Open Problem: The Sample Complexity of Multi-Distribution Learning for VC Classes

Pranjal Awasthi
Google Research, Mountain View, CA, USA

Nika Haghtalab
University of California, Berkeley, CA, USA

Eric Zhao
University of California, Berkeley, CA, USA

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Multi-distribution learning is a natural generalization of PAC learning to settings with multiple data distributions. There remains a significant gap between the known upper and lower bounds for PAC-learnable classes. In particular, though we understand the sample complexity of learning a VC dimension \( d \) class on \( k \) distributions to be 
\[
O\left(\varepsilon^{-2}\ln(k)(d + k) + \min\left\{\varepsilon^{-1}dk, \varepsilon^{-4}\ln(k)d\right\}\right),
\]
the best lower bound is 
\[
\Omega\left(\varepsilon^{-2}(d + k\ln(k))\right).
\]
We discuss recent progress on this problem and some hurdles that are fundamental to the use of game dynamics in statistical learning.

Keywords: PAC learning, multi-distribution learning, distributional robustness, learning in games.

1. Introduction

The pervasive need for robustness, fairness, and multi-agent welfare in learning processes has led to the development of learning paradigms whose performance hold under multiple distributions and scenarios. Multi-distribution learning, or MDL, is a setting introduced by [HJZ22] to address these needs and unify several existing frameworks and applications, such as notions of min-max fairness [MSS19, AAK+22], group distributionally robust optimization [SKHL20], and collaborative learning [BHPQ17]. MDL is a generalization of the agnostic learning paradigms [Val84, BEHW89] to multiple data distributions. In this setting, given a set of distributions \( \mathcal{D} = \{D_1, \ldots, D_k\} \) supported on \( X \times Y \), loss function \( \ell \), and a hypothesis class \( \mathcal{H} \), the goal of MDL is to find a (possibly randomized) hypothesis \( h \) where

\[
\max_{D \in \mathcal{D}} \mathcal{L}_D(h) \leq \varepsilon + \min_{h^* \in \mathcal{H}} \max_{D \in \mathcal{D}} \mathcal{L}_D(h^*), \quad \text{where} \quad \mathcal{L}_D(h) := \mathbb{E}_{(x,y) \sim D}[\ell(h_i(x,y))].
\]  

Such an \( h \) is called an \( \varepsilon \)-optimal solution to the MDL problem \((\mathcal{D}, \mathcal{H})\) and we denote \( \text{OPT} := \min_{h^* \in \mathcal{H}} \max_{D \in \mathcal{D}} \mathcal{L}_D(h^*) \). Our open problem concerns the sample complexity of MDL.

Problem Statement. Consider an example oracle \( \mathcal{E}_i \) for each distribution \( D_i \in \mathcal{D} \), which once queried returns an independent sample \((x,y) \sim D_i\). The optimal sample complexity of MDL is the smallest total number of queries issued to examples oracles, in a possibly adaptive fashion, that is sufficient for learning an \( \varepsilon \)-optimal solution. Formally, a multi-distribution learning algorithm at each iteration \( t = 1, 2, \ldots \), chooses an index \( i(t) \in [k] \), queries \( \mathcal{E}_{i(t)} \) to sample an instance \((x^{(t)}, y^{(t)})\) and, upon termination, returns a (possibly randomized) solution \( h \). We use the shorthands \( z^{(t)} = (x^{(t)}, y^{(t)}, i^{(t)}) \), \( Z = X \times Y \times [k] \), and \( Z^* \) to denote a sequence \( z^{(1)}, z^{(2)}, \ldots \) of any size.

© 2023 P. Awasthi, N. Haghtalab & E. Zhao.
Definition 1 (Multi-Distribution Learnability) We say a hypothesis class $\mathcal{H}$ is multi-distribution learnable with sample complexity $m_{\mathcal{H}} : (0,1)^2 \times \mathbb{N} \rightarrow \mathbb{N}$ if there exists functions $A_{\mathcal{H}} : Z^* \rightarrow [k]$ and $A_h : Z^* \rightarrow \Delta(\mathcal{Y})^X$ where the following holds: for every $(\varepsilon, \delta) \in (0,1), k \in \mathbb{N}$, and set of $k$ distributions $D$ over $X \times Y$, by letting $i(t) = A_{\mathcal{H}}(z^{(1)}, \ldots, z^{(t-1)})$ for $t \in [m_{\mathcal{H}}(\varepsilon, \delta, k)]$, with probability at least $1 - \delta$, the solution $h = A_h(z^{(1)}, \ldots, z^{(m)})$ is $\varepsilon$-optimal, i.e., satisfying (1).

Problem 1 What is the optimal sample complexity of MDL? Are hypothesis classes $\mathcal{H}$ with VC dimension $d$ multi-distribution learnable with a sample complexity of $O\left(\varepsilon^{-2} \ln(k) d + k \ln(k/\delta)\right)$?

Recalling that the sample complexity of agnostic learning is $m_{\mathcal{H}}(\varepsilon, \delta, 1) \in \Theta(\varepsilon^{-2}(d + \ln(1/\delta)))$ [SB14], one hopes to avoid paying the $\Omega(k \cdot m_{\mathcal{H}}(\varepsilon, \delta/k, 1))$ samples necessary to independently learn each of the $k$ data distributions. This is why our conjectured sample complexity avoids a dependence on $dk$ and has an optimal $\varepsilon^{-2}$ dependence. Existing results, however, have fallen short of meeting both of these requirements and traded off lack of dependence on $dk$ with the optimal dependence on $\varepsilon$, as shown in rows 1 and 2 of Table 1. On the other hand, the optimal sample complexity of MDL has been rightly characterized for finite hypothesis classes in row 3 (and more generally those of finite Littlestone dimension or Bregman diameter [HJZ22]) and obtains optimal $\varepsilon^{-2} \ln(|\mathcal{H}|)$ dependence. The best lower bound, row 4, leaves a logarithmic gap with the conjectured upper bound. Near-optimal bounds are known for realizable settings where OPT = 0 (row 5) and personalized settings where one can produce a different hypothesis for each distribution (row 6).

Table 1: Best known bounds on the sample complexity of MDL for hypothesis classes with VC dimension $d$. $\tilde{O}$ hides double-log factors and an additive factor of $\varepsilon^{-2} k \ln(k/\delta)$.

<table>
<thead>
<tr>
<th>Bound</th>
<th>Assumption</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $\tilde{O}(\varepsilon^{-2} \ln(k) d + \varepsilon^{-1} dk \log(d/\varepsilon))$</td>
<td>N/A</td>
<td>[HJZ22]</td>
</tr>
<tr>
<td>2. $\tilde{O}(\varepsilon^{-4} \ln(k) (d + \ln(1/\delta \varepsilon)))$</td>
<td>N/A</td>
<td>See full version.</td>
</tr>
<tr>
<td>3. $\tilde{O}(\varepsilon^{-2} \ln(</td>
<td>\mathcal{H}</td>
<td>))$</td>
</tr>
<tr>
<td>4. $\Omega(\varepsilon^{-2} (d + k \ln(\min{d, k}/\delta)))$</td>
<td>N/A</td>
<td>[HJZ22]</td>
</tr>
<tr>
<td>5. $O(\ln(k) \varepsilon^{-1} (d \ln(1/\varepsilon) + k \ln(k/\delta)))$</td>
<td>OPT = 0</td>
<td>[CZZ18, NZ18]</td>
</tr>
<tr>
<td>6. $\tilde{O}(\ln(k) \varepsilon^{-2} (d \ln(d/\varepsilon) + k \ln(k/\delta)))$</td>
<td>Personalized</td>
<td>See full version.</td>
</tr>
</tbody>
</table>

Broad Applications. One of the motivating application of MDL is collaborative learning, where multiple stakeholders (representing $D_i$) collaborate in training a model that provides high performance for each stakeholder [BHPQ17, NZ18, CZZ18, BHPS21]. The sample complexity of MDL thus quantifies the value of collaboration in learning: whereas our conjectured upper bound would imply that collaboration reduces the amount of data needed by a $\ln(k)/k$ factor, existing bounds only imply a $\min\{\ln(k)/k\varepsilon^2, \varepsilon\}$ factor reduction.

Another application of MDL is to Group distributionally robust optimization (DRO) which concerns learning a model with performance guarantees for many deployment environments [SKHL20, SRKL20]. MDL sample complexity bounds quantify the cost of obtaining this robustness, a question of growing interest and which has been studied in terms of finite-sum convergence [CH22,
ACJ+21] and sample complexity [HJZ22]. Our conjectured upper bound would extend these favorable results to VC classes by only increasing the sample complexity logarithmically.

MDL also captures notions of min-max fairness in learning, which concerns prioritizing the well-being of the worst-off subgroup and has applications in federated learning [MSS19] and equity [AAK+22]. Min-max fair learning has mainly been studied in settings with presampled datasets, where an inevitable sample complexity lower bound of $\Omega(dk/\varepsilon^2)$ arises as one cannot adaptively choose distributions to sample from. The sample complexity of MDL thus captures how min-max fairness can be attained at less cost by adapting one’s data collection strategy on the fly.

2. Overview of Current Approaches

Multi-distribution learning can be formulated as the zero-sum game between a “learner” who chooses hypotheses $h \in \mathcal{H}$ and an “adversary” whose chooses indices $i \in [k]$, with the payoff function $\mathcal{L}_{D_i}(h)$. Importantly, for any mixed-strategy $\varepsilon$-min-max equilibrium $(p, q) \in \Delta(\mathcal{H}) \times \Delta_k$, the randomization map $p$ is a $2\varepsilon$-optimal solution. All existing multi-distribution learning algorithms can be expressed as finding a $\varepsilon$-equilibrium using no-regret dynamics (see [HJZ22] for an overview).

**Game dynamics.** Formally, a game dynamic is a $T$-iteration process where, at each $t \in [T]$, a learner chooses hypothesis $h^{(t)}(x) \in \mathcal{H}$ with a no-regret algorithm and an adversary chooses a distribution $i^{(t)} \in [k]$ with a (semi-)bandit algorithm. The learner estimates its current cost function $h \mapsto \mathcal{L}_{D_{i^{(t)}}}(h)$ by sampling $N_{\text{learn}}$ datapoints from $\mathcal{E}_{X_{i^{(t)}}}$, while the adversary estimates its cost function $i \mapsto -\mathcal{L}_{D_i}(h^{(t)})$ by, for $N_{\text{adv}}$ choices of $i \in [k]$, sampling a datapoint from each $\mathcal{E}_i$. The random mapping $p$ where $p(x) = \text{Uniform}(h^{(1)(x)}, \ldots, h^{(t)(x)})$ is a $2\varepsilon$-optimal solution.

**Different instantiations.** Every result in Table 1 can be obtained by instantiating this game dynamics template. Row 3 can be obtained by setting $N_{\text{learn}} = N_{\text{adv}} = 1$, $T \propto \varepsilon^{-2}(\ln(|\mathcal{H}|) + k \ln(k/\delta))$, having the learner choose $h^{(t)}$ with Hedge and the adversary choose $i^{(t)}$ with Exp3-IX [Neu15]. Row 1 can be obtained with the same algorithm but first creating an offline $\varepsilon$-covering of the class $\mathcal{H}$ on each data distribution $D_i \in \mathcal{D}$, using $O(d/\varepsilon)$ samples per distribution. Row 2 can be obtained by setting $N_{\text{adv}} = k$, $N_{\text{learn}} \propto \varepsilon^{-2}(d + \ln(1/\delta\varepsilon))$, $T \propto \varepsilon^{-2}\ln(k/\delta)$, having the learner choose $h^{(t)}$ to be the (approximate) risk minimizer of the current cost function and the adversary choose $i^{(t)}$ with a high-probability variant of ELP [MS11]; in contrast to the prior upper bound, this bound uses an algorithm that iterates fewer times but samples more at each iteration.

**Personalization.** We can pinpoint the challenge of negotiating trade-offs between different data distributions as the primary difficulty of handling infinite classes. Consider the personalized setting where, during inference time, $\mathcal{A}_h(z^{(1)}, \ldots, z^{(m)})$ can return a different hypothesis $h_i$ for each distribution $D_i$. This assumes away the difficulty of combining hypotheses that are each near-optimal for different distributions. As we show in the full version of this paper, the conjectured sample complexity bound of $O(\ln(k)\varepsilon^{-2}(d \ln(d/\varepsilon) + k \ln(k/\delta)))$ can be obtained in the personalized setting (Row 6 of Table 1) by running the Row 1 algorithm $\ln(k)$ times, at each round limiting the adversary to playing within a small region of the simplex $\Delta_k$ that we can efficiently cover $\mathcal{H}$ on.

2.1. Existing Challenges

**Adaptive coverings.** A potential approach to closing the gap with the conjectured sample complexity bound is to find a method of adaptively covering the hypothesis class $\mathcal{H}$. Whereas Row 1
was obtained by taking a naive offline $\epsilon$-covering of $\mathcal{H}$ on all $k$ distributions, Row 2 was obtained by an algorithm that (implicitly) $\epsilon$-covers the class $\mathcal{H}$ on $O(\ln(k)\epsilon^{-2})$ adaptive choices of $D_t \in \mathcal{D}$. It is unclear whether a covering of lower resolution can be used, or if it is possible to only cover $\mathcal{H}$ on $O(\ln(k))$ choices of distributions $D_t \in \mathcal{D}$. We also note that it is not the size of the $\epsilon$-covering of $k$ distributions, i.e., $k\epsilon^{-O(d)}$, that is the bottleneck, but rather the number of samples needed to create such a cover. In contrast, the personalized setting decided in an online fashion what distributions need to be covered and it only covers $\mathcal{H}$ on $O(\ln(k))$ choice of (mixture) distributions from $\mathcal{D}$.

**Agnostic-to-realizable.** Another potential tool is an agnostic-to-realizable reduction [HKLM22], since nearly-optimal sample complexity bounds are known for realizable settings where $\text{OPT} = 0$ [BHPQ17, CZZ18, NZ18]. This technique has had success in related problems, such as the closely related adversarial PAC learning problem [MHS19]. Unfortunately, because multi-distribution learning involves online decision-making—determining which example oracles to call—the usual reduction of testing all possible labelings of observed datapoints is intractable.

**Bounding regret.** Game dynamics algorithms rely on the learner achieving a low regret on the sequence of distributions chosen by the adversary. However, with VC classes, even when all distributions share a Bayes classifier, an oblivious adversary can force the learner to suffer regret linear in $k$. It is therefore necessary to reason about the adversary’s behavior to bound the regret of the learner. This is atypical; game dynamics proofs usually bound each player’s regret independently.

**Proposition 2** Consider an algorithm $A$ that, given distributions $D_1, \ldots, D_T$, draws only $N$ datapoints in total and returns a sequence of hypotheses $h_1, \ldots, h_k$ where each $h_t$ is trained only on datapoints sampled from $D_1, \ldots, D_t$. There exists a sequence $D_1, \ldots, D_T$ with only $k$ distinct members, where $\mathbb{E}[T^{-1} \sum_{t \in [T]} L_{D_t}(h_t)] - \min_{h^* \in \mathcal{H}} T^{-1} \sum_{t \in [T]} L_{D_t}(h^*) \in \Omega(\sqrt{dk/N})$.

### 3. Intermediate Open Problems

**Lower Bounds.** We believe a $\ln(k)d$ factor is missing from the best known sample complexity lower bound of $\Theta(\epsilon^{-2}(d + k\ln(\min \{k, d\} / \delta)))$. The absence of a $\ln(k)d$ term would be significant as it would imply that, when VC dimension dominates sample complexity, handling more data distributions comes effectively for free. Interestingly, this $\ln(k)$ factor does not appear in the upper bound when the complexity of $\mathcal{H}$ is characterized by Littlestone dimension, perhaps due to the stronger compression guarantees for online-learnable classes. A $\ln(k)d$ term would also shed light on compression schemes for VC classes [LW86]; a lower bound of $\Theta(\ln(k)d + k)$ would lend evidence against the existence of $O(\mathcal{VC}(\mathcal{H}))$-size compression schemes.

**Problem 2** Is the sample complexity of multi-distribution learning in $\Omega(\log(k)d)$?

**Proper learning.** All existing multi-distribution learning algorithms with fast sample complexity rates produce either a randomized hypothesis $h \in \Delta(\mathcal{H})$ or an improper hypothesis resulting from taking a majority vote. An open question is whether impropriety is necessary for fast rates.

**Problem 3** What is the sample complexity of proper multi-distribution learning?

**Oracle-efficient learning.** For oracle-efficient algorithms, that is an algorithm only accessing $\mathcal{H}$ through an ERM oracle [DHL+20], only the sample complexity bound from Row 2 in Table 1 is known. An open question is whether there exists a statistical-computational trade-off for MDL.

**Problem 4** What is the sample complexity of oracle-efficient multi-distribution learning?
REFERENCES


[HKLM22] Max Hopkins, Daniel M. Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In Po-Ling Loh and Maxim Raginsky, editors, Proceedings of


