Is Multi-Distribution Learning as Easy as PAC Learning: Sharp Rates with Bounded Label Noise

Rafael Hanashiro RAFAH@MIT.EDU
Abhishek Shetty SHETTY@MIT.EDU
Patrick Jaillet JAILLET@MIT.EDU
MIT

Editors: Steve Hanneke and Tor Lattimore

Abstract

Towards understanding the statistical complexity of learning from heterogeneous sources, we study the problem of multi-distribution learning. Given k data sources, the goal is to output a classifier for each source by exploiting shared structure to reduce sample complexity. We focus on the bounded label noise setting to determine whether the fast $1/\epsilon$ rates achievable in single-task learning extend to this regime with minimal dependence on k. Surprisingly, we show that this is not the case. We demonstrate that learning across k distributions inherently incurs slow rates scaling with k/ϵ^2, even under constant noise levels, unless each distribution is learned separately. A key technical contribution is a structured hypothesis-testing framework that captures the statistical cost of certifying near-optimality under bounded noise—a cost we show is unavoidable in the multi-distribution setting.

Finally, we prove that when competing with the stronger benchmark of each distribution's optimal Bayes error, the sample complexity incurs a *multiplicative* penalty in k. This establishes a *statistical* separation between random classification noise and Massart noise, highlighting a fundamental barrier unique to learning from multiple sources.

Keywords: Multi-distribution learning, Random Classification Noise, Massart noise, Hypothesis testing

1. Introduction

Data heterogeneity inherent in learning from multiple sources is a fundamental challenge in modern machine learning. In federated learning, data resides across numerous devices or institutions (Kairouz et al., 2021); in algorithmic fairness, models must perform well across diverse demographic groups (Sagawa et al., 2020); in domain adaptation, one must reason about distinct but related distributions (Ben-David et al., 2010); and in language modeling, the composition of pretraining data sources is critical. These applications have sparked significant interest in understanding how shared structure can be leveraged to minimize data requirements, with the ideal goal being that learning across heterogeneous sources incurs only a mild statistical overhead.

This motivates the study of *Multi-Distribution Learning (MDL)*, where a learner must perform well across k distributions simultaneously. Remarkably, in both the realizable and agnostic settings, MDL admits sample complexities of $\tilde{\Theta}\left(\frac{d+k}{\epsilon}\right)$ and $\tilde{\Theta}\left(\frac{d+k}{\epsilon^2}\right)$, respectively, for classes of VC dimension d (Blum et al., 2017; Haghtalab et al., 2022; Zhang et al., 2024b; Peng, 2024). This reflects only an *additive* overhead in k relative to single-distribution learning, suggesting that in these classical settings, learning from heterogeneous sources is essentially "as easy as PAC learning."

However, the realizable and agnostic settings represent extremes of label quality. A more realistic scenario involves labels that are corrupted, but not entirely adversarially. This is captured by the *bounded noise* model (Massart and Nédélec, 2006), which assumes there exists a function $f^* \in \mathcal{F}$ such that[1] $P\left(f^*\left(X\right) \neq Y | X = x\right) \leq \eta < 1/2$ for all $x \in \mathcal{X}$. Here, f^* is the *Bayes classifier* and attains the minimum possible error. In the standard single-distribution PAC setting (Valiant, 1984), empirical risk minimization

1. We will often write $P\left((X,Y) \in A\right)$ as shorthand for $\mathbb{P}_{(X,Y)\sim P}\left((X,Y) \in A\right)$.

(ERM) achieves a sample complexity of $\tilde{O}\left(\frac{d}{\epsilon(1-2\eta)}\right)$. Notably, if we treat η as constant, this recovers the fast $1/\epsilon$ realizable rate. Given the precedent that MDL incurs only a small overhead in k, one might hope that these fast rates persist under label noise as well.

In this work, we answer this question in the negative: we show that under bounded label noise, the cost of handling multiple distributions can be significantly higher than in the single-distribution setting.

Notation. We define $[n] \coloneqq \{1, 2, \ldots, n\}$ and $[\![n]\!] \coloneqq \{0, 1, \ldots, n\}$. We use $\lesssim$ and $\gtrsim$ to denote inequalities up to universal constants, and $\tilde{O}, \tilde{\Omega}, \tilde{\Theta}$ to hide logarithmic factors. For a distribution P over $\mathcal{X} \times \{0, 1\}$ and a function $f : \mathcal{X} \to \{0, 1\}$, we define the classification error $\operatorname{err}(f; P) \coloneqq P(f(X) \neq Y)$.

1.1. Problem Setup

We assume sample access to k distributions $P_1, \ldots, P_k$ over $\mathcal{X} \times \{0, 1\}$, and assume there exists a *shared* Bayes classifier $f^* \in \mathcal{F}$ such that $P_i(f^*(X) \neq Y | X = x) \leq \eta_i < 1/2$ for each $i \in [k]$ and $x \in \mathcal{X}$. We assume that the noise upper bounds $\eta_1, \ldots, \eta_k$ are known, but the optimal errors $\eta_i^* \coloneqq \operatorname{err}(f^*; P_i) \leq \eta_i$ are unknown. Note that setting all η_i to 0 recovers the standard realizable setting, whereas removing the existence assumption on f^* yields the agnostic regime. Our goal is to adaptively sample from the P_i and output decisions $\hat{f}_1, \ldots, \hat{f}_k : \mathcal{X} \to \{0, 1\}$ that compete with prescribed benchmarks $\mathsf{OPT}_1, \ldots, \mathsf{OPT}_k$, in the sense that

$$\mathbb{P}\left(\max_{i \in [k]} \left\{\operatorname{err}\left(\hat{f}_i; P_i\right) - \mathsf{OPT}_i\right\} \leq \epsilon\right) \geq 1 - \delta \tag{1}$$

This is known as the *personalized* setting, as opposed to *centralized*, since we are allowed to output a different decision per distribution. We focus on three separate regimes.

RCN. Under *random classification noise*, the pointwise error is constant, $P_i(f^*(X) \neq Y | X = x) = \eta_i$, implying the optimal error is exactly η_i. Motivated by this, we introduce the benchmark:

$$\mathsf{OPT}_i = \eta_i \tag{MDL-RCN}$$

Here, the learner aims to compete with the known upper bounds η_i, even though the true optimal errors η_i^* may be strictly smaller. We retain distinct indices η_i for completeness, but note that their heterogeneity does not fundamentally alter the hardness of the problem; one may effectively treat them as equal.

Minimax. The standard benchmark in MDL is the minimax risk $\eta^* \coloneqq \max_{i \in [k]} \eta_i^*$, corresponding to the best worst-case error a single hypothesis can achieve across all distributions. We adopt this as our second benchmark:

$$\mathsf{OPT}_i = \eta^* \tag{MDL-MM}$$

Massart. Finally, we consider the most fine-grained objective, where the learner must compete with the true optimal error of each distribution individually:

$$\mathsf{OPT}_i = \eta_i^* \tag{MDL-Mass}$$

Unlike the minimax setting, this requires the learner to adapt to the noise level of each distribution. Evidently, this variant is at least as hard as the other two, as the learner targets the strictest benchmark for each instance.

For any regime, we say that an algorithm $\mathcal{A}$ has sample complexity $T_{\mathcal{A}} : (0, 1)^{2k+2} \to \mathbb{N}$ if, given any instance $(\mathcal{F}, P_{1:k}, \eta_{1:k}^*, \eta_{1:k}, \epsilon, \delta)$, it satisfies (1) using at most $T_{\mathcal{A}}(\eta_{1:k}^*, \eta_{1:k}, \epsilon, \delta)$ samples in total across all distributions. We omit parameters from the input when there is no dependence on them. For convenience, we additionally define $\eta \coloneqq \max_{i \in [k]} \eta_i$.

Multi-Distribution Learning. MDL was introduced by (Blum et al., 2017), who established the tight realizable sample complexity $\tilde{\Theta}\left(\frac{d+k}{\epsilon}\right)$ in both the personalized and centralized settings, with later work sharpening log factors (Nguyen and Zakynthinou, 2018; Chen et al., 2018). For centralized agnostic MDL, (Haghtalab et al., 2022) gave a tight rate $\tilde{\Theta}\left(\frac{\log|\mathcal{F}|+k}{\epsilon^2}\right)$ for finite classes and proved a VC lower bound of $\tilde{\Omega}\left(\frac{d+k}{\epsilon^2}\right)$, which was later matched (up to logs) by (Zhang et al., 2024b; Peng, 2024). In the personalized regime, (Deng and Qiao, 2024) showed that if different distributions are realized by distinct classifiers, a lower bound of $\Omega\left(kd/\epsilon\right)$ applies. We circumvent this barrier by assuming a *shared* Bayes classifier.

1.2. Overview and Contributions

We provide a near-complete characterization of the sample complexity for MDL under bounded label noise. Along the way, we introduce a structured hypothesis testing problem that may be of independent interest.

Upper Bounds (Section 2). We develop a meta-algorithm (Algorithm 1) for the first two variants, following the iterative approach of (Blum et al., 2017). In each round, the algorithm learns at least half of the active distributions and identifies them via a test subroutine. Under (MDL-RCN), the target η_i are known, allowing for direct optimality certification and yielding the upper bound $\tilde{O}\left(\frac{d}{\epsilon(1-2\eta)}+\sum_{i=1}^{k}\frac{\epsilon+\eta_i}{\epsilon^2}\right)$ (Theorem 3). The minimax variant is more subtle, as the benchmark η^* is unknown. To address this, we devise a procedure that estimates η^* in increments and incurs only a logarithmic overhead. This results in an (MDL-MM) upper bound of $\tilde{O}\left(\frac{d}{\epsilon(1-2\eta)}+\frac{k(\epsilon+\eta^*)}{\epsilon^2}\right)$ (Theorem 4).

In both settings, the first term reflects the statistical cost of learning, while the second captures the testing complexity. Since the latter scales quadratically with ϵ, it can be preferable to learn each distribution independently, which would require $\tilde{O}\left(\sum_{i=1}^{k}\frac{d}{\epsilon(1-2\eta_i)}\right)$ samples. Algorithm 1 explicitly compares these regimes (via an initial check) and switches to per-distribution learning whenever it is more sample-efficient. Note that in the realizable case ($\eta_i = 0$ for all i), we recover the rate $\tilde{O}\left(\frac{d+k}{\epsilon}\right)$.

The testing component arises naturally from the learning objective, as the ability to learn an optimal classifier intuitively implies the ability to verify its performance. However, this certification incurs a statistical cost of $\tilde{O}\left(\frac{\epsilon+\eta}{\epsilon^2}\right)$ to reliably distinguish between noise rates η and $\eta+\epsilon$. Establishing the necessity of this cost constitutes the more challenging direction. Our lower bounds confirm that it is not an artifact of analysis, but a fundamental barrier in learning under bounded noise.

Structured Hypothesis Testing (Section 3). As discussed, the ability to test whether a classifier is optimal is intrinsic to the learning problem. To formalize this task, we introduce *Structured Hypothesis Testing* (SHT): given a fixed function, decide whether it is ϵ-optimal under RCN with known noise $\eta = 1/4$.

This problem admits two natural strategies: (i) directly thresholding the empirical errors, or (ii) first learning f^* to compare against the candidate. The former succeeds with $\tilde{O}\left(1/\epsilon^2\right)$ samples, while the latter requires $\tilde{O}\left(d/\epsilon\right)$. We demonstrate that this trade-off is fundamental by establishing a matching lower bound, yielding the optimal sample complexity $\tilde{\Theta}\left(\min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$ (Theorems 8 and 9).

We then extend this paradigm to *Multi-Distribution Structured Hypothesis Testing* (MSHT), where the goal is to solve k simultaneous instances of (SHT) across distributions $P_1, \ldots, P_k$ that share a Bayes classifier with fixed noise rate $\eta = 1/4$. We prove that the naive strategy of testing each distribution separately is, in fact, optimal. To establish the lower bound, we reduce (SHT) to (MSHT) by embedding a hard (SHT) instance into a specific P_i, while setting the remaining distributions to the null hypothesis. Since the index i is unknown, the (MSHT) algorithm is forced to sample sufficiently from all distributions to locate and solve the embedded instance. Crucially, we construct the (SHT) hard instance such that extending it to multiple distributions preserves the VC dimension of the underlying hypothesis class. This yields the final sample complexity $\tilde{\Theta}\left(k \cdot \min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$ (Theorems 9 and 10).

MDL Lower Bounds (Section 4). As discussed, testing optimality is a necessary condition for learning. In Lemma 12, we formalize this by showing that any MDL algorithm can be used to solve (MSHT) with an additional cost of only $\tilde{O}(k/\epsilon)$ samples. Consequently, in the fixed-noise regime where $\eta_i^* = \eta_i = 1/4$ for all $i \in [k]$, every MDL variant requires $\Omega\left(d/\epsilon + k \cdot \min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$ samples (Theorem 13). When $d \lesssim 1/\epsilon$, this becomes $\Omega(dk/\epsilon)$—no better than learning each distribution separately. This establishes optimality of our algorithms for (MDL-RCN) and (MDL-MM), and answers our motivating question in the negative.

RCN and Massart Separation (Section 5). For (MDL-Mass), where $\hat{f}_i$ must compete with the *unknown* Bayes error η_i^*, we prove the stronger lower bound $\Omega\left(d/\epsilon + k \cdot \min\left\{1/\epsilon^2, d/\epsilon\right\} + k\sqrt{d}/\epsilon\right)$ (Theorem 15) assuming that all $\eta_i \leq 0.49$. Relative to the fixed-noise regime, this reveals an additional $k\sqrt{d}/\epsilon$ penalty: certifying near-optimality against an unknown baseline is strictly harder. Consequently, while RCN and Massart noise are statistically comparable in the single-distribution setting, they become separable in MDL. This separation has long been known computationally and is closely tied to distribution shift (Chen et al., 2020). We interpret the present statistical separation in MDL as shedding new light on this phenomenon.

2. MDL Upper Bounds

In this section, we develop strategies for (MDL-RCN) and (MDL-MM). Let d denote the VC dimension of the hypothesis class $\mathcal{F}$. Algorithm 1 outlines the general template for our approach. At a high level, the learner proceeds in round t as follows:

- **Active Set Maintenance:** Maintain a set of distributions $\mathcal{U}^{(t-1)}$ that still need to be learned, initialized with the full set $\mathcal{U}^{(0)} \coloneqq \{P_1, \ldots, P_k\}$.
- **Statistical Learning:** Learn a hypothesis $f^{(t)}$ with respect to the uniform mixture over the current active set, denoted by $\bar{P}_{\mathcal{U}^{(t-1)}}$, using an ERM oracle (Line 6):

$$\mathrm{ERM}_{\mathcal{F}}(S) \in \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \widehat{\mathrm{err}}(f; S) \coloneqq \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{I}\{f(x) \neq y\} \right\}$$

- **Testing:** Invoke a variant-specific Test subroutine to identify the distributions on which $f^{(t)}$ performs well. These distributions are removed from $\mathcal{U}^{(t-1)}$ to form the next active set $\mathcal{U}^{(t)}$ (Lines 7–8).

We will show that the testing component can require $\Omega\left(k/\epsilon^2\right)$ samples. Due to this high testing overhead, it is more efficient to learn each distribution separately whenever $d \ll 1/\epsilon$. To address this, Algorithm 1 incorporates a preliminary check to determine if separate learning yields better complexity.

2.1. Statistical Learning

Under label noise bounded by η, ERM achieves (ϵ, δ)-learning with sample complexity

$$T_{\mathsf{SL}}(\eta, \epsilon, \delta) \coloneqq C_{\mathsf{SL}} \cdot \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon(1 - 2\eta)}$$

where C_{SL} is a universal constant. We provide further details in Appendix C and refer the reader to (Boucheron et al., 2005) for a comprehensive treatment. We leverage this guarantee to determine the sample size for the statistical learning component of Algorithm 1. Specifically, to ensure that the hypothesis $f^{(t)}$ is ϵ-optimal with respect to the mixture $\bar{P}_{\mathcal{U}^{(t-1)}}$, we rely on the following result.

Algorithm 1 MDL Meta-Algorithm

Input: Class $\mathcal{F}$, distributions $P_1, \ldots, P_k$, iteration count T, parameters $(\epsilon, \epsilon', \delta, \delta') \in (0,1)^4$, condition Cond, tester function Test.

Output: $\hat{f}_1, \ldots, \hat{f}_k$.

if Cond **then** ▷ Learn each distribution separately.
  for $i = 1, \ldots, k$ **do** $\hat{f}_i \leftarrow \mathrm{ERM}_{\mathcal{F}}(S_i)$ where $S_i \overset{iid}{\sim} P_i$ is of size $|S_i| = T_{\mathsf{SL}}(\eta_i, \epsilon, \delta/k)$
else
  $\mathcal{U}^{(0)} \leftarrow \{P_1, \ldots, P_k\}$
  for $t = 1, \ldots, T$ **do**
    $f^{(t)} \leftarrow \mathrm{ERM}_{\mathcal{F}}\left(S^{(t)}\right)$ where $S^{(t)} \overset{iid}{\sim} \bar{P}_{\mathcal{U}^{(t-1)}}$ is of size $\left|S^{(t)}\right| = T_{\mathsf{SL}}(\eta, \epsilon', \delta')$
    $\mathcal{U}^{(t)} \leftarrow \mathrm{Test}\left(f^{(t)}, \mathcal{U}^{(t-1)}\right)$
    for $i \in \mathcal{U}^{(t-1)} \backslash \mathcal{U}^{(t)}$ **do** $\hat{f}_i \leftarrow f^{(t)}$
  end
end

Lemma 1 *Let $\mathcal{U} \subset \{P_1, \ldots, P_k\}$ and let $\bar{P}_{\mathcal{U}} = \frac{1}{|\mathcal{U}|}\sum_{i \in \mathcal{U}} P_i$. Then, ERM $\hat{f} = \mathrm{ERM}_{\mathcal{F}}(S)$ on a sample $S \overset{iid}{\sim} \bar{P}_{\mathcal{U}}$ of size $|S| = T_{\mathsf{SL}}(\eta, \epsilon, \delta)$ satisfies*

$$\mathbb{P}\left(\mathrm{err}\left(\hat{f}; \bar{P}_{\mathcal{U}}\right) \le \epsilon + \frac{1}{|\mathcal{U}|}\sum_{i \in \mathcal{U}} \eta_i^*\right) \ge 1 - \delta$$

Applying this to Algorithm 1, we conclude that for any iteration $t \in [T]$, with probability $\ge 1 - \delta'$, we have that $\frac{1}{\left|\mathcal{U}^{(t-1)}\right|}\sum_{i \in \mathcal{U}^{(t-1)}}\left(\mathrm{err}\left(f^{(t)}; P_i\right) - \eta_i^*\right) \le \epsilon'$. Now, let $k^{(t)} := \left|\left\{i \in \mathcal{U}^{(t-1)} : \mathrm{err}\left(f^{(t)}; P_i\right) - \eta_i^* \ge 2\epsilon'\right\}\right|$ denote the number of distributions in $\mathcal{U}^{(t-1)}$ on which $f^{(t)}$ performs poorly. Then,

$$2\epsilon' \cdot \frac{k^{(t)}}{\left|\mathcal{U}^{(t-1)}\right|} \le \frac{1}{\left|\mathcal{U}^{(t-1)}\right|}\sum_{i \in \mathcal{U}^{(t-1)}}\left(\mathrm{err}\left(f^{(t)}; P_i\right) - \eta_i^*\right) \le \epsilon' \Longrightarrow k^{(t)} \le \frac{\left|\mathcal{U}^{(t-1)}\right|}{2}$$

In other words, $f^{(t)}$ is $2\epsilon'$-optimal for at least half of the distributions in $\mathcal{U}^{(t-1)}$ with high probability.

2.2. Testing

The statistical learning step guarantees $f^{(t)}$ is good on at least half of $\mathcal{U}^{(t-1)}$, but does not identify which ones. The next lemma shows we can find them by estimating empirical errors from fresh samples: by sampling $\tilde{O}\left(\frac{\epsilon+\nu}{\epsilon^2}\right)$ times from P, we can reliably distinguish whether the error $\mathrm{err}(f; P)$ is ϵ-close to a target level ν.

Lemma 2 *Let P be a distribution over $\mathcal{X} \times \{0,1\}$, $f : \mathcal{X} \to \{0,1\}$ be some function, and $\nu \ge 0$ be a constant. Let $S \overset{iid}{\sim} P$ be a sample of size $|S| \ge T_{\mathsf{T}}(\nu, \epsilon, \delta) := 16\log\left(\frac{2}{\delta}\right)\frac{\epsilon+8\nu}{\epsilon^2}$. Then, with probability at least $1 - \delta$, the following holds:*

$$\mathrm{err}(f; P) \le \frac{\epsilon}{8} + \nu \Longrightarrow \widehat{\mathrm{err}}(f; S) \le \frac{\epsilon}{2} + \nu \Longrightarrow \mathrm{err}(f; P) \le \epsilon + \nu$$

2.3. (MDL-RCN) Upper Bound

When our goal is to compete with the noise upper bounds η_i, the testing problem is simple because the target thresholds are known. Fix a round $t \in [T]$. Recall from the statistical learning step that with probability at

least $1-\delta'$, for at least half of the distributions in $\mathcal{U}^{(t-1)}$, we have $\operatorname{err}\left(f^{(t)}; P_i\right) \leq 2\epsilon' + \eta_i^* \leq 2\epsilon' + \eta_i$. Since each η_i is known, we can verify this condition directly. For each distribution $i \in \mathcal{U}^{(t-1)}$, we draw a sample $S_i^{(t)} \overset{iid}{\sim} P_i$ of size $\left|S_i^{(t)}\right| = T_{\mathsf{T}}\left(\eta_i, \epsilon, \delta''\right)$. Applying Lemma 2 together with a union bound, we guarantee that with probability at least $1 - k\delta''$, the following implications hold for all i:

$$\operatorname{err}\left(f^{(t)}; P_i\right) \leq \frac{\epsilon}{8} + \eta_i \Longrightarrow \widehat{\operatorname{err}}\left(f^{(t)}; S_i^{(t)}\right) \leq \frac{\epsilon}{2} + \eta_i \Longrightarrow \operatorname{err}\left(f^{(t)}; P_i\right) \leq \epsilon + \eta_i$$

We set $\epsilon' = \epsilon/16$ and define the update rule as follows: eliminate distribution i if and only if its empirical error satisfies $\widehat{\operatorname{err}}\left(f^{(t)}; S_i^{(t)}\right) \leq \epsilon/2 + \eta_i$. This strategy guarantees two key properties:

- **(Progress)** at least half of the distributions in $\mathcal{U}^{(t-1)}$ (specifically those with $\operatorname{err}\left(f^{(t)}; P_i\right) \leq \epsilon/8 + \eta_i$) will satisfy the empirical condition and be eliminated.
- **(Correctness)** any eliminated distribution is guaranteed to satisfy $\operatorname{err}\left(f^{(t)}; P_i\right) \leq \epsilon + \eta_i$.

Thus, in each round t, we successfully learn and remove at least half of the active distributions with probability $1 - \delta' - k\delta''$. Appropriately tuning the parameters results in Algorithm 2.

Algorithm 2 RCN-Test

Input: Decision $f^{(t)}$, distribution set $\mathcal{U}^{(t-1)}$.
Output: $\mathcal{U}^{(t)}$.

1 **for** $i \in \mathcal{U}^{(t-1)}$ **do** Sample $S_i^{(t)} \overset{iid}{\sim} P_i$ of size $\left|S_i^{(t)}\right| = T_{\mathsf{T}}\left(\eta_i, \epsilon, \frac{\delta}{2kT}\right)$

2 $\mathcal{U}^{(t)} \leftarrow \left\{i \in \mathcal{U}^{(t-1)} : \widehat{\operatorname{err}}\left(f^{(t)}; S_i^{(t)}\right) > \epsilon/2 + \eta_i\right\}$

Theorem 3 ((MDL-RCN) upper bound) *Let* $T = \lceil \log_2 k \rceil$, $\epsilon' = \frac{\epsilon}{16}$, $\delta' = \frac{\delta}{2T}$, *and* Test = RCN-Test. *Additionally, define the condition* $\mathrm{Cond} = \{T_{joint} > T_{sep}\}$, *where*

$$T_{\mathsf{joint}} := \log^2\left(\frac{k}{\delta}\right)\left(\frac{d\log\left(1/\epsilon\right)}{\epsilon\left(1-2\eta\right)} + \sum_{i=1}^{k}\frac{\epsilon+\eta_i}{\epsilon^2}\right) \quad \textit{and} \quad T_{\mathsf{sep}} := \sum_{i=1}^{k}\frac{d\log\left(1/\epsilon\right)+\log\left(k/\delta\right)}{\epsilon\left(1-2\eta_i\right)}$$

Then, Algorithm 1 solves (MDL-RCN) *with sample complexity* $T_{\mathsf{RCN}}\left(\eta_{1:k}, \epsilon, \delta\right) = O\left(\min\left\{T_{\mathsf{joint}}, T_{\mathsf{sep}}\right\}\right)$.

In the special case where the noise upper bound is uniform across distributions—that is, $\eta_i = \eta$ for all $i \in [k]$—the sample complexity bound from Theorem 3 simplifies (ignoring logarithmic factors) to $\tilde{O}\left(\min\left\{\frac{d}{\epsilon(1-2\eta)} + \frac{k(\epsilon+\eta)}{\epsilon^2}, \frac{dk}{\epsilon(1-2\eta)}\right\}\right)$.

2.4. (MDL-MM) Upper Bound

When our goal is to compete with the *unknown* maximum noise rate $\eta^* = \max_{i\in[k]} \eta_i^*$, the previous testing strategy is inapplicable because we lack the explicit thresholds required for RCN-Test.

Fix a round $t \in [T]$. The statistical learning component guarantees that with probability at least $1 - \delta'$, at least half of the distributions $i \in \mathcal{U}^{(t-1)}$ satisfy $\operatorname{err}\left(f^{(t)}; P_i\right) \leq 2\epsilon' + \eta_i^* \leq 2\epsilon' + \eta^*$. If η^* was known, we could simply set $\epsilon' = \epsilon/16$ and invoke RCN-Test using the uniform threshold $\epsilon/2 + \eta^*$ for all $i \in [k]$. Since η^* is unknown, we proceed by guessing its value in small increments.

Let ν denote a candidate guess for η^*. We set $\epsilon' = \epsilon/32$, ensuring that with probability at least $1 - \delta'$, half of the active distributions satisfy $\operatorname{err}\left(f^{(t)}; P_i\right) \leq \epsilon/16 + \eta^*$. Lemma 2 guarantees that by drawing a

sample $S_i^{(t)} \overset{iid}{\sim} P_i$ of size $\left|S_i^{(t)}\right| = T_{\mathsf{T}}(\nu, \epsilon/2, \delta'')$, the following holds with probability $1 - k\delta''$: for every $i \in [k]$,

$$\operatorname{err}\left(f^{(t)}; P_i\right) \le \frac{\epsilon}{16} + \nu \Longrightarrow \widehat{\operatorname{err}}\left(f^{(t)}; S_i^{(t)}\right) \le \frac{\epsilon}{4} + \nu \Longrightarrow \operatorname{err}\left(f^{(t)}; P_i\right) \le \frac{\epsilon}{2} + \nu$$

Based on this, if we remove all $i \in \mathcal{U}^{(t-1)}$ (i.e., output $f^{(t)}$ for P_i) such that $\widehat{\operatorname{err}}\left(f^{(t)}; S_i^{(t)}\right) \le \epsilon/4 + \nu$, we ensure the following:

- **(Progress)** any distribution with true error $\operatorname{err}\left(f^{(t)}; P_i\right) \le \epsilon/16 + \nu$ will be removed.
- **(Correctness)** any removed distribution satisfies $\operatorname{err}\left(f^{(t)}; P_i\right) \le \epsilon/2 + \nu$.

Starting from $\nu = 0$, we increase the guess in steps of $\epsilon/2$ and apply the test at each step until at least half of the active distributions are eliminated. By Progress, this loop terminates after at most $2\eta^*/\epsilon + 2$ iterations, and by Correctness every eliminated distribution is indeed learned. See Algorithm 3 for details.

Algorithm 3 MM-Test

Input: Decision $f^{(t)}$, distribution set $\mathcal{U}^{(t-1)}$.
Output: $\mathcal{U}^{(t)}$.

$\mathcal{U}^{(t)} \leftarrow \mathcal{U}^{(t-1)}$; $\nu \leftarrow 0$
for $i \in \mathcal{U}^{(t-1)}$ **do** $S_i^{(t)} \leftarrow \emptyset$
while $\left|\mathcal{U}^{(t)}\right| > \left|\mathcal{U}^{(t-1)}\right|/2$ **do**
  for $i \in \mathcal{U}^{(t)}$ **do**
    $S_i^{(t)} \leftarrow S_i^{(t)} \cup \tilde{S}_i^{(t)}$ where $\tilde{S}_i^{(t)} \overset{iid}{\sim} P_i$ is of size $\left|\tilde{S}_i^{(t)}\right| = T_{\mathsf{T}}\left(\nu, \epsilon/2, \frac{\delta}{4(\eta/\epsilon+1)kT}\right) - \left|S_i^{(t)}\right|$
    if $\widehat{\operatorname{err}}\left(f^{(t)}; S_i^{(t)}\right) \le \epsilon/4 + \nu$ **then** Remove i from $\mathcal{U}^{(t)}$
  end
  $\nu \leftarrow \nu + \epsilon/2$
end

Theorem 4 ((MDL-MM) upper bound) *Let* $T = \lceil \log_2 k \rceil$, $\epsilon' = \frac{\epsilon}{32}$, $\delta' = \frac{\delta}{2T}$, Cond $= \emptyset$ *and* Test $=$ MM-Test. *Then, Algorithm 1 solves* (MDL-MM) *with sample complexity*

$$T_{\mathsf{MM}}\left(\eta^*_{1:k}, \eta_{1:k}, \epsilon, \delta\right) = O\left(\log^2\left(\frac{k\left(\eta/\epsilon + 1\right)}{\delta}\right)\left(\frac{d\log\left(1/\epsilon\right)}{\epsilon\left(1-2\eta\right)} + \frac{k\left(\epsilon + \eta^*\right)}{\epsilon^2}\right)\right)$$

Remark 5 (Condition for separate learning) *For* (MDL-MM)*, we do not explicitly enforce a switching condition for separate learning because the bound in Theorem 4 depends on the unknown* η^**. However, if we use the upper bound* η *as a proxy for* η^* *and define* Cond $= \left\{\frac{d}{\epsilon(1-2\eta)} + \frac{k(\epsilon+\eta)}{\epsilon^2} > \sum_{i=1}^k \frac{d}{\epsilon(1-2\eta_i)}\right\}$*, then we achieve a sample complexity of* $\tilde{O}\left(\sum_{i=1}^k \frac{d}{\epsilon(1-2\eta_i)}\right)$ *if* Cond *holds, and* $\tilde{O}\left(\frac{d}{\epsilon(1-2\eta)} + \frac{k(\epsilon+\eta^*)}{\epsilon^2}\right)$ *otherwise.*

3. Structured Hypothesis Testing

As a precursor to establishing the MDL lower bound, we first investigate a related problem. Throughout this section, we fix the noise level to $\eta = 1/4$ to focus our analysis on the other problem parameters.

Let $\mathcal{F}$ be a hypothesis class with VC-dimension d. Let P be a distribution over $\mathcal{X} \times \{0, 1\}$ such that the Bayes classifier f^* belongs to $\mathcal{F}$ and satisfies the noise condition $P\left(f^*(X) \neq Y | X = x\right) = 1/4$ for all

$x \in \mathcal{X}$. This corresponds to the standard RCN setting with fixed, known noise. We are given a query function $f : \mathcal{X} \to \{0,1\}$ (not necessarily in $\mathcal{F}$) and tasked with determining whether f is ϵ-optimal with respect to P. Specifically, the algorithm must draw i.i.d. samples from P and output a decision $\mathsf{D} \in \{\mathsf{YES}, \mathsf{NO}\}$ that satisfies the following conditions with probability at least $1-\delta$:

$$\begin{aligned} &\text{If } \operatorname{err}(f;P) \le 1/4 + \epsilon/12, \text{ output } \mathsf{D} = \mathsf{YES}. \\ &\text{If } \operatorname{err}(f;P) \ge 1/4 + \epsilon, \text{ output } \mathsf{D} = \mathsf{NO}. \end{aligned} \tag{SHT}$$

We refer to this problem as *Structured Hypothesis Testing (SHT)*. We say that an algorithm $\mathcal{A}$ for (SHT) has sample complexity $T_{\mathcal{A}} : (0,1)^2 \to \mathbb{N}$ if, for any valid instance $(\mathcal{F}, P, f, \epsilon, \delta)$, it satisfies the conditions above using at most $T_{\mathcal{A}}(\epsilon, \delta)$ samples.

3.1. (SHT) Upper Bounds

In this section, we present two distinct strategies for solving the (SHT) problem.

Agnostic Testing. Since the goal is to test whether $P(f(X) \neq Y)$ is ϵ-close to $1/4$, we can simply disregard the supervised nature of the data and work directly with the empirical errors $(\mathbb{I}\{f(X_t) \neq Y_t\})_{t=1}^{T}$. The problem then reduces to distinguishing between two Bernoulli biases separated by a gap of $\Theta(\epsilon)$, which requires a sample complexity of $\Theta\left(\log(1/\delta)/\epsilon^2\right)$.

Lemma 6 (Testing via empirical errors) *Let $f : \mathcal{X} \to \{0,1\}$ be a fixed function, and let $S \overset{iid}{\sim} P$ be a sample of size $|S| \ge T_{\mathsf{C}}(\epsilon,\delta) := \frac{72\log(2/\delta)}{\epsilon^2}$. Then, with probability at least $1-\delta$, the following holds:*

$$\operatorname{err}(f;P) \le 1/4 + \epsilon/12 \implies \widehat{\operatorname{err}}(f;S) \le 1/4 + \epsilon/6 \implies \operatorname{err}(f;P) \le 1/4 + \epsilon$$

Learning-Augmented Testing. An alternative approach is to first learn a reference hypothesis $\hat{f}$ that satisfies $\operatorname{err}\left(\hat{f};P\right) \le 1/4 + \epsilon$ with high probability. This can be achieved by performing ERM on a sample of size $O(d\log(1/\delta)/\epsilon)$. With $\hat{f}$ in hand, we can draw an additional sample to efficiently test the quality of a candidate function f. In the following lemma, we establish the validity of this approach for a general noise level η, as this broader result will be instrumental in later sections.

Lemma 7 (From learning to testing) *Let P be a distribution over $\mathcal{X} \times \{0,1\}$ such that for some $f^* : \mathcal{X} \to \{0,1\}$, we have that $P(f^*(X) \neq Y | X = x) \le \eta$ for all $x \in \mathcal{X}$. Let $\eta^* := P(f^*(X) \neq Y)$ denote the optimal error. Suppose we know a reference hypothesis $\hat{f} : \mathcal{X} \to \{0,1\}$ (possibly random) that satisfies*

$$\mathbb{P}\left(\operatorname{err}\left(\hat{f};P\right) \le \eta^* + \frac{\epsilon}{48}\right) \ge 1 - \frac{\delta}{3}$$

Let $f : \mathcal{X} \to \{0,1\}$ be a fixed candidate function, and let $S \overset{iid}{\sim} P$ be a sample drawn independently of $\hat{f}$, of size $|S| \ge T_{\mathsf{L}}(\eta,\epsilon,\delta) := \frac{96}{\epsilon(1-2\eta)}\log\left(\frac{6}{\delta}\right)$. Then, with probability at least $1-\delta$, the following holds:

$$\operatorname{err}(f;P) \le \eta^* + \epsilon/12 \implies \left|\widehat{\operatorname{err}}(f;S) - \widehat{\operatorname{err}}\left(\hat{f};S\right)\right| \le \epsilon/3 \implies \operatorname{err}(f;P) \le \eta^* + \epsilon$$

In the context of (SHT), we apply Lemma 7 with $\eta = \eta^* = 1/4$. This yields a sample size of $T_{\mathsf{L}}\left(\frac{1}{4},\epsilon,\delta\right) = \frac{192}{\epsilon}\log\left(\frac{6}{\delta}\right)$. By combining the agnostic and learning-augmented strategies into a single procedure (Algorithm 4), we obtain the following guarantee.

Theorem 8 ((SHT) upper bound) *Algorithm 4 solves the (SHT) problem with sample complexity $T_{\mathsf{SHT}}(\epsilon,\delta) = O\left(\log(1/\delta)\min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$.*

This bound highlights a natural trade-off: in the low-precision regime where $\epsilon \gg 1/d$, the agnostic testing approach ($1/\epsilon^2$) is superior. However, in the high-precision regime where $\epsilon \ll 1/d$, exploiting the supervised structure of the data via the learning-augmented approach (d/ϵ) yields an improvement.

Algorithm 4 SHT

Input: Function class $\mathcal{F}$, distribution P, function $f : \mathcal{X} \to \{0,1\}$, parameters $(\epsilon, \delta) \in (0,1)^2$.
Output: Decision D.

if $d \geq 1/\epsilon$ **then** ▷ Agnostic Testing.
  Sample $S \overset{iid}{\sim} P$ of size $|S| = T_{\mathsf{C}}(\epsilon, \delta)$
  if $\widehat{\mathrm{err}}(f; S) \leq 1/4 + \epsilon/6$ **then** $\mathsf{D} \leftarrow \mathsf{YES}$ **else** $\mathsf{D} \leftarrow \mathsf{NO}$
else ▷ Learning-Augmented Testing.
  $\hat{f} \leftarrow \mathrm{ERM}_{\mathcal{F}}(S)$ where $S \overset{iid}{\sim} P$ is of size $|S| = T_{\mathsf{SL}}(1/4, \epsilon/48, \delta/3)$
  Sample $S' \overset{iid}{\sim} P$ of size $|S'| = T_{\mathsf{L}}(1/4, \epsilon, \delta)$
  if $\left|\widehat{\mathrm{err}}(f; S') - \widehat{\mathrm{err}}\left(\hat{f}; S'\right)\right| \leq \epsilon/3$ **then** $\mathsf{D} \leftarrow \mathsf{YES}$ **else** $\mathsf{D} \leftarrow \mathsf{NO}$
end

3.2. (SHT) Lower Bound

Recall that under the RCN assumption, ERM requires only $\tilde{O}(d/\epsilon)$ samples, a significant improvement over the $\tilde{O}\left(d/\epsilon^2\right)$ complexity of the fully agnostic setting. It is natural to ask whether the testing problem (SHT) can similarly be solved with $\tilde{O}(1/\epsilon)$ samples. The fast rate for ERM relies on a critical variance bound: for any distribution P satisfying the RCN condition under constant noise, we have that (see Appendix C)

$$\mathrm{Var}_P\left(\mathbb{I}\{f(X) \neq Y\} - \mathbb{I}\{f^*(X) \neq Y\}\right) \lesssim \mathrm{err}(f; P) - \mathrm{err}(f^*; P)$$

That is, we can control the variance when we *offset* by f^*. Bernstein's inequality then ensures that

$$|\mathrm{err}(f; P) - \mathrm{err}(f^*; P) - (\widehat{\mathrm{err}}(f; S) - \widehat{\mathrm{err}}(f^*; S))| \lesssim \frac{1}{|S|} + \mathrm{err}(f; P) - \mathrm{err}(f^*; P)$$

with high probability over a sample $S \overset{iid}{\sim} P$. If we had access to the quantity $|\widehat{\mathrm{err}}(f; S) - \widehat{\mathrm{err}}(f^*; S)|$, we could perform the test efficiently using the reasoning of Lemma 7. In fact, we showed that having access to an ϵ-optimal hypothesis (a proxy for f^*) allows us to test with just $\tilde{O}(1/\epsilon)$ additional samples.

However, a fundamental obstacle remains: we do not know f^*. Although we know its population error is $\mathrm{err}(f^*; P) = 1/4$, we cannot readily use its empirical error $\widehat{\mathrm{err}}(f^*; S)$ as a reference point because the concentration of $|\mathrm{err}(f^*; P) - \widehat{\mathrm{err}}(f^*; S)|$ is slow, scaling with $\tilde{O}\left(1/\sqrt{|S|}\right)$. Notably, the ERM sample size of $\tilde{O}(d/\epsilon)$ ensures uniform convergence of the excess risks (the offset counterparts), but not of the absolute errors $|\mathrm{err}(f; P) - \widehat{\mathrm{err}}(f; S)|$.

In the following result, we prove that the trade-off achieved by Algorithm 4 is optimal. Specifically, we show that the sample complexity is lower-bounded by the minimum of the agnostic rate and the learning-augmented rate, yielding a tight rate of $\tilde{\Theta}\left(\min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$.

Theorem 9 ((SHT) lower bound) *Let $\delta \leq 1/4$. Any algorithm $\mathcal{A}$ that solves (SHT) requires a sample size of $T_{\mathcal{A}}(\epsilon, \delta) \geq \frac{3}{16}\log(1 + \log 2)\min\left\{1/\epsilon^2, d/\epsilon\right\}$.*

Proof sketch of Theorem 9 We will show the bound $\Omega(d/\epsilon)$ for $d \leq \frac{1}{2\epsilon}$. When $d > \frac{1}{2\epsilon}$, we can simply choose a shattered set of size $\frac{1}{2\epsilon}$, so that the lower bound $\Omega\left(1/\epsilon^2\right)$ holds.

Hence, assume that $d \leq \frac{1}{2\epsilon}$ and suppose that X is supported on $\mathcal{X} = [\![d^2]\!]$. Let $p_X(0) = 1 - 2d\epsilon$ and $p_X(x) = 2\epsilon/d$ for each $x \in \left[\!\left[d^2\right]\!\right]$. Consider the hypotheses (H_0) $f^* = f_0$, where $f_0 := x \mapsto 0$ is the constant zero function, and (H_1) f^* equals 1 precisely on a random subset $R \subset \left[\!\left[d^2\right]\!\right]$ of size $|R| = d$. Note that we never need to shatter a set of size larger than d. Under H_1, $\mathbb{P}(Y = 1) = (1 - 2d\epsilon) \cdot \frac{1}{4} + \left(d^2 - d\right) \cdot \frac{2\epsilon}{d} \cdot \frac{1}{4} +$

$d \cdot \frac{2\epsilon}{d} \cdot \frac{3}{4} = \frac{1}{4} + \epsilon$. In particular, f_0 has errors $1/4$ under H_0 and $1/4 + \epsilon$ under H_1, so that an (SHT) learner should be able to distinguish both hypotheses. Using Ingster's method, we can show that testing H_0 vs. H_1 under a uniform mixture over subsets R requires $\Omega(d/\epsilon)$ samples. ■

3.3. Multi-Distribution SHT

We now extend the (SHT) framework to the multi-distribution setting. Let $P_1, \ldots, P_k$ be distributions over $\mathcal{X} \times \{0,1\}$ that all satisfy the RCN condition $P_i(f^*(X) \neq Y | X = x) = 1/4$, for all $x \in \mathcal{X}$, with respect to a *shared* unknown $f^* \in \mathcal{F}$. We are given candidate functions $f_1, \ldots, f_k : \mathcal{X} \to \{0,1\}$ (not necessarily in $\mathcal{F}$), and our objective is to determine whether $\mathrm{err}(f_i; P_i)$ is ϵ-optimal for every $i \in [k]$. More precisely, the algorithm draws a total of T samples from the distributions and outputs a decision vector $(\mathsf{D}_1, \ldots, \mathsf{D}_k) \in \{\mathsf{YES}, \mathsf{NO}\}^k$. We require that for each $i \in [k]$, with probability at least $1 - \delta$,

$$\begin{aligned} &\text{If } \mathrm{err}(f_i; P_i) \leq 1/4 + \epsilon/12, \text{ output } \mathsf{D}_i = \mathsf{YES}. \\ &\text{If } \mathrm{err}(f_i; P_i) \geq 1/4 + \epsilon, \text{ output } \mathsf{D}_i = \mathsf{NO}. \end{aligned} \tag{MSHT}$$

We say that an algorithm $\mathcal{A}$ for (MSHT) has sample complexity $T_{\mathcal{A}} : (0,1)^2 \to \mathbb{N}$ if it satisfies this guarantee for any valid instance $(\mathcal{F}, P_{1:k}, f_{1:k}, \epsilon, \delta)$ using at most $T_{\mathcal{A}}(\epsilon, \delta)$ total samples.

Note that we only require each D_i correct w.p. $\geq 1 - \delta$. This is to avoid a $\log k$ factor in the reduction to (MDL-RCN) (see Lemma 12); simultaneous correctness can be ensured via a union bound. To solve the (MSHT) problem, we apply the strategy of Algorithm 4 to each distribution individually. For completeness, this is detailed in Algorithm 6, and its sample complexity is established in Theorem 10.

Theorem 10 ((MSHT) upper bound) *Algorithm 6 solves the* (MSHT) *problem with sample complexity* $T_{\mathsf{MSHT}}(\epsilon, \delta) = O\left(k \log(1/\delta) \min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$.

Testing each distribution separately results in a k-fold increase in sample complexity relative to (SHT). We now show that this linear scaling is unavoidable in the worst case–any (MSHT) solver must essentially perform independent tests–so the tight sample complexity is $\tilde{\Theta}\left(k \cdot \min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$.

Theorem 11 (Multi-Distribution SHT lower bound) *Let $\delta \leq 0.01$ and $d \geq 8\epsilon$. Any algorithm $\mathcal{A}$ that solves* (MSHT) *requires a sample size of $T_{\mathcal{A}}(\epsilon, \delta) \geq \frac{0.015k}{4} \min\left\{1/\epsilon^2, d/\epsilon\right\}$.*

Proof sketch of Theorem 11 Fix $\epsilon \in (0,1)$. Again, we focus on the setting where $d \leq \frac{1}{2\epsilon}$ and aim to show the lower bound $\Omega(kd/\epsilon)$. This immediately implies the $\Omega\left(k/\epsilon^2\right)$ bound when $d > \frac{1}{2\epsilon}$.

In essence, we will construct distributions whose X-marginals will be supported on consecutive disjoint sets of size $d^2 + 1$, each with the same weights as in the (SHT) lower bound. We then consider the problem of testing (H_0) $f^* = f_0$ vs. (H_1) f^* equals 1 precisely on a size-d subset on the support of one of the distributions. Note that this only requires shattering a set of size d. We then show that the learner must sample $\Omega(d/\epsilon)$ times from each distribution by constructing an (SHT) algorithm that simulates the (MSHT) one. ■

4. MDL Lower Bound

As in Section 3, we establish the lower bound under RCN noise $\eta_i = \eta_i^* = 1/4$ for all $i \in [k]$, in which case all MDL variants coincide. The key observation is that (MSHT) is *necessary* for MDL: any MDL algorithm can be used as a black box to obtain proxies for the optimal classifier and thereby solve (MSHT) with only $O(\log(1/\delta)/\epsilon)$ additional samples per distribution. We formalize this in Algorithm 5 and analyze it below.

Lemma 12 (MDL to MSHT upper bound) *Algorithm 5, under an MDL algorithm $\mathcal{A}$ with sample complexity $T_{\mathcal{A}}$, solves the* (MSHT) *problem with sample complexity $T_{\mathcal{A}\to\mathsf{MSHT}}(\epsilon, \delta) = T_{\mathcal{A}}(\epsilon/48, \delta/3) + \frac{192k \log(6/\delta)}{\epsilon}$.*

Algorithm 5 MDL to MSHT

Input: Function class $\mathcal{F}$, distributions $P_1, \dots, P_k$, functions $f_1, \dots, f_k : \mathcal{X} \to \{0,1\}$, parameters $(\epsilon, \delta) \in (0,1)^2$, MDL algorithm $\mathcal{A}$ with sample complexity $T_{\mathcal{A}}$.

Output: Decision vector $(\mathsf{D}_1, \dots, \mathsf{D}_k)$.

$\left(\hat{f}_1, \dots, \hat{f}_k\right) \leftarrow \mathcal{A}$ where we run $\mathcal{A}$ for $T_{\mathcal{A}}(\epsilon/48, \delta/3)$ rounds.

for $i = 1, \dots, k$ **do**

  Sample $S_i \overset{iid}{\sim} P_i$ of size $|S_i| = T_{\mathsf{L}}(1/4, \epsilon, \delta)$

  if $\left|\widehat{\mathrm{err}}(f_i; S_i) - \widehat{\mathrm{err}}\left(\hat{f}_i; S_i\right)\right| \le \epsilon/3$ **then** $\mathsf{D}_i \leftarrow \mathsf{YES}$ **else** $\mathsf{D}_i \leftarrow \mathsf{NO}$

end

Recall that our proposed strategies for (MDL-RCN) and (MDL-MM) achieve a sample complexity of $\tilde{O}\left(\min\left\{dk/\epsilon, d/\epsilon + k/\epsilon^2\right\}\right)$. We now show that this rate is optimal, up to logarithmic factors.

Theorem 13 (MDL lower bound) *Let $\delta \le 0.01/3$, $d \ge 384\epsilon$ and $\min\{d, 1/\epsilon\} \ge 4 \cdot 10^7$. Any MDL algorithm $\mathcal{A}$ requires a sample size of $T_{\mathcal{A}}(\epsilon, \delta) = \Omega\left(d/\epsilon + k \cdot \min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$.*

5. (MDL-Mass) Lower Bound

We now study the hardest variant, (MDL-Mass). The lower bound from Section 4, $\Omega\left(d/\epsilon + k \cdot \min\left\{1/\epsilon^2, d/\epsilon\right\}\right)$, becomes $\Omega(d/\epsilon + k/\epsilon^2)$ when $d \gtrsim 1/\epsilon$, losing a multiplicative dependence on d. We show that (MDL-Mass) admits a stronger lower bound with a $k\sqrt{d}$ factor (even for large d) via the same reduction template: define the auxiliary test (SHT-Mass) and argue that any (MDL-Mass) solver must solve it for each distribution.

5.1. Massart Testing

We begin by defining an auxiliary testing problem. Consider a pair of random variables $(X, Y) \in [\![d]\!] \times \{0,1\}$, where the PMF of X is given by $p_X = (1 - \epsilon, \epsilon/d, \dots, \epsilon/d)$ and the conditional distributions are $Y|X = x \sim \mathrm{Ber}(q_x)$. The vector of biases $\mathbf{q} = (q_x)_{x=0}^{d}$ is *unknown*, but constrained such that $q_0 \in [0, 0.49]$ and $q_x \in [0, 0.49] \cup [0.51, 1]$ for all $x \in [d]$. Let

$$\Delta_{\mathbf{q}} := \frac{\epsilon}{d} \sum_{x \in [\![d]\!] : q_x \ge 1/2} (2q_x - 1)$$

The goal is to draw i.i.d. (X, Y) pairs and determine which of the following two hypotheses holds:

$$\begin{aligned} &H_0\text{: If } q_x = 0.465 \text{ for all } x \in [\![d]\!], \text{ output } \mathsf{YES} \text{ with probability at least } 1 - \delta. \\ &H_1\text{: If } \Delta_{\mathbf{q}} \ge 0.3\epsilon, \text{ output } \mathsf{NO} \text{ with probability at least } 1 - \delta. \end{aligned} \tag{SHT-Mass}$$

In the analysis below, we demonstrate that solving (SHT-Mass) requires a polynomial dependence on d, even when $d \gg 1/\epsilon$, standing in sharp contrast to the behavior of (SHT). We emphasize that the moment matching argument requires varying the noise levels, preventing us from fixing the noise as in the (SHT) setup. Consequently, this construction is specific to (MDL-Mass) and does not extend to the other variants.

Theorem 14 (SHT-Mass lower bound) *Let $\delta \le 1/4$ and $d \ge 1750$. Any algorithm $\mathcal{A}$ that solves (SHT-Mass) requires a sample size of $T_{\mathcal{A}}(\epsilon, \delta) \ge \frac{\sqrt{d}}{2\sqrt{2}\epsilon}$.*

Proof sketch of Theorem 14 Our ultimate goal is to construct bias vector $\mathbf{q}_h$, for each $h \in \{0,1\}$, that falls under hypothesis H_h and such that $\mathbf{q}_0$ and $\mathbf{q}_1$ are difficult to distinguish. To start, we set the 0th coordinate of each to 0.465, so that the only informative samples are those in which $X \in [d]$. We then construct priors μ_h and sample the remaining biases i.i.d. from μ_h in a way that the resulting $\mathbf{q}_h$ is in H_h with high probability. More importantly, we also make sure that the first moments of μ_0 and μ_1 match, which ensures that Y has the same mean under both hypotheses. Since the marginal p_X is independent of the hypothesis, it is thus necessary to exploit the relationship between X and Y. This can only be accomplished once we observe repeated values of X in $[d]$, which takes $\Omega\left(\sqrt{d}\right)$ samples. The $\Omega\left(\sqrt{d}/\epsilon\right)$ bound then follows from the fact that it takes $O\left(1/\epsilon\right)$ samples to obtain a single point in $[d]$. ■

Limitations of the (SHT-Mass) construction. Higher-order moment matching and related polynomial techniques could improve the $\sqrt{d}$ bound (Wu and Yang, 2020; Canonne, 2022), but this particular construction cannot yield linear-in-d hardness. A constant $\Delta_{\mathbf{q}}$ gap forces a constant fraction of heavy coordinates under H_1, so a tester can reject H_0 by locating a single heavy coordinate; because the light/heavy bias gap is constant, this can be done with sublinear-in-d samples. Although other hard instances might achieve linear dependence, translating them into an MDL lower bound via a similar reduction is delicate, since we must preserve a fixed VC dimension across multiple distributions (here enforced by our choice of H_0).

5.2. Lower Bound

Next, we construct an MDL instance that inherits the (SHT-Mass) difficulty. Let the covariate space be $\mathcal{X} \coloneqq [\![kd + k - 1]\!]$, which we partition into k disjoint blocks of size $d+1$. We denote the ith block by $\mathcal{X}_i \coloneqq \{(i-1)(d+1) + x : x \in [\![d]\!]\}$ for each $i \in [k]$. We define the hypothesis class $\mathcal{F}$ to be the set of binary functions consisting of the zero function f_0 and all functions that equal 1 precisely on a subset of $\mathcal{X}_i \setminus \{(i-1)(d+1)\}$ (i.e., we exclude the first coordinate) for some $i \in [k]$. Note that $\mathrm{VCdim}(\mathcal{F}) = d$.

Next, we define distributions $P_1, \ldots, P_k$ over $\mathcal{X} \times \{0,1\}$ where the marginal PMF of X under P_i is given by $(1-\epsilon, \epsilon/d, \ldots, \epsilon/d)$ on $\mathcal{X}_i$ and is zero elsewhere. Furthermore, we require the label noise to satisfy the Massart constraint with respect to some target $f^* \in \mathcal{F}$; namely, $\eta_i^*(x) \coloneqq P_i(f^*(X) \neq Y | X = x) \leq 0.49$ for all $x \in \mathcal{X}$ and $i \in [k]$. The next theorem shows that any (MDL-Mass) algorithm in this setup must incur an additional $\Omega\left(k\sqrt{d}/\epsilon\right)$ cost; together with Theorem 13, this yields the combined lower bound $\Omega\left(d/\epsilon + k \cdot \min\left\{1/\epsilon^2, d/\epsilon\right\} + k\sqrt{d}/\epsilon\right)$ under $\eta_i \leq 0.49$ for all $i \in [k]$.

Theorem 15 ((MDL-Mass) lower bound) *Let $\delta \leq 0.1/3$ and $d \geq 7 \cdot 10^{10}$. Any algorithm $\mathcal{A}$ that solves (MDL-Mass) in this setup requires a sample size of $T_{\mathcal{A}}(\epsilon, \delta) = \Omega\left(k\sqrt{d}/\epsilon\right)$.*

6. Discussion

We study multi-distribution learning under bounded label noise through three benchmarks (known noise bounds, minimax error, and unknown optimal errors): for the first two we give near-optimal upper bounds $\tilde{O}\left(\frac{d}{\epsilon(1-2\eta)} + \sum_{i=1}^{k} \frac{\epsilon+\eta_i}{\epsilon^2}\right)$ and $\tilde{O}\left(\frac{d}{\epsilon(1-2\eta)} + \frac{k(\epsilon+\eta^*)}{\epsilon^2}\right)$ and matching lower bounds via a reduction to multi-distribution testing (ruling out $\tilde{O}(\frac{d+k}{\epsilon})$ even at constant noise), while for the hardest objective we prove a stronger lower bound $\Omega(k\sqrt{d}/\epsilon)$ by moment matching, revealing a multiplicative penalty and a separation between RCN and Massart noise.

Future Directions. Our work leaves several compelling avenues for future research: (i) *Centralized MDL:* do comparable rates hold when the learner must output a single hypothesis $\hat{f}$ that performs well on all distributions (Haghtalab et al., 2022)? (ii) *Noise-dependent lower bounds:* our lower bounds are proved in a

constant-noise regime, and it remains open to characterize the optimal complexity as an explicit function of the noise rate. (iii) *Closing the gap for* (MDL-Mass)*:* a tight characterization remains unknown; resolving the precise dependence on d and k likely requires new lower-bound constructions (see Section 5).

References

Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 53–65. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/abernethy22a.html.

Dana Angluin and Philip Laird. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, April 1988.

Pranjal Awasthi, Nika Haghtalab, and Eric Zhao. Open problem: The sample complexity of multi-distribution learning for VC classes. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5943–5949. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/awasthi23a.html.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/186a157b2992e7daed3677ce8e9fe40f-Paper.pdf.

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

Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, London, England, February 2016.

Tom Bylander. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the Seventh Annual Conference on Computational Learning Theory*, COLT '94, page 340–347, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916557. doi: 10.1145/180139.181176. URL https://doi.org/10.1145/180139.181176.

Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends® in Communications and Information Theory*, 19(6):1032–1198, 2022. ISSN 1567-2190. doi: 10.1561/0100000114. URL http://dx.doi.org/10.1561/0100000114.

Yair Carmon and Danielle Hausler. Distributionally robust optimization via ball oracle acceleration. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, page 35866–35879. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e90b00adc3ba130eb2510d93ba3ff250-Paper-Conference.pdf.

Jiecao Chen, Qin Zhang, and Yuan Zhou. Tight bounds for collaborative PAC learning via multiplicative weights. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/ed519dacc89b2bead3f453b0b05a4a8b-Paper.pdf.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Classification under misspecification: Halfspaces, generalized linear models, and evolvability. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8391–8403. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5f8b73c0d4b1bf60dd7173b660b87c29-Paper.pdf.

Yuyang Deng and Mingda Qiao. Collaborative learning with different labeling functions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10530–10552. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/deng24d.html.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/358aee4cc897452c00244351e4d91f69-Paper.pdf.

Cynthia Dwork, Lunjia Hu, and Han Shao. How many domains suffice for domain generalization? a tight characterization via the domain shattering dimension. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=inN0WrBJVc.

Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 406–419. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/02917acec264a52a729b99d9bc857909-Paper-Conference.pdf.

Nika Haghtalab, Omar Montasser, and Mingda Qiao. Sample-adaptivity tradeoff in on-demand sampling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=gaHjGx1cMh.

Rafael Hanashiro and Patrick Jaillet. Distribution-dependent rates for multi-distribution learning. In *37th International Conference on Algorithmic Learning Theory*, 2025. URL https://openreview.net/forum?id=lj5jvHZLuq.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. ISSN 0890-5401. doi: https://doi.org/10.1016/0890-5401(92)90010-D. URL https://www.sciencedirect.com/science/article/pii/089054019290010D.

Yu I Ingster and I A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics. Springer, New York, NY, 2003 edition, October 2002.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22 (4):807–837, 1993. doi: 10.1137/0222052. URL https://doi.org/10.1137/0222052.

Kasper Green Larsen, Omar Montasser, and Nikita Zhivotovskiy. Derandomizing multi-distribution learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 94246–94264. Curran Associates, Inc., 2024. doi: 10.52202/079017-2989. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ab63d1eb181e920273504411fe0942dc-Paper-Conference.pdf.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *Ann. Stat.*, 27(6):1808–1829, December 1999.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing*. Cambridge University Press, jan 31 2005. ISBN 9780521835404. URL https://books.google.com/books/about/Probability_and_Computing.html?hl=&id=0bAYl6d7hvkC.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/mohri19a.html.

Huy Nguyen and Lydia Zakynthinou. Improved algorithms for collaborative PAC learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3569df159ec477451530c4455b2a9e86-Paper.pdf.

Quan M. Nguyen, Nishant A Mehta, and Cristóbal A Guzmán. Beyond minimax rates in group distributionally robust optimization via a novel notion of sparsity. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6kGTxbn4Qf.

Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust language modeling. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1432. URL https://aclanthology.org/D19-1432/.

Binghui Peng. The sample complexity of multi-distribution learning. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4185–4204. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/peng24b.html.

Yury Polyanskiy and Yihong Wu. *Information theory*. Cambridge University Press, Cambridge, England, January 2025.

Nicholas Rittler and Kamalika Chaudhuri. Agnostic multi-group active learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 1100–1118. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/03b1043052700b1a471996b0baf309d4-Paper-Conference.pdf.

Guy N Rothblum and Gal Yona. Multi-group agnostic pac learnability. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9107–9115. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/rothblum21a.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning*. Cambridge University Press, Cambridge, England, May 2014.

Tasuku Soma, Khashayar Gatmiry, Sharut Gupta, and Stefanie Jegelka. Near-optimal algorithms for group distributionally robust optimization and beyond, 2025. URL https://arxiv.org/abs/2212.13669.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Yihong Wu and Pengkun Yang. Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends® in Communications and Information Theory*, 17(4):402–586, 2020. ISSN 1567-2190. doi: 10.1561/0100000095. URL http://dx.doi.org/10.1561/0100000095.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lXuByUeHhd.

Dingzhi Yu, Yunuo Cai, Wei Jiang, and Lijun Zhang. Efficient algorithms for empirical group distributionally robust optimization and beyond. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57384–57414. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24a.html.

Chengbo Zang, Mehmet Kerem Turkcan, Gil Zussman, Zoran Kostic, and Javad Ghaderi. Adaptive data collection for robust learning across multiple distributions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=n8fJdB6GMZ.

Chicheng Zhang and Yihan Zhou. Towards fundamental limits for active multi-distribution learning. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 6041–6090. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/zhang25c.html.

Lijun Zhang, Haomin Bai, Peng Zhao, Tianbao Yang, and Zhi-Hua Zhou. Stochastic approximation approaches to group distributionally robust optimization and beyond, 2024a. URL https://arxiv.org/abs/2302.09267.

Zihan Zhang, Wenhao Zhan, Yuxin Chen, Simon S Du, and Jason D Lee. Optimal multi-distribution learning. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 5220–5223. PMLR, 30 Jun–03 Jul 2024b. URL https://proceedings.mlr.press/v247/zhang24b.html.

Appendix A. Related Work

Multi-Distribution Learning. Multi-distribution learning was introduced by (Blum et al., 2017) under the name *collaborative PAC learning*. In the realizable case, they proved a tight sample complexity of $\tilde{\Theta}\left(\frac{d+k}{\epsilon}\right)$ under both personalized and centralized settings.

In the centralized realizable setting, (Nguyen and Zakynthinou, 2018; Chen et al., 2018) improved the logarithmic dependence, and (Nguyen and Zakynthinou, 2018) also gave strategies that compete with a constant multiple of the minimax error.

For centralized agnostic MDL, (Haghtalab et al., 2022) proposed a general algorithmic framework based on playing online learning strategies against each other to reach an approximate Nash equilibrium. They obtained a tight rate of $\tilde{\Theta}\left(\frac{\log|\mathcal{F}|+k}{\epsilon^2}\right)$ for finite classes and, for hypothesis classes of VC dimension d, an upper bound of $\tilde{O}\left(\frac{dk}{\epsilon}+\frac{d+k}{\epsilon^2}\right)$; for the VC case they also proved a lower bound of $\tilde{\Omega}\left(\frac{d+k}{\epsilon^2}\right)$.

Subsequently, (Awasthi et al., 2023) asked whether this VC lower bound is tight for centralized agnostic MDL. They also provided an algorithm for personalized agnostic MDL with complexity $\tilde{O}\left(\frac{d+k}{\epsilon^2}\right)$ and an algorithm for centralized agnostic MDL with complexity $\tilde{O}\left(\frac{d}{\epsilon^4}+\frac{k}{\epsilon^2}\right)$. This question was later answered affirmatively by (Zhang et al., 2024b; Peng, 2024), who gave algorithms for centralized agnostic MDL with sample complexity $\tilde{O}\left(\frac{d+k}{\epsilon^2}\right)$, matching the lower bound up to logarithmic factors.

Closest to our setting, (Deng and Qiao, 2024) studied a weaker form of realizability where a small subset of classifiers achieves small error: they assume there exist $f_1^*,\dots,f_m^* \in \mathcal{F}$ such that $\max_{i\in[k]}\min_{j\in[m]}\operatorname{err}\left(f_j^*;P_i\right) \le \epsilon$. They gave an algorithm with sample complexity $\tilde{O}\left(\frac{md}{\epsilon}+\frac{k}{\epsilon}\right)$ and showed it is tight up to log factors. In particular, when different distributions are realized by different classifiers, a learner must in the worst case learn each distribution separately; in contrast, we assume a *shared* Bayes classifier, a structural distinction that enables information sharing not possible in the setting of (Deng and Qiao, 2024).

Related Learning Settings. The MDL framework has also been studied under a variety of structural assumptions—including convexity (Sagawa et al., 2020; Abernethy et al., 2022; Haghtalab et al., 2022; Soma et al., 2025; Zhang et al., 2024a; Yu et al., 2024; Carmon and Hausler, 2022; Zang et al., 2025), sparsity (Nguyen et al., 2025), and bounded suboptimality (Hanashiro and Jaillet, 2025)—as well as through alternative lenses such as label complexity (Zhang and Zhou, 2025; Rittler and Chaudhuri, 2023) and the role of adaptivity in sample complexity (Haghtalab et al., 2025). Beyond these theoretical directions, distributionally robust objectives have also been adopted in language models to improve performance across diverse data sources (Oren et al., 2019; Xie et al., 2023).

On the computational side, a complementary line of work studies the cost of making statistically optimal procedures efficient: since the optimal strategies are inherently randomized, (Larsen et al., 2024) showed that derandomizing them is computationally hard. Related notions of learning across heterogeneous groups have also been investigated in other learning frameworks (Mohri et al., 2019; Dwork et al., 2025; Rothblum and Yona, 2021).

PAC Learning. Statistical learning theory has long been developed through the lens of PAC learning (Valiant, 1984; Haussler, 1992). Classical formulations focus either on the realizable case (zero noise) or on the fully agnostic case (no assumptions on the data-generating process). To model intermediate regimes, subsequent works (Angluin and Laird, 1988; Kearns and Li, 1993; Mammen and Tsybakov, 1999) introduced explicit assumptions on the label noise, which can often lead to faster learning rates; see (Boucheron et al., 2005) for background.

In this work, we focus on the RCN and Massart noise models. From a statistical perspective, the distinction between the two is minor; both admit fast rates. Computationally, however, the Massart generalization proves substantially more challenging. A polynomial-time algorithm for RCN was given by (Bylander, 1994), whereas a distribution-independent equivalent for Massart noise remained elusive until much later (Di-

akonikolas et al., 2019). Furthermore, current polynomial-time approaches generally compete only with the noise upper bound, and (Chen et al., 2020) established a super-polynomial statistical query SQ lower bound for competing with the Bayes error.

Appendix B. Concentration Inequalities

We record a few standard concentration inequalities that will be used throughout. For background, we refer to (Boucheron et al., 2016).

Lemma 16 (Concentration inequalities) *Let $Z_1, \ldots, Z_T$ be i.i.d. random variables and define the empirical mean $\bar{Z} \coloneqq \frac{1}{T}\sum_{t=1}^{T} Z_t$.*

- *(Hoeffding) If $Z_1 \in [a, b]$, then*

$$\mathbb{P}\left(\left|\bar{Z} - \mathbb{E}\left[Z_1\right]\right| \leq (b-a)\sqrt{\frac{1}{2T}\log\left(\frac{2}{\delta}\right)}\right) \geq 1 - \delta$$

- *(Bernstein) If $\left|Z_1 - \mathbb{E}\left[Z_1\right]\right| \leq b$, then*

$$\mathbb{P}\left(\left|\bar{Z} - \mathbb{E}\left[Z_1\right]\right| \leq \frac{2b}{3T}\log\left(\frac{2}{\delta}\right) + \sqrt{\frac{2\operatorname{Var}\left(Z_1\right)}{T}\log\left(\frac{2}{\delta}\right)}\right) \geq 1 - \delta$$

Appendix C. Learning under Massart Noise

Let P be a distribution over $(X, Y) \in \mathcal{X} \times \{0, 1\}$ such that

$$\eta^*(x) \coloneqq P\left(f^*(X) \neq Y | X = x\right) \leq \eta < \frac{1}{2} \quad \forall x \in \mathcal{X}$$

for some $f^* : \mathcal{X} \to \{0, 1\}$. This is the Massart noise condition with rate η. We denote the Bayes error by $\eta^* \coloneqq P\left(f^*(X) \neq Y\right)$. The following result characterizes the pointwise error of a classifier in terms of f^*.

Lemma 17 *For any $f : \mathcal{X} \to \{0, 1\}$,*

$$P\left(f(X) \neq Y | X = x\right) - \eta^*(x) = \left(1 - 2\eta^*(x)\right)\mathbb{I}\left\{f(x) \neq f^*(x)\right\}$$

Proof of Lemma 17 This follows from

$$\begin{aligned} P\left(f(X) \neq Y | X = x\right) &= \eta^*(x)\,\mathbb{I}\left\{f(x) = f^*(x)\right\} + \left(1 - \eta^*(x)\right)\mathbb{I}\left\{f(x) \neq f^*(x)\right\} \\ &= \eta^*(x)\left(1 - \mathbb{I}\left\{f(x) \neq f^*(x)\right\}\right) + \left(1 - \eta^*(x)\right)\mathbb{I}\left\{f(x) \neq f^*(x)\right\} \\ &= \eta^*(x) + \left(1 - 2\eta^*(x)\right)\mathbb{I}\left\{f(x) \neq f^*(x)\right\} \end{aligned}$$

■

The bounded noise assumption implies that $1 - 2\eta^*(x) \geq 1 - 2\eta$ for any $x \in \mathcal{X}$. Then, if we integrate the equality of Lemma 17 with respect to marginal P_X and rearrange, we obtain

$$P\left(f(X) \neq f^*(X)\right) \leq \frac{P\left(f(X) \neq Y\right) - \eta^*}{1 - 2\eta}$$

The key idea behind the Massart fast rate is that we can bound the following variance:

$$\begin{aligned}\operatorname{Var}_P\left(\mathbb{I}\left\{f\left(X\right)\neq Y\right\}-\mathbb{I}\left\{f^*\left(X\right)\neq Y\right\}\right) &\leq \mathbb{E}_P\left[\left|\mathbb{I}\left\{f\left(X\right)\neq Y\right\}-\mathbb{I}\left\{f^*\left(X\right)\neq Y\right\}\right|\right]\\ &= P\left(f\left(X\right)\neq f^*\left(X\right)\right)\\ &\leq \frac{P\left(f\left(X\right)\neq Y\right)-\eta^*}{1-2\eta}\end{aligned}$$

Suppose that we are trying to learn a finite function class $\mathcal{F}$ such that $f^*\in\mathcal{F}$. Let $\left(X_t,Y_t\right)_{t=1}^T \overset{iid}{\sim} P$ be a sample and consider the empirical minimizer

$$\hat{f}\in\operatorname*{argmin}_{f\in\mathcal{F}}\left\{\frac{1}{T}\sum_{t=1}^T\mathbb{I}\left\{f\left(X_t\right)\neq Y_t\right\}\right\}$$

Applying Bernstein's inequality on random variables $\mathbb{I}\left\{f\left(X_t\right)\neq Y_t\right\}-\mathbb{I}\left\{f^*\left(X_t\right)\neq Y_t\right\}$ and taking a union bound over $\mathcal{F}$, we can conclude that

$$P\left(\hat{f}\left(X\right)\neq Y\right)-\eta^*\lesssim\frac{\log\left(\left|\mathcal{F}\right|/\delta\right)}{n\left(1-2\eta\right)}$$

In other words, $\hat{f}$ is ϵ-optimal provided that $n\gtrsim\frac{\log(|\mathcal{F}|/\delta)}{\epsilon(1-2\eta)}$. To extend this idea to general VC classes, we refer to (Boucheron et al., 2005).

Appendix D. Proofs of Section 2

D.1. Proof of Lemma 1

Note that the mixture also satisfies the bounded noise assumption:

$$\bar{P}_{\mathcal{U}}\left(f^*\left(X\right)\neq Y|X=x\right)=\sum_{i\in\mathcal{U}}\alpha_i\left(x\right)P_i\left(f^*\left(X\right)\neq Y|X=x\right)\leq\eta$$

where $\alpha_i\coloneqq\frac{p_i^X}{\sum_j p_j^X}$ and p_i^X is the density of the X-marginal of P_i with respect to some dominating measure. Then, the ERM solution $\hat{f}$ on $T_{\mathsf{SL}}\left(\eta,\epsilon,\delta\right)$ samples from $\bar{P}_{\mathcal{U}}$ ensures that

$$\bar{P}_{\mathcal{U}}\left(\hat{f}\left(X\right)\neq Y\right)\leq\epsilon+\bar{P}_{\mathcal{U}}\left(f^*\left(X\right)\neq Y\right)=\epsilon+\frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}}\eta_i^*$$

with probability $\geq 1-\delta$.

D.2. Proof of Lemma 2

For convenience, let us rewrite the sample size lower bound as $|S|\geq 2\log\left(\frac{2}{\delta}\right)\frac{\epsilon/8+\nu}{(\epsilon/8)^2}$. By Bernstein's inequality, we know that

$$\left|\operatorname{err}\left(f;P\right)-\widehat{\operatorname{err}}\left(f;S\right)\right|\leq\sqrt{\frac{2\operatorname{err}\left(f;P\right)\log\left(2/\delta\right)}{|S|}}+\frac{2\log\left(2/\delta\right)}{|S|}\leq\frac{\epsilon}{8}\sqrt{\frac{\operatorname{err}\left(f;P\right)}{\epsilon/8+\nu}}+\frac{\epsilon}{8}$$

with probability $1-\delta$. Suppose that $\operatorname{err}\left(f;P\right)\leq\epsilon/8+\nu$. Then,

$$\widehat{\operatorname{err}}\left(f;S\right)\leq\operatorname{err}\left(f;P\right)+\frac{\epsilon}{8}\sqrt{\frac{\operatorname{err}\left(f;P\right)}{\epsilon/8+\nu}}+\frac{\epsilon}{8}\leq\frac{\epsilon}{2}+\nu$$

Now, suppose that $\operatorname{err}(f;P) > \epsilon + \nu$ and define $g(p) \coloneqq p - a\sqrt{\frac{p}{a+b}}$. Then $g'(p) = 1 - \frac{a}{2\sqrt{p(a+b)}}$. In other words, g is increasing for $a \le 2\sqrt{p(a+b)} \iff p \ge \frac{a^2}{4(a+b)}$. Specializing to $p = \epsilon + \nu$, $a = \epsilon/8$ and $b = \nu$, we see that this condition is clearly ensured:

$$\epsilon + \nu \ge \frac{\epsilon}{8} > \frac{(\epsilon/8)^2}{4(\epsilon/8 + \nu)}$$

Since $\operatorname{err}(f;P) > \epsilon + \nu$ by assumption, we can again apply Bernstein to obtain

$$\begin{aligned}
\widehat{\operatorname{err}}(f;S) &\ge \operatorname{err}(f;P) - \frac{\epsilon}{8}\sqrt{\frac{\operatorname{err}(f;P)}{\epsilon/8+\nu}} - \frac{\epsilon}{8} \\
&> \epsilon + \nu - \frac{\epsilon}{8}\sqrt{\frac{\epsilon+\nu}{\epsilon/8+\nu}} - \frac{\epsilon}{8} \\
&> \frac{\epsilon}{2} + \nu
\end{aligned}$$

where we used the fact that $\sqrt{\frac{\epsilon+\nu}{\epsilon/8+\nu}} \le \sqrt{8}$ and $7/8 - 1/\sqrt{8} > 1/2$.

D.3. Proof of Theorem 3

By the reasoning above, we know that we succeed with probability $1-\delta$ under the specified parameters. Since we learn at least half of the distributions on every round, we only require $T = \lceil \log_2 k \rceil$ rounds. On every round, ERM requires $T_{\mathsf{SL}}\left(\eta, \frac{\epsilon}{16}, \frac{\delta}{2T}\right)$ samples. For testing, each P_i is sampled at most $T_{\mathsf{T}}\left(\eta_i, \epsilon, \frac{\delta}{2kT}\right)$ times on every round. Hence, the total sample size required is

$$\begin{aligned}
T_{\mathsf{RCN}}(\eta_{1:k}, \epsilon, \delta) &\le T\left(T_{\mathsf{SL}}\left(\eta, \frac{\epsilon}{16}, \frac{\delta}{2T}\right) + \sum_{i=1}^{k} T_{\mathsf{T}}\left(\eta_i, \epsilon, \frac{\delta}{2kT}\right)\right) \\
&\lesssim (\log k)\left(\frac{d\log(1/\epsilon) + \log\left(\frac{\log k}{\delta}\right)}{\epsilon(1-2\eta)} + \log\left(\frac{k}{\delta}\right)\sum_{i=1}^{k}\frac{\epsilon+\eta_i}{\epsilon^2}\right) \\
&\le \log^2\left(\frac{k}{\delta}\right)\left(\frac{d\log(1/\epsilon)}{\epsilon(1-2\eta)} + \sum_{i=1}^{k}\frac{\epsilon+\eta_i}{\epsilon^2}\right)
\end{aligned}$$

D.4. Proof of Theorem 4

We begin by describing Algorithm 3 in detail. We start with the candidate guess $\nu = 0$ and draw samples $S_i^{(t,1)} \overset{iid}{\sim} P_i$ of size $\left|S_i^{(t,1)}\right| = T_{\mathsf{T}}(0, \epsilon/2, \delta'')$ for each active distribution and eliminate those satisfying $\widehat{\operatorname{err}}\left(f^{(t)}; S_i^{(t,1)}\right) \le \epsilon/4$. If this step results in the removal of at least half of $\mathcal{U}^{(t-1)}$, we terminate the current round t. Crucially, the Correctness property ensures that any removed distribution satisfies $\operatorname{err}\left(f^{(t)}; P_i\right) \le \epsilon/2 \le \epsilon + \eta^*$, guaranteeing that we only output good hypotheses. Furthermore, if the true noise rate is indeed $\eta^* = 0$, the Progress property guarantees that this round terminates. By union bounding with the ERM guarantee, all of the above holds with probability $1 - \delta' - k\delta''$.

If the first check fails to eliminate half of the distributions, we proceed with the updated guess $\nu = \epsilon/2$. Rather than discarding the previous data, we augment the existing samples: for each remaining distribution i, we draw a fresh sample $\tilde{S}_i^{(t,2)} \overset{iid}{\sim} P_i$ of size $\left|\tilde{S}_i^{(t,2)}\right| = T_{\mathsf{T}}(\epsilon/2, \epsilon/2, \delta'') - \left|S_i^{(t,1)}\right|$. We then define $S_i^{(t,2)} = S_i^{(t,1)} \cup \tilde{S}_i^{(t,2)}$ as the aggregated sample set, which now satisfies the required size $\left|S_i^{(t,2)}\right| = T_{\mathsf{T}}(\epsilon/2, \epsilon/2, \delta'')$.

We eliminate any distribution satisfying $\widehat{\mathrm{err}}\left(f^{(t)}; S_i^{(t,2)}\right) \leq \epsilon/4 + \epsilon/2$, terminating the round if at least half of $\mathcal{U}^{(t-1)}$ has been removed. Observe that this threshold guarantees the removal of any i where $\mathrm{err}\left(f^{(t)}; P_i\right) \leq \epsilon/16 + \epsilon/2$, due to Progress. Consequently, if the true noise rate lies in the range $\eta^* \in (0, \epsilon/2]$, the round terminates. Furthermore, Correctness ensures that any removed distribution satisfies $\mathrm{err}\left(f^{(t)}; P_i\right) \leq \epsilon/2 + \epsilon/2 \leq \epsilon + \eta^*$. Taking a union bound over both testing sub-rounds, all of the above holds with probability at least $1 - \delta' - 2k\delta''$.

We continue iteratively, testing the hypotheses that η^* lies within consecutive intervals $\{0\}, (0, \epsilon/2], (\epsilon/2, \epsilon]$, and so on. Suppose that the true noise rate falls in the interval $\eta^* \in ((m-1)\,\epsilon/2, m\epsilon/2]$ and the procedure has not yet terminated in the first m iterations. Then in the $(m+1)$th iteration, we set the guess $\nu = m\epsilon/2$. By assumption, we know that $m \leq 2\eta^*/\epsilon + 1$ and $\eta^* \leq m\epsilon/2 \leq \eta^* + \epsilon/2$. For each remaining distribution i, we sample $\tilde{S}_i^{(t,m+1)} \overset{iid}{\sim} P_i$ such that the aggregated sample $S_i^{(t,m+1)} = S_i^{(t,m)} \cup \tilde{S}_i^{(t,m+1)}$ has size

$$\left|S_i^{(t,m+1)}\right| = T_{\mathsf{T}}\left(m\epsilon/2, \epsilon/2, \delta''\right) = 32\log\left(\frac{2}{\delta''}\right)\frac{\epsilon + 16m\epsilon/2}{\epsilon^2} \leq 32\log\left(\frac{2}{\delta''}\right)\frac{9\epsilon + 16\eta^*}{\epsilon^2}$$

We then remove any distribution satisfying $\widehat{\mathrm{err}}\left(f^{(t)}; S_i^{(t,m+1)}\right) \leq \epsilon/4 + m\epsilon/2$. Once again, the known properties yield the following:

- **Progress:** We are guaranteed to remove any i where $\mathrm{err}\left(f^{(t)}; P_i\right) \leq \epsilon/16 + m\epsilon/2$. Since we know that $\mathrm{err}\left(f^{(t)}; P_i\right) \leq \epsilon/16 + \eta^* \leq \epsilon/16 + m\epsilon/2$ for at least half of $\mathcal{U}^{(t-1)}$, the procedure is guaranteed to terminate at this step.
- **Correctness:** Any distribution i removed in this iteration satisfies $\mathrm{err}\left(f^{(t)}; P_i\right) \leq \frac{\epsilon}{2} + \frac{m\epsilon}{2} = \epsilon + \frac{(m-1)\epsilon}{2} \leq \epsilon + \eta^*$.

Consequently, the testing phase for round t terminates in at most $m + 1 \leq 2\eta^*/\epsilon + 2$ iterations. Crucially, every eliminated distribution i satisfies the target guarantee $\mathrm{err}\left(f^{(t)}; P_i\right) \leq \epsilon + \eta^*$. By a union bound, this result holds with probability at least

$$1 - \delta' - (m+1)\,k\delta'' \geq 1 - \delta' - 2\left(\eta^*/\epsilon + 1\right)k\delta'' \geq 1 - \delta' - 2\left(\eta/\epsilon + 1\right)k\delta''$$

By tuning δ' and δ'' appropriately, we know that we can guarantee success with probability at least $1 - \delta$. Again, we learn half of the distributions on every round, so that $T = \lceil \log_2 k \rceil$ rounds suffices. On each round, ERM uses

$$T_{\mathsf{SL}}\left(\eta, \frac{\epsilon}{32}, \frac{\delta}{2T}\right) = O\left(\frac{d + \log\left(\frac{\log k}{\delta}\right)}{\epsilon\left(1 - 2\eta\right)}\right)$$

samples, and testing requires at most

$$T_{\mathsf{T}}\left(\eta^* + \frac{\epsilon}{2}, \frac{\epsilon}{2}, \frac{\delta}{4\left(\eta/\epsilon + 1\right)kT}\right) = 32\log\left(\frac{8\left(\eta/\epsilon + 1\right)kT}{\delta}\right)\frac{9\epsilon + 16\eta^*}{\epsilon^2}$$

samples per distribution. Putting these together, we obtain a total sample complexity upper bound of

$$\begin{aligned} T_{\mathsf{MM}}\left(\eta^*_{1:k}, \eta_{1:k}, \epsilon, \delta\right) &\lesssim (\log k)\left(\frac{d + \log\left(\frac{\log k}{\delta}\right)}{\epsilon\left(1 - 2\eta\right)} + \log\left(\frac{\left(\eta/\epsilon + 1\right)k}{\delta}\right)\frac{k\left(\epsilon + \eta^*\right)}{\epsilon^2}\right) \\ &\lesssim \log^2\left(\frac{\left(\eta/\epsilon + 1\right)k}{\delta}\right)\left(\frac{d}{\epsilon\left(1 - 2\eta\right)} + \frac{k\left(\epsilon + \eta^*\right)}{\epsilon^2}\right) \end{aligned}$$

Appendix E. Proofs of Section 3

E.1. Proof of Lemma 6

Hoeffding's inequality tells us that, with probability at least $1-\delta$,

$$\left|\operatorname{err}\left(f;P\right)-\widehat{\operatorname{err}}\left(f;S\right)\right|\leq\sqrt{\frac{\log\left(2/\delta\right)}{2\left|S\right|}}\leq\frac{\epsilon}{12}$$

Under this high-probability event, we immediately obtain our claim:

$$\begin{aligned}\widehat{\operatorname{err}}\left(f;S\right)\leq\frac{1}{4}+\frac{\epsilon}{6}&\Longrightarrow\operatorname{err}\left(f;P\right)\leq\frac{1}{4}+\frac{\epsilon}{6}+\frac{\epsilon}{12}\leq\frac{1}{4}+\epsilon\\ \widehat{\operatorname{err}}\left(f;S\right)>\frac{1}{4}+\frac{\epsilon}{6}&\Longrightarrow\operatorname{err}\left(f;P\right)>\frac{1}{4}+\frac{\epsilon}{6}-\frac{\epsilon}{12}=\frac{1}{4}+\frac{\epsilon}{12}\end{aligned}$$

E.2. Proof of Lemma 7

From Appendix C, we know that

$$\operatorname{Var}_P\left(\mathbb{I}\left\{g\left(X\right)\neq Y\right\}-\mathbb{I}\left\{f^*\left(X\right)\neq Y\right\}\right)\leq\frac{\operatorname{err}\left(g;P\right)-\eta^*}{1-2\eta}$$

for any function $g:\mathcal{X}\to\{0,1\}$. Bernstein's inequality then tells us that, with probability $\geq 1-\delta/3$,

$$\begin{aligned}\left|\operatorname{err}\left(g;P\right)-\eta^*-\left(\widehat{\operatorname{err}}\left(g;S\right)-\widehat{\operatorname{err}}\left(f^*;S\right)\right)\right|&\leq\frac{4}{3\left|S\right|}\log\left(\frac{6}{\delta}\right)+\sqrt{\frac{2\left(\operatorname{err}\left(g;P\right)-\eta^*\right)}{\left(1-2\eta\right)\left|S\right|}\log\left(\frac{6}{\delta}\right)}\\ &\leq\frac{\epsilon}{48}+\sqrt{\frac{\epsilon}{48}\left(\operatorname{err}\left(g;P\right)-\eta^*\right)}\\ &\leq\frac{3\epsilon}{96}+\frac{\operatorname{err}\left(g;P\right)-\eta^*}{2}\end{aligned}$$

where we used AM-GM in the last inequality. Since $\hat{f}$ and S are independent, we can then conclude that $\hat{f}$ and f^* are close in empirical error: with probability $\geq 1-2\delta/3$,

$$\left|\widehat{\operatorname{err}}\left(\hat{f};S\right)-\widehat{\operatorname{err}}\left(f^*;S\right)\right|\leq\frac{3\epsilon}{96}+\frac{3}{2}\left(\operatorname{err}\left(\hat{f};P\right)-\eta^*\right)\leq\frac{3\epsilon}{48}$$

Let us take a union bound of this event with the Bernstein bound above on our function of interest f, so that they simultaneously occur with probability $\geq 1-\delta$. Under this event, we then have that

$$\begin{aligned}&\left|\widehat{\operatorname{err}}\left(f;S\right)-\widehat{\operatorname{err}}\left(\hat{f};S\right)\right|\leq\frac{\epsilon}{3}\\ &\quad\Longrightarrow\left|\widehat{\operatorname{err}}\left(f;S\right)-\widehat{\operatorname{err}}\left(f^*;S\right)\right|\leq\left|\widehat{\operatorname{err}}\left(f;S\right)-\widehat{\operatorname{err}}\left(\hat{f};S\right)\right|+\left|\widehat{\operatorname{err}}\left(\hat{f};S\right)-\widehat{\operatorname{err}}\left(f^*;S\right)\right|\leq\frac{19\epsilon}{48}\\ &\quad\Longrightarrow\operatorname{err}\left(f;P\right)-\eta^*\leq\frac{3\epsilon}{48}+2\left|\widehat{\operatorname{err}}\left(f;S\right)-\widehat{\operatorname{err}}\left(f^*;S\right)\right|\leq\frac{41\epsilon}{48}\end{aligned}$$

and

$$\begin{aligned}&\operatorname{err}\left(f;P\right)-\eta^*\leq\frac{\epsilon}{12}\\ &\quad\Longrightarrow\left|\widehat{\operatorname{err}}\left(f;S\right)-\widehat{\operatorname{err}}\left(f^*;S\right)\right|\leq\frac{3\epsilon}{96}+\frac{3}{2}\left(\operatorname{err}\left(f;P\right)-\eta^*\right)\leq\frac{5\epsilon}{32}\\ &\quad\Longrightarrow\left|\widehat{\operatorname{err}}\left(f;S\right)-\widehat{\operatorname{err}}\left(\hat{f};S\right)\right|\leq\left|\widehat{\operatorname{err}}\left(f;S\right)-\widehat{\operatorname{err}}\left(f^*;S\right)\right|+\left|\widehat{\operatorname{err}}\left(f^*;S\right)-\widehat{\operatorname{err}}\left(\hat{f};S\right)\right|\leq\frac{21\epsilon}{96}\end{aligned}$$

E.3. Proof of Theorem 8

When $d \geq 1/\epsilon$, Lemma 6 immediately implies correctness of the decision, and the sample size is

$$T_{\mathsf{C}}(\epsilon, \delta) = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$$

Now, suppose that $d < 1/\epsilon$. With $T_{\mathsf{SL}}(1/4, \epsilon/48, \delta/3)$ samples, we know that ERM guarantees $\mathrm{err}\left(\hat{f}; P\right) \leq 1/4 + \epsilon/48$ with probability at least $1 - \delta/3$. Then, correctness of our decision follows from Lemma 7. In this case, the sample size obtained is

$$T_{\mathsf{SL}}\left(\frac{1}{4}, \frac{\epsilon}{48}, \frac{\delta}{3}\right) + T_{\mathsf{L}}\left(\frac{1}{4}, \epsilon, \delta\right) = O\left(\frac{d}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$$

Combining both conditions yields the desired sample complexity.

E.4. Proof of Theorem 9

To prove the hardness of (SHT), we first construct a preliminary base testing problem and establish a lower bound for it.

Theorem 18 *Let $d \in \mathbb{N}$ and $\epsilon \in (0,1)$ be such that $d\epsilon \leq 1$. Define a random variable $X \in [\![d^2]\!]$ with PMF*

$$p_X(x) = \begin{cases} 1 - d\epsilon, & x = 0 \\ \epsilon/d, & x \in [d^2] \end{cases}$$

In addition, let $f^* : [\![d^2]\!] \to \{0,1\}$ be an underlying function, and $Y \in \{0,1\}$ be such that

$$\mathbb{P}(f^*(X) \neq Y | X = x) = 1/4 \quad \forall x \in [\![d^2]\!]$$

We sample i.i.d. pairs (X, Y) and must ensure that:

H_0: If $f^* = f_0$, output YES *with probability at least $3/4$.*

H_1: If f^* equals 1 precisely on a size-d subset, output NO *with probability at least $3/4$.* (SHT-Base)

This problem requires a sample size of

$$T \geq \frac{3d}{4\epsilon}\log(1 + \log 2)$$

Proof of Theorem 18 Let P_0^T be the joint distribution over $(X_t, Y_t)_{t=1}^T$ under H_0 and, similarly, P_R^T the joint under H_1 and subset $R \subset [d^2]$ where f^* equals 1. We also use lower-case p to denote the corresponding PMFs. With abuse of notation, let YES and NO denote the event that a successful algorithm outputs each decision. Let R be chosen uniformly at random from all subsets of size d. The learner must ensure that

$$\frac{3}{2} \leq P_0^T(\mathsf{YES}) + \mathbb{E}_R\left[P_R^T(\mathsf{NO})\right] \leq 1 + \mathrm{TV}\left(P_0^T, \mathbb{E}_R\left[P_R^T\right]\right)$$

In other words, we need

$$\mathrm{TV}\left(P_0^T, \mathbb{E}_R\left[P_R^T\right]\right) \geq \frac{1}{2}$$

To upper bound the TV distance, we will work with the χ^2 divergence and apply Ingster's method (Ingster and Suslina, 2002; Polyanskiy and Wu, 2025):

$$\chi^2\left(\mathbb{E}_R\left[P_R^T\right] \| P_0^T\right) + 1 = \mathbb{E}_{R,R'}\left[G^T\left(R, R'\right)\right]$$

where R and R' are i.i.d. subsets and

$$\begin{aligned} G\left(R, R'\right) &\coloneqq \mathbb{E}_{(X,Y)\sim P_0}\left[\frac{p_R\left(X,Y\right)}{p_0\left(X,Y\right)} \cdot \frac{p_{R'}\left(X,Y\right)}{p_0\left(X,Y\right)}\right] \\ &= \sum_{x \in [\![d^2]\!]} p_0\left(x\right) \sum_{y \in \{0,1\}} \frac{p_R\left(y|x\right) p_{R'}\left(y|x\right)}{p_0\left(y|x\right)} \\ &= 1 - d\epsilon + \frac{\epsilon}{d} \sum_{x \in [d^2]} \underbrace{\left[\frac{4}{3} p_R\left(0|x\right) p_{R'}\left(0|x\right) + 4 p_R\left(1|x\right) p_{R'}\left(1|x\right)\right]}_{=:(\star)} \end{aligned}$$

To evaluate this sum, let us consider two possibilities:

- For $x \notin R \cap R'$, at least one of the conditional PMFs will be $\mathrm{Ber}\left(1/4\right)$. Let the other one be $\mathrm{Ber}\left(q\right)$. Then,

$$(\star) = \frac{4}{3} \cdot \frac{3}{4} \cdot (1-p) + 4 \cdot \frac{1}{4} \cdot p = 1$$

- For $x \in R \cap R'$,

$$(\star) = \frac{4}{3} \cdot \left(\frac{1}{4}\right)^2 + 4 \cdot \left(\frac{3}{4}\right)^2 = \frac{7}{3}$$

As a result, we obtain

$$G\left(R, R'\right) = 1 - d\epsilon + \frac{\epsilon}{d}\left(d^2 - \left|R \cap R'\right| + \frac{7}{3}\left|R \cap R'\right|\right) = 1 + \frac{4\epsilon}{3d}\left|R \cap R'\right|$$

Plugging this back into the χ^2 formula yields

$$\chi^2\left(\mathbb{E}_R\left[P_R^T\right] \| P_0^T\right) + 1 = \mathbb{E}_{R,R'}\left[\left(1 + \frac{4\epsilon}{3d}\left|R \cap R'\right|\right)^T\right] \leq \mathbb{E}_{R,R'}\left[\exp\left(\frac{4\epsilon T}{3d}\left|R \cap R'\right|\right)\right]$$

Next, we observe that the random variable $|R \cap R'|$ follows a hypergeometric distribution: conditioned on R, we can see $|R \cap R'|$ as the number of "successes" when we draw $|R'| = d$ times without replacement from a population of size d^2 that contains exactly $|R| = d$ successes. That is, $|R \cap R'| \, | R \sim \mathrm{HG}\left(N = d^2, K = d, n = d\right)$, which is independent of R, so that $|R \cap R'| \sim \mathrm{HG}\left(N = d^2, K = d, n = d\right)$. Our goal is then to bound the MGF of this hypergeometric, which is smaller than the MGF of its binomial counterpart:

$$\begin{aligned} \mathbb{E}_{R,R'}\left[\exp\left(\lambda\left|R \cap R'\right|\right)\right] &\leq \mathbb{E}\left[\exp\left(\lambda \, \mathrm{Bin}\left(n = d, p = 1/d\right)\right)\right] \\ &= \left(1 - \frac{1}{d} + \frac{e^\lambda}{d}\right)^d \\ &= \exp\left(d \log\left(1 + \frac{e^\lambda - 1}{d}\right)\right) \\ &\leq \exp\left(e^\lambda - 1\right) \end{aligned}$$

Finally, we can combine everything to obtain

$$\begin{aligned}\mathrm{TV}\left(P_0^T, \mathbb{E}_R\left[P_R^T\right]\right) &\leq \frac{1}{2}\sqrt{\chi^2\left(\mathbb{E}_R\left[P_R^T\right] \middle\| P_0^T\right)} \\ &\leq \frac{1}{2}\sqrt{\mathbb{E}_{R,R'}\left[\exp\left(\frac{4\epsilon T}{3d}\left|R \cap R'\right|\right)\right] - 1} \\ &\leq \frac{1}{2}\sqrt{\exp\left(\exp\left(\frac{4\epsilon T}{3d}\right) - 1\right) - 1}\end{aligned}$$

Since the TV distance has to be larger than $1/2$, we can then conclude that

$$T \geq \frac{3d}{4\epsilon}\log\left(1 + \log 2\right)$$

■

Remark 19 (External randomness) *The lower bound of Theorem 18 still applies when the learner has access to an external source of randomness; i.e., it applies to randomized learners.*

We can readily prove our original claim using Theorem 18. Suppose that $d \leq \frac{1}{2\epsilon}$ and let $\mathcal{F} \subset \{0,1\}^{[\![d^2]\!]}$ be the class of $f = f_0$ along with all functions f that equal 1 precisely on a subset $R \subset \left[d^2\right]$ of size d. Note that $\mathrm{VCdim}\left(\mathcal{F}\right) = d$.

Let $\mathcal{A}$ be an (SHT) algorithm with sample complexity $T_{\mathcal{A}}$. We will construct a learner for (SHT-Base) with parameters $(d, 2\epsilon)$. To that end, let our data-generating distribution P be according to one of the hypotheses, and suppose that $\mathcal{A}$ wishes to test if $f = f_0$ is ϵ-optimal. Its error is given by $\mathrm{err}\left(f; P\right) = \mathbb{P}\left(Y = 1\right)$, so that we can make the following observations:

- Under H_0, we have that $f = f^*$ and, thus, $\mathrm{err}\left(f; P\right) = 1/4$.
- Under H_1 and subset R, the functions f and f^* disagree precisely on R, so that

$$\mathrm{err}\left(f; P\right) = \left(1 - 2d\epsilon\right) \cdot \frac{1}{4} + \left(d^2 - d\right) \cdot \frac{2\epsilon}{d} \cdot \frac{1}{4} + d \cdot \frac{2\epsilon}{d} \cdot \frac{3}{4} = \frac{1}{4} + \epsilon$$

Hence, if we run $\mathcal{A}$ on $T_{\mathcal{A}}\left(\epsilon, 1/4\right)$ samples, we output YES precisely under H_0 with probability $3/4$. In other words, we are able to distinguish H_0 and H_1. The lower bound of Theorem 18 then implies that $\mathcal{A}$ requires a sample size of at least

$$T_{\mathcal{A}}\left(\epsilon, 1/4\right) \geq \frac{3d}{8\epsilon}\log\left(1 + \log 2\right)$$

E.5. Proof of Theorem 10

Algorithm 6 MSHT

Input: Function class $\mathcal{F}$, distributions $P_1, \ldots, P_k$, functions $f_1, \ldots, f_k : \mathcal{X} \to \{0,1\}$, parameters $(\epsilon, \delta) \in (0,1)^2$.

Output: Decision vector $(\mathsf{D}_1, \ldots, \mathsf{D}_k)$.

if $d \geq 1/\epsilon$ **then**
  for $i = 1, \ldots, k$ **do**
    Sample $S_i \overset{iid}{\sim} P_i$ of size $|S_i| = T_{\mathsf{C}}(\epsilon, \delta)$
    if $\widehat{\mathrm{err}}(f_i; S_i) \leq 1/4 + \epsilon/6$ **then** $\mathsf{D}_i \leftarrow \mathsf{YES}$ **else** $\mathsf{D}_i \leftarrow \mathsf{NO}$
  end
else
  for $i = 1, \ldots, k$ **do**
    $\hat{f}_i \leftarrow \mathrm{ERM}_{\mathcal{F}}(S_i)$ where $S_i \overset{iid}{\sim} P_i$ is of size $|S_i| = T_{\mathsf{SL}}(1/4, \epsilon/48, \delta/3)$
    Sample $S_i' \overset{iid}{\sim} P_i$ of size $|S_i'| = T_{\mathsf{L}}(1/4, \epsilon, \delta)$
    if $\left|\widehat{\mathrm{err}}(f_i; S_i') - \widehat{\mathrm{err}}\left(\hat{f}_i; S_i'\right)\right| \leq \epsilon/3$ **then** $\mathsf{D}_i \leftarrow \mathsf{YES}$ **else** $\mathsf{D}_i \leftarrow \mathsf{NO}$
  end
end

This proof is the natural extension of Theorem 8 to multiple distributions. Note that we do not require any union bounds due to the objective. When $d \geq 1/\epsilon$, we similarly apply Lemma 6 to conclude correctness. The sample size is

$$k \cdot T_{\mathsf{C}}(\epsilon, \delta) = O\left(\frac{k \log(1/\delta)}{\epsilon^2}\right)$$

When $d < 1/\epsilon$, Lemma 7 along with a union bound over $i \in [k]$ again ensures correctness, with a sample size of

$$k\left[T_{\mathsf{SL}}\left(\frac{1}{4}, \frac{\epsilon}{48}, \frac{\delta}{3}\right) + T_{\mathsf{L}}\left(\frac{1}{4}, \epsilon, \delta\right)\right] = O\left(\frac{kd \log(1/\delta)}{\epsilon}\right)$$

E.6. Proof of Theorem 11

In the next result, we construct an (MSHT) instance that suffers from the $\Omega(kd/\epsilon)$ lower bound when $d \leq \frac{1}{2\epsilon}$, thereby proving our desired hardness claim.

Theorem 20 *Let $d \in \mathbb{N}$ and $\epsilon \in (0,1)$ be such that $d \geq 8\epsilon$ and $2d\epsilon \leq 1$. Define covariate space $\mathcal{X} := [\![kd^2 + k - 1]\!]$, composed of k disjoint blocks each of size $d^2 + 1$, and denote the ith block by $\mathcal{X}_i := \left\{(i-1)\left(d^2+1\right) + x : x \in [\![d^2]\!]\right\}$. Let $\mathcal{F}$ be the class of binary-valued functions on $[\![kd^2 + k - 1]\!]$ consisting of*

- *The $f = f_0$ function.*
- *All functions f that equal 1 precisely on a size-d subset of $\mathcal{X}_i \setminus \left\{(i-1)\left(d^2+1\right)\right\}$ (we exclude the first coordinate) for some $i \in [k]$.*

Note that this only requires shattering a set of size at most d. Now, define distributions $P_1, \ldots, P_k$ over $\mathcal{X} \times \{0,1\}$ with marginal PMFs

$$P_i\left(X=x\right)=\begin{cases}1-2d\epsilon, & x=(i-1)\left(d^2+1\right)\\ 2\epsilon/d, & x\in\left\{(i-1)\left(d^2+1\right)+1,\ldots,i\left(d^2+1\right)-1\right\}\end{cases}$$

and label noise

$$P_i\left(f^*\left(X\right)\neq Y|X=x\right)=\frac{1}{4}\quad\forall x\in\mathcal{X},i\in[k]$$

for some unknown $f^*\in\mathcal{F}$.

Any (MSHT) algorithm $\mathcal{A}$, with confidence parameter $\delta=0.01$, that tests the functions $f_1=\cdots=f_k=f_0$ on this instance requires a sample size of

$$T_{\mathcal{A}}\left(\epsilon,0.01\right)\geq\frac{0.015kd}{2\epsilon}$$

Proof of Theorem 20 Consider the following hypotheses:

H_0: $f^*=f_0$.

H_1: f^* equals 1 precisely on a size-d subset of $\mathcal{X}_i\backslash\left\{(i-1)\left(d^2+1\right)\right\}$ for some $i\in[k]$.

As in the proof of Theorem 9, we note that

- Under H_0, we have that $\mathrm{err}\left(f_i;P_i\right)=1/4$ for every $i\in[k]$.
- Under H_1 and distribution i, we have that $\mathrm{err}\left(f_i;P_i\right)=1/4+\epsilon$ and $\mathrm{err}\left(f_j;P_j\right)=1/4$ for every $j\neq i$.

Hence, $\mathcal{A}$ must be able to distinguish both hypotheses. Define T_i to be the number of times that P_i is sampled, which is a random quantity since the algorithm can be adaptive. Let $\mathbb{P}_0$ be the probability law under H_0.

Fix $i\in[k]$. We will show, by contradiction, that $T_i\gtrsim d/\epsilon$ with constant probability under H_0. Assume to the contrary that

$$\mathbb{P}_0\left(T_i\leq\frac{0.05d}{2\epsilon}\right)\geq 0.7$$

We will construct an (SHT-Base) algorithm $\mathcal{A}_s$ by simulating $\mathcal{A}$. Suppose that we are given sample access to a distribution P according to one of the (SHT-Base) hypotheses. Let $p_X=(1-2d\epsilon,2\epsilon/d,\ldots,2\epsilon/d)$ denote the PMF over $[\![d^2]\!]$ from Theorem 18 (with ϵ scaled by 2). Consider the following strategy:

1. Run $\mathcal{A}$ on parameters $(\epsilon,\delta)=(\epsilon,0.1)$. When it samples distribution j,
   - If $j=i$, sample $(X,Y)\sim P$.
   - If $j\neq i$, sample $X\sim p_X$ and $Y\sim\mathrm{Ber}\left(1/4\right)$.

   Return the data point $\left(X+(j-1)\left(d^2+1\right),Y\right)$. The shift on X ensures that it lives in $\mathcal{X}_j$. In other words, we set f^* to 0 outside of $\mathcal{X}_i$ and set P_i to be the unknown (SHT-Base) instance. Hence, H_0 and H_1 in both coincide.
2. We terminate whenever the first of the following occurs:

- If $\mathcal{A}$ has sampled distribution i more than $\frac{0.05d}{2\epsilon}$ times, output YES with probability 0.24.
- If $\mathcal{A}$ terminates, output D_i.

Note that this process requires at most $\frac{0.05d}{2\epsilon} + 1 \leq \frac{0.3d}{2\epsilon}$ (since $d \geq 8\epsilon$ by assumption) samples from P, which beats the lower bound of Theorem 18. Nevertheless, under the chosen parameters, $\mathcal{A}_s$ indeed succeeds. In what follows, we use $\mathbb{P}_0$ and $\mathbb{P}_1$ to denote the probability law under H_0 and H_1, respectively. When the probability does not depend on the instance, we drop the subscript. Next, we analyze the output of $\mathcal{A}_s$ under each hypothesis.

H_0: Suppose that the null hypothesis H_0 is true. Then,

$$\begin{aligned}
\mathbb{P}_0\left(\mathcal{A}_s = \mathsf{YES}\right) &= \mathbb{P}_0\left(\left\{T_i > \frac{0.05d}{2\epsilon}\right\} \cap \{\mathcal{A}_s = \mathsf{YES}\}\right) + \mathbb{P}_0\left(\left\{T_i \leq \frac{0.05d}{2\epsilon}\right\} \cap \{\mathsf{D}_i = \mathsf{YES}\}\right) \\
&= \mathbb{P}_0\left(T_i > \frac{0.05d}{2\epsilon}\right) \mathbb{P}\left(\mathcal{A}_s = \mathsf{YES} \middle| T_i > \frac{0.05d}{2\epsilon}\right) \\
&\qquad + \mathbb{P}_0\left(\mathsf{D}_i = \mathsf{YES}\right) - \mathbb{P}_0\left(\left\{T_i > \frac{0.05d}{2\epsilon}\right\} \cap \{\mathsf{D}_i = \mathsf{YES}\}\right) \\
&\geq \mathbb{P}_0\left(T_i > \frac{0.05d}{2\epsilon}\right) \mathbb{P}\left(\mathcal{A}_s = \mathsf{YES} \middle| T_i > \frac{0.05d}{2\epsilon}\right) + \mathbb{P}_0\left(\mathsf{D}_i = \mathsf{YES}\right) - \mathbb{P}_0\left(T_i > \frac{0.05d}{2\epsilon}\right) \\
&= \mathbb{P}_0\left(\mathsf{D}_i = \mathsf{YES}\right) - \mathbb{P}_0\left(T_i > \frac{0.05d}{2\epsilon}\right) \mathbb{P}\left(\mathcal{A}_s = \mathsf{NO} \middle| T_i > \frac{0.05d}{2\epsilon}\right) \\
&\geq 0.99 - 0.3 \cdot 0.76 \\
&= 0.762
\end{aligned}$$

H_1: Suppose that the alternative hypothesis H_1 is true. Then,

$$\begin{aligned}
\mathbb{P}_1\left(\mathcal{A}_s = \mathsf{NO}\right) &= \mathbb{P}_1\left(\left\{T_i > \frac{0.05d}{2\epsilon}\right\} \cap \{\mathcal{A}_s = \mathsf{NO}\}\right) + \mathbb{P}_1\left(\left\{T_i \leq \frac{0.05d}{2\epsilon}\right\} \cap \{\mathsf{D}_i = \mathsf{NO}\}\right) \\
&= \mathbb{P}_1\left(T_i > \frac{0.05d}{2\epsilon}\right) \mathbb{P}\left(\mathcal{A}_s = \mathsf{NO} \middle| T_i > \frac{0.05d}{2\epsilon}\right) \\
&\qquad + \mathbb{P}_1\left(\mathsf{D}_i = \mathsf{NO}\right) - \mathbb{P}_1\left(\left\{T_i > \frac{0.05d}{2\epsilon}\right\} \cap \{\mathsf{D}_i = \mathsf{NO}\}\right) \\
&\geq \mathbb{P}_1\left(T_i > \frac{0.05d}{2\epsilon}\right) \mathbb{P}\left(\mathcal{A}_s = \mathsf{NO} \middle| T_i > \frac{0.05d}{2\epsilon}\right) + \mathbb{P}_1\left(\mathsf{D}_i = \mathsf{NO}\right) - \mathbb{P}_1\left(T_i > \frac{0.05d}{2\epsilon}\right) \\
&= \mathbb{P}_1\left(\mathsf{D}_i = \mathsf{NO}\right) - \mathbb{P}_1\left(T_i > \frac{0.05d}{2\epsilon}\right) \mathbb{P}\left(\mathcal{A}_s = \mathsf{YES} \middle| T_i > \frac{0.05d}{2\epsilon}\right) \\
&\geq 0.99 - 0.24 \\
&= 0.75
\end{aligned}$$

Note that we only required D_i succeeding with high probability; this allows us to avoid union bounds over $i \in [k]$. Since $\mathcal{A}_s$ always succeeds with a sample size smaller than the lower bound of Theorem 18, we have thus shown that

$$\mathbb{P}_0\left(T_i > \frac{0.05d}{2\epsilon}\right) > 0.3$$

which we can readily convert into an in-expectation bound:

$$\mathbb{E}_0\left[T_i\right] \geq \mathbb{E}_0\left[T_i \mathbb{I}\left\{T_i > \frac{0.05d}{2\epsilon}\right\}\right] \geq \frac{0.05d}{2\epsilon} \mathbb{P}_0\left(T_i > \frac{0.05d}{2\epsilon}\right) > \frac{0.015d}{2\epsilon}$$

Since this must hold for every $i \in [k]$, we can conclude that

$$T_{\mathcal{A}}(\epsilon, 0.01) \geq \mathbb{E}_0\left[\sum_{i=1}^{k} T_i\right] \geq \frac{0.015kd}{2\epsilon}$$

■

Appendix F. Proofs of Section 4

F.1. Proof of Lemma 12

Correctness of $\mathcal{A}$ implies that with probability at least $1 - \delta/3$, for all $i \in [k]$, $\operatorname{err}\left(\hat{f}_i; P_i\right) \leq 1/4 + \epsilon/48$. Lemma 7 then shows that D_i is correct with probability $1 - \delta$, for each $i \in [k]$.

F.2. Proof of Theorem 13

The $\Omega(d/\epsilon)$ term follows from the single-distribution learning lower bound (e.g., see Theorem 6.8 of (Shalev-Shwartz and Ben-David, 2014)). For the second term, we can apply Lemma 12 to an MDL algorithm $\mathcal{A}$ and the (MSHT) lower bound of Theorem 11 to conclude that

$$T_{\mathcal{A}}\left(\epsilon, \frac{0.01}{3}\right) + \frac{4k\log(600)}{\epsilon} = T_{\mathcal{A}\to\mathsf{MSHT}}(48\epsilon, 0.01) \geq \frac{0.015k}{9216}\min\left\{\frac{1}{\epsilon^2}, \frac{d}{\epsilon}\right\}$$

Rearranging under the lower bound on $\min\{d, 1/\epsilon\}$ then yields the claim.

Appendix G. Proofs of Section 5

In this section, we prove the (MDL-Mass) lower bound of $\Omega\left(k\sqrt{d}/\epsilon\right)$. We begin by establishing a preliminary lower bound for the testing problem (SHT-Mass), and then show how it implies the (MDL-Mass) lower bound.

G.1. Proof of Theorem 14

We will apply the technique of *Poissonization*: instead of sampling a deterministic number of times T, we will sample $N \sim \operatorname{Pois}(T)$ pairs (X_t, Y_t) and must ensure the same guarantee as (SHT-Mass). We will first show a lower bound for the Poissonized variant and subsequently relate it back to the original setup.

Let us define the count of $(x, 1)$ and $(x, 0)$ observations:

$$N_x \coloneqq \sum_{t=1}^{N} \mathbb{I}\{X_t = x, Y_t = 1\} \quad \text{and} \quad M_x \coloneqq \sum_{t=1}^{N} \mathbb{I}\{X_t = x, Y_t = 0\}$$

We will work with random $\mathbf{q}$, independent of N, in which case we can ensure that

$$N_x|\mathbf{q} \sim \operatorname{Pois}(Tp_X(x)\, q_x) \quad \text{and} \quad M_x|\mathbf{q} \sim \operatorname{Pois}(Tp_X(x)\,(1 - q_x))$$

and are independent conditional on $\mathbf{q}$. Also, note that $N_x + M_x = \sum_{t=1}^{N} \mathbb{I}\{X_t = x\} \sim \operatorname{Pois}(Tp_X(x))$ and is independent of $\mathbf{q}$.

The next step is to construct appropriate biases for each hypothesis. For H_0, we will simply define the constant vector $\mathbf{q}_0 \coloneqq (0.465, \ldots, 0.465)$. For H_1, we define the prior $\mu_1 \coloneqq 0.75 \cdot \operatorname{Ber}(0.62)$ and set

$\mathbf{q}_1 \sim \delta_{0.465} \times \mu_1^d$. In other words, the 0th coordinate is always 0.465, and the rest are sampled i.i.d. from μ_1. One important consequence of this construction is that the first moments match:

$$\mathbb{E}_{q\sim\mu_1}[q] = 0.75 \cdot 0.62 = 0.465 = \mathbb{E}_{q\sim\mu_0}[q]$$

where $\mu_0 \coloneqq \delta_{0.465}$.

While $\mathbf{q}_0$ falls under H_0 with probability 1, we can only ensure a high-probability guarantee for $\mathbf{q}_1$. Since

$$\mathbb{E}[\Delta_{\mathbf{q}_1}] = \frac{\epsilon}{d} \cdot d \cdot 0.62 \cdot 0.5 = 0.31\epsilon$$

we can apply Hoeffding's inequality (note that $\Delta_{\mathbf{q}_1} \in [0, \epsilon/2]$) to conclude that

$$\mathbb{P}(\Delta_{\mathbf{q}_1} \geq 0.3\epsilon) = \mathbb{P}(\Delta_{\mathbf{q}_1} - \mathbb{E}[\Delta_{\mathbf{q}_1}] \geq -0.01\epsilon) \geq 1 - \exp(-0.0008d) \geq \frac{3}{4}$$

Under hypothesis $h \in \{0, 1\}$, our decision is based on observations $O_h \coloneqq \left(N, (X_t, Y_t)_{t=1}^N\right) \sim P_h^O$. Our objective requires that $P_0^O(\mathsf{YES}) \geq 3/4$, since the bias vector is deterministic under our constructed H_0. Furthermore, defining the event $A \coloneqq \{\Delta_{\mathbf{q}_1} \geq 0.3\epsilon\}$, we must also satisfy

$$P_1^O(\mathsf{YES}) = \underbrace{P_1^O(\mathsf{YES}|A)}_{\leq 1/4} P_1^O(A) + P_1^O(\mathsf{YES}|A^c) \underbrace{P_1^O(A^c)}_{\leq 1/4} \leq \frac{1}{2}$$

As a consequence, we get the lower bound

$$\mathrm{TV}\left(P_0^O, P_1^O\right) \geq P_0^O(\mathsf{YES}) - P_1^O(\mathsf{YES}) \geq \frac{1}{4}$$

To further bound the TV distance, we instead switch our attention to the counts $C_h \coloneqq (M_x, N_x)_{x=0}^d \sim P_h^C$. To see why they suffice, note that C_h is a deterministic function of O_h. Moreover, we simply pass C_h through a Markov kernel that is *independent* of $\mathbf{q}_h$ to obtain O_h (choose an ordering of the (X_t, Y_t) uniformly at random). This implies that

$$\mathrm{TV}\left(P_0^O, P_1^O\right) = \mathrm{TV}\left(P_0^C, P_1^C\right)$$

Next, we upper bound the right-hand side.

Lemma 21 *We have that*

$$\mathrm{TV}\left(P_0^C, P_1^C\right) \leq \frac{T^2\epsilon^2}{2d}$$

Proof of Lemma 21 We begin with the observation that the coordinates of C_h are independent, and the 0th coordinate of both C_0 and C_1 is the same by construction. Let $\tilde{C}_h \sim P_h^{\tilde{C}}$ denote coordinates $[d]$ of C_h, so that $\mathrm{TV}\left(P_0^C, P_1^C\right) = \mathrm{TV}\left(P_0^{\tilde{C}}, P_1^{\tilde{C}}\right)$. To bound this, we note that $\tilde{C}_h \overset{iid}{\sim} P_{\mu_h}$, where

$$P_q \coloneqq \mathrm{Pois}\left(\frac{T\epsilon q}{d}\right) \times \mathrm{Pois}\left(\frac{T\epsilon(1-q)}{d}\right) \quad \text{and} \quad P_{q\sim\mu} \coloneqq \mathbb{E}_\mu[P_q]$$

Using subadditivity of TV, we then have that $\mathrm{TV}\left(P_0^{\tilde{C}}, P_1^{\tilde{C}}\right) \leq d\,\mathrm{TV}(P_{\mu_0}, P_{\mu_1})$. To upper bound the right-hand side, let $\lambda = T\epsilon/d$ and note that

$$P_q(n, m) = \left(\frac{(\lambda q)^n e^{-\lambda q}}{n!}\right)\left(\frac{(\lambda(1-q))^m e^{-\lambda(1-q)}}{m!}\right) = \frac{\lambda^{n+m} q^n (1-q)^m e^{-\lambda}}{n!m!}$$

In particular,

$$P_q(0,0) = e^{-\lambda}, \quad P_q(1,0) = \lambda e^{-\lambda} q, \quad P_q(0,1) = \lambda e^{-\lambda}(1-q)$$

Since $\mathbb{E}_{\mu_0}[q] = \mathbb{E}_{\mu_1}[q]$, we then know that

$$P_{\mu_0}(n,m) = P_{\mu_1}(n,m) \quad \forall (n,m) \in \{(0,0),(1,0),(0,1)\}$$

As a result, we get that

$$\begin{aligned}
\mathrm{TV}(P_{\mu_0}, P_{\mu_1}) &= \frac{1}{2} \sum_{n+m \geq 2} |P_{\mu_0}(n,m) - P_{\mu_1}(n,m)| \\
&\leq \frac{1}{2} [P_{\mu_0}(N+M \geq 2) + P_{\mu_1}(N+M \geq 2)] \\
&= \mathbb{P}(\mathrm{Pois}(\lambda) \geq 2)
\end{aligned}$$

where in the last line we used the fact that $N + M \sim \mathrm{Pois}(\lambda)$ is independent of q. Lastly, we must bound the Poisson tail above. To start, note that

$$\mathbb{P}(\mathrm{Pois}(\lambda) \geq 2) = 1 - \mathbb{P}(\mathrm{Pois}(\lambda) = 0) - \mathbb{P}(\mathrm{Pois}(\lambda) = 1) = 1 - e^{-\lambda}(1+\lambda)$$

Define $g(\lambda) \coloneqq e^{-\lambda}(1+\lambda)$ and note that $g'(\lambda) = -\lambda e^{-\lambda}$. Then,

$$1 - e^{-\lambda}(1+\lambda) = g(0) - g(\lambda) = \int_0^\lambda -g'(t)\,dt = \int_0^\lambda te^{-t}dt \leq \int_0^\lambda tdt = \frac{\lambda^2}{2}$$

Putting everything together, we conclude that

$$\mathrm{TV}\left(P_0^C, P_1^C\right) \leq \frac{d\lambda^2}{2} = \frac{T^2\epsilon^2}{2d}$$

■

From Lemma 21, we can establish a lower bound on the Poissonized setting:

$$\frac{T^2\epsilon^2}{2d} \geq \frac{1}{4} \Longrightarrow T \geq \frac{\sqrt{d}}{\sqrt{2}\epsilon}$$

To obtain a similar bound for our original problem, we rely on the following Poisson tail bound (Mitzenmacher and Upfal, 2005): for $\lambda \geq 12\log(2/\delta)$,

$$\mathbb{P}\left(\mathrm{Pois}(\lambda) \in \left[\frac{\lambda}{2}, \frac{3\lambda}{2}\right]\right) \geq 1 - \delta$$

Suppose that a tester $\mathcal{A}$ solves (SHT-Mass) with at most T samples. Consider the following strategy $\mathcal{A}_P$ for the Poissonized variant:

- Sample $N \sim \mathrm{Pois}(2T')$, where $T' = \max\{T, 14\}$.
- If $N \geq T$, run $\mathcal{A}$ on the first T samples and output its answer. Otherwise, decide based on a fair coin flip.

Since $2T' \geq 28 \geq 12 \log(2/0.2)$, the Poisson tail bound implies that

$$\mathbb{P}(N \geq T) \geq \mathbb{P}\left(N \in \left[T', 3T'\right]\right) \geq 0.8$$

Let $\mathbb{P}_h$ denote the probability measure under any instance of hypothesis $h \in \{0,1\}$, and define the event $B \coloneqq \{N \geq T\}$. Then,

$$\begin{aligned}
\mathbb{P}_0(\mathcal{A}_P = \mathsf{YES}) &= \mathbb{P}_0(\mathcal{A} = \mathsf{YES}|B)\,\mathbb{P}(B) + \mathbb{P}_0(\mathcal{A}_P = \mathsf{YES}|B^c)\,\mathbb{P}(B^c) \\
&= \mathbb{P}_0(\mathcal{A}_P = \mathsf{YES}|B^c) + \mathbb{P}(B)\left(\mathbb{P}_0(\mathcal{A} = \mathsf{YES}|B) - \mathbb{P}_0(\mathcal{A}_P = \mathsf{YES}|B^c)\right) \\
&\geq 0.5 + 0.8\,(0.9 - 0.5) \\
&\geq 0.75
\end{aligned}$$

By a symmetric argument, $\mathbb{P}_0(\mathcal{A}_P = \mathsf{NO}) \geq 0.75$. In other words, we solve the Poissonized problem and its lower bound implies that

$$2T' \geq \frac{\sqrt{d}}{\sqrt{2}\epsilon} \Longrightarrow T \geq \frac{\sqrt{d}}{2\sqrt{2}\epsilon}$$

where we used the assumption $d \geq 1750$, so that $\frac{\sqrt{d}}{\sqrt{2}\epsilon} > 28$ and, thus, $T' = T$.

G.2. Proof of Theorem 15

To start, we note that our construction requires that f^* always equals 0 on the first coordinate of each $\mathcal{X}_i$. Then, if we apply Lemma 17 on f_0, we can see that

$$\Delta_i \coloneqq \operatorname{err}(f_0; P_i) - \eta_i^* = \sum_{x \in \mathcal{X}: f^*(x) = 1} (1 - 2\eta_i^*(x))\, P_i(X = x) = \frac{\epsilon}{d} \sum_{x \in \mathcal{X}_i : q_x^i \geq 1/2} \left(2q_x^i - 1\right)$$

where $q_x^i \coloneqq P_i(Y = 1|X = x)$.

Consider the setting where $f^* = f_0$ and $\eta_i^*(x) = q_x^i = 0.465$ for all $x \in \mathcal{X}$ and $i \in [k]$. Let $\mathbb{P}_0$ denote the probability law under this environment. We will show that any successful (MDL-Mass) algorithm must sample $\Omega\left(\sqrt{d}/\epsilon\right)$ times from each distribution under $\mathbb{P}_0$ with high probability.

To do this, we will solve an (SHT-Mass) hard instance P by simulating a (MDL-Mass) algorithm $\mathcal{A}$. Here, P is a distribution over (X, Y) where $p_X = (1 - \epsilon, \epsilon/d, \ldots, \epsilon/d)$ on $[\![d]\!]$ and $Y|X = x \sim \operatorname{Ber}(q_x)$. Recall our testing objective:

- H_0: $q_x = 0.465$ for all $x \in [\![d]\!]$.
- H_1: $\Delta_{\mathbf{q}} = \frac{\epsilon}{d} \sum_{x \in [\![d]\!] : q_x \geq 1/2} (2q_x - 1) \geq 0.3\epsilon$.

Fix some $i \in [k]$ and let T_i be number of times that $\mathcal{A}$ samples P_i. We will show, by contradiction, that $T_i \gtrsim \sqrt{d}/\epsilon$ with high probability under $\mathbb{P}_0$. Let $B_i \coloneqq \left\{T_i \leq 0.05\sqrt{d}/\epsilon\right\}$ and assume to the contrary that

$$\mathbb{P}_0(B_i) \geq 0.85$$

We will construct an (SHT-Mass) $\mathcal{A}_s$ strategy as follows:

1. Run $\mathcal{A}$ on parameters $(\eta, \epsilon, \delta) = (0.49, \epsilon/192, 0.1/3)$. When it samples distribution $j \in [k]$,
   - If $j = i$, sample $(X, Y) \sim P$.

- If $j \neq i$, sample $X \sim p_X$ and $Y \sim \mathrm{Ber}\,(0.465)$.

To ensure that $X \in \mathcal{X}_j$, we shift X by $(j-1)(d+1)$ before returning it to $\mathcal{A}$. In other words, we set P_i to P, with an appropriate shift, and set f^* to 0 outside of $\mathcal{X}_i$. By construction, this aligns with our MDL setup.

2. Terminate when the first of the following occurs:

   - If T_i exceeds $0.05\sqrt{d}/\epsilon$, output NO.
   - If $\mathcal{A}$ terminates and outputs $\hat{f}_i$, sample $S \overset{iid}{\sim} P$ of size

$$|S| = T_{\mathsf{L}}\,(0.49, \epsilon/4, 0.1) = \frac{384}{0.02\epsilon}\log(60) \leq \frac{80000}{\epsilon}$$

   shift the X's appropriately, and output YES if and only if event

$$E_i \coloneqq \left\{\left|\widehat{\mathrm{err}}\,(f_0; S) - \widehat{\mathrm{err}}\left(\hat{f}_i; S\right)\right| \leq \frac{\epsilon}{12}\right\}$$

   occurs.

Note that the suboptimality gaps coincide: $\Delta_i = \Delta_{\mathbf{q}}$. In addition, correctness of $\mathcal{A}$ implies, via Lemma 7, that with probability at least 0.9,

- Under H_0, we have that $\Delta_i = 0$, so that E_i occurs.
- Under H_1, we have that $\Delta_i \geq 0.3\epsilon$, so that E_i^c occurs.

Importantly, this statement is *unconditional*; that is, under no assumption of B_i occurring. Let $\mathbb{P}_i$ denote the probability law under H_i, where we note that $\mathbb{P}_0$ coincides with our earlier definition.

H_0:

$$\mathbb{P}_0\,(\mathcal{A}_s = \mathsf{YES}) = \mathbb{P}_0\,(B_i \cap E_i) = \mathbb{P}_0\,(E_i) - \mathbb{P}_0\,(B_i^c \cap E_i) \geq \mathbb{P}_0\,(E_i) - \mathbb{P}_0\,(B_i^c) \geq 0.75$$

H_1:

$$\mathbb{P}_1\,(\mathcal{A}_s = \mathsf{NO}) = \mathbb{P}_1\,(B_i^c) + \mathbb{P}_1\,(B_i \cap E_i^c) = \mathbb{P}_1\,(B_i^c) + \mathbb{P}_1\,(E_i^c) - \mathbb{P}_1\,(B_i^c \cap E_i^c) \geq \mathbb{P}_1\,(E_i^c) \geq 0.9$$

That is, we solve (SHT-Mass) with a sample size of at most

$$\frac{0.05\sqrt{d}}{\epsilon} + 1 + \frac{80000}{\epsilon} \leq \frac{\sqrt{d}}{2\sqrt{2}\epsilon}$$

beating the lower bound of Theorem 14. This contradiction ensures that

$$\mathbb{P}_0\left(T_i > \frac{0.05\sqrt{d}}{\epsilon}\right) > 0.15 \implies \mathbb{E}_0\,[T_i] \geq \mathbb{E}_0\left[T_i \mathbb{I}\left\{T_i > \frac{0.05\sqrt{d}}{\epsilon}\right\}\right] \geq \frac{0.0075\sqrt{d}}{\epsilon}$$

Since this holds for each $i \in [k]$, we finally get that

$$T_{\mathcal{A}}\left(\frac{\epsilon}{192}, \frac{0.1}{3}\right) \geq \mathbb{E}_0\left[\sum_{i=1}^{k} T_i\right] \geq \frac{0.0075k\sqrt{d}}{\epsilon}$$