Proceedings of Machine Learning Research vol 247:1–4, 2024 37th Annual Conference on Learning Theory

Optimal Multi-Distribution Learning

Zihan Zhang ZZ5478@PRINCETON.EDU
Princeton University, Princeton, NJ 08544, USA.

Wenhao Zhan WZ3993@PRINCETON.EDU
Princeton University, Princeton, NJ 08544, USA.

Yuxin Chen YUXINC@WHARTON.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104, USA.

Simon S. Du SSDU@CS.WASHINGTON.EDU
University of Washington, Seattle, WA 98195, USA.

Jason D. Lee JASONLEE@PRINCETON.EDU
Princeton University, Princeton, NJ 08544, USA.

Editors: Shipra Agrawal and Aaron Roth

Abstract

Multi-distribution learning (MDL), which seeks to learn a shared model that minimizes the worst-case risk across k distinct data distributions, has emerged as a unified framework in response to the evolving demand for robustness, fairness, multi-group collaboration, etc. Achieving data-efficient MDL necessitates adaptive sampling, also called on-demand sampling, throughout the learning process. However, there exist substantial gaps between the state-of-the-art upper and lower bounds on the optimal sample complexity. Focusing on a hypothesis class of Vapnik–Chervonenkis (VC) dimension d, we propose a novel algorithm that yields an ε-optimal randomized hypothesis with a sample complexity on the order of $\frac{d+k}{\varepsilon^2}$ (modulo log factor), matching the best-known lower bound. Our algorithmic ideas and theory are further extended to accommodate Rademacher classes. The proposed algorithms are oracle-efficient, which access the hypothesis class solely through an empirical risk minimization oracle. We also establish the necessity of randomization, revealing a large sample size barrier when only deterministic hypotheses are permitted. These findings resolve three open problems presented in COLT 2023 (i.e., Awasthi et al. (2023, Problems 1, 3 and 4)).[1]

Keywords: multi-distribution learning; on-demand sampling; game dynamics; VC classes; Rademacher classes; oracle efficiency

Acknowledgement

We thank Eric Zhao for answering numerous questions about the open problems. YC is supported in part by the Alfred P. Sloan Research Fellowship, the NSF grants CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773. JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994.

1. Extended abstract. Full version appears as [arXiv:2312.05134, v4].

References

Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. International Conference on Machine Learning, pages 53–65, 2022.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. Theory of Computing, 8(1):121–164, 2012.

Hilal Asi, Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Stochastic bias-reduced gradient methods. Advances in Neural Information Processing Systems, 34:10810–10822, 2021.

Pranjal Awasthi, Nika Haghtalab, and Eric Zhao. Open problem: The sample complexity of multi-distribution learning for VC classes. In Conference on Learning Theory (COLT), volume 195, pages 5943–5949, July 2023.

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

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 International Conference on Machine Learning, pages 1005–1014, 2021a.

Avrim Blum, Shelby Heinecke, and Lev Reyzin. Communication-aware collaborative learning. In AAAI Conference on Artificial Intelligence, volume 35, pages 6786–6793, 2021b.

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.

Peter Bühlmann and Nicolai Meinshausen. Magging: maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE, 104(1):126–135, 2015.

Yair Carmon and Danielle Hausler. Distributionally robust optimization via ball oracle acceleration. Advances in Neural Information Processing Systems, 35:35866–35879, 2022.

Jiecao Chen, Qin Zhang, and Yuan Zhou. Tight bounds for collaborative PAC learning via multiplicative weights. Advances in Neural Information Processing Systems, 31, 2018.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. Advances in Neural Information Processing Systems, 33:15111–15122, 2020.

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

Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. Journal of the ACM (JACM), 67(5):1–57, 2020.

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

David A Freedman. On tail probabilities for martingales. the Annals of Probability, pages 100–118, 1975.

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.

Zijian Guo. Statistical inference for maximin effects: Identifying stable associations across multiple studies. Journal of the American Statistical Association, pages 1–17, 2023.

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

Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In Advances in Neural Information Processing Systems, 2023.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In International Conference on Machine Learning, pages 1929–1938, 2018.

Elad Hazan. Introduction to online convex optimization. MIT Press, 2022.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In International Conference on Machine Learning, pages 1939–1948, 2018.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In International Conference on Machine Learning, pages 2029–2037, 2018.

Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In IEEE/CVF International Conference on Computer Vision, pages 4551–4560, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. Advances in neural information processing systems, 21, 2008.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2018.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In International Conference on Machine Learning, pages 4615–4625, 2019.

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

Binghui Peng. The sample complexity of multi-distribution learning. arXiv preprint arXiv:2312.04027, 2023.

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

Tim Roughgarden. Twenty lectures on algorithmic game theory. Cambridge University Press, 2016.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In International Conference on Machine Learning, pages 8346–8356, 2020.

Shai Shalev-Shwartz. Online learning and online convex optimization. Foundations and Trends in Machine Learning, 4(2):107–194, 2012.

Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

Christopher J Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In International Conference on Machine Learning, pages 21633–21657, 2022.

J v. Neumann. Zur theorie der gesellschaftsspiele. Mathematische annalen, 100(1):295–320, 1928.

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

Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the VC-dimension of a learning machine. Neural computation, 6(5):851–876, 1994.

Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge university press, 2019.

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

Xin Xiong, Zijian Guo, and Tianxi Cai. Distributionally robust transfer learning. arXiv preprint arXiv:2309.06534, 2023.

Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In International Conference on Learning Representations, 2020.

Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In Conference on Learning Theory (COLT), pages 3858–3904, 2022.

Sicheng Zhao, Bo Li, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. arXiv preprint arXiv:2002.12169, 2020.