each iteration, achieving competitive label complexity bounds relative to the instance-optimal solution; Balcan et al. (2007); Balcan and Long (2013); Zhang and Chaudhuri (2014); Huang et al. (2015) refines the disagreement-based active learning idea by learning from carefully constructed subsets of the disagreement regions; under some regression realizability assumptions, Cesa-Bianchi et al. (2009); Dekel et al. (2012); Agarwal (2013); Krishnamurthy et al. (2019); Zhu and Nowak (2022); Sekhari et al. (2023) uses rigorous uncertainty quantification on examples' Bayes optimal labels to guide label queries.

3. Problem Definition and Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label spaces, respectively, with $\mathcal{Y} = \{-1, 1\}$ (binary classification). Let $\mathcal{H}$ be a hypothesis class on $\mathcal{X} \to \mathcal{Y}$, and for any $h \in \mathcal{H}$, define its 0-1 loss on an example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ as $\ell(h, (x, y)) = I(h(x) \neq y)$, where $I(A)$ is the indicator of event A. We use $\ln$ to denote natural logarithm and $\log$ to denote logarithm with base 2. A multi-distribution learning instance $\mathcal{D}$ is a collection of k distributions $(D_1, \ldots, D_k)$ over $\mathcal{X} \times \mathcal{Y}$. The error of h on a specific distribution D is given by $L(h, D) = \mathbb{E}_{(x,y)\sim D}\ell(h, (x, y))$; when the context is clear, we write $L_i(h)$ for $L(h, D_i)$ and define h's (worst-case) *multi-distribution error* as $L_{\mathcal{D}}(h) = \max_{i\in[k]} L(h, D_i)$. When the context is clear, we drop the subscript and abbreviate $L_{\mathcal{D}}(h)$ as $L(h)$. For any distribution w on $[k]$, define the mixture distribution $D_w = \sum_{i\in[k]} w(i)D_i$ with corresponding error $L(h, w) = \mathbb{E}_{(x,y)\sim D_w}\ell(h, (x, y))$. For any subset of hypotheses $V \subseteq \mathcal{H}$, its disagreement region is defined as $\mathrm{DIS}(V) := \{x \in \mathcal{X} : \exists h_1, h_2 \in V \text{ with } h_1(x) \neq h_2(x)\}$, and its complement (the agreement region) is $\mathrm{AGR}(V) := \mathcal{X} \setminus \mathrm{DIS}(V)$. Given V and an example $x \in \mathrm{AGR}(V)$, we slightly abuse notation and denote by $V(x)$ the prediction of any classifier $h \in V$ on x, which is independent of the choice of h. Given a distribution D, the disagreement metric between hypotheses h and h' is $\rho_D(h, h') = \Pr_{x\sim D}[h(x) \neq h'(x)]$, and the disagreement ball centered at h with radius r is $\mathbf{B}_D(h, r) = \{h' \in \mathcal{H} : \rho_D(h, h') \leq r\}$. We abbreviate $\rho_i(h, h') := \rho_{D_i}(h, h')$, and define $\rho(h, h') := \max_{i\in[k]} \rho_i(h, h')$; it can be seen that ρ is also a metric. It is well-known that triangle and reverse triangle inequalities hold, i.e., for any h, h' and $i \in [k]$,

$$\left|L_i(h) - L_i(h')\right| \leq \rho_i(h, h') \leq L_i(h) + L_i(h'), \quad \left|L(h) - L(h')\right| \leq \rho(h, h') \leq L(h) + L(h').$$

The disagreement coefficient of a reference hypothesis h^* (with respect to $\mathcal{H}$ and D) (Hanneke, 2007) is given by

$$\theta_{D,\mathcal{H},h^*}(r_0) = \sup_{r\geq r_0} \frac{\Pr_{x\sim D}[x \in \mathrm{DIS}(\mathbf{B}_D(h^*, r))]}{r}$$

(abbreviated as θ_D, or θ_i when $D = D_i$). It is well-known that $\theta_D(r_0) \leq \frac{1}{r_0}$ and can be much smaller (Hanneke, 2007; Friedman, 2009; Wang, 2011). We use $\theta_{\max}(r_0) := \max_{i\in[k]} \theta_i(r_0)$ to denote the maximum θ_i among all distributions. The star number of $\mathcal{H}$ with respect to a reference hypothesis h^* (Hanneke and Yang, 2015) is defined as the largest integer $\mathfrak{s}$ for which there exist points $x_1, \ldots, x_{\mathfrak{s}} \in \mathcal{X}$ and hypotheses $h_1, \ldots, h_{\mathfrak{s}} \in \mathcal{H}$ satisfying $h_i(x_i) \neq h^*(x_i)$ and $h_i(x_j) = h^*(x_j)$ for all $j \neq i$; $\mathfrak{s}$ is defined as ∞ if the size of such point set can be arbitrarily large.

We consider active multi-distribution learning in the PAC setting (Rittler and Chaudhuri, 2023; Hanneke, 2014; Haghtalab et al., 2022). For each distribution D_i, the learner has access to two oracles which it can query interactively: an example oracle EX_i that independently draws an unlabeled example from D_i's marginal distribution over $\mathcal{X}$ $D_{i,\mathcal{X}}$, and a labeling oracle $\mathcal{O}_i$ that returns a label $y \sim \Pr_{D_i}(y \mid x)$ for any queried example x. The goal is to output a hypothesis $\hat{h}$ such that $L\left(\hat{h}\right) \leq \nu + \varepsilon$, with probability at least $1 - \delta$, using as few label queries as possible, where $(h^*, \nu) = (\arg\min_{h\in\mathcal{H}} L(h), \min_{h\in\mathcal{H}} L(h))$ is the best hypothesis in $\mathcal{H}$ and its multi-distribution error, respectively. The total number of queries made by the learner to any of the $\mathcal{O}_i$, as a function of ε and δ, is called its *label complexity*.

4. Large ε Regime

We start with studying the setting that there is small amount of label noise in the MDL instance $\mathcal{D}$, specifically, the target excess error ε is much larger than the optimal multi-distribution error ν:

Assumption 1 *The target error $\varepsilon \geq 100\nu$, where we recall that $\nu = \min_{h\in\mathcal{H}} L(h)$.*

This setting captures the realizable setting where there exists some $h^* \in \mathcal{H}$ such that $L_i(h) = 0$ for all $i \in [k]$, and is more general in that it allows ν to be a constant factor smaller than ε. In this setting, state-of-the-art algorithms (Rittler and Chaudhuri, 2023, Algorithm 2 and Theorem 3) has a label complexity of $O\left(dk\theta_{\max}(\varepsilon)\ln\frac{1}{\varepsilon}\right)$; albeit novel, it is not clear if this bound is the best we can hope for. For example, in the realizable case where $\nu = 0$, when $\theta_{\max}(\varepsilon) = O\left(\frac{1}{\varepsilon}\right)$, the bound translates to $\widetilde{O}\left(\frac{dk}{\varepsilon}\right)$, which is worse than the state-of-the-art passive MDL sample complexity $\widetilde{O}(\frac{d+k}{\varepsilon})$ (Blum et al., 2017). This motivates our question: can we design an active MDL algorithm with label complexity no worse than their passive counterparts?

To tackle this question, an naive approach is to perform single-distribution active learning on the average distribution $\overline{D} = \frac{1}{k}\sum_{i=1}^{k} D_i$ with target excess error $\frac{\varepsilon}{k}$. By standard guarantees in single-distribution active learning, this yields an algorithm that works in the realizable setting, with a label complexity of $O(\theta_{\overline{D}}(\frac{\varepsilon}{k})\cdot d\cdot\ln\frac{k}{\varepsilon})$. While simple and general, we show that its $\theta_{\overline{D}}(\frac{\varepsilon}{k})$ dependence can have a hidden k factor in general.

Proposition 1 *For any $k \in \mathbb{N}$ and $\varepsilon > 0$, there exist a MDL instance $\mathcal{D} = (D_i)_{i=1}^k$ and a hypothesis class $\mathcal{H}$ such that $\theta_{\overline{D}}(\frac{\varepsilon}{k}) \geq k\theta_{\max}(\varepsilon)$.*

The proof of Proposition 1 can be found in Appendix A. Importantly, it shows that the above naive approach still leads to an undesirable $O(kd)$ dependence in label complexity, in the worst case.

To bypass this $O(kd)$ barrier, we next propose an algorithm, Algorithm 1, with a label complexity of $O\left((k+d)\theta_{\max}(\varepsilon)\ln\frac{1}{\varepsilon}\right)$. In sharp contrast with the previous approach that reduces active MDL to single-distribution active learning in one shot, Algorithm 1 reduces active MDL to a series of passive MDL problems with decreasing target excess error, similar to phased and robust versions of disagreement-based active learning (Cohn et al., 1994; Hanneke, 2014; Hsu, 2010).

Algorithm 1 maintains a version space, $V_n \subset \mathcal{H}$ and aims to shrink it iteratively (step 5). Similar to disagreement-based active learning for the single-distribution setting (Hanneke, 2014), it queries the label of an example whenever it lies in the disagreement region of V_n. Specifically, at iteration n, it uses the algorithm of Blum et al. (2017) to perform passive multi-distribution learning on MDL instance $\mathcal{D}_n = (D_{i,n})_{i\in[k]}$, whose individual distributions' probability mass functions (PMFs) are defined as[1]:

$$D_{i,n}(x,y) = D_i(x,y)I(x\in\mathrm{DIS}(V_{n-1})) + D_i(x)I(y = V_{n-1}(x))I(x\in\mathrm{AGR}(V_{n-1})), \quad (1)$$

where we recall that $V_{n-1}(x)$ denotes the unanimous prediction of classifiers in V_{n-1} on $x \in \mathrm{AGR}(V_{n-1})$.

1. More generally, when $\mathcal{X}$ is not necessarily discrete, define $D_{i,n}$ as:

$$\Pr_{D_{i,n}}[(x,y)\in A] := \Pr_{D_i}[x\in\mathrm{DIS}(V_{n-1})\wedge(x,y)\in A] + \Pr_{D_i}[x\in\mathrm{AGR}(V_{n-1})\wedge(x,V_{n-1}(x))\in A],$$

for any measurable $A \subset \mathcal{X}\times\mathcal{Y}$. In this case, Proposition 9 can be proved similarly.

A random sample (x, y) from $D_{i,n}$ can be obtained in a label-efficient manner as in Algorithm 5 in Appendix A.1: first, use example oracle EX_i to draw $x \sim D_{i,X}$ (step 1); if $x \in \mathrm{DIS}(V_{n-1})$, query the labeling oracle $\mathcal{O}_i$ for label y (step 3); otherwise $x \in \mathrm{AGR}(V_{n-1})$, we infer label y to be the prediction of any $h \in V_{n-1}$ on x (step 5). As we will see, Algorithm 1 maintains the invariant that $h^* \in V_n$, and thus the inferred label equals $h^*(x)$, which maintains a favorable bias for PAC learning the original distributions (Hsu, 2010, Lemma 5.2; see also Lemma 25). It can be readily seen that each call has expected label cost $\Pr_{D_i}[x \in V_{n-1}]$ (Proposition 9 in Appendix A.1).

Algorithm 1 Distribution-dependent active multi-distribution learning, large ε regime

Require: Target error $\varepsilon > 0$, failure probability $\delta > 0$

1: Initialization: $V_0 \leftarrow \mathcal{H}$, $n_0 \leftarrow \lceil \log \frac{1}{\varepsilon} \rceil$, $\varepsilon_n = 2^{-n}$, $\delta_n = \frac{\delta}{2n^2}$
2: **for** $n = 1, .., n_0$ **do**
3: Define $D_{i,n}$ as in Eq. (1), for all $i \in [k]$.
4: $h_n \leftarrow \textsc{Passive-MDL}(V_{n-1}, (D_{i,n})_{i\in[k]}, \varepsilon_n, \delta_n)$, where we choose PASSIVE-MDL to be the passive multi-distribution learning algorithm of Blum et al. (2017), and use Algorithm 5 to obtain iid samples from $D_{i,n}$.
5: Update version space $V_n \leftarrow \{h \in V_{n-1} : \rho(h, h_n) \le 2\varepsilon_n\}$
6: **return** $\hat{h}$, an arbitrary classifier from V_{n_0}.

We now present the guarantee of Algorithm 1:

Theorem 2 *Suppose Assumption 1 holds. If Algorithm 1 takes into target error ε and failure probability δ, then with probability $1 - \delta$, (1) its output classifier $\hat{h}$ is such that $L(\hat{h}) \le \nu + \varepsilon$, (2) it queries $O\left((d + k)\theta_{\max}(\varepsilon) \ln \frac{1}{\varepsilon}\right)$ labels.*

Our theorem implies a $O\left((d + k)\theta_{\max}(\varepsilon) \ln \frac{1}{\varepsilon}\right)$ label complexity upper bound for active MDL. In Section 6 below, we show a $\Omega(k\theta_{\max}(\varepsilon))$ information-theoretic label complexity lower bound; this, combined with the $\Omega(d\theta(\varepsilon))$ lower bound in Hanneke (2014), shows that our label complexity upper bound is unimprovable in general, up to a $O\left(\ln \frac{1}{\varepsilon}\right)$ factor.

The proof of Thoerem 2 can be found in Appendix A. Its key idea is as follows: Algorithm 1 iteratively shrinks the version spaces V_n and maintains two invariants with high probability: first, $h^* \in V_n$; second, for all $h \in V_n$, $\max_{i\in[k]} \rho_i(h, h^*) = \rho(h, h^*) \le O(\varepsilon_n)$. The first invariant ensures that the distributions $D_{i,n}$ have favorable biases (as mentioned above), ensuring the final PAC learning guarantee. The second invariant generalizes similar claims in the analyses of single-distribution active learning (e.g. Hanneke, 2014; Hsu, 2010); it ensures that the disagreement region of V_n are have small probabilities (i.e., $\le \theta_{\max}(\varepsilon) \cdot \varepsilon_n$) under all distributions $(D_i)_{i=1}^k$. To this end, it learns h_n such that it has small errors (i.e., $\le \varepsilon_n$) under all $D_{i,n}$ simulatenously (see Lemma 10).

5. Active MDL in Small ε Regime

We next move on to the more challenging small ε regime, where ν, the noise level in the distributions $(D_i)_{i=1}^k$, is large compared with target excess error ε. As mentioned in the introduction section, state-of-the-art result of Rittler and Chaudhuri (2023) generalizes the idea of robust single-distribution active learning, achieving a label complexity of $\widetilde{O}\left(\frac{\nu^2}{\varepsilon^2} kd\theta_{\max}(\nu)^2 + \frac{k}{\varepsilon^2}\right)$. Again, this label complexity may sometimes be worse than passive learning: for example, when

$\theta_{\max}(\nu) = O(\frac{1}{\nu})$, the bound becomes $O(\frac{kd}{\varepsilon^2})$, which is higher than state-of-the-art passive multi-distribution learning sample complexity $O(\frac{k+d}{\varepsilon^2})$ (Zhang et al., 2024; Peng, 2024).

Although Algorithm 1 has near-optimal label complexity when $\varepsilon \geq O(\nu)$, it cannot be directly applied to the small-ε regime that $\varepsilon \leq O(\nu)$. Indeed, the version space construction of Algorithm 1 (step 5) ensures the invariant $h^* \in V_n$ only when the target excess error ε is much larger than ν. We further illustrate the challenge with an example, showing the necessity of label querying in the agreement region of the hypothesis class:

Example 1 *Let $\mathcal{H} = \{h_1, h_2\}$. Suppose D_1 is a distribution over $X_1 \subseteq \mathcal{X}$ with $L_1(h_1) = 0$ and $L_1(h_2) = 2\nu'$. Also, let $D_2 = \nu' D_{X_2} + (1-\nu')D_{X_2'}$, where D_{X_2} and $D_{X_2'}$ are distributions over X_2 and X_2', respectively, with $X_2 \subseteq \mathrm{DIS}(\mathcal{H})$ and $X_2' \subseteq \mathrm{AGR}(\mathcal{H})$. Assume further that X_1, X_2, and X_2' are pairwise disjoint, and that $L(h_1, D_{X_2}) = 1$ while $L(h_2, D_{X_2}) = 0$. Now consider two cases for setting the error in the agreement region:*

(a) If we set $(1-\nu')L(h_1, D_{X_2'}) = (1-\nu')L(h_2, D_{X_2'}) \leq \nu' - \varepsilon$, then $L_2(h_1) \leq 2\nu' - \varepsilon$ and $L_2(h_2) \leq \nu' - \varepsilon$, so that h_1 is the only valid output when the target excess error is ε.

(b) If we set $(1-\nu')L(h_1, D_{X_2'}) = (1-\nu')L(h_2, D_{X_2'}) \geq \nu' + \varepsilon$, then $L_2(h_1) \geq 2\nu' + \varepsilon$, making h_2 the only valid output when the target excess error is ε.

Note that this example is valid only when $\varepsilon \ll \nu$ (since $\nu' - \varepsilon \geq 0$), which aligns with our discussion.

In view of this, we present an algorithm, namely Algorithm 2, for active MDL in the small ε regime. Our approach consists of two stages. In stage one, we run Algorithm 1 with a target excess error of $\varepsilon = 100\nu$, obtaining a version space V_0 that contains h^* with small radius, in the sense that $\max_{h \in V_0} \rho(h, h^*) \leq O(\nu)$. From Theorem 2, this stage costs $O(\theta_{\max}(\nu)(d+k)\ln\frac{1}{\nu})$ label queries. After obtaining V_0, we introduce a second stage that relies on estimating the error of h^* in the agreement region for each of the k distributions. In particular, it first draws iid samples $(S_i)_{i=1}^k$ in the agreement region and use their empirical distributions as surrogates of $D_i|_{\mathrm{AGR}(V_0)}$ (step 5), and then runs the passive MDL algorithm on distributions $(D_i')_{i=1}^k$, which can be viewed as surrogates of $(D_i)_{i=1}^k$ (step 6). Sampling from D_i' can be done in a label-efficient manner, as we detail in Algorithm 3: with probability $1 - \Pr_{D_i}[x \in \mathrm{DIS}(V_0)]$, we draw samples from $D_i|_{\mathrm{DIS}(V_0)}$ by making new label queries; otherwise, we draw an example uniformly at random from the already-labeled dataset S_i.

The following theorem gives the correctness and label complexity of Algorithm 2.

Theorem 3 *If Algorithm 2 is run with target error ε and failure probability δ, then with probability at least $1-\delta$: (1) its output classifier $\hat{h}$ satisfies $L\left(\hat{h}\right) \leq \nu + \varepsilon$, and (2) its label complexity is at most:*

$$O\left(\left(\frac{k\nu}{\varepsilon^2} + \frac{\theta_{\max}\left(100\nu\right)(d+k)(\nu+\varepsilon)^2}{\varepsilon^2}\right) \cdot \mathrm{polylog}\left(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta}\right)\right)$$

The proof of Theorem 3 and the auxiliary lemmas are provided in Appendix B. On a high level, the second stage's label cost dominates that of the first stage. In the second stage, we choose the number of samples n_0 for S_i so that we estimate h^*'s error in $\mathrm{AGR}(V_0)$ with precision $O(\varepsilon)$, which allows us to apply the sample complexity bound of the passive MDL algorithm; specifically $n_0 = \widetilde{O}(\frac{\nu}{\varepsilon^2})$ is enough, resulting in the $\frac{k\nu}{\varepsilon^2}$ term in the label complexity.

Algorithm 2 Distribution-dependent Active Multi-distribution Learning, Small ε Regime

Require: Target excess error $\varepsilon > 0$, optimal error ν, failure probability $\delta > 0$.

1: Run Algorithm 1 with target excess error $\varepsilon' = 100\nu$ and failure probability $\delta' = \frac{1}{6}\delta$; let h' be the output.

2: Initialize $V_0 \leftarrow \mathcal{H}' = \{h \in \mathcal{H} : \mathrm{dist}(h, h') \le 2\varepsilon'\}$ and $n_0 = \left\lceil \frac{100(\varepsilon+\nu)}{\varepsilon^2} \ln \frac{k}{\delta'} \right\rceil$.

3: **for** $i = 1, \ldots, k$ **do**

4: Sample n_0 samples from D_i conditioned on $x \in \mathrm{AGR}(V_0)$ and query their labels; denote the sample set as $S_i = \{(x_1, y_1), \ldots, (x_{n_0}, y_{n_0})\}$.

5: Define the empirical distribution $D_{S_i}(x, y) = \frac{1}{n_0} \sum_{(x',y') \in S_i} I(x' = x \text{ and } y' = y)$.

6: Define the surrogate distribution D_i' as

$$D_i'(x, y) = I\left(x \in \mathrm{DIS}(V_0)\right) D_i(x, y) + \Pr_{x \sim D_{i,\mathcal{X}}}[x \in \mathrm{AGR}(V_0)]\, D_{S_i}(x, y).$$

7: **return** $\hat{h} \leftarrow$ PASSIVE-MDL $\left(V_0, \{D_i'\}_{i \in [k]}, \frac{\varepsilon}{2}, \frac{\delta}{6}\right)$, where we choose PASSIVE-MDL to be the `MDL-Hedge-VC` algorithm of Zhang et al. (2024) (with different hyperparameters; see Section 5.1), and use Algorithm 3 to obtain iid samples from D_i'.

Algorithm 3 Sampling from D_i'

Require: Samples $(S_i)_{i=1}^k$ drawn from $(D_i)_{i=1}^k$, and version space V_0.

1: Use the example oracle EX_i to sample $x \sim D_{i,\mathcal{X}}$.

2: **if** $x \in \mathrm{DIS}(V_0)$ **then**

3: Set $y \leftarrow \mathcal{O}_i(x)$.

4: **else**

5: Uniformly sample (x, y) from S_i.

6: **return** (x, y).

Furthermore, suppose PASSIVE-MDL makes a total of n_1 calls to any of the sampling procedure from D_i''s; by Chernoff bound, it will make an additional $n_1 \max_i \Pr_{D_i}(x \in \mathrm{DIS}(V_0))$ label queries. Note that if we were to directly apply the passive sample complexity upper bound $n_1 \le \widetilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$ from Zhang et al. (2024), we would obtain a label complexity bound of $\widetilde{O}\left(\frac{k\nu}{\varepsilon^2} + \frac{\theta_{\max}(100\nu)(d+k)\nu}{\varepsilon^2}\right)$. Such bound still does not fully exploit the benefit of having small ν: for example, when $\nu = O(\varepsilon)$ and $\theta_{\max}(100\nu) = O(\frac{1}{\nu})$, this gives a $\widetilde{O}(\frac{d+k}{\varepsilon^2})$ label complexity, while the naive $\widetilde{O}(\frac{kd(\nu+\varepsilon)}{\varepsilon^2})$ sample complexity baseline we mentioned in Section 1 evaluates to $\widetilde{O}(\frac{kd}{\varepsilon})$, which can be significantly better for small ε. Motivated by this, we refine the sample complexity bound of the passive MDL algorithm by Zhang et al. (2024) and use it to establish an improved $\widetilde{O}\left(\frac{k\nu}{\varepsilon^2} + \frac{\theta_{\max}(100\nu)(d+k)(\nu+\varepsilon)^2}{\varepsilon^2}\right)$ label complexity as stated in Theorem 3.

5.1. Refined Sample Complexity Bounds for Passive MDL

We refine the sample complexity of `MDL-Hedge-VC`, the passive MDL algorithm of Zhang et al. (2024, Algorithm 1) from $\widetilde{O}(\frac{d+k}{\varepsilon^2})$ to $\widetilde{O}(\frac{(d+k)(\varepsilon+\nu)}{\varepsilon^2})$, by choosing different hyperparameters η, T

and T_1. For completeness, we give a brief recap of `MDL-Hedge-VC` in Appendix C. This is analogous to refined sample complexity analysis of empirical risk minimization from $\widetilde{O}(\frac{d}{\varepsilon^2})$ to $\widetilde{O}(\frac{d(\nu+\varepsilon)}{\varepsilon^2})$ in the single-distribution PAC learning setting (Boucheron et al., 2005; Vapnik, 1982, Section 5). When both ν and ε are $\ll 1$, Our new bound is tighter. We summarize our refined algorithm and guarantees below, with proofs deferred to Appendix C:

Theorem 4 *Set* $\varepsilon_1 = \frac{\varepsilon}{100}$ *and* $\eta = \frac{\varepsilon_1}{100(\varepsilon_1+\nu)}$. *Further, set* $T = 20000\left(\frac{1}{\varepsilon_1} + \frac{\nu}{\varepsilon_1^2}\right)\ln\left(\frac{k}{\delta\varepsilon}\right)$ *and* $T_1 = 4000\left(\frac{1}{\varepsilon_1} + \frac{\nu}{\varepsilon_1^2}\right)\left(k\ln\left(\frac{k}{\varepsilon}\right) + d\ln\left(\frac{kd}{\varepsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)$. *Then the randomized hypothesis* h^{final} *returned by Algorithm 6 satisfies* $L(h^{\text{final}}) \le \nu + \varepsilon$ *with probability at least* $1-\delta$, *provided the total sample size exceeds*

$$O\left(\frac{(d+k)(\nu+\varepsilon)}{\varepsilon^2}\cdot \text{polylog}(k, d, 1/\varepsilon, 1/\delta)\right).$$

Notably, our bound provides a smooth transition between the $\widetilde{O}\left(\frac{d+k}{\varepsilon^2}\right)$ bound in Zhang et al. (2024) for the agnostic setting and the $\widetilde{O}\left(\frac{d+k}{\varepsilon}\right)$ bound in Blum et al. (2017) for the realizable setting. At a high level, the improvement in the sample complexity bounds is achieved via better concentration bounds. We employ Bernstein-style concentration bounds instead of Hoeffding-style bounds, which yield a tighter analysis. One caveat is that since we have increased the step size η, a more delicate analysis is required to bound the norm of the weight vector $\|\overline{w}^T\|_1$, which is crucial in establishing the sample complexity bounds.

6. Lower Bounds

We present our lower bounds in this section, for realizable and agnostic cases, respectively. Our lower bounds are information-theoretic in nature, and apply to algorithms we present in previous sections.

6.1. Lower Bound in the Realizable Setting

We first prove a $\Omega(k\theta_{\max})$ lower bound in the realizable setting. As mentioned in Section 4, this lower bound in conjunction with the $\Omega(d\theta_{\max})$ lower bound in the single distribution setting (Hanneke, 2014) shows that the label complexity given in Theorem 2 is not improvable in general.

Theorem 5 *For any* $k \ge 2, d \ge 1, \vartheta \ge 1, \varepsilon \in (0, \frac{1}{2\vartheta})$, *and hypothesis class* $\mathcal{H}$ *whose* VC-dim$(\mathcal{H}) \le d$ *and star number is* $\ge k\vartheta$, *and proper active learning algorithm* A, *there exists a problem instance* $\mathcal{D} = (D_i)_{i\in[k]}$ *such that: (1)* $\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$; *(2)* $\theta_{\max}(\varepsilon) \le \vartheta$; *(3) unless* A *queries* $\Omega(k\vartheta)$ *labels, the probability that* A *returns an* ε*-optimal hypothesis is at most* 0.7.

In proving Theorem 5, we construct an example with $k\theta_{\max}$ points and k distributions, each span a disjoint section of the feature space, which any algorithm has to query at least a constant fraction to get information.

6.2. Lower Bound in the Agnostic Setting

We next prove a $\Omega\left(\frac{k\nu}{\varepsilon^2}\right)$ label copmlexity lower bound for all proper active MDL learners:

Theorem 6 *For any $k \geq 2, d \geq 2, \varepsilon > 0, \frac{1}{2} \geq \nu > 0$ with $\nu \geq 8\varepsilon$, and a hypothesis class $\mathcal{H}$ that contains h_1, h_2 that agrees on at least two examples and disagrees on k examples with $\text{VC-dim}(\mathcal{H}) \leq d$, any proper active learning algorithm A, there exists a problem instance $(D_i)_{i\in[k]}$ such that: (1) $\min_{h\in\mathcal{H}} L(h, \mathcal{D}) \leq \nu$; (2) unless A queries $\Omega\left(k\frac{\nu}{\varepsilon^2}\right)$ labels, the probability that A returns an ε-optimal hypothesis is at most 0.9.*

Theorem 2 demonstrates the necessity of the $k\frac{\nu}{\varepsilon^2}$ term of label complexity in Theorem 3, since Algorithm 2 is a proper learning algorithm. This formalizes the intuition in Rittler and Chaudhuri (2023, Example 1) and our Example 1 that, in sharp contrast to active single-distribution learning, sampling from the disagreement region only may not be enough for active multi-distribution learning. Indeed, our lower bound instances highlight the need in querying the labels in the agreement region of the hypothesis class, ensuring the returned classifier balances its worst-case error across all k distributions (see Appendix D for the proof).

In light of our upper bound in Section 4, we observe that such $\Omega(k\frac{\nu}{\varepsilon^2})$ lower bound cannot appear in the large ε regime, i.e., when $\nu \leq \frac{\varepsilon}{100}$. This shows an interesting "phase transition" behavior of the fundamental label complexity of active multi-distribution learning, which does not appear in classical PAC active learning (Hanneke, 2014; Raginsky and Rakhlin, 2011). We conjecture that similar lower bounds can be established for improper learning algorithms, and leave it as an interesting open question.

7. Active MDL Algorithms in Distribution-free Settings

Theorems 2 and 3 provide useful algorithmic results on active MDL with improved label complexity; even though optimality properties have been established, their bounds can sometimes be suboptimal when considering other problem-dependent complexity measures. One important distribution-free quantity that characterizes the complexity of active learning is the *star number* of the hypothesis class $\mathcal{H}$ (Hanneke and Yang, 2015), denoted as $\mathfrak{s}$ (recall its definition in Section 3). It is known that for any distribution D and $r > 0$, $\mathfrak{s} \geq \theta_D(r)$ (Hanneke and Yang, 2015), and this bound can sometimes be tight. Translating our distribution-dependent upper bounds Theorems 2 and 3 in terms of star number, we obtain that Algorithms 1 and 2 have label complexities $O\left(\mathfrak{s}(d+k)\ln\frac{1}{\epsilon}\right)$ and $O\left(\mathfrak{s}(d+k)\left(1+\frac{\nu^2}{\varepsilon^2}\right)+\frac{d\nu}{\varepsilon^2}\right)$, in the small and large ε regimes, respectively. How tight are these bounds?

We start with making the simple observation that these direct translations can sometimes result in suboptimal label complexity bounds. Specifically, in the realizable setting ($\nu = 0$), consider the naive algorithm that performs disagreement-based active learning over the average distribution $\overline{D} = \frac{1}{k}\sum_{i=1}^k D_i$ with target error $\frac{\varepsilon}{k}$. This ensures that we return $\hat{h}$ such that $\frac{1}{k}\sum_{i=1}^k L_i(\hat{h}) \leq \frac{\varepsilon}{k}$, and therefore $L_i(\hat{h}) \leq \varepsilon$. Its label complexity is at most $\widetilde{O}\left(\mathfrak{s}\ln\frac{k}{\varepsilon}\right)$ (Wiener et al., 2015), much better than $O\left(\mathfrak{s}(d+k)\ln\frac{1}{\epsilon}\right)$ provided by Algorithm 1. However, when generalizing to the nonrealizable setting ($\nu > 0$), this reduction can only yield PAC guarantees when $\nu \lesssim \frac{\varepsilon}{k}$, and in combination with state-of-the-art distribution-free agnostic active learning guarantees (Hanneke and Yang, 2015, Theorem 8), this gives a label complexity of $\widetilde{O}\left(\mathfrak{s}d\,\text{polylog}(\frac{1}{\varepsilon})\right)$. This motivates our question: can we design active MDL algorithms with sharp label complexity guarantees in terms of star number, that smoothly interpolates between realizable and nonrealizable regimes?

We answer this question in the positive in this section by designing an active MDL algorithm, Algorithm 4 with label complexity $\widetilde{O}(\mathfrak{s}\ln\frac{1}{\varepsilon})$ when the target error $\varepsilon \geq (k+d)\nu$. Similar to

Algorithm 4 Distribution-free active multi-distribution learning, large ε regime

Require: Target error $\varepsilon > 0$, failure probability $\delta > 0$

1: Initialization: $V_0 \leftarrow \mathcal{H}$, $n_0 \leftarrow \lceil \log \frac{(d+k)}{\mathfrak{s}\varepsilon} \rceil$, $\varepsilon_n = 2^{-n}$, $\delta_n = \frac{\delta}{2n^2}$, and $f_0 \equiv 0$.
2: **for** $n = 1, .., n_0$ **do**
3: Define $D_{i,n}$ as:

$$D_{i,n}(x, y) = D_i(x, y)I(f_{n-1}(x) = 0) + I(y = f_{n-1}(x))I(f_{n-1}(x) \neq 0),$$

for all $i \in [k]$.
4: **if** $n < n_0$ **then**
5: $f_n \leftarrow$ PASSIVE-RPU-MDL$(\mathcal{H}, \{D_{i,n}\}_{i\in[k]}, \varepsilon_n, \delta_n)$
6: **else**
7: // In this case, $\Pr(f_n(x) = 0) \leq \frac{\mathfrak{s}}{(d+k)}\varepsilon$
8: **return** $\hat{h} \leftarrow$ PASSIVE-MDL$(\mathcal{H}, \{D_{i,n}\}_{i\in[k]}, \varepsilon_n, \delta_n)$.

distribution-free algorithms for single-distribution active learning (Wiener et al., 2015; Kane et al., 2017), Algorithm 4 progressively learns Reliably and Probably Useful (RPU) classifiers f_n (Rivest and Sloan, 1988; Hopkins et al., 2020) with larger coverages. RPU classifiers are those that have an extra option of outputting 0, indicating "I don't know"; we now give its formal definition:

Definition 7 *A classifier $f : \mathcal{X} \to \{-1, +1, 0\}$ is said to be ξ-Reliable and Probably Useful (RPU) with respect to $h^* : \mathcal{X} \to \{-1, +1\}$ and distribution D, if it is simultaneously:*

1. *Reliable:* $\Pr[f(x) \neq 0, f(x) \neq h^*(x)] = 0$.
2. *Probably useful:* $\Pr[f_n(x) = 0] \leq \xi$.

Additionally, f is said to be ξ-RPU with respect to h^* and $\mathcal{D} = (D_i)_{i=1}^k$, if it is ξ-RPU with respect to h^*, D_i for all i's.

Specifically, we design a subprocedure PASSIVE-RPU-MDL (Algorithm 7 in Appendix E.1) to iteractively refine such RPU classifiers using label queries (step 5), such that the abstention probabilities of f_n shrink exponentially in n, uniformly across all distributions $\{D_i\}_{i=1}^k$. Taking advantage of active learning, at iteration n we only make label queries in the abstention region $\{x : f_{n-1}(x) = 0\}$ (see Definition of $D_{i,n}$ in step 3).

At the last epoch n_0, we have access to an RPU classifier f_{n_0-1} with abstention probability $O(\frac{\mathfrak{s}}{d+k}\varepsilon)$ in all D_i's. Were we to directly covert f_{n_0-1} to a binary classifier by predicting arbitrarily in its abstention region, we would get a somewhat undesirable $O(\frac{\mathfrak{s}}{(d+k)}\varepsilon)$ excess error guarantee. Our key observation here is that, with a constant factor overhead of label cost, we can do a better RPU-to-PAC conversion by reusing the PASSIVE-MDL procedure in the preceding sections. Specifically, with $O(\mathfrak{s})$ labels, PASSIVE-MDL on $D_{i,n}$'s outputs a classifier $\hat{h}$ with excess error ε (step 8).

We present the performance guarantees of Algorithm 4 in the theorem below:

Theorem 8 *Suppose $\varepsilon \geq 100(k + d)\nu$. If Algorithm 4 takes into target error ε and failure probability δ, then with probability $1 - \delta$, (1) its output classifier $\hat{h}$ is such that $L(\hat{h}) \leq \nu + \varepsilon$, (2) it queries $O\left(\mathfrak{s} \ln \frac{1}{\varepsilon}\right)$ labels.*

The proof of Theorem 8 can be found in Appendix E. Its main ideas are twofold. First, we argue that with $\widetilde{O}(\mathfrak{s})$ label queries, at every iteration n, PASSIVE-RPU-MDL computes f_n whose abstention region is half the size of those of f_{n-1}, measured in all D_i's. Due to the presence of label noise, PASSIVE-RPU-MDL needs to be designed in a robust way; specifically it relies on our design of a new single-distribution RPU algorithm, ROBUST-RPU-LEARN (Algorithm 8), that can tolerate adversarial label noise. Second, we argue that at the last iteration, calling PASSIVE-MDL makes a total of $O(\frac{d+k}{\varepsilon})$ samples to any of the D_i's; since we only need to query f_{n_0-1}'s abstention regions, this results in a factor of $O(\frac{\mathfrak{s}}{d+k}\varepsilon)$ label savings, yielding a final label complexity of $O(\mathfrak{s}\ln\frac{1}{\varepsilon})$.

For the single-distribution setting that $k = 1$, Algorithm 4 achieves a label complexity of $O(\mathfrak{s}\ln\frac{1}{\varepsilon})$ when the target error $\varepsilon \geq O(d\nu)$. To the best of our knowledge, this result already improves over the state of the art $O(d\mathfrak{s})$ (Hanneke and Yang, 2015, Theorem 8), which may be of independent interest.

8. Conclusion

In this paper, we establish novel and tighter label complexity bounds for active MDL, significantly improving the dependence on kd to an additive $(k + d)$. Specifically, we develop algorithms and lower bounds in both the large and small ε regimes, achieving distribution-dependent and distribution-free bounds that are sometimes tight. Our results bridge the gap between MDL and active PAC learning, refining previous bounds and providing a more nuanced understanding of label complexity in multi-distribution settings. All our algorithms we present are proper, and our $\Omega(\frac{k\nu}{\varepsilon^2})$ lower bound applies to proper learning only; it would be interesting to investigate whether improper learning has an benefit in improving label complexity. Our passive and active learning algorithms requires the knowledge of ν; it would be nice to design adaptive algorithms without such knowledge. Another promising direction is to analyze a wider variety of noise settings beyond agnostic with optimal error ν, such as the ones studied in Hanneke and Yang (2015). We are also interested in designing computationally efficient versions of our algorithms, perhaps by utilizing regression-based active learning (e.g., Zhu and Nowak, 2022; Sekhari et al., 2023).

Acknowledgments

We thank the anonymous COLT reviewers for their constructive comments. We thank Eric Zhao for helpful communications about the results in Haghtalab et al. (2022). We thank Steve Hanneke for helpful conversations about some preliminary results in this paper. We thank Zihan Zhang for confirming a technical detail in Zhang et al. (2024). CZ would like to thank Nick Rittler and Kamalika Chaudhuri for sparking interest in the active multi-distribution learning problem. CZ acknowledges support from the University of Arizona FY23 Eighteenth Mile TRIF Funding. YZ was supported by the NSF CAREER Award CCF-1751040 and by the NSF AI Institute for Foundations of Machine Learning (IFML). YZ thanks his advisor, Eric Price, for his encouragement and guidance. .

References

Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228. PMLR, 2013.

Nir Ailon, Ron Begleiter, and Esther Ezra. Active learning using smooth relative regret approximations with applications. In *Conference on Learning Theory*, pages 19–1. JMLR Workshop and Conference Proceedings, 2012.

Pranjal Awasthi, Nika Haghtalab, and Eric Zhao. Open problem: The sample complexity of multidistribution learning for vc classes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5943–5949. PMLR, 2023.

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316. PMLR, 2013.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72, 2006.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pages 335–342. Omnipress, 2008.

Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Nicolo Cesa-Bianchi, Claudio Gentile, and Francesco Orabona. Robust bounds for classification via selective sampling. In *Proceedings of the 26th annual international conference on machine learning*, pages 121–128, 2009.

Jiecao Chen, Qin Zhang, and Yuan Zhou. Tight bounds for collaborative pac learning via multiplicative weights. *Advances in neural information processing systems*, 31, 2018.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. *Advances in neural information processing systems*, 17, 2004.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18, 2005.

Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 20, 2007.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1):2655–2697, 2012.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122, 2020.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28:133–168, 1997.

Eric Friedman. Active learning for smooth problems. In *COLT*, 2009.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35:406–419, 2022.

Rafael Hanashiro and Patrick Jaillet. Distribution-dependent rates for multi-distribution learning. *arXiv preprint arXiv:2312.13130*, 2023.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pages 353–360, 2007.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Steve Hanneke. The star number and eluder dimension: Elementary observations about the dimensions of disagreement. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2308–2359. PMLR, 2024.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(1): 3487–3602, 2015.

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

Daniel Joseph Hsu. *Algorithms for active learning*. PhD thesis, UC San Diego, 2010.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Daniel M Kane, Shachar Lovett, Shay Moran, and Jiapeng Zhang. Active classification with comparison queries. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 355–366. IEEE, 2017.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 341–352, 1992.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 20(65): 1–50, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR, 2019.

Huy Nguyen and Lydia Zakynthinou. Improved algorithms for collaborative pac learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Robert D Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011.

Binghui Peng. The sample complexity of multi-distribution learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4185–4204. PMLR, 2024.

Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. *Advances in Neural Information Processing Systems*, 24, 2011.

Nicholas Rittler and Kamalika Chaudhuri. Agnostic multi-group active learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Ronald L Rivest and Robert Sloan. Learning complicated concepts reliably and usefully. In *Proceedings of the Seventh AAAI National Conference on Artificial Intelligence*, pages 635–640, 1988.

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

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Selective sampling and imitation learning via online regression. *Advances in Neural Information Processing Systems*, 36, 2023.

Christopher Tosh and Sanjoy Dasgupta. Diameter-based active learning. In *International Conference on Machine Learning*, pages 3444–3452. PMLR, 2017.

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

Vladimir Vapnik. Estimation of dependences based on empirical data, 1982.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Liwei Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12(7), 2011.

Zhenyu Wang, Peter Bühlmann, and Zijian Guo. Distributionally robust machine learning with multi-source data. *arXiv preprint arXiv:2309.02211*, 2023.

Yair Wiener, Steve Hanneke, and Ran El-Yaniv. A compression technique for analyzing disagreement-based active learning. *J. Mach. Learn. Res.*, 16:713–745, 2015.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Zihan Zhang, Wenhao Zhan, Yuxin Chen, Simon S Du, and Jason D Lee. Optimal multi-distribution learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5220–5223. PMLR, 2024.

Yihan Zhou and Eric Price. A competitive algorithm for agnostic active learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Yinglun Zhu and Robert Nowak. Efficient active learning with abstention. *Advances in Neural Information Processing Systems*, 35:35379–35391, 2022.

Appendix A. Deferred Materials for Section 4

A.1. Sampling Algorithm

We formally present the sampling algorithm used in Algorithm 1 as the following.

Algorithm 5 Sampling from $D_{i,n}$

Require: Version space V_n
  Query EX_i to sample $x \sim D_{i,X}$
  if $x \in \mathrm{DIS}(V_n)$ **then**
    $y \leftarrow \mathcal{O}_i(x)$
  else
    $y \leftarrow h(x)$, where h is an arbitrary classifier in V_n
  return (x, y)

Then Proposition 9 gives the expected label cost of each call of Algorithm 5 and the proof is given in Appendix A.2:

Proposition 9 *Each call to Algorithm 5 returns (x, y), an independent sample drawn from $D_{i,n}$, and has an expected label cost of $\Pr_{D_i}[x \in V_{n-1}]$.*

A.2. Proofs of Propositions

Proof [Proof of Proposition 1] Let $\mathcal{X} = \{x_0, x_1, \ldots, x_k\}$, and $\mathcal{H} = \{h^*, h_1, \ldots, h_k\}$, such that $h^* \equiv -1$, and for every $i \in [k]$, h_i is defined as:

$$h_i(x_j) = \begin{cases} +1 & j = i \\ -1 & \text{otherwise} \end{cases}$$

For every $i \in [k]$, we construction distribution D_i with PMF:

$$D_i(x, y) = (1 - \varepsilon) I(x = x_0, y = -1) + \varepsilon I(x = x_i, y = +1).$$

It can be readily checked that for every $i \in [k]$, for $r \geq \varepsilon$, $\mathbf{B}_{D_i}(h^*, r) = \mathcal{H}$, and $\mathrm{DIS}(\mathcal{H}) = \{x_1, \ldots, x_k\}$. Therefore, for every $i \in [k]$

$$\theta_i(\varepsilon) = \sup_{r \geq \varepsilon} \frac{\Pr_{D_i}[x \in \mathrm{DIS}(\mathbf{B}_{D_i}(h^*, r))]}{r} = \sup_{r \geq \varepsilon} \frac{\varepsilon}{r} = 1.$$

We now calculate $\theta_{\bar{D}}(\frac{\varepsilon}{k})$. Note that $\bar{D}$ has PMF

$$\bar{D}(x, y) = (1 - \varepsilon) I(x = x_0, y = -1) + \sum_{i=1}^{k} \frac{\varepsilon}{k} I(x = x_i, y = +1).$$

For $r \geq \frac{\varepsilon}{k}$, $\mathbf{B}_{\bar{D}}(h^*, r) = \mathcal{H}$, and $\mathrm{DIS}(\mathcal{H}) = \{x_1, \ldots, x_k\}$. Therefore,

$$\theta_{\bar{D}}(\frac{\varepsilon}{k}) = \sup_{r \geq \frac{\varepsilon}{k}} \frac{\Pr_{\bar{D}}[x \in \mathrm{DIS}(\mathbf{B}_{\bar{D}}(h^*, r))]}{r} = \sup_{r \geq \frac{\varepsilon}{k}} \frac{\varepsilon}{r} = k.$$

Therefore, for this MDL instance $\mathcal{D}$ and hypothesis class $\mathcal{H}$, $\theta_{\bar{D}}(\frac{\varepsilon}{k}) = k \max_{i\in[k]} \theta_i(\varepsilon)$. ■

Proof [Proof of Proposition 9] Denote by (x, y) the random example returned by Algorithm 5. Fix any x_0, y_0:

$$\Pr[x = x_0, y = y_0] = \Pr[x = x_0] \Pr[y = y_0 \mid x = x_0]$$

Now, according to step 1, $\Pr[x = x_0] = D_i(x_0)$. For $\Pr[y = y_0 \mid x = x_0]$, it is equal to $D_i(y_0 \mid x_0)$ according to step 3, and $I(y_0 = V_{n-1}(x_0))$ according to step 5. Therefore,

$$\begin{aligned}\Pr[x = x_0, y = y_0] =& D_i(x_0) \left(D_i(y_0 \mid x_0) I(x_0 \in \mathrm{DIS}(V_{n-1})) + I(y_0 = V_{n-1}(x_0)) I(x_0 \in \mathrm{AGR}(V_{n-1}))\right) \\ =& D_{i,n}(x_0, y_0).\end{aligned}$$

This completes the proof of the first part. For the second part, we note that we make a query to oracle $\mathcal{O}_i$ if $x \in \mathrm{DIS}(V_n)$, which happens with probability $\Pr_{D_i}[x \in \mathrm{DIS}(V_n)]$. ■

A.3. PASSIVE-MDL and its Guarantees

Algorithm 1 uses a passive multi-distribution learning algorithm PASSIVE-MDL as input. Therein, we choose PASSIVE-MDL to be the collaborative PAC learning algorithm of Blum et al. (2017), which we recall has the following guarantee:

Lemma 10 (Blum et al. (2017), Theorem D.2) *Suppose hypothesis class $\mathcal{H}$ and distributions $(\mu_1, \ldots, \mu_k)$ are such that*

$$\min_{h\in\mathcal{H}} \max_{i\in[k]} L(h, \mu_i) \le \eta.$$

In addition, the target error $\zeta \ge 100\eta$. Then, PASSIVE-MDL(ζ, δ) *satisfies that: (1) with probability $1 - \delta$, it outputs a classifier $\hat{h}$ such that*

$$\max_{i\in[k]} L\left(\hat{h}, \mu_i\right) \le \zeta,$$

(2) the total number of times it samples from any of the D_i's is $\widetilde{O}\left(\frac{d+k}{\zeta}\right)$.

A.4. Proof of Theorem 2

Proof [Proof of Theorem 2] Denote by E_n the success event in Lemma 10 when calling PASSIVE-MDL for iteration n; the lemma implies that $\Pr(E_n) \ge 1 - \delta_n$. Define $E := \cap_{n=1}^{n_0} E_n$. By a union bound, $\Pr(E) \ge 1 - \delta$. We henceforth condition on event E holding.

We will next prove by induction on n that: (1) $h^* \in V_n$; (2) for all $h \in V_n$, $\rho(h, h^*) \le 4\varepsilon_n$.

Base case. For $n = 0$, $V_0 = \mathcal{H}$. $h^* \in V_0$ trivially holds, and for all $h \in \mathcal{H}$, $\rho(h, h^*) \le 1 \le 4\varepsilon_0$.

Inductive case. Suppose the inductive claim holds for iteration $n-1$, specifically $h^* \in V_{n-1}$.

For round n, by the definition of E_n, h_n output by PASSIVE-MDL is such that for every $i \in [k]$, $L(h_n, D_{i,n}) \le \varepsilon_n$. In addition, for h^*, we also have that for every $i \in [k]$,

$$L(h^*, D_{i,n}) = \Pr_{D_n}[h^*(x) \neq y, x \in \mathrm{AGR}(V_{n-1})] \le \nu \le \varepsilon_n.$$

Therefore, by triangle inequality, for every i,

$$\rho_i(h_n, h^*) \le L(h_n, D_{i,n}) + L(h^*, D_{i,n}) \le 2\varepsilon_n.$$

Hence, taking the maximum over all $i \in [k]$, we have $\rho(h_n, h^*) \le 2\varepsilon_n$, implying that $h^* \in V_n$.

Additionally, for all $h \in V_n$, $\rho(h, h_n) \le 2\varepsilon_n$. By triangle inequality, $\rho(h, h^*) \le \rho(h, h_n) + \rho(h^*, h_n) \le 4\varepsilon_n$. Together, these show that the inductive claim holds for iteration n. This completes the induction.

Applying the claim with $n = n_0$, $\hat{h}$ is such that $\rho(\hat{h}, h^*) \le 4\varepsilon_{n_0} \le \varepsilon$. Therefore,

$$L(\hat{h}) \le L(h^*) + \rho(\hat{h}, h^*) \le \nu + \varepsilon,$$

establishing the PAC guarantee.

We now analyze the label complexity of Algorithm 1. For each iteration n, Lemma 10 implies that the number of samples to any of $D_{i,n}$ is at most $m_n = \widetilde{O}(\frac{k+d}{\varepsilon_n})$, which also upper bounds the number of calls to Algorithm 5. By Proposition 9 and Chernoff bound (Lemma 20), with probability $1 - \delta_n$, the number of label queries N_n at iteration n is at most

$$N_n \le O\left(m_n \max_i \Pr_{D_i}[x \in \mathrm{DIS}(V_n)] + \ln\frac{1}{\delta_n}\right) \le O\left(\theta_{\max}(\varepsilon) \cdot (k+d)\right).$$

Summing over all rounds $n \in [n_0]$, the total number of label queries throughout is at most

$$O\left(\theta_{\max}(\varepsilon)(k+d)\ln\frac{1}{\varepsilon}\right).$$

■

Appendix B. Deferred Materials for Section 5

B.1. Auxiliary Lemmas

First we show that D'_i is a valid distribution and that Algorithm 3 correctly samples according to D'_i.

Lemma 11 *For every $i \in [k]$, the surrogate distribution D'_i is a valid distribution, and Algorithm 3 samples according to D'_i.*

Proof To show that D'_i is a valid distribution, we must verify that its total mass is 1, i.e., $\int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}D'_i(x,y) = 1$.

By definition,

$$\int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}D'_i(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} I\left(x \in \mathrm{DIS}(V_0)\right) \mathrm{d}D_i(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} \Pr_{x\sim D_{i,\mathcal{X}}}[x \in \mathrm{AGR}(V_0)]\, \mathrm{d}D_{S_i}(x,y)$$
$$= \mathbb{E}_{(x,y)\sim D_i}\left[I\left(x \in \mathrm{DIS}(V_0)\right)\right] + \Pr_{x\sim D_{i,\mathcal{X}}}[x \in \mathrm{AGR}(V_0)]\, \mathbb{E}_{(x,y)\sim D_{S_i}}[1].$$

Since $\mathbb{E}_{(x,y)\sim D_{S_i}}[1] = 1$ and because $\Pr_{x\sim D_{i,\mathcal{X}}}[x \in \mathrm{DIS}(V_0)] + \Pr_{x\sim D_{i,\mathcal{X}}}[x \in \mathrm{AGR}(V_0)] = 1$, it follows that $\int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}D'_i(x,y) = 1$. Next, observe that in Algorithm 3:

- If $x \in \mathrm{DIS}(V_0)$, the algorithm queries $\mathcal{O}_i(x)$ so that the sampled pair (x,y) is drawn with probability $D_i(x,y)$.
- Otherwise, when $x \in \mathrm{AGR}(V_0)$, the algorithm uniformly samples (x,y) from S_i; hence the probability of obtaining (x,y) is $\Pr_{x\sim D_{i,\mathcal{X}}}[x \in \mathrm{AGR}(V_0)]\, D_{S_i}(x,y)$.

This exactly matches the definition of D'_i, so the algorithm correctly samples from D'_i. ■

Next, we show that the empirical distribution D_{S_i} approximates the conditional distribution $D_{i|\,\mathrm{AGR}(V_0)}$ well.

Lemma 12 *Assume that $h^* \in V_0$. If we set $n_0 = \frac{100(\varepsilon+\nu)}{\varepsilon^2} \ln \frac{k}{\delta'}$, then with probability at least $1-\delta'$, for every $i \in [k]$ and every $h \in V_0$,*

$$\left|L(h, D'_i) - L(h, D_i)\right| \le \frac{\varepsilon}{4}.$$

Proof First notice that by definition, for every h,

$$\left|L(h, D'_i) - L(h, D_i)\right|$$
$$= \left|\Pr_{x\sim D_{i,\mathcal{X}}}[x \in \mathrm{AGR}(V_0)]\, L(h, D_{S_i}) - \Pr_{(x,y)\sim D_i}[x \in \mathrm{AGR}(V_0) \text{ and } h(x) \neq y]\right|.$$

For every $(x_j, y_j) \in S_i$, let $Z_j = \Pr_{x\sim D_{i,\mathcal{X}}} L(h, D_{(x_j,y_j)})$ where $D_{(x_j,y_j)} = I(x = x_j \text{ and } y = y_j)$ is the singleton distribution. Let $Z = \sum_{j=1}^{n_0} Z_j = \Pr_{x\sim D_{i,\mathcal{X}}}[x \in \mathrm{AGR}(V_0)]\, L(h, D_{S_i})$ and

$$\mathbb{E}[Z] = \Pr_{(x,y)\sim D_i}[x \in \mathrm{AGR}(V_0) \text{ and } h(x) \neq y]$$
$$\Pr_{(x,y)\sim D_i}[x \in \mathrm{AGR}(V_0) \text{ and } h^*(x) \neq y]$$
$$\le \nu,$$

where the first equality comes from the definition of Z and the second equality comes from the assumption that $h^* \in V_0$. Then from Bernstein's Inequality (Theorem 21), if we draw $n_0 = \frac{100(\varepsilon+\nu)}{\varepsilon^2} \ln \frac{k}{\delta'}$ samples from $D_{i|\,\mathrm{AGR}(V_0)}$, then for any i, with probability at least $1 - \frac{\delta'}{k}$,

$$\left|L(h, D'_i) - L(h, D_i)\right| \le \frac{\varepsilon}{4}.$$

Applying the union bound over all k distributions and the proof finishes. ■

B.2. Proof of Theorem 3

Correctness: Recall that we have verified the distributions $\{D'_i\}_{i\in[k]}$ are well-defined and that Algorithm 3 correctly samples from D'_i (Lemma 11) , and $h^* = \arg\min_{h\in\mathcal{H}} \max_{i\in[k]} L(h, D_i)$. By Theorem 2, with probability at least $1-\delta/6$, V_0 contains h^* and the diameter of V_0 with respect metric ρ is at most 400ν. Moreover, Theorem 4 guarantees that with probability at least $1-\delta/6$,

$$\max_{i\in[k]} L\left(\hat{h}, D'_i\right) \le \min_{h\in V_0} \max_{i\in[k]} L\left(h, D'_i\right) + \frac{\varepsilon}{2}. \tag{2}$$

Let $i^*_{\max} = \arg\max_{i\in[k]} L(h^*, D'_i)$. Then by Lemma 12, with probability at least $1-\delta/6$,

$$\min_{h\in V_0} \max_{i\in[k]} L\left(h, D'_i\right) \le \max_{i\in[k]} L\left(h^*, D'_i\right) \le L\left(h^*, D_{i^*_{\max}}\right) + \frac{\varepsilon}{4} \le \nu + \frac{\varepsilon}{4}. \tag{3}$$

Similarly, let $\hat{i}_{\max} = \arg\max_{i\in[k]} L\left(\hat{h}, D_i\right)$ and observe that

$$\max_{i\in[k]} L\left(\hat{h}, D_i\right) \le L\left(\hat{h}, D'_{\hat{i}_{\max}}\right) + \frac{\varepsilon}{4} \le \max_{i\in[k]} L\left(\hat{h}, D'_i\right) + \frac{\varepsilon}{4}. \tag{4}$$

Combining inequalities (4), (2), and (3) with the union bound yields

$$L\left(\hat{h}\right) \le \nu + \varepsilon,$$

with probability at least $1-\delta/2$.

Sample Complexity: We abbreviate $\theta_{\max}(100\nu)$ as $\theta_{\max}$. Theorem 2 implies that the stage one of Algorithm 3 (step 1) requires $\widetilde{O}\left(\theta_{\max}(d+k)\right)$ samples with probability at least $1-\frac{\delta}{6}$. In stage two of Algorithm 3 (steps 2 to 7), we query labels only in the disagreement region of V_0 during the execution of PASSIVE-MDL. Since $V_0 \subseteq \mathbf{B}_D(h^*, 400\nu)$ and by the definition of $\theta_{\max}$, the probability that a single sample lands in the disagreement region is at most $400\theta_{\max}\nu \le 400\theta_{\max}\cdot(\nu+\varepsilon)$. By Theorem 4, PASSIVE-MDL samples

$$O\Big(\frac{(d+k)(\nu+\varepsilon)}{\varepsilon^2}\cdot \operatorname{polylog}\Big(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta}\Big)\Big)$$

examples from any of the D'_i's; applying the Chernoff bound (Lemma 20) gives that, with probability at least $1-\frac{\delta}{3}$, we query

$$O\Big(\theta_{\max}(\nu+\varepsilon)\cdot\frac{(d+k)(\nu+\varepsilon)}{\varepsilon^2}\cdot \operatorname{polylog}\Big(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta}\Big)\Big)$$

fresh labels. Additionally, step 4 in Algorithm 2 samples $k\cdot n_0$ points to construct the surrogate distributions, which requires

$$O\Big(\frac{k(\nu+\varepsilon)}{\varepsilon^2}\cdot \operatorname{polylog}\Big(k, \frac{1}{\varepsilon}\Big)\Big)$$

labels. Taking the sum and applying the union bound, we conclude that, with probability at least $1-\delta/3$, the overall label complexity is

$$O\Big(\Big(\frac{k(\nu+\varepsilon)}{\varepsilon^2} + \frac{\theta_{\max}(d+k)(\nu+\varepsilon)^2}{\varepsilon^2}\Big)\cdot \operatorname{polylog}\Big(k, d, \frac{1}{\varepsilon}, \frac{1}{\delta}\Big)\Big).$$

Conclusion: Taking a union bound over the above events, with probability at least $1-\delta$ both the correctness and the label complexity guarantees hold. This completes the proof.

Appendix C. Deferrd Materials for Section 5.1

C.1. Overview

We follow Zhang et al. (2024)'s analysis of their `MDL-Hedge-VC` algorithm and Theorem 1, as well as their notations closely. As mentioned in Section 5.1, we will choose the hyperparameters in line 2 differently as in Theorem 4, which is different from Zhang et al. (2024)'s original setting.

`MDL-Hedge-VC` solves passive multi-distribution learning by simulating a two-player zero-sum game similar to Haghtalab et al. (2022), where the row player chooses classifier $h^t \in \mathcal{H}$ and the column player chooses weight $w^t \in \Delta^{k-1}$ over the k distributions. Specifically, at each round t, the column player chooses w^t in a no-regret manner, and the row player chooses h^t as an approximate best response. For completeness, we replicate it in Algorithm 6.

We prove enhanced versions of their three main lemmas, after which the proof of Theorem 4 follows identically as in Section 5 of Zhang et al. (2024). The first lemma states that at each iteration t, h^t is roughly the best hypothesis under the distribution D_{w^t}.

Lemma 13 (Enhanced Lemma 1 from Zhang et al. (2024)) *With probability at least $1-\frac{\delta}{4}$,*

$$L(h^t, w^t) \le \min_{h\in\mathcal{H}} L(h, w^t) + \varepsilon_1$$

holds for all $1 \le t \le T$, where h^t (resp. w^t) is the hypothesis (resp. weight vector) computed in round t of Algorithm 6, with the modifications stated in Theorem 4.

The second key lemma states that, under Lemma 13, the hypothesis returned by Algorithm 1 in Zhang et al. (2024) has good performance, thereby establishing the correctness of the algorithm.

Lemma 14 (Enhanced Lemma 2 from Zhang et al. (2024)) *Suppose that lines 6–11 in Algorithm 6 give a hypothesis h^t satisfying $L\big(h^t, w^t\big) \le \min_{h\in\mathcal{H}} L\big(h, w^t\big) + \varepsilon_1$ in the t-th round for each $1 \le t \le T$ and that we choose the hyperparameters as stated in Theorem 4. With probability exceeding $1-\frac{\delta}{4}$, the hypothesis h^{final} output by Algorithm 1 is ε-optimal in the sense that*

$$L\big(h^{\text{final}}\big) \le \nu + \varepsilon.$$

The last key lemma helps bound the ℓ_1-norm of the weight vector $\overline{w}^t$ updated by `MDL-Hedge-VC` (step 12). It is crucial in controlling the algorithm's sample complexity (see step 14).

Lemma 15 (Enhanced Lemma 3 from Zhang et al. (2024)) *Assume that lines 6–11 in Algorithm 6 returns a hypothesis h^t satisfying $L(h^t, w^t) \le \min_{h\in\mathcal{H}} L(h, w^t) + \varepsilon_1$ in the t-th round for each $1 \le t \le T$. If we choose the hyperparameters as stated in Theorem 4, then with probability at least $1-\frac{\delta}{4}$, for all t,*

$$\|\overline{w}^t\|_1 \le O\left(\ln^8\left(\frac{k}{\delta\varepsilon}\right)\right).$$

Algorithm 6 `MDL-Hedge-VC` (Zhang et al., 2024)

Input: labeled data distributions $(D_1, \ldots, D_k)$, hypothesis class $\mathcal{H}$, target excess error ε, failure probability δ
Hyperparameters: step size η, number of rounds T, auxiliary excess error level ε_1, auxiliary sample size T_1
Initialization: weight $W_i^1 = 1$ for all $i \in [k]$, $\hat{w}_i^0 = 0$ and $n_i^0 = 0$, $\mathcal{S} = \emptyset$
for $t = 1, \ldots, T$ **do**
  Set $w^t = (w_1^t, \ldots, w_k^t)$, where $w_i^t = \frac{W_i^t}{\sum_j W_j^t}$ and $\hat{w}^t = (\hat{w}_1^t, \ldots, \hat{w}_k^t)$, where $\hat{w}_i^t = \hat{w}_i^{t-1}, \forall i$
  if there exists $j \in [k]$ such that $w_j^t \geq 2\hat{w}_j^{t-1}$ **then**
    $\hat{w}_i^t \leftarrow \max(w_i^t, \hat{w}_i^{t-1})$, for all $i \in [k]$
    for $i = 1, \ldots, k$ **do**
      $n_i \leftarrow \lceil T_1 \hat{w}_i^t \rceil$
      draw $n_i^t - n_i^{t-1}$ independent samples from D_i, and add these samples to $\mathcal{S}$
  Compute $h^t \leftarrow \arg\min_{h \in \mathcal{H}} \hat{L}^t(h, w^t)$, where

$$\hat{L}^t(h, w^t) = \sum_{i=1}^{k} \frac{w_i^t}{n_i^t} \sum_{j=1}^{n_i^t} \ell(h, (x_{i,j}, y_{i,j})),$$

  here, $(x_{i,j}, y_{i,j})$ denotes the j-th example from D_i's samples in $\mathcal{S}$.
  $\bar{w}_i^t \leftarrow \max_{\tau \in \{1,\ldots,t\}} w_i^\tau$, for all $i \in [k]$
  for $i = 1, \ldots, k$ **do**
    Draw from D_i $\lceil k\bar{w}_i^t \rceil$ independent samples $\left\{(x_{i,j}^t, y_{i,j}^t)\right\}_{j=1}^{\lceil k\bar{w}_i^t \rceil}$, and set

$$\hat{r}_i^t = \frac{1}{\lceil k\bar{w}_i^t \rceil} \sum_{j=1}^{\lceil k\bar{w}_i^t \rceil} \ell(h^t, (x_{i,j}^t, y_{i,j}^t))$$

    Update weight $W_i^{t+1} \leftarrow W_i^t e^{\eta \hat{r}_i^t}$.
return a hypothesis h^{final} that predicts by following a hypothesis uniformly at random from $\left\{h^t\right\}_{t=1}^{T}$

C.2. Proof of Lemma 13

We closely follow Zhang et al. (2024)'s proof for their Lemma 14. The major difference is that we enhance their step 1 to Bernstein-style bound. Let's recall that $\ell(h,(x,y))$ is the 0-1 loss of hypothesis h on the feature-label pair (x,y).

Step 1: concentration bounds for any fixed $n=\{n_i\}_{i=1}^k$ and $w\in\Delta(k)$. We first prove the following claim that establishes a fine-grained uniform concentration of weighted empirical losses to their expectations. Below, we use $\hat{L}(h,w):=\sum_{i=1}^k w_i\frac{1}{n_i}\sum_{j=1}^{n_i}\ell\left(h,(x_{i,j},y_{i,j})\right)$ to denote the empirically importance-weighted 0-1 loss of h.

Claim 1 *For fixed $n=\{n_i\}_{i=1}^k$ such that $\max_i n_i\le T_1$ and $w\in\Delta(k)$ (where $\Delta(k)$ denotes the k-dimensional probability simplex),*

$$\Pr\left[\max_{h\in\mathcal{H}}\frac{\left(\hat{L}(h,w)-L(h,w)\right)^2}{4\left(\hat{L}(h,w)+L(h,w)\right)}\ge\varepsilon'\right]\le(2kT_1+1)^d\exp\left(-\frac{\epsilon'}{\max\limits_{i\in[k]}\frac{w_i}{n_i}}\right).$$

Note that the denominator term $\hat{L}(h,w)+L(h,w)$ depends on h, which allows loss concentration to be tighter for h's that have smaller losses. This generalizes classical relative VC inequalities (Vapnik, 1982) to losses that are weighted averages over samples from multiple distributions.

Proof We denote samples $S:=\{(x_{i,j},y_{i,j})\}_{i\in[k],j\in[n_i]}$ and denote a set of "ghost" samples $S^+:=\left\{\left(x_{i,j}^+,y_{i,j}^+\right)\right\}_{i\in[k],j\in[n_i]}$ independently drawn from the same distribution as S. For any $h\in\mathcal{H}$, conditioned on samples $(x_{i,j},y_{i,j})$ and $\left(x_{i,j}^+,y_{i,j}^+\right)$, the random variable

$$\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right),$$

(where $\varepsilon_{i,j}\sim\mathrm{Uniform}(\{-1,+1\})$'s are independent Rademacher random variables), is sub-Gaussian with parameter $\left(\ell\left(h,(x_{i,j},y_{i,j})\right)+\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)$ (Vershynin, 2018, Example 2.5.8). Then

$$\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\tag{5}$$

is a sub-Gaussian random variable with parameter

$$\sigma^2(h,S,S^+)=\sum_{i=1}^k\frac{w_i^2}{n_i}\left(\hat{\mathbb{E}}_i[\ell]+\hat{\mathbb{E}}_i^+[\ell]\right),$$

where $\hat{\mathbb{E}}_i[\ell]=\frac{1}{n_i}\sum_{j=1}^{n_i}\ell\left(h,(x_{i,j},y_{i,j})\right)$ and $\hat{\mathbb{E}}_i^+[\ell]=\frac{1}{n_i}\sum_{j=1}^{n_i}\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)$ (Vershynin, 2018, Proposition 2.6.1).

Therefore, (5) also has ψ_2-Orlicz norm at most $4\sigma^2(h,S,S^+)$ for some constant $c>0$. Applying (Vershynin, 2018, Lemma 2.7.6), the random variable

$$\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2$$

has ψ_1-Orlicz norm at most $4\sigma^2(h, S, S^+)$. This implies that

$$\mathop{\mathbb{E}}_{\varepsilon} \exp\left(\frac{1}{4\sigma^2(h,S,S^+)}\left(\sum_{i=1}^{k}\frac{w_i}{n_i}\sum_{j=1}^{n_i}\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2\right)\leq 2$$

Let $\mathcal{C} \coloneqq (S, S^+)$. Further, let's define $H_{\min,\mathcal{C}} \subseteq \mathcal{H}$ to be the minimum-cardinality subset of $\mathcal{H}$ that results in the same labeling outcome as $\mathcal{H}$ when applied to the unlabeled examples of $\mathcal{C}$, namely, $H_{\min,\mathcal{C}}(\mathcal{C}) = \mathcal{H}(\mathcal{C})$ and $|H_{\min,\mathcal{C}}| = |\mathcal{H}(\mathcal{C})|$. Then if we take the maximum over all hypothesis and take expectations over S, S^+, we have

$$\mathbb{E}\max_{h\in\mathcal{H}}\exp\left(\frac{1}{4\sigma^2(h,S,S^+)}\left(\sum_{i=1}^{k}\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2\right) \tag{6}$$

$$=\mathop{\mathbb{E}}_{S,S^+}\mathop{\mathbb{E}}_{\varepsilon}\left[\max_{h\in\mathcal{H}}\exp\left(\frac{1}{4\sigma^2(h,S,S^+)}\left(\sum_{i=1}^{k}\frac{w_i}{n_i}\sum_{j=1}^{n_i}\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2\right)\middle|\mathcal{C}\right] \tag{7}$$

$$\leq\mathop{\mathbb{E}}_{S,S^+}\mathop{\mathbb{E}}_{\varepsilon}\left[\sum_{h\in\mathcal{H}_{\min,\mathcal{C}}}\exp\left(\frac{1}{4\sigma^2(h,S,S^+)}\left(\sum_{i=1}^{k}\frac{w_i}{n_i}\sum_{j=1}^{n_i}\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2\right)\middle|\mathcal{C}\right] \tag{8}$$

$$=\mathop{\mathbb{E}}_{S,S^+}\left[\sum_{h\in\mathcal{H}_{\min,\mathcal{C}}}\mathop{\mathbb{E}}_{\varepsilon}\left[\exp\left(\frac{1}{4\sigma^2(h,S,S^+)}\left(\sum_{i=1}^{k}\frac{w_i}{n_i}\sum_{j=1}^{n_i}\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2\right)\middle|\mathcal{C}\right]\right] \tag{9}$$

$$\leq 2\cdot(2kT_1+1)^d, \tag{10}$$

where the last step is by Sauer's Lemma, $|\mathcal{H}_{\min,\mathcal{C}}| \leq (2kT_1+1)^d$, where we use that $\mathcal{H}$ has VC dimensional at most d, and $\mathcal{C}$ has at most $2kT_1$ examples.

As a result,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S}\max_{h\in\mathcal{H}}\exp\left(\frac{\left(\hat{L}(h,w)-L(h,w)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\hat{L}(h,w)+L(h,w)\right)}\right)\\
&\mathop{\mathbb{E}}_{S}\max_{h\in\mathcal{H}}\exp\left(\frac{\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\ell\left(h,(x_{i,j},y_{i,j})\right)-L(h,w)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\ell\left(h,(x_{i,j},y_{i,j})\right)+L(h,w)\right)}\right)\\
=&\mathop{\mathbb{E}}_{S}\max_{h\in\mathcal{H}}\exp\left(\frac{\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\mathop{\mathbb{E}}_{S^+}\left[\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right]\right)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)+\mathop{\mathbb{E}}_{S^+}\left[\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right]\right)\right)}\right)\\
\le&\mathop{\mathbb{E}}_{S}\max_{h\in\mathcal{H}}\exp\mathop{\mathbb{E}}_{S^+}\left(\frac{\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)+\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)}\right)\\
\le&\mathop{\mathbb{E}}_{S}\max_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{S^+}\exp\left(\frac{\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)+\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)}\right)\\
\le&\mathop{\mathbb{E}}_{S}\max_{h\in\mathcal{H}}\exp\left(\frac{\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)+\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)}\right)\\
\le&\mathop{\mathbb{E}}_{S,S^+}\left[\max_{h\in\mathcal{H}}\exp\left(\frac{\left(\sum_{i=1}^k\frac{w_i}{n_i}\sum_{j=1}^{n_i}\varepsilon_{i,j}\left(\ell\left(h,(x_{i,j},y_{i,j})\right)-\ell\left(h,\left(x_{i,j}^+,y_{i,j}^+\right)\right)\right)\right)^2}{4\sigma^2(h,S,S^+)}\right)\right]\\
\le&2\cdot(2kT_1+1)^d\,.
\end{aligned}
$$

In the above, for the first and second inequality, we used Jensen's inequality twice (e^x and $\frac{(a-x)^2}{c(b+x)}$ are both convex). Then in the third inequality, we moved the expectation outside of the max function using Jensen's inequality. In the fourth inequality, we used the fact that

$$
\sigma^2(h,S,S^+)=\sum_{i=1}^k\frac{w_i^2}{n_i}\left(\hat{\mathbb{E}}_i[\ell]+\hat{\mathbb{E}}_i^+[\ell]\right)\le\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\sum_{i=1}^k w_i\left(\hat{\mathbb{E}}_i[\ell]+\hat{\mathbb{E}}_i^+[\ell]\right),
$$

and the last inequality is from Eq. (10).

As a result, by applying Markov's inequality, we have

$$\begin{aligned}
&\Pr\left[\max_{h\in\mathcal{H}}\frac{\left(\hat{L}(h,w)-L(h,w)\right)^2}{4\left(\hat{L}(h,w)+L(h,w)\right)}\geq\varepsilon'\right]\\
=&\Pr\left[\exp\max_{h\in\mathcal{H}}\frac{\left(\hat{L}(h,w)-L(h,w)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\hat{L}(h,w)+L(h,w)\right)}\geq\exp\left(\frac{\varepsilon'}{\max_{i\in[k]}\frac{w_i}{n_i}}\right)\right]\\
\leq&\Pr\left[\max_{h\in\mathcal{H}}\exp\left(\frac{\left(\hat{L}(h,w)-L(h,w)\right)^2}{4\left(\max_{i\in[k]}\frac{w_i}{n_i}\right)\left(\hat{L}(h,w)+L(h,w)\right)}\right)\geq\exp\left(\frac{\varepsilon'}{\max_{i\in[k]}\frac{w_i}{n_i}}\right)\right]\\
\leq&2\left(2kT_1+1\right)^d\cdot\exp\left(-\frac{\varepsilon'}{\max_{i\in[k]}\frac{w_i}{n_i}}\right).
\end{aligned}$$

■

Step 2: uniform concentration bounds over epsilon-nets w.r.t. n and w. First we recall some notations in Zhang et al. (2024) as below.

- We use $\Delta_{\varepsilon_2}(k)\subseteq\Delta(k)$ to denote an ε_2-net of the probability simplex $\Delta(k)$ — namely, for any $x\in\Delta(k)$, there exists a vector $x_0\in\Delta_{\varepsilon_2}(k)$ obeying $\|x-x_0\|_\infty\leq\varepsilon_2$. We shall choose $\Delta_{\varepsilon_2}(k)$ properly so that
$$|\Delta_{\varepsilon_2}(k)|\leq(1/\varepsilon_2)^k.$$
- Define the following set
$$\mathcal{B}=\left\{n=\{n_i\}_{i=1}^k,w=\{w_i\}_{i=1}^k\;\middle|\;\frac{n_i}{w_i}\geq\frac{T_1}{2},\;n_i\in[0,T_1]\cap\mathbb{N},\forall i\in[k],w\in\Delta_{\varepsilon_1/(16k)}(k)\right\},$$
by the constraints on n_i's and the definition of $\Delta_{\varepsilon_1/(16k)}(k)$,
$$|\mathcal{B}|\leq T_1^k\cdot\left(\frac{16k}{\varepsilon_1}\right)^k.$$

It is clear from the definition of $\mathcal{B}$ that any (n,w) in $\mathcal{B}$ satisfies that

$$\max_{i\in[k]}\frac{w_i}{n_i}\leq\frac{2}{T_1}.$$

Then by taking a union bound over all (n,w)'s in $\mathcal{B}$, we get

$$\begin{aligned}
&\Pr\left[\exists(n,w)\in\mathcal{B},\max_{h\in\mathcal{H}}\frac{\left(\hat{L}(h,w)-L(h,w)\right)^2}{4\left(\hat{L}(h,w)+L(h,w)\right)}\geq\varepsilon'\right]\\
\leq&4\left(16kT_1/\varepsilon_1\right)^k\left(2kT_1+1\right)^d\exp\left(-\frac{T_1\varepsilon'}{4}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\Pr\left[\exists(n,w)\in\mathcal{B}, \max_{h\in\mathcal{H}} \frac{\left(\hat{L}(h,w)-L(h,w)\right)^2}{4(\hat{L}(h,w)+L(h,w))} \geq \varepsilon'\right] \\
=&\Pr\left[\exists(n,w)\in\mathcal{B}, h\in\mathcal{H} \text{ such that } \left|\hat{L}(h,w)-L(h,w)\right| \geq \sqrt{4\varepsilon'\left(\hat{L}(h,w)+L(h,w)\right)}\right] \\
\geq&\Pr\left[\exists(n,w)\in\mathcal{B}, h\in\mathcal{H} \text{ such that } \left|\hat{L}(h,w)-L(h,w)\right| \geq 2\left(4\varepsilon'+\sqrt{4\varepsilon'\min\left\{\hat{L}(h,w),L(h,w)\right\}}\right)\right],
\end{aligned}$$

where the last inequality is from Lemma 23.

Putting it together, we have

$$\begin{aligned}
&\Pr\left[\exists(n,w)\in\mathcal{B}, h\in\mathcal{H} \text{ such that } \left|\hat{L}(h,w)-L(h,w)\right| \geq 2\left(4\varepsilon'+\sqrt{4\varepsilon'\min\left\{\hat{L}(h,w),L(h,w)\right\}}\right)\right] \\
\leq& 4\left(16kT_1/\varepsilon_1\right)^k\left(2kT_1+1\right)^d\exp\left(-\frac{T_1\varepsilon'}{4}\right).
\end{aligned}$$

Step 3: concentration bounds w.r.t. n^t and w^t. By the definition of $\mathcal{B}$, one can always find $\left(n^t,\tilde{w}^t\right)\in\mathcal{B}$ satisfying

$$\|w^t-\tilde{w}^t\|_1 \leq k\|w^t-\tilde{w}^t\|_\infty \leq \frac{\varepsilon_1}{16}.$$

As a result, $\left|L(h,w^t)-L(h,\tilde{w}^t)\right| \leq \frac{\varepsilon_1}{16}$ and $\left|\hat{L}(h,w^t)-\hat{L}(h,\tilde{w}^t)\right| \leq \frac{\varepsilon_1}{16}$, where $\hat{L}(h,w^t) := \sum_{i=1}^k w_i^t \frac{1}{n_i^t}\sum_{j=1}^{n_i^t}\ell\left(h,(x_{i,j},y_{i,j})\right)$, and $\hat{L}(h,\tilde{w}^t) := \sum_{i=1}^k \tilde{w}_i^t \frac{1}{n_i^t}\sum_{j=1}^{n_i^t}\ell\left(h,(x_{i,j},y_{i,j})\right)$.

Therefore, for any $\varepsilon' \leq \frac{\varepsilon_1}{64}$,

$$\begin{aligned}
&\Pr\left[\exists t\in[T], h\in\mathcal{H}, \left|\hat{L}(h,w^t)-L(h,w^t)\right| \geq 2\left(\varepsilon'+\sqrt{\varepsilon'\min\left\{\hat{L}(h,w^t),L(h,w^t)\right\}}\right)+\frac{\varepsilon_1}{4}\right] \\
\leq&\Pr\left[\exists t\in[T], h\in\mathcal{H}, \left|\hat{L}(h,\tilde{w}^t)-L(h,\tilde{w}^t)\right| \geq 2\left(\varepsilon'+\sqrt{\varepsilon'\min\left\{\hat{L}(h,\tilde{w}^t),L(h,\tilde{w}^t)\right\}}\right)\right] \\
\leq&\Pr\left[\exists t\in[T], (n,w)\in\mathcal{B}, h\in\mathcal{H}, \left|\hat{L}(h,w)-L(h,w)\right| \geq 2\left(\varepsilon'+\sqrt{\varepsilon'\min\left\{\hat{L}(h,w),L(h,w)\right\}}\right)\right] \\
\leq& 4T\left(16kT_1/\varepsilon_1\right)^k\left(2kT_1+1\right)^d\exp\left(-\frac{T_1\varepsilon'}{4}\right),
\end{aligned}$$

where the first inequality also uses that $\sqrt{\varepsilon'\min\left\{\hat{L}(h,\tilde{w}^t),L(h,\tilde{w}^t)\right\}}-\sqrt{\varepsilon'\min\left\{\hat{L}(h,w^t),L(h,w^t)\right\}} \leq \sqrt{\varepsilon'\cdot\frac{\varepsilon_1}{8}} \leq \frac{\varepsilon_1}{8}$, and the second inequality is because $\left(n^t,\tilde{w}^t\right)\in\mathcal{B}$, and last inequality is from Step 2.

Step 4: putting things together. We take $\varepsilon' = \frac{\varepsilon_1^2}{64(\varepsilon_1+\nu)} \le \frac{\varepsilon_1}{64}$, and apply Step 3 with $T_1 = 4000\left(\frac{1}{\varepsilon_1}+\frac{\nu}{\varepsilon_1^2}\right)\left(k\ln\left(\frac{k}{\varepsilon}\right)+d\ln\left(\frac{kd}{\varepsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right)$. This implies that that failure probability in Step 3 is at most

$$4T\left(8kT_1/\varepsilon_1\right)^k\left(2kT_1+1\right)^d\exp\left(-10\left(k\ln\left(k/\varepsilon_1\right)+d\ln\left(kd/\varepsilon_1\right)+\ln\left(1/\delta\right)\right)\right) \le \frac{\delta}{2}.$$

To summarize, we showed that, with probability $1-\delta/2$,

$$\left|\hat{L}(h,w^t)-L(h,w^t)\right| \le 2\left(\varepsilon'+\sqrt{\varepsilon'\min\left\{\hat{L}(h,w),L(h,w)\right\}}\right)+\frac{\varepsilon_1}{4} \tag{11}$$

holds simultaneously for all $t\in[T]$ and all $h\in\mathcal{H}$. Let $h'=\arg\min_{h\in\mathcal{H}}L(h,w^t)$; by its definition, $L(h',w^t)\le\min_{h\in\mathcal{H}}\max_{i\in[k]}L(h,D_i)=\nu$. Therefore, repeatedly applying Eq. (11), we get:

$$\begin{aligned}
L(h^t,w^t) &\le \hat{L}(h^t,w^t)+2\varepsilon'+2\sqrt{\varepsilon'\hat{L}(h^t,w^t)}+\frac{\varepsilon_1}{4}\\
&\le \hat{L}(h',w^t)+2\varepsilon'+2\sqrt{\varepsilon'\hat{L}(h',w^t)}\\
&\le \hat{L}(h',w^t)+2\varepsilon'+2\sqrt{\varepsilon'\left(L(h',w^t)+2\left(\varepsilon'+\sqrt{\varepsilon' L(h',w^t)}\right)+\frac{\varepsilon_1}{4}\right)}\\
&\le L(h',w^t)+4\varepsilon'+2\sqrt{\varepsilon'\left(L(h',w^t)+2\left(\varepsilon'+\sqrt{\varepsilon' L(h',w^t)}\right)+\frac{\varepsilon_1}{4}\right)}+2\sqrt{\varepsilon' L(h',w^t)}+\frac{\varepsilon_1}{4}\\
&= \min_{h\in\mathcal{H}}L(h,w^t)+4\varepsilon'+2\sqrt{\varepsilon'\left(L(h',w^t)+2\left(\varepsilon'+\sqrt{\varepsilon' L(h',w^t)}+\frac{\varepsilon_1}{4}\right)\right)}+2\sqrt{\varepsilon' L(h',w^t)}+\frac{\varepsilon_1}{4}.
\end{aligned}$$

Recall that $L(h',w^t)\le\nu$, we have

$$\sqrt{\varepsilon' L(h',w^t)} \le \frac{\varepsilon_1}{8}.$$

Combining this with the previous bounds yields

$$L(h^t,w^t) \le \min_{h\in\mathcal{H}}L(h,w^t)+\varepsilon_1.$$

C.3. Proof of Lemma 14

We replace the usage of Azuma-Hoeffding in the proof of Lemma 2 in Zhang et al. (2024) with Freedman's inequality, resulting in a stronger bound. Let $\mathcal{F}_i$ denote all of the information up to iteration i. Then for any $i\in[k]$, let

$$X_i^j = \sum_{t=1}^{j}\left(\hat{r}_i^t - L_i\left(h^t\right)\right),$$

where $\hat{r}_i^t$ is defined in step 14 in Algorithm 6. Then the sequence $\left\{X_i^t\right\}_{t=1}^T$ is a martingale because

$$\mathbb{E}\left[X_i^{j+1}-X_i^j\middle|\mathcal{F}_j\right] = \mathbb{E}_j\left[\hat{r}_i^j - L_i\left(h^j\right)\right] = 0,$$

where $\mathbb{E}_j$ denote the conditional expectation on $\mathcal{F}_j$. The last equality is because $\hat{r}_i^j$ is an unbiased estimator of $L_i\left(h^j\right)$ conditioned on $\mathcal{F}_j$. Let $\left(\sigma_i^j\right)^2$ denote the conditional variance of the martingale difference $X_i^j - X_i^{j-1}$. Then

$$\left(\sigma_i^j\right)^2 \leq \mathbb{E}_j\left[\left(\hat{r}_i^j - L_i\left(h^j\right)\right)^2\right] \leq \mathbb{E}_j\left[(\hat{r}_i^j)^2\right] \leq \mathbb{E}_j\left[\hat{r}_i^j\right] = L_i\left(h^j\right),$$

so

$$\sum_{t=1}^{T}\left(\sigma_i^t\right)^2 \leq \sum_{t=1}^{T} L_i\left(h^t\right).$$

Let $\delta' \coloneqq \frac{\delta}{4(T+k+1)\log T}$, then by Freedman's inequality (Theorem 22) and taking the union bound over all i, we see that with probability at least $1-(2k\log T)\delta'$, for all i,

$$\left|\sum_{t=1}^{T}\hat{r}_i^t - \sum_{t=1}^{T} L_i\left(h^t\right)\right| \leq 4\sqrt{\ln(1/\delta')\sum_{t=1}^{T} L_i\left(h^t\right)} + 2\ln\left(1/\delta'\right). \tag{12}$$

Similarly, $Y^j = \sum_{t=1}^{j}\left(\langle w^t, \hat{r}^t\rangle - L\left(h^t, w^t\right)\right)$ is also a martingale. Moreover, the sum of its conditional variance is also upper bounded by $\sum_{t=1}^{T} L\left(h^t, w^t\right)$ by the same argument. Then again by applying Freedman's inequality, we get with probability at least $1-2\log T\cdot\delta'$,

$$\left|\sum_{t=1}^{T}\langle\hat{r}_i^t, w^t\rangle - \sum_{t=1}^{T} L\left(h^t, w^t\right)\right| \leq 4\sqrt{\ln(1/\delta')\sum_{t=1}^{T} L\left(h^t, w^t\right)} + 2\ln\left(1/\delta'\right). \tag{13}$$

Taking the union bound, with probability at least $1-(2k+2)\log T\cdot\delta'$, equations (12) and (13) both hold. We then resort to standard analysis for the Hedge algorithm (Freund and Schapire, 1997). We recall that W^t is the unnormalized weight vector at iteration t. Direct calculation gives

$$\begin{aligned}
\ln\left(\frac{\sum_{i=1}^{k} W_i^{t+1}}{\sum_{i=1}^{k} W_i^t}\right) &\overset{\text{(i)}}{=} \ln\left(\sum_{i=1}^{k} w_i^t \exp(\eta\hat{r}_i^t)\right) \\
&\overset{\text{(ii)}}{\leq} \ln\left(\sum_{i=1}^{k} w_i^t\left(1+\eta\hat{r}_i^t + \eta^2(\hat{r}_i^t)^2\right)\right) \\
&\leq \ln\left(1+\eta\sum_{i=1}^{k} w_i^t\hat{r}_i^t + \eta^2\sum_{i=1}^{k} w_i^t(\hat{r}_i^t)^2\right) \\
&\leq \eta\sum_{i=1}^{k} w_i^t\hat{r}_i^t + \eta^2\langle w^t, \hat{r}^t\rangle.
\end{aligned} \tag{14}$$

Here, (i) is valid since $w_i^t = \frac{W_i^t}{\sum_j W_j^t}$ and $W_i^{t+1} = W_i^t\exp(\eta\hat{r}_i^t)$ (cf. lines 5 and 15 of Algorithm 1); (ii) arises from the elementary inequality $e^x \leq 1+x+x^2$ for $x\in[0,1]$ as well as the facts that

$\eta \leq 1$ and $|\hat{r}_i^t| \leq 1$. Summing the inequality (14) over all t and rearranging terms, we are left with

$$\begin{aligned}
\eta \sum_{t=1}^{T} \langle w^t, \hat{r}^t \rangle &\geq \sum_{t=1}^{T} \left\{ \ln \left(\frac{\sum_{i=1}^{k} W_i^{t+1}}{\sum_{i=1}^{k} W_i^t} \right) - \eta^2 \langle w^t, \hat{r}^t \rangle \right\} \\
&= \ln \left(\frac{\sum_{i=1}^{k} W_i^{T+1}}{\sum_{i=1}^{k} W_i^1} \right) - \eta^2 \sum_{t=1}^{T} \langle w^t, \hat{r}^t \rangle \\
&= \ln \left(\sum_{i=1}^{k} W_i^{T+1} \right) - \ln \left(\sum_{i=1}^{k} W_i^1 \right) - \eta^2 \sum_{t=1}^{T} \langle w^t, \hat{r}^t \rangle \\
&\geq \max_{1 \leq i \leq k} \ln(W_i^{T+1}) - \ln(k) - \eta^2 \sum_{t=1}^{T} \langle w^t, \hat{r}^t \rangle \\
&= \eta \max_{1 \leq i \leq k} \sum_{t=1}^{T} \hat{r}_i^t - \ln(k) - \eta^2 \sum_{t=1}^{T} \langle w^t, \hat{r}^t \rangle,
\end{aligned}$$

where the penultimate line makes use of $W_i^1 = 1$ for all $i \in [k]$, and the last line holds since $\ln(W_i^{T+1} \exp(\eta \hat{r}_i^t)) \geq \eta \hat{r}_i^t$. Dividing both sides by η yields

$$\sum_{t=1}^{T} \langle w^t, \hat{r}^t \rangle \geq \max_{i \in [k]} \sum_{t=1}^{T} \hat{r}_i^t - \frac{\ln(k)}{\eta} - \eta \sum_{t=1}^{T} \langle w^t, \hat{r}^t \rangle.$$

Combining the above inequality with (12) and (13), we have shown that with probability at least $1 - (2k+2) \log T \cdot \delta'$,

$$\begin{aligned}
&\sum_{t=1}^{T} L(h^t, w^t) \\
&\geq \max_{i \in [k]} \sum_{t=1}^{T} L_i(h^t) \\
&- \left(\frac{\ln(k)}{\eta} + \eta \sum_{t=1}^{T} L(h^t, w^t) + 4 \sqrt{\ln(1/\delta') \max_{i \in [k]} \sum_{t=1}^{T} L_i(h^t)} + 5 \sqrt{\ln(1/\delta') \sum_{t=1}^{T} L(h^t, w^t)} + 5 \ln(1/\delta') \right).
\end{aligned}$$

Applying Lemma 24 with $A = \max_{i \in [k]} \sum_{t=1}^{T} L_i(h^t)$, $B = 5\sqrt{2 \ln(1/\delta')}$, $C = \sum_{t=1}^{T} L(h^t, w^t)$ and $D = \frac{\ln(k)}{\eta} + \eta \sum_{t=1}^{T} L(h^t, w^t) + 5 \ln(1/\delta')$ so we get

$$\begin{aligned}
\max_{i \in [k]} \sum_{t=1}^{T} L_i(h^t) \leq \sum_{t=1}^{T} L(h^t, w^t) + 50 \ln(1/\delta') + \frac{3}{2} \left(\frac{\ln(k)}{\eta} + \eta \sum_{t=1}^{T} L(h^t, w^t) + 5 \ln(1/\delta') \right) \\
+ 5 \sqrt{2 \ln(1/\delta') \sum_{t=1}^{T} L(h^t, w^t)}.
\end{aligned}$$

From the assumption of Lemma 14, $\sum_{t=1}^{T} L(h^t, w^t) \leq T(\nu + \varepsilon_1)$. As a result, we have

$$\max_{i\in[k]} \sum_{t=1}^{T} L_i(h^t) \leq \sum_{t=1}^{T} L(h^t, w^t) + \left(\frac{3\ln(k)}{2\eta} + \frac{3}{2}\eta T(\nu+\varepsilon_1) + 58\ln(1/\delta') + 5\sqrt{2\ln(1/\delta')T(\nu+\varepsilon_1)}\right) \tag{15}$$

$$\leq T(\nu+\varepsilon_1) + \left(\frac{3\ln(k)}{2\eta} + \frac{3}{2}\eta T(\nu+\varepsilon_1) + 58\ln(1/\delta') + 5\sqrt{2\ln(1/\delta')T(\nu+\varepsilon_1)}\right). \tag{16}$$

By the definition of h^{final} (average over all T rounds) and substitute in $\varepsilon_1 = \frac{1}{100}\varepsilon$, $T = \frac{2000(\varepsilon_1+\nu)\ln\left(\frac{k}{\delta\varepsilon}\right)}{\varepsilon_1^2}$ and $\eta = \frac{\varepsilon_1}{100(\varepsilon_1+\nu)}$, we have that each of the four terms in the parenthesis above is at most $\varepsilon/5$, and thus,

$$\max_{i\in[k]} L_i(h^{\text{final}}) = \max_{i\in[k]} \frac{1}{T}\sum_{t=1}^{T} L_i(h^t) \leq \nu + \varepsilon$$

with probability at least $1 - (2k+2)\log T \cdot \delta'$. Since $\delta' = \frac{\delta}{4(T+k+1)\log T}$, the proof finishes.

C.4. Proof of Lemma 15

Again, our proof closely follows that of Lemma 3 in Zhang et al. (2024), with the only relevant change being the new step size $\eta = \frac{\varepsilon_1}{100(\varepsilon_1+\nu)}$. The entire argument is based on Lemma 13 in Zhang et al. (2024), and since this lemma holds as long as $\eta \leq \frac{1}{20}$, the derivation from Lemma 13 in Zhang et al. (2024) to Lemma 15 is identical to that presented in Section B.3 of Zhang et al. (2024). Therefore, our main focus will be on proving Lemma 13 in Zhang et al. (2024), which we restate below.

Lemma 16 (Lemma 13 from Zhang et al. (2024)) *Assume that lines 6–11 in Algorithm 6 returns a hypothesis h^t satisfying $L(h^t, w^t) \leq \min_{h\in\mathcal{H}} L(h, w^t) + \varepsilon_1$ in the t-th round for each $1 \leq t \leq T$. Then with probability exceeding $1 - 8T^4k\delta'$,*

$$|\mathcal{W}_j| \leq 8\cdot 10^7 \cdot \left((\log\left(\frac{1}{\eta(\nu+\epsilon_1)}\right)+1)^2\log^2(k)(\ln(k)+\ln(1/\delta'))^3(\log(T)+1)\right)\cdot 2^j \tag{17}$$

holds for all $1 \leq j \leq \bar{j}$, with $\bar{j}$ and $\mathcal{W}_j$ defined in (67) and (68a) in Zhang et al. (2024).

We derive Lemma 16 from Lemmas 14, 15, 16 and 17 in Zhang et al. (2024). An examination of the proof of their Lemma 14 shows that it still holds as long as $\eta \leq \frac{1}{20}$, which is satisfied by our choice of η. Their Lemmas 15 and 16 continue to hold as they are independent of the choice of η. Therefore, first we enhance Lemma 17 in Zhang et al. (2024) as below, where the major differences are: first, our enhanced version has an extra multiplicative factor of $\frac{1}{\nu+\varepsilon}$ on the right hand side of Eq. (19); second, we change the definition of v^t from $L(h^t, w^t) - \nu$ to $L(h^t, w^t) - \bar{\nu}$, where

$$\bar{\nu} = \min_{p\in\Delta(\mathcal{H})} \max_i \mathbb{E}_{h\sim p} L(h, D_i),$$

and $\Delta(\mathcal{H})$ is the set of probability distributions over $\mathcal{H}$. Note that $\bar{\nu} \leq \nu$ always holds, while they are not equal in general.

Lemma 17 (Enhanced Lemma 17 from Zhang et al. (2024)) *Let* $j_{\max} = \lfloor \log(1/\eta(\nu+\varepsilon_1)) \rfloor + 1$. *Assume that lines 6–11 in Algorithm 6 returns a hypothesis* h^t *satisfying* $L(h^t, w^t) \le \min_{h\in\mathcal{H}} L(h, w^t) + \varepsilon_1$ *in the* t*-th round for each* $1 \le t \le T$. *Suppose* (t_1, t_2) *is a* (p, q, x)*-segment (Zhang et al., 2024, Definition 1) satisfying* $p \ge 2q > 0$. *Then,*

$$t_2 - t_1 \ge \frac{x}{2\eta}. \tag{18}$$

Moreover, if

$$\frac{qx^2}{50(\log(1/\eta(\nu+\varepsilon_1))+1)^2} \ge \frac{1}{k},$$

holds, then with probability exceeding $1 - 6T^4 k\delta'$, *at least one of the following two claims holds:*

1. *The length of the segment satisfies*

$$t_2 - t_1 \ge \frac{qx^2}{200(\log(1/\eta(\nu+\varepsilon_1))+1)^2\eta^2(\nu+\varepsilon_1)}. \tag{19}$$

2. *The quantities* $\{v^t = L\left(h^t, w^t\right) - \bar{\nu}\}$ *obey*

$$4\sum_{\tau=t_1}^{t_2-1}(-v^\tau + \varepsilon_1) \ge \frac{qx^2}{100(\log(1/\eta(\nu+\varepsilon_1))+1)^2\eta}.$$

Proof Eq. (18) follows an indentical proof as Zhang et al. (2024, Section C.4, Part 1).

For the second claim, we would like to improve Eq. (114) in Step 4 of Zhang et al. (2024, Section C.4, Part 2) to the following form with probability $1 - \delta'$:

$$D_{\mathrm{KL}}(w^t \parallel w^{t_2}) \le 8(t_2 - t)\eta\varepsilon_1 - 4(t_2 - t)\eta v^t + 2\eta\sqrt{\frac{(t_2 - t)\nu \ln\frac{2}{\delta'}}{k}} + 68\eta \ln\frac{2}{\delta'}. \tag{20}$$

Bounding the KL Divergence. First, in their Step 1, let us avoid using inequalities and write down the following equation:

$$D_{\mathrm{KL}}(w^t \parallel w^{t_2}) = \langle w^t, \ln\frac{w^t}{w^{t_2}}\rangle = \langle w^t, \ln\frac{\frac{W^t}{Z^t}}{\frac{W^{t_2}}{Z^{t_2}}}\rangle = \ln\frac{Z^{t_2}}{Z^t} - \eta\sum_{\tau=t}^{t_2-1}\langle \hat{r}^\tau, w^t\rangle,$$

where Z^t is the normalizing factor of the weight vector W^t. In addition, their Step 2 can be refined to

$$\begin{aligned}
\ln\frac{Z^{t_2}}{Z^t} &= \sum_{\tau=t}^{t_2-1}\ln\left(\sum_{i=1}^k w_i^\tau \exp\left(\eta\hat{r}_i^\tau\right)\right) \\
&\le \sum_{\tau=t}^{t_2-1}\ln\left(\sum_{i=1}^k w_i^\tau + \sum_{i=1}^k w_i^\tau\left(\eta\hat{r}_i^\tau\right) + 2\sum_{i=1}^k w_i^\tau\eta^2\left(\hat{r}_i^\tau\right)^2\right) \\
&\le \sum_{\tau=t}^{t_2-1}\ln\left(1 + \eta\sum_{i=1}^k w_i^\tau\hat{r}_i^\tau + 2\eta^2\sum_{i=1}^k w_i^\tau\hat{r}_i^\tau\right) \\
&\le (\eta + 2\eta^2)\sum_{\tau=t}^{t_2-1}\langle\hat{r}^\tau, w^\tau\rangle
\end{aligned}$$

Combining the above two, we have

$$D_{\mathrm{KL}}(w^t \parallel w^{t_2}) \leq (\eta + 2\eta^2) \sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^\tau \rangle - \eta \sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^t \rangle \tag{21}$$

$$= \eta \left(\sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^\tau \rangle - \sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^t \rangle \right) + 2\eta^2 \sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^\tau \rangle \tag{22}$$

Let us now apply concentration bounds (Freedman's inequality, Theorem 22) on $\sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^\tau \rangle$ and $\sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^t \rangle$ respectively. First,

$$\begin{aligned} \mathrm{Var}[\langle \hat{r}^\tau, w^\tau \rangle \mid \mathcal{F}_\tau] &\leq \sum_{i=1}^{k} (w_i^\tau)^2 \cdot \frac{\mathrm{Var}_{(x,y)\sim D_i}[\ell(h^\tau, (x,y))]}{k \bar{w}_i^\tau} \\ &\leq \frac{1}{k} \sum_{i=1}^{k} w_i^\tau \ell(h^\tau, D_i) = \frac{1}{k} \ell(h^\tau, w^\tau) \end{aligned}$$

Therefore, Freedman's inequality (Theorem 22) with $X_t = \langle \hat{r}^\tau, w^\tau \rangle$, $b = 1$ implies that with probability $1 - \log T \cdot \delta'/2$,

$$\sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^\tau \rangle \leq \sum_{\tau=t}^{t_2-1} L(h^\tau, w^\tau) + 4\sqrt{\frac{\ln \frac{2}{\delta'}}{k} \sum_{\tau=t}^{t_2-1} L(h^\tau, w^\tau)} + \ln \frac{2}{\delta'} \tag{23}$$

$$\leq (t_2 - t)(\bar{\nu} + \varepsilon_1) + 4\sqrt{\frac{\ln \frac{2}{\delta'}}{k} (t_2 - t)(\bar{\nu} + \varepsilon_1)} + \ln \frac{2}{\delta'}, \tag{24}$$

where the second inequality uses the assumption that $L(h^\tau, w^\tau) \leq \min_{h \in \mathcal{H}} L(h, w^t) + \varepsilon_1$, which in turn $= \min_{p \in \Delta(\mathcal{H})} \mathbb{E}_{h \sim p} \left[L(h, w^t) \right] + \varepsilon_1 \leq \bar{\nu} + \varepsilon_1$.

As a side result, the above inequality combined with AM-GM inequality implies that

$$\sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^\tau \rangle \leq 2(t_2 - t)(\bar{\nu} + \varepsilon_1) + 5 \ln \frac{2}{\delta'} \tag{25}$$

Similarly, with probability $1 - \log T \cdot \delta'/2$,

$$-\sum_{\tau=t}^{t_2-1} \langle \hat{r}^\tau, w^t \rangle \leq -\sum_{\tau=t}^{t_2-1} L(h^\tau, w^t) + 4\sqrt{\frac{\ln \frac{2}{\delta'}}{k} \sum_{\tau=t}^{t_2-1} L(h^\tau, w^t)} + \ln \frac{2}{\delta'}. \tag{26}$$

Here, we know that for all τ, $L(h^\tau, w^t) \geq \min_{h \in \mathcal{H}} L(h^\tau, w^t) \geq L(h^t, w^t) - \varepsilon_1$ by the assumption of the lemma. Thus,

$$\sum_{\tau=t}^{t_2-1} L(h^\tau, w^t) \geq (t_2 - t)(L(h^t, w^t) - \varepsilon_1) = (t_2 - t)(\bar{\nu} + v^t - \varepsilon_1),$$

where we recall $v^t = L(h^t, w^t) - \bar{\nu}$. Now we do a case analysis based on the value of $(t_2 - t)(\bar{\nu} + v^t - \varepsilon)$:

- If $(t_2-t)(\bar{\nu}+v^t-\varepsilon_1) \geq \frac{16\ln\frac{2}{\delta'}}{k}$, observe that $f(z) = -z+\sqrt{Az}$ is monotonically decreasing for $z \in [A, +\infty)$, we have that

$$-\sum_{\tau=t}^{t_2-1}\langle \hat{r}^\tau, w^t\rangle \leq -(t_2-t)(\bar{\nu}+v^t-\varepsilon_1) + \sqrt{\frac{16\ln\frac{2}{\delta'}}{k}(t_2-t)(\bar{\nu}+v^t-\varepsilon_1)} + \ln\frac{2}{\delta'}$$

where the second inequality uses the fact that $v^t \leq \varepsilon_1$. Combining this with inequality (24), we have

$$\begin{aligned}\eta\left(\sum_{\tau=t}^{t_2-1}\langle \hat{r}^\tau, w^\tau\rangle - \sum_{\tau=t}^{t_2-1}\langle \hat{r}^\tau, w^t\rangle\right) \leq& \eta\left((t_2-t)(2\varepsilon_1-v^t) + 8\sqrt{\frac{\ln\frac{2}{\delta'}}{k}(t_2-t)(\bar{\nu}+\varepsilon_1)} + 2\ln\frac{2}{\delta'}\right)\\ \leq& \eta\left((t_2-t)(6\varepsilon_1-v^t) + 8\sqrt{\frac{\ln\frac{2}{\delta'}}{k}(t_2-t)\bar{\nu}} + 6\ln\frac{2}{\delta'}\right),\end{aligned}$$

where the second inequality uses the elementary fact that $\sqrt{A+B} \leq \sqrt{A}+\sqrt{B}$ and AM-GM inequality that $8\sqrt{\frac{\ln\frac{2}{\delta'}}{k}(t_2-t)\varepsilon_1} \leq 4(t_2-t)\varepsilon_1 + \frac{\ln\frac{2}{\delta'}}{k}$.

in addition, by Eq. (25),

$$\begin{aligned}2\eta^2\sum_{\tau=t}^{t_2-1}\langle \hat{r}^\tau, w^\tau\rangle \leq& 4\eta^2(t_2-t)(\bar{\nu}+\varepsilon_1) + 10\eta^2\ln\frac{2}{\delta'}\\ \leq& \eta(t_2-t)\varepsilon_1 + \eta\ln\frac{2}{\delta'},\end{aligned}$$

where the second inequality uses that $\eta = \frac{\varepsilon_1}{100(\nu+\varepsilon_1)} \leq \frac{1}{100}$. Plugging these bounds into Eq. (22), we get the desired inequality (20).

- Otherwise, $(t_2-t)(\bar{\nu}+v^t-\varepsilon) < \frac{16\ln\frac{2}{\delta'}}{k}$. Intuitively, this means that canceling out the $(t_2-t)\nu$ factors, i.e., the leading terms in Eqs. (24) and (26), is no longer important. In this case, continuing inequality (24), we have:

$$\begin{aligned}\sum_{\tau=t}^{t_2-2}\langle \hat{r}^\tau, w^\tau\rangle \leq& 2(t_2-t)(\bar{\nu}+\varepsilon) + 2\ln\frac{2}{\delta'} && \text{(AM-GM)}\\ \leq& 2(t_2-t)(2\varepsilon-v^t) + 34\ln\frac{2}{\delta'} && \text{(assumption for this case)}\end{aligned}$$

Using inequality (21) and dropping the (nonnegative) second term, we have

$$\begin{aligned}D_{\mathrm{KL}}(w^t \parallel w^{t_2}) \leq& 2\eta\cdot\sum_{\tau=t}^{t_2-2}\langle \hat{r}^\tau, w^\tau\rangle\\ \leq& 4\eta\cdot(t_2-t)(2\varepsilon-v^t) + 68\eta\ln\frac{2}{\delta'}\end{aligned}$$

Final Steps Let's recall that $T = \widetilde{O}\left(\frac{\varepsilon+\nu}{\varepsilon^2}\right)$, $\eta = \tilde{O}\left(\sqrt{\frac{1}{T(\varepsilon+\nu)}}\right) = O\left(\frac{\varepsilon}{\varepsilon+\nu}\right)$, $\varepsilon_1 = O(\varepsilon)$ and $j_{\max} = \left\lfloor \log \frac{1}{\eta(\nu+\varepsilon_1)} \right\rfloor + 1$. Due to the slight change in $j_{\max}$, Eq. (116) of Zhang et al. (2024) now needs to be modified to: there exists some $1 \le \tilde{j} \le j_{\max}$ such that

$$\ln\left(\frac{y_{\tilde{j}+1}}{y_{\tilde{j}}}\right) \ge \frac{x}{\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1}$$

For for this $\tilde{j}$, by the stronger KL divergence upper bound (Eq. (20)), equation (119) in Zhang et al. (2024) becomes the following:

$$\tau_{\tilde{j}+1} - \tau_{\tilde{j}} \gtrsim \min\left(\frac{qx^2}{\left(\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^2} \min\left\{ \frac{1}{\eta\varepsilon_1}, \frac{2^{\tilde{j}-1}}{\eta} \right\}, \frac{kq^2x^4}{\eta^2\nu \log\frac{1}{\delta'}\left(\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^4} \right) \tag{27}$$

$$\gtrsim \frac{qx^2}{\left(\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^2} \cdot \frac{2^{\tilde{j}-1}}{\eta} \tag{28}$$

Here, the second inequality holds since

$$\frac{qx^2 \cdot 2^{\tilde{j}-1}}{\eta} \lesssim \frac{qx^2}{\eta\varepsilon_1} \lesssim \frac{kq^2x^4}{\eta^2\nu \log\frac{1}{\delta'}\left(\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^2}$$

where the first inequality uses that $2^{\tilde{j}-1} \le 2^{j_{\max}-1} \le \frac{1}{\eta(\nu+\epsilon)}$ and the second uses the assumption that $\frac{qx^2}{50(\log \frac{1}{\eta(\nu+\varepsilon_1)}+1)^2} \ge \frac{1}{k}$ and $\eta\nu \lesssim \varepsilon_1$.

Then Case 1 ($\tilde{j} = j_{\max}$) in Step 5 of Section C.4 in Zhang et al. (2024) becomes

$$t_2 - t_1 \ge \tau_{\tilde{j}+1} - \tau_{\tilde{j}} \gtrsim \frac{qx^2}{\left(\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^2} \cdot \frac{2^{j_{\max}-1}}{\eta} \gtrsim \frac{qx^2}{200\left(\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^2 \eta^2(\nu+\varepsilon)}$$

Moreover, Case 2 ($1 \le \tilde{j} \le j_{\max} - 1$) in Step 5 of Section C.4 in Zhang et al. (2024) stays the same. This completes the proof of the lemma. ■

We now proceed to prove Lemma 16 using Lemma 17.

Proof Define $\left\{\mathcal{V}_j^n\right\}_{n=1}^N$ and $\{(\hat{s}_n, \hat{e}_n)\}_{n=1}^N$ the disjoint expert subsets and time intervals guaranteed to exist by Lemma 16 of Zhang et al. (2024).

Using our new Lemma 17 (which is a strengthening of their Lemma 17), the inequality (79) in Zhang et al. (2024) can be strengthened to

$$T\eta(\nu+\varepsilon) + \left\{ 4T\varepsilon_1 + 4\sum_{t=1}^{T}\left(-v^t\right) \right\} \ge (\nu+\varepsilon)\sum_{n=1}^{N}(\hat{e}_n - \hat{s}_n)\,\eta + 4\sum_{n=1}^{N}\sum_{\tau=\hat{s}_n}^{\hat{e}_n-1}\left(-v^\tau + \varepsilon_1\right) \tag{29}$$

$$\gtrsim \frac{2^{-j}\sum_{n=1}^{N}|\mathcal{V}_j^n|}{\log^2(k)(\log \frac{1}{\eta(\nu+\varepsilon_1)} + 1)^2 \cdot \eta} \tag{30}$$

where the first inequality uses that $-v^\tau + \varepsilon_1 \geq 0$ for all τ and dropping those terms whose time steps do not lie in $[\hat{s}_n, \hat{e}_n)$; the second inequality is from our new Lemma 17.

In addition, we strengthen their Eq. (81) and bound $\sum_{t=1}^{T}(-v^t)$ by the following:

$$\begin{aligned}\sum_{t=1}^{T}(-v^t) =& \bar{\nu}T - \sum_{t=1}^{T} L(h^t, w^t) \\ \leq& \max_{i\in[k]} \sum_{t=1}^{T} L_i(h^t) - \sum_{t=1}^{T} L(h^t, w^t) \\ \leq& \frac{3\ln(k)}{2\eta} + \frac{3}{2}\eta T(\nu+\varepsilon_1) + 58\ln(1/\delta') + 5\sqrt{2\ln(1/\delta')T(\nu+\varepsilon_1)} \\ \leq& T\varepsilon\end{aligned}$$

where the first inequality is since $\bar{\nu} = \min_{p\in\Delta(\mathcal{H})}\max_i \mathbb{E}_{h\sim p} L(h, D_i) \leq \max_{i\in[k]} \frac{1}{T}\sum_{t=1}^{T} L_i(h^t)$, and the second inequality is from Eq. (15).

As a result, inequality (82) in Zhang et al. (2024) becomes

$$\begin{aligned}\frac{\sum_{n=1}^{N}\left|\mathcal{V}_j^n\right|}{2^j} &\lesssim \left(\eta\left(T\varepsilon\right) + \eta\left(T\eta(\nu+\varepsilon) + T\varepsilon\right)\right)\cdot \log^2(k)\left(\log\frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^2 \\ &\lesssim \ln\left(\frac{k}{\delta'\varepsilon}\right)\cdot\log^2(k)\left(\log\frac{1}{\eta(\nu+\varepsilon_1)} + 1\right)^2\end{aligned}$$

as we desired. The rest of the proof then follows. ■

Appendix D. Deferred materials from Section 6

Proof [Proof of Theorem 5] Since $\mathcal{H}$ has star number $k\vartheta$, there exists some $h_0, h_1, \ldots, h_{k\vartheta} \in \mathcal{H}$ and a set of $k\vartheta$ examples $X = \{x_1, \ldots, x_{k\vartheta}\}$, such that: for every $i \in [k\vartheta]$, $h_i(x_i) \neq h_0(x_i)$, and $h_i(x_j) = h_0(x_j)$ for all $j \neq i$.

For $1 \leq i \leq k$, let

$$X^i = \left\{x_{(i-1)\vartheta+1}, \cdots, x_{i\vartheta}\right\},$$

Thus,

$$X = \bigcup_{1\leq i\leq k} X^i,$$

in other words, $(X^i)_{i=1}^k$ form a partition of the set X. For each $1 \leq i \leq k$ and $1 \leq j \leq \vartheta$, let D_j^i be the distribution with marginal uniform on X^i and the labels of all examples are consistent with $h_{(i-1)\vartheta+j}$. Also, let D_0^i be the distribution with marginal uniform on X^i in which the labels of all examples are consistent with h_0.

It can be checked that the disagreement coefficient of any $D \in \left\{D_j^i : i \in [k], j \in \{0, 1, \ldots, \vartheta\}\right\}$ is at most ϑ; to see this, note that for any D above, $\mathbf{B}_D(h, r) = \{h\}$ unless $r \geq \frac{1}{\vartheta}$, and thus $\Pr\left[\mathrm{DIS}(\mathbf{B}_D(h, r))\right] \leq \vartheta r$ for all $r > 0$.

Define a family of $k\vartheta$ MDL instances $\mathcal{D}_j^i$, $i \in [k]$, $j \in [\vartheta]$ as follows:

$$\mathcal{D}_j^i = (D_0^1, \ldots, D_0^{j-1}, D_i^j, D_0^{j+1}, \ldots, D_0^k)$$

By the previous paragraph, we know that for every $\mathcal{D}_j^i$, $\theta_{\max}(\varepsilon)$ is at most ϑ. In addition, it can be easily seen that $L_{\mathcal{D}_j^i}(h_{(i-1)\vartheta+j}) = 0$, thus $\min_{h\in\mathcal{H}} L_{\mathcal{D}_j^i}(h) = 0$. We next show a simple property that all $\mathcal{D}_j^i$'s are sufficiently apart, in that no classifier can be simultaneously near-optimal in any pair of them:

Claim 2 *For any classifier $\hat{h}$, and any $(i_1, j_1) \neq (i_2, j_2)$, $L_{\mathcal{D}_{j_1}^{i_1}}(\hat{h}) \leq \varepsilon$ and $L_{\mathcal{D}_{j_2}^{i_2}}(\hat{h}) \leq \varepsilon$ cannot hold simultaneously.*

Proof $L_{\mathcal{D}_{j_1}^{i_1}}(\hat{h}) \leq \varepsilon$ implies that

$$\Pr_{x\sim\mathrm{Unif}(X^{i_1})}\left[\hat{h}(x) = h_0(x), x = x_{\vartheta(i_1-1)+j_1}\right] \leq \Pr_{(x,y)\sim D_{j_1}^{i_1}}\left[\hat{h}(x) \neq y\right] \leq \varepsilon. \tag{31}$$

On the other hand, $L_{\mathcal{D}_{j_2}^{i_2}}(\hat{h}) \leq \varepsilon$ implies that

$$\Pr_{x\sim\mathrm{Unif}(X^{i_1})}\left[\hat{h}(x) \neq h_0(x), x = x_{\vartheta(i_1-1)+j_1}\right] \leq \Pr_{(x,y)\sim(\mathcal{D}_{j_2}^{i_2})_{i_1}}\left[\hat{h}(x) \neq y\right] \leq \varepsilon. \tag{32}$$

Here, $(\mathcal{D}_{j_2}^{i_2})_{i_1}$ is the i_1-th distribution of problem instance $\mathcal{D}_{j_2}^{i_2}$, which is $D_{j_2}^{i_1}$ if $i_1 = i_2$, and is $D_0^{i_1}$ if $i_1 \neq i_2$; in either case, the distribution has $x_{\vartheta(i_1-1)+j_1}$ agreeing with h_0.

Adding up Eqs. (31) and (32) and using triangle inequality, we have

$$\frac{1}{\vartheta} = \Pr_{x\sim\mathrm{Unif}(X^{i_1})}(x = x_{\vartheta(i_1-1)+j_1}) \leq 2\varepsilon.$$

which contradicts with the assumption that $\varepsilon < \frac{1}{2\vartheta}$. ■

For any algorithm A, consider an algorithm A_0 that chooses label queries identical to A except that every label query made by A_0 on example x returns $h_0(x)$. Observe that the distribution of the transcript of A_0 is the same on every problem instance $\mathcal{D}_j^i$ because different problem instances has the same set of marginal distributions over $\mathcal{X}$.

Next, we show that if A makes fewer than $\frac{k\vartheta}{20}$ queries, then it fails to solve a family of two problem instances with good probability. Denote by $\hat{h}$ the classifier returned by A after $\frac{k\vartheta}{20}$ queries. Let $d(x)$ denote the expected number of queries for each $x \in \mathcal{X}$ during the first $\frac{k\vartheta}{20}$ queries of A_0. Therefore,

$$\sum_{x\in X} d(x) \leq \frac{k\vartheta}{20}.$$

By the Pigeonhole Principle, at least half of x in X satisfies $d(x) \leq \frac{1}{10}$. Let $x_{(i_1-1)\vartheta+j_1}$ and $x_{(i_2-1)\vartheta+j_2}$ be two such points. Then, by Markov's inequality and the union bound, the probability that A_0 queries either of these points is at most 0.2. Consider the problem instances $\mathcal{D}$ chosen uniformly from $\left\{\mathcal{D}_{j_1}^{i_1}, \mathcal{D}_{j_2}^{i_2}\right\}$, denote by event E that A does not query point $x_{(i_1-1)\vartheta+j_1}$ or $x_{(i_2-1)\vartheta+j_2}$, then we claim that

$$\Pr\left[L_{\mathcal{D}}\left(\hat{h}\right) \leq \varepsilon \mid E\right] \leq 0.5,$$

where the randomness comes from both the internal randomness of algorithm A and the distribution of the problem instance $\mathcal{D}$. To see this, note that conditioned on the event A doesn't query $x_{(i_1-1)\vartheta+j_1}$ or $x_{(i_2-1)\vartheta+j_2}$, the posterior distribution of $\mathcal{D}$ after all $\frac{k\vartheta}{20}$ queries is still uniform over $\left\{\mathcal{D}_{j_1}^{i_1}, \mathcal{D}_{j_2}^{i_2}\right\}$. Therefore, the left hand side is

$$\frac{1}{2}\left(\Pr\left[L_{\mathcal{D}_{j_1}^{i_1}}\left(\hat{h}\right) \leq \varepsilon \mid E\right] + \Pr\left[L_{\mathcal{D}_{j_2}^{i_2}}\left(\hat{h}\right) \leq \varepsilon \mid E\right]\right),$$

which is at most 0.5 because of Lemma 2.

Notice that the behavior of A_0 and A is identical until A queries some x and encounters a label $-h_0(x)$. Therefore,

$$\Pr\left[A \text{ doesn't query } x_{(i_1-1)\vartheta+j_1} \text{ or } x_{(i_2-1)\vartheta+j_2}\right] = \Pr\left[A_0 \text{ doesn't query } x_{(i_1-1)\vartheta+j_1} \text{ or } x_{(i_2-1)\vartheta+j_2}\right].$$

Consequently, the success probability of A is at most

$$\begin{aligned}&\Pr\left[L_{\mathcal{D}}\left(\hat{h}\right) \leq \varepsilon \mid A \text{ doesn't query } x_{(i_1-1)\vartheta+j_1} \text{ or } x_{(i_2-1)\vartheta+j_2}\right] \\ &+ \Pr\left[A_0 \text{ queries } x_{(i_1-1)\vartheta+j_1} \text{ or } x_{(i_2-1)\vartheta+j_2}\right] \leq 0.5 + 0.2 = 0.7\end{aligned}$$

Therefore, there exists at least one instance in $\{\mathcal{D}_{j_1}^{i_1}, \mathcal{D}_{j_2}^{i_2}\}$ that A cannot solve with fewer than $\frac{k\vartheta}{20}$ queries, with probability exceeding 0.7. ■

Proof [Proof of Theorem 6] We construct the problem instance as follows. Let the hypothesis class $\{h_1, h_2\}$ be the two hypotheses satisfying the assumption in the theorem. Let $\mathcal{X} = X_{\mathrm{DIS}} \cup X_{\mathrm{AGR}}$, where:

- $X_{\mathrm{DIS}} = \{x_1, x_2\}$ is the disagreement region: for all $x \in X_{\mathrm{DIS}}$, $h_1(x) \neq h_2(x)$.
- $X_{\mathrm{AGR}} = \{z_1, \cdots, z_k\}$ is the agreement region: for all $x \in X_{\mathrm{AGR}}$, $h_1(x) = h_2(x)$.

We next define several building blocks for constructing our distributions. Define

$$D_{x_1}(x, y) = I(x = x_1)I(y = h_1(x_1)) \quad \text{and} \quad D_{x_2}(x, y) = I(x = x_2)I(y = h_2(x_2)).$$

Also, let

$$D_{z_1}(x, y) = I(x = z_1)I(y = h_1(z_1)),$$

and for $2 \leq i \leq k$,

$$D_{z_i}(x, y) = I(x = z_i)\left(\left(1 - \frac{\nu - 4\varepsilon}{2 - \nu}\right) I\left(y = h_1(z_i)\right) + \frac{\nu - 4\varepsilon}{2 - \nu} I\left(y \neq h_1(z_i)\right)\right),$$

and

$$D'_{z_i}(x, y) = I(x = z_i)\left(\left(1 - \frac{\nu + 4\varepsilon}{2 - \nu}\right) I\left(y = h_1(z_i)\right) + \frac{\nu + 4\varepsilon}{2 - \nu} I\left(y \neq h_1(z_i)\right)\right),$$

for $2 \leq i \leq k$. Next, define the distributions

$$D_1 = \nu D_{x_1} + (1 - \nu)D_{z_1},$$

and for $2 \le i \le k$,

$$D_i = \frac{\nu}{2} D_{x_2} + \left(1 - \frac{\nu}{2}\right) D_{z_i},$$

and

$$D'_i = \frac{\nu}{2} D_{x_2} + \left(1 - \frac{\nu}{2}\right) D'_{z_i}.$$

From the construction, we obtain the following table for the error of h_1 and h_2 on the distributions under consideration:

Error	D_1	D_2	$\dots$	D_k	D'_2	$\dots$	D'_k
h_1	0	$\nu - 2\varepsilon$	$\dots$	$\nu - 2\varepsilon$	$\nu + 2\varepsilon$	$\dots$	$\nu + 2\varepsilon$
h_2	ν	$\frac{\nu}{2} - 2\varepsilon$	$\dots$	$\frac{\nu}{2} - 2\varepsilon$	$\frac{\nu}{2} + 2\varepsilon$	$\dots$	$\frac{\nu}{2} + 2\varepsilon$

Table 3: Error of h_1 and h_2 on the constructed distributions

Define our central MDL problem instance as $\mathcal{D} = \{D_i\}_{i \in [k]}$, and define $k-1$ auxiliary MDL problem instances such that $\mathcal{D}_i$ is $\mathcal{D}$ with the ith distribution D_i replaced by D'_i for $2 \le i \le k$. Suppose algorithm A uses a label budget of n and guarantees that under all instances $\mathcal{Q} \in \{\mathcal{D}, \mathcal{D}_2, \dots, \mathcal{D}_k\}$,

$$\Pr_{A,\mathcal{Q}}\left[L_{\mathcal{Q}}(\hat{h}) \le \min_{h \in \mathcal{H}} L_{\mathcal{Q}}(h) + \varepsilon\right] \ge 0.9.$$

Note that the event

$$\left\{L_{\mathcal{D}}(\hat{h}) \le \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\right\} = \left\{\hat{h} = h_1\right\},$$

and for all $i \in \{2, \dots, k\}$,

$$\left\{L_{\mathcal{D}_i}(\hat{h}) \le \min_{h \in \mathcal{H}} L_{\mathcal{D}_i}(h) + \varepsilon\right\} = \left\{\hat{h} = h_2\right\}.$$

Therefore, for all $2 \le i \le k$,

$$\Pr_{A,\mathcal{D}}\left[\hat{h} = h_2\right] + \Pr_{A,\mathcal{D}_i}\left[\hat{h} = h_1\right] \le 0.2.$$

By the Bretagnolle-Huber inequality, it follows that

$$\frac{1}{2}\exp\Big(-D_{\mathrm{KL}}\left(\mathbb{P}_{A,\mathcal{D}} \| \mathbb{P}_{A,\mathcal{D}_i}\right)\Big)$$

bounds the total variation distance between the joint distributions $\mathbb{P}_{A,\mathcal{D}}$ and $\mathbb{P}_{A,\mathcal{D}_i}$, where these denote the joint distribution of the interaction transcript $(x_1, y_1, \dots, x_n, y_n)$ between A and the problem instances $\mathcal{D}$ and $\mathcal{D}_i$, respectively. This implies that

$$D_{\mathrm{KL}}(\mathbb{P}_{A,\mathcal{D}} \| \mathbb{P}_{A,\mathcal{D}_i}) \ge \ln \frac{5}{2}.$$

By the divergence decomposition lemma for interactive learning algorithms (Lattimore and Szepesvári, 2020, Lemma 15.1),

$$D_{\mathrm{KL}}\left(\mathbb{P}_{A,\mathcal{D}} \| \mathbb{P}_{A,\mathcal{D}_i}\right) = \mathbb{E}_{A,\mathcal{D}}\Big[n_i D_{\mathrm{KL}}\left(\mathbb{P}_{Y_i} \Big\| \mathbb{P}_{Y'_i}\right)\Big],$$

where n_i is the number of label queries to z_i, $\mathbb{P}_{Y_i}$ is z_i's label distribution under D_i, and $\mathbb{P}_{Y_i'}$ is z_i's label distribution under D_i'. Let $p = \frac{\nu - 4\varepsilon}{2-\nu}$ and $q = \frac{\nu + 4\varepsilon}{2-\nu}$, for each i, from Lemma 27, we have

$$D_{\mathrm{KL}}\left(\mathbb{P}_{Y_i} \middle\| \mathbb{P}_{Y_i'}\right) = \int_p^q \frac{x-p}{x(1-x)} \mathrm{d}x.$$

Because $0 < 8\varepsilon \leq \nu \leq \frac{1}{2}$,

$$x \geq p = \frac{\nu - 4\varepsilon}{2-\nu} \geq \frac{\nu/2}{2} \geq \frac{\nu}{4},$$

and

$$1 - x \geq 1 - q \geq \frac{2 - \frac{3}{2}\nu}{2-\nu} \geq \frac{1}{2}.$$

As a result, $x(1-x) \geq \frac{\nu}{8}$ for all $x \in [p, q]$. Thus,

$$D_{\mathrm{KL}}(\mathbb{P}_{A,\mathcal{D}} \| \mathbb{P}_{A,\mathcal{D}_i}) \leq \int_p^q \frac{x-p}{\nu/8} \mathrm{d}x = \frac{8}{\nu} \int_p^q (x-p)\, \mathrm{d}x = \frac{4(q-p)^2}{\nu}.$$

Note that $q - p = \frac{4\varepsilon}{2-\nu}$ so

$$\frac{4(q-p)^2}{\nu} = \frac{4}{\nu}\left(\frac{4\varepsilon}{2-\nu}\right)^2 \leq \frac{64\varepsilon^2}{\nu},$$

where the last inequality comes from the assumption $\nu \leq \frac{1}{2}$. This implies that for all $i \in \{2, \ldots, k\}$,

$$\mathbb{E}_{A,\mathcal{D}}[n_i] \geq \frac{\nu}{64\varepsilon^2} \ln \frac{5}{2}.$$

Taking the summation over all i, we conclude that the expected number of queries made by A on $\mathcal{D}$ is at least

$$(k-1) \cdot \frac{\nu}{64\varepsilon^2} \ln \frac{5}{2}.$$

■

Appendix E. Deferred materials from Section 7

E.1. PASSIVE-RPU-MDL and its guarantees

We first present PASSIVE-RPU-MDL (Alg. 7), a key subprocedure used by our distribution-free active learning algorithm Alg. 4; it can be viewed as a collaborative RPU learning algorithm robust to label noise. It takes a collection of distributions $(\mu_i)_{i\in[k]}$ that are approximately realizable by h^*, and tries to find a ξ-RPU classifier for some target $\xi > 0$ (recall definition 7). To this end, it calls a subprocedure ROBUST-RPU-LEARN (Alg. 8) that learns a RPU classifier for a single distribution (step 4). Our algorithmic idea is largely inspired by Blum et al. (2017) for the ordinary collaborative PAC learning problem: calling ROBUST-RPU-LEARN on uniform mixtures of subsets of distributions in $(\mu_i)_{i\in[k]}$ (denoted by N_r, step 3); as soon as our trained RPU classifier has low abstention probability on some μ_j, we remove μ_j from the subset (step 5). After learning RPU

Algorithm 7 PASSIVE-RPU-MDL

Require: Hypothesis class $\mathcal{H}$ with start number $\mathfrak{s}$, distributions $(\mu_i)_{i\in[k]}$, target reliability ξ, target confidence δ

$N_1 \leftarrow [k]$, $r \leftarrow 1$

while $N_r \neq \emptyset$ **do**

Define $\widetilde{\mu}_r := \frac{1}{|N_r|}\sum_{i\in N_r}\mu_i$

Let $f^r \leftarrow$ ROBUST-RPU-LEARN$(\mathcal{H}, \widetilde{\mu}_r, \frac{\xi}{2}, \frac{\delta}{2})$

$N_{r+1} \leftarrow N_r \setminus G_r$, where $G_r = \{i : \Pr_{D_i}(f^r(x) = 0) \leq \xi\}$

$r \leftarrow r + 1$

return $\hat{f}$, such that

$$\hat{f}(x) = \begin{cases} 0, & \forall r, f^r(x) = 0 \\ f^c(x), & c = \min\{r : f^r(x) \neq 0\} \end{cases}$$

classifiers that cover all of the $(\mu_i)_{i\in[k]}$, we return $\hat{f}$, which makes a ± 1 prediction whenever one of the learned f^r's makes a ± 1 prediction (step 7).

ROBUST-RPU-LEARN (Alg. 8) is our procedure for robust RPU learning for a single distribution in the presence of noise. Different from the previous RPU learning algorithms (Wiener et al., 2015; Kane et al., 2017; Hopkins et al., 2020), it is able to tolerate adversarial label noise. To this end, it samples multiple subsamples S_i's and builds a RPU classifier f_i for each S_i. Noting that some S_i's may be inconsistent with h^*, causing the output f_i to be unreliable, we take a thresholded majority vote of the f_i's, denoted as $\hat{f}$ (step 9): $\hat{f}$ abstains on x if not enough f_i's output a binary prediction on it; otherwise we predict the majority binary label of the f_i's. We present in the following lemma about the sample efficiency and noise tolerance of ROBUST-RPU-LEARN:

Lemma 18 *Suppose hypothesis class $\mathcal{H}$ and distribution μ is such that: there exists $h^* \in \mathcal{H}$ such that $L(h^*, \mu) \leq \eta$, and the target reliability $\xi \geq 100\mathfrak{s}\eta$, then* ROBUST-RPU$(\xi, \delta)$ *is such that with probability $1 - \delta$: (1) it outputs $\hat{f}$ that is ξ-RPU with respect to (h^*, μ); (2) the total number of iid examples sampled from μ is $\widetilde{O}(\frac{\mathfrak{s}}{\xi}\ln\frac{1}{\delta})$.*

With this, we now present the sample complexity and noise tolerance guarantee of PASSIVE-RPU-MDL in Lemma 19. The proofs of both lemmas can be found at the end of this subsection.

Lemma 19 *Suppose hypothesis class $\mathcal{H}$ and distributions $\mu_1, \ldots, \mu_k$ are such that there exists h^* such that*

$$\max_{i\in[k]} L(h^*, \mu_i) \leq \eta,$$

and the target reliability $\xi \geq 100\mathfrak{s}\eta$, then PASSIVE-RPU-MDL$(\xi, \delta)$ *is such that with probability $1 - \delta$: (1) it outputs $\hat{f}$ that is ξ-RPU with respect to $(h^*, (\mu_i)_{i=1}^k)$; (2) The total number of examples sampled from any of the μ_i's is $\widetilde{O}(\frac{\mathfrak{s}}{\xi})$.*

We prove Lemma 18 first.

Proof [Proof of Lemma 18] Define event E_i as:

$$E_i := \left\{ f_i \text{ is } \frac{\xi}{2}\text{-RPU with respect to } (h^*, \mu) \right\}$$

Algorithm 8 ROBUST-RPU-LEARN

Require: Hypothesis class $\mathcal{H}$ with star number $\mathfrak{s}$, distribution μ, target reliability ξ, target confidence δ

Let $N \leftarrow 60\lceil \ln \frac{1}{\delta} \rceil$

for $i = 1, \ldots, N$: **do**

Let $S_i \leftarrow$ Sample $n = O\left(\frac{\mathfrak{s}+\ln \frac{N}{\delta}}{\xi}\right)$ iid examples from μ

define $V_i := \{h \in \mathcal{H} : h(x) = y, \text{for all } (x, y) \in S_i\}$

if $V_i = \emptyset$ **then**

Define $f_i \equiv 0$.

{In this case, we know that sample S_i has been corrupted; we choose $f_i \equiv 0$ as a convention}

else

Define $f_i : \mathcal{X} \to \{-1, +1, 0\}$ such that

$$f_i(x) = \begin{cases} 0 & x \in \mathrm{DIS}(V_i) \\ V_i(x) & \text{otherwise,} \end{cases}$$

return $\hat{f}$, defined as:

$$\hat{f}(x) = \begin{cases} 0 & \sum_{i=1}^{N} I(f_i(x) \neq 0) \leq \frac{N}{5} \\ \mathrm{sign}(\sum_{i=1}^{N} f_i(x)) & \text{otherwise} \end{cases}$$

We first prove a claim that shows that E_i happens with constant probability.

Claim 3 $\Pr[E_i] \geq \frac{9}{10}$.

Proof For a set of labeled examples S, let $V[h^*, S]$ denote the set of hypotheses in $\mathcal{H}$ consistent with h^* on S.

Define $F_i = \{\forall (x, y) \in S_i : h^*(x) = y\}$ and $G_i = \left\{\Pr_\mu[x \in \mathrm{DIS}(V[h^*, S])] \leq \frac{\xi}{2}\right\}$. It can be seen that $E_i \supseteq F_i \cap G_i$. Indeed, when $F_i \cap G_i$ happens, for all $(x, y) \in S_i$, $h^*(x) = y$ and therefore,

$$f_i(x) \neq 0 \implies f_i(x) = h^*(x)$$

is true. In addition, $V_i = V[h^*, S_i]$. Since when G_i happens, $\Pr_\mu[x \in \mathrm{DIS}(V[h^*, S_i])] \leq \frac{\xi}{2}$, we also have $\Pr_\mu[x \in \mathrm{DIS}(V_i)] \leq \frac{\xi}{2}$.

We now lower bound $\Pr[F_i]$ and $\Pr[G_i]$ respectively.

- By union bound, $\Pr[F_i] \geq 1 - n\eta \geq \frac{39}{40}$.
- Meanwhile, by (Wiener et al., 2015, Lemma 8) (see Lemma 26, also (Hanneke, 2024, Appendix E.1)), $\Pr[G_i] \geq 1 - \frac{1}{40} = \frac{39}{40}$.

Therefore, by union bound,

$$\Pr[E_i] \geq 1 - \Pr\left[F_i^C\right] - \Pr\left[G_i^C\right] \geq \frac{19}{20}.$$

■

We now continue the proof of Lemma 18. Define

$$\mathcal{I} := \left\{ i \in [N] : f_i \text{ is } \frac{\xi}{2}\text{-RPU with respect to } (h^*, \mu) \right\}$$

Using Claim 3 and applying Chernoff bound, we have that with probability $1-\delta$,

$$|\mathcal{I}| \geq \frac{9N}{10}. \tag{33}$$

We henceforth condition on Eq. (33) happening.

We now prove that $\hat{f}$ is reliable. By the definition of $\hat{f}(x)$, when $\hat{f}(x) \neq 0$,

$$\sum_{i=1}^{N} I(f_i(x) \neq 0) \geq \frac{N}{5} + 1.$$

Out of those i's such that $f_i(x) \neq 0$, by Eq. (33), at most $\frac{N}{10}$ of them disagree with $h^*(x)$. Therefore, at least half of the f_i's that predict ± 1 agree with h^*. Formally, $\hat{f}(x) = \text{sign}(\sum_{i=1}^{N} f_i(x)) = h^*(x)$.

To prove that $\hat{f}$ is ξ-probably useful, we have:

$$\begin{aligned}
\Pr_\mu\left[\hat{f}(x) = 0\right] &= \Pr_\mu\left[\sum_{i=1}^{N} I(f_i(x) \neq 0) \leq \frac{N}{5}\right] \\
&= \Pr_\mu\left[\sum_{i=1}^{N} I(f_i(x) = 0) \geq \frac{4N}{5}\right] \\
&\leq \Pr_\mu\left[\sum_{i\in\mathcal{I}} I(f_i(x) = 0) \geq \frac{7N}{10}\right] \\
&\leq \frac{10}{7N}\left(\sum_{i\in\mathcal{I}} \mathbb{E}_\mu\left[I(f_i(x) = 0)\right]\right) \\
&\leq \frac{10}{7N} \cdot N\frac{\xi}{2} \leq \xi,
\end{aligned}$$

where the first inequality is from that $|\mathcal{I}| \geq \frac{9N}{10}$; the second inequality is Markov's inequality; the third inequality is from that for all $i \in \mathcal{I}$, f_i's are $\frac{\xi}{2}$-probably useful.

Finally, for item (2), the total number of samples drawn from μ is $N \cdot n = O(\frac{\mathfrak{s}}{\xi} \ln \frac{1}{\delta})$. ■

Proof [Proof of Lemma 19] For each iteration r, there exists an event E_r, in which f^r returned by ROBUST-RPU-LEARN$(\mathcal{H}, \widetilde{\mu}_r, \xi, \frac{\delta}{2})$ is ξ-RPU with respect to $(h^*, \tilde{\mu}_r)$. Denote by $E = \cap_{r=1}^{R} E_r$. By union bound, $\Pr(E) \geq 1-\delta$.

We henceforth condition on E happening. We first show that $N_{R+1} = \emptyset$. At each iteration r, f^r is ξ-RPU with respect to $(h^*, \tilde{\mu}_r)$, which implies that

$$\mathbb{E}_{i\sim\text{Unif}(N_r)}\left[\Pr_{\mu_i}\left[f^r(x) = 0\right]\right] \leq \frac{\xi}{2}$$

Markov's inequality yields that

$$\Pr_{i\sim\mathrm{Unif}(N_r)}\left[\Pr_{\mu_i}\left[f^r(x)=0\right]\geq\xi\right]\leq\frac{1}{2},$$

which is equivalent to $|G_r|\geq\frac{|N_r|}{2}$. This implies that $|N_{r+1}|\geq|N_r|-|G_r|\leq\frac{|N_r|}{2}$, and thus after at most $\lceil\log k\rceil$ iterations N_r will become empty.

By the reliability of f^r for all r, we have that whenever $\hat{f}(x)\neq 0$, $\hat{f}(x)=f^c(x)=h^*(x)$, proving the reliability of $\hat{f}$.

To show that $\hat{f}$ is ξ-probably useful with respect to $(\mu_i)_{i=1}^k$, we first observe that $\{f(x)=0\}\subseteq\{f^r(x)=0\}$. Now, for every $i\in[k]$, denote by r_i the index r such that $i\in G_r$. Therefore, $\Pr_{\mu_i}\left[f(x)=0\right]\leq\Pr_{\mu_i}\left[f^{r_i}(x)=0\right]\leq\xi$.

For item (2), it follows from Lemma 18 that the total number of examples sampled from any of the $\widetilde{\mu}_r$ at iteration r is $\widetilde{O}(\frac{\mathfrak{s}}{\xi})$, and thus the total number of samples across $\lceil\log k\rceil$ iterations is $\widetilde{O}(\frac{\mathfrak{s}}{\xi})$. ■

E.2. Proof of Theorem 8

Denote by event E_n that the success event in Lemma 19 holds at iteration n; that lemma implies that $\Pr(E_n)\geq 1-\delta_n$. Define $E:=\cap_{n=1}^{n_0}E_n$. By a union bound, $\Pr(E)\geq 1-\delta$. We henceforth condition on event E holding.

We will next prove by induction on n that for all $n\in[n_0-1]$, f_n is 2^{-n}-RPU with respect to h^* and $(D_i)_{i=1}^k$.

Base case. For $n=0$, we have $f_0\equiv 0$, which is 1-RPU with respect to h^* and $(D_i)_{i=1}^k$ trivially.

Inductive case. Suppose the inductive claim holds for iteration f_{n-1}. Then, each distribution $D_{i,n}$ can be equivalently expressed as:

$$D_{i,n}(x,y)=D_i(x,y)I(f_{n-1}(x)=0)+I(y=h^*(x))I(f_{n-1}(x)\neq 0).$$

Therefore, for every $i\in[k]$, $L(h^*,D_{i,n})\leq L(h^*,D_i)\leq\nu$. We also have that for $n\leq n_0$, $\varepsilon_n\geq 2^{-n_0-1}\geq\frac{\mathfrak{s}\varepsilon}{d+k}\geq 100\mathfrak{s}\nu$; combining this with the fact that E_n happens, we have that f_n is 2^{-n}-RPU with respect to h^* and $(D_{i,n})_{i=1}^k$.

Since D_i and $D_{i,n}$ have the same marginal distribution over $\mathcal{X}$, f_n is also 2^{-n} RPU with respect to h^* and $(D_i)_{i=1}^k$; this completes the induction.

We next analyze the algorithm in iteration n_0. Recall that when E happens, we have a classifier f_{n_0-1} that is $2^{-n_0-1}\leq\frac{\mathfrak{s}\varepsilon}{d+k}$-RPU with respect to h^* and $(D_i)_{i=1}^k$.

Define event F as the success event in Lemma 10 when calling PASSIVE-MDL. Since for every $i\in[k]$, $L(h^*,D_{i,n_0})\leq L(h^*,D_i)\leq\nu$, we have that $\hat{h}$, the output of PASSIVE-MDL, satisfies that

$$\max_{i\in[k]}L(\hat{h},D_{i,n})\leq\varepsilon,\tag{34}$$

and therefore, applying Lemma 25 gives us that for every $i\in[k]$,

$$L(\hat{h},D_i)\leq L(h^*,D_i)+L(\hat{h},D_{i,n})\leq\nu+\varepsilon.$$

This concludes the proof of the excess error guarantee.

We now turn to analyzing the label complexity of Algorithm 4.

- For the iterations $n \in [n_0 - 1]$, we have by Lemma 19 that the number of examples sampled from any of the $D_{i,n}$ by PASSIVE-RPU-MDL is at most $\frac{\mathfrak{s}}{\varepsilon_n}$, and by a Chernoff bound, the number of label queries made at iteration n is at most

$$N_n \leq O\left(\frac{\mathfrak{s}}{\varepsilon_n} \cdot \varepsilon_n\right) = O(\mathfrak{s}).$$

- For the last iteration n_0, Lemma 10 guarantees that the number of samples to any of the D_{i,n_0} made by PASSIVE-MDL is at most $O(\frac{d+k}{\varepsilon})$, which, combined with the fact that

$$\max_{i\in[k]} \Pr_{D_i}[f_n(x) = 0] \leq \frac{\mathfrak{s}}{d+k}\varepsilon$$

and Chernoff bound, makes

$$N_{n_0} = O\left(\mathfrak{s} + \ln\frac{1}{\delta}\right)$$

label queries.

In summary, the total number of label queries made by Algorithm 4 is at most $\sum_{n=1}^{n_0} N_n = O(\mathfrak{s} \ln \frac{1}{\varepsilon})$.

Appendix F. Additional Theorems and Lemmas

Theorem 20 (Chernoff Bound) *Suppose $X_1, \ldots, X_n$ is a collection of i.i.d. Bernoulli random variables with mean p. Then*

$$\Pr\left[\sum_{i=1}^{n} X_i \geq 2np + 2\ln\frac{1}{\delta}\right] \leq \delta.$$

Theorem 21 (Bernstein's Inequality) *For independent random variables $X_1, \ldots, X_n$ with $|X_i - \mathbb{E}[X_i]| \leq M$ almost surely, and letting $S_n = \sum_{i=1}^n X_i$ and $\sigma^2 = \sum_{i=1}^n Var(X_i)$, then for any $t > 0$:*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{2Mt}{3}}\right).$$

Theorem 22 (Freedman's Inequality, Bartlett et al. (2008)'s version) *Suppose $X_1, \ldots, X_T$ is a martingale difference sequence with $|X_t| \leq b$. Let*

$$V = \sum_{t=1}^{T} \mathrm{Var}(X_t \mid X_1, \ldots, X_{t-1}).$$

be the sum of the conditional variances of X_t's. Then we have, for any $\delta < 1/e$ and $T \geq 4$,

$$\Pr\left[\sum_{t=1}^{T} X_t > 4\sqrt{V\ln(1/\delta)} + 2b\ln(1/\delta)\right] \leq \log(T)\delta.$$

Lemma 23 *Let $A, B, C \geq 0$, if $|A - C| \leq B\sqrt{A + C}$, then $|A - C| \leq 2\left(B^2 + B\min\left\{\sqrt{C}, \sqrt{A}\right\}\right)$.*

Proof Taking square on both sides and manipulating to get the following

$$A^2 - \left(2C + B^2\right)A + C^2 - B^2C \leq 0.$$

By the formula of quadratic equations, we have

$$\begin{aligned} A &\leq \frac{(2C + B^2) + \sqrt{(2C + B^2)^2 - 4(C^2 - B^2C)}}{2} \\ &\leq \frac{(2C + B^2) + \sqrt{B^4 + 8B^2C}}{2} \\ &\leq \frac{(2C + B^2) + B^2 + 2\sqrt{2}B\sqrt{C}}{2} \\ &\leq C + B^2 + \sqrt{2}B\sqrt{C}, \end{aligned}$$

where the third inequality comes from $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$.

As a result,

$$\begin{aligned} (A - C)^2 &\leq B^2(A + C) \\ &\leq B^2\left(2C + B^2 + B\sqrt{2C}\right) \\ &\leq 2B^2\left(2C + B^2\right), \end{aligned}$$

where the last step comes from $ab \leq a^2 + b^2$. Taking square root on both sides and applying $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ again and we get

$$|A - C| \leq 2\left(B\sqrt{C} + B^2\right).$$

Due to symmetry of A and C in the condition, we also have

$$|A - C| \leq 2\left(B\sqrt{A} + B^2\right).$$

As a result,

$$|A - C| \leq 2\left(B^2 + B\min\left\{\sqrt{A}, \sqrt{C}\right\}\right).$$

■

Lemma 24 *Let $A, B, C, D \geq 0$, if $A - C \leq B\sqrt{A + C} + D$, then $A - C \leq \frac{3}{2}D + \frac{3}{2}B^2 + \sqrt{2}B\sqrt{C}$.*

Proof Moving D to the LHS, taking square on both sides and manipulating to get the following

$$A^2 - 2CA - 2DA - B^2A + 2DC + C^2 + D^2 - B^2C \leq 0.$$

By the formula of quadratic equations, we have

$$A \leq \frac{2C + 2D + B^2 + \sqrt{\left(-2C - 2D - B^2\right)^2 - 4\left(C^2 + 2DC + D^2 - CB^2\right)}}{2}.$$

Let $S = \left(-2C - 2D - B^2\right)^2 - 4\left(C^2 + 2DC + D^2 - CB^2\right)$, then let's upper bound $\sqrt{S}$. By simply expanding S, we have

$$\begin{aligned} S &= 4C^2 + 4D^2 + B^4 + 8CD + 4CB^2 + 4DB^2 - 4C^2 - 8CD - 4D^2 + 4CB^2 \\ &= B^4 + 8CB^2 + 4DB^2 \\ &= B^2\left(B^2 + 8C + 4D\right). \end{aligned}$$

Therefore,

$$\sqrt{S} \le B\sqrt{B^2 + 8C + 4D} \le B\left(B + 2\sqrt{2}\sqrt{C} + 2\sqrt{D}\right) \le B^2 + 2\sqrt{2}B\sqrt{C} + B^2 + D,$$

where the second inequality comes from $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$ for $a, b, c \ge 0$ and the last inequality is AM-GM. Putting everything together, we have

$$A \le C + \frac{3}{2}D + \frac{3}{2}B^2 + \sqrt{2}B\sqrt{C}.$$

Moving C to the LHS and we proved the lemma. ■

Lemma 25 (Favorable bias, Hsu (2010), Lemma 5.2) *Suppose we have a distribution D over $\mathcal{X} \times \mathcal{Y}$, and a classifier h^* and a region $R \subset \mathcal{X}$. Define the "favorably biased" distribution $\widetilde{D}$:*

$$\widetilde{D}(x, y) = D_i(x, y)I(x \in R) + I(y = h^*(x))I(x \notin R).$$

Then for any classifier h,

$$L(h, D) - L(h^*, D) \le L(h, \widetilde{D}) - L(h^*, \widetilde{D}).$$

Lemma 26 (Wiener et al. (2015), Lemma 8) *Suppose we have S, n iid samples from distribution D over $\mathcal{X} \times \mathcal{Y}$ realizable by hypothesis h^* in hypothesis class $\mathcal{H}$. Denote by $V = \{h : h \text{ is consistent with } S\}$, then with probability $1 - \delta$,*

$$\Pr_D\left[x \in \mathrm{DIS}(V)\right] \le \frac{10\mathfrak{s}\ln\frac{en}{\mathfrak{s}} + 4\ln\frac{2}{\delta}}{n},$$

where $\mathfrak{s}$ is the star number of $\mathcal{H}$.

Lemma 27 *For any $p, q \in (0, 1)$, the KL divergence between $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$ admits the integral representation*

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(p) \| \mathrm{Bern}(q)\big) = \int_p^q \frac{x - p}{x(1 - x)}\, dx.$$

Proof Define

$$F(q) = D_{\mathrm{KL}}\big(\mathrm{Bern}(p) \| \mathrm{Bern}(q)\big) = p\ln\frac{p}{q} + (1 - p)\ln\frac{1 - p}{1 - q}.$$

Then

$$F'(q) = -\frac{p}{q} + \frac{1-p}{1-q} = \frac{q-p}{q(1-q)}.$$

Since $F(p) = 0$, integrating from p to q yields

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(p)\|\mathrm{Bern}(q)\big) = F(q) - F(p) = \int_p^q F'(x)\,dx = \int_p^q \frac{x-p}{x(1-x)}\,dx.$$

■