Proceedings of Machine Learning Research vol 195:1–11, 2023 36th Annual Conference on Learning Theory

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

Pranjal Awasthi PRANJALAWASTHI@GOOGLE.COM
Google Research, Mountain View, CA, USA

Nika Haghtalab NIKA@BERKELEY.EDU
University of California, Berkeley, CA, USA

Eric Zhao ERIC.ZH@BERKELEY.EDU
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 $O(\varepsilon^{-2}\ln(k)(d+k)+\min\{\varepsilon^{-1}dk,\varepsilon^{-4}\ln(k)d\})$, the best lower bound is $\Omega(\varepsilon^{-2}(d+k\ln(k)))$. 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 $\mathcal{X}\times\mathcal{Y}$, loss function ℓ, and a hypothesis class $\mathcal{H}$, the goal of MDL is to find a (possibly randomized) hypothesis h where

$$\max_{D\in\mathcal{D}}\mathcal{L}_D(h)\leq\varepsilon+\min_{h^*\in\mathcal{H}}\max_{D\in\mathcal{D}}\mathcal{L}_D(h^*),\ \text{where }\mathcal{L}_D(h)\coloneqq\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(h,(x,y))\right]. \tag{1}$$

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

Problem Statement. Consider an example oracle EX_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 ε-optimal solution. Formally, a multi-distribution learning algorithm at each iteration $t=1,2,\ldots$, chooses an index $i^{(t)}\in[k]$, queries $\mathrm{EX}_{i^{(t)}}$ to sample an instance $(x^{(t)},y^{(t)})$ and, upon termination, returns a (possibly randomized) solution h. We use the shorthands $z^{(t)}=(x^{(t)},y^{(t)},i^{(t)})$, $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}\times[k]$, and $\mathcal{Z}^*$ to denote a sequence $z^{(1)},z^{(2)},\ldots$ of any size.

Definition 1 (Multi-Distribution Learnability) *We say a hypothesis class $\mathcal{H}$ is multi-distribution learnable with sample complexity $m_{\mathcal{H}} : (0,1)^2 \times \mathbb{N} \to \mathbb{N}$ if there exists functions $\mathcal{A}_s : \mathcal{Z}^* \to [k]$ and $\mathcal{A}_h : \mathcal{Z}^* \to \Delta(\mathcal{Y})^{\mathcal{X}}$ where the following holds: for every $(\varepsilon, \delta) \in (0,1)$, $k \in \mathbb{N}$, and set of k distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, by letting $i^{(t)} = \mathcal{A}_s(z^{(1)}, \ldots, z^{(t-1)})$ for $t \in [m_{\mathcal{H}}(\varepsilon, \delta, k)]$, with probability at least $1 - \delta$, the solution $h = \mathcal{A}_h(z^{(1)}, \ldots, z^{(m)})$ is ε-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 ε^{-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 ε, 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 $\mathrm{OPT} = 0$ (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. $\widetilde{O}$ hides double-log factors and an additive factor of $\varepsilon^{-2}k\ln(k/\delta)$.

	Bound	Assumption	Citation
1.	$\widetilde{O}(\varepsilon^{-2}\ln(k)d + \varepsilon^{-1}dk\log(d/\varepsilon))$	N/A	[HJZ22]
2.	$\widetilde{O}(\varepsilon^{-4}\ln(k)(d + \ln(1/\delta\varepsilon))$	N/A	See full version.
3.	$\widetilde{O}(\varepsilon^{-2}\ln(\lvert\mathcal{H}\rvert))$	N/A	[HJZ22]
4.	$\Omega(\varepsilon^{-2}(d + k\ln(\min\{d,k\}/\delta)))$	N/A	[HJZ22]
5.	$O(\ln(k)\varepsilon^{-1}(d\ln(1/\varepsilon) + k\ln(k/\delta)))$	$\mathrm{OPT} = 0$	[CZZ18, NZ18]
6.	$\widetilde{O}(\ln(k)\varepsilon^{-2}(d\ln(d/\varepsilon) + k\ln(k/\delta)))$	Personalized	See full version.

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 $\min\left\{\ln(k)/k\varepsilon^2, \varepsilon\right\}$ 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 ε-min-max equilibrium $(p, q) \in \Delta(\mathcal{H}) \times \Delta_k$, the randomized map p is a 2ε-optimal solution. All existing multi-distribution learning algorithms can be expressed as finding a ε-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 $h^{(t)} \in \mathcal{H}$ 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_{learn} datapoints from $\text{EX}_{i^{(t)}}$, while the adversary estimates its cost function $i \mapsto -\mathcal{L}_{D_i}(h^{(t)})$ by, for N_{adv} choices of $i \in [k]$, sampling a datapoint from each EX_i. The random mapping p where $p(x) = \text{Uniform}(h^{(1)}(x), \dots, h^{(t)}(x))$ is a 2ε-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 ε-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, $\mathcal{A}_h(z^{(1)}, \dots, z^{(m)})$ 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 $\widetilde{O}(\ln(k)\varepsilon^{-2}(d\ln(d/\varepsilon) + k\ln(k/\delta)))$ 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 Δ_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 ε-covering of $\mathcal{H}$ on all k distributions, Row 2 was obtained by an algorithm that (implicitly) ε-covers the class $\mathcal{H}$ on $O(\ln(k)\varepsilon^{-2})$ adaptive choices of $D_i \in \mathcal{D}$. 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 $D_i \in \mathcal{D}$. We also note that it is not the size of the ε-covering of k distributions, i.e., $k\varepsilon^{-O(d)}$, 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 $\mathrm{OPT} = 0$ [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 $\mathcal{A}$ that, given distributions $D_1, \dots, D_T$, draws only N datapoints in total and returns a sequence of hypotheses $h_1, \dots, h_k$ where each h_t is trained only on datapoints sampled from $D_1, \dots, D_t$. There exists a sequence $D_1, \dots, D_T$ with only k distinct members, where $\mathbb{E}[T^{-1} \sum_{t\in[T]} \mathcal{L}_{D_t}(h_t)] - \min_{h^*\in\mathcal{H}} T^{-1} \sum_{t\in[T]} \mathcal{L}_{D_t}(h^*) \in \Omega(\sqrt{dk/N})$.*

3. Intermediate Open Problems

Lower Bounds. We believe a $\ln(k)d$ factor is missing from the best known sample complexity lower bound of $\Theta(\varepsilon^{-2}(d+k\ln(\min\{k,d\}/\delta)))$. 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 $O(\mathrm{VC}(\mathcal{H}))$-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 improperness 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

[AAK+22] 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 Szepesvári, Gang Niu, and Sivan Sabato, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 53–65. PMLR, 2022.

[ACJ+21] Hilal Asi, Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Stochastic bias-reduced gradient methods. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, pages 10810–10822, 2021.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[BHPQ17] 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 30*, pages 2392–2401. Curran Associates, Inc., 2017.

[BHPS21] Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for one, or all for all: equilibria and optimality of collaboration in federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 1005–1014. PMLR, 2021.

[CH22] Yair Carmon and Danielle Hausler. Distributionally robust optimization via ball oracle acceleration. In *Advances in Neural Information Processing Systems*, 2022.

[CZZ18] 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 31*, pages 3602–3611. Curran Associates, Inc., 2018.

[DHL+20] Miroslav Dudik, Nika Haghtalab, Haipeng Luo, Robert E. Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. *J. ACM*, 67(5):26:1–26:57, 2020.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[HJZ22] Nika Haghtalab, Michael I. Jordan, and Eric Zhao. On-Demand Sampling: Learning Optimally from Multiple Distributions. *CoRR*, abs/2210.12529, 2022.

[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* *the Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pages 3015–3069. PMLR, 2022.

[HW86] D Haussler and E Welzl. Epsilon-nets and simplex range queries. In *Proceedings of the Second Annual Symposium on Computational Geometry*, SCG '86, page 61–71. Association for Computing Machinery, 1986.

[LW86] Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986.

[MHS19] Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 2019.

[MS11] Shie Mannor and Ohad Shamir. From bandits to experts: on the value of side-observations. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

[MSS19] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 2019.

[Neu15] Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3168–3176, 2015.

[NJLS09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[NZ18] Huy L. 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 31*, pages 7642–7650. Curran Associates, Inc., 2018.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[SKHL20] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview, 2020.

[SRKL20] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and

Aarti Singh, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 2020.

[Val84] Leslie G. Valiant. A theory of the learnable. In *Proceedings of the Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 436–445. ACM, 1984.