Majorizing Measures, Sequential Complexities, and Online Learning

Adam Block ABLOCK@MIT.EDU
MIT

Yuval Dagan
MIT

Alexander Rakhlin
MIT

Editors: Mikhail Belkin and Samory Kpotufe

One[1] of the primary goals of learning theory is to understand how the sample complexity of learning depends on the complexity of the model. Classically, much of the research focused on the case where data are drawn independently from some population distribution. The mismatch between sample and population inevitably leads to questions of uniform convergence. To answer such questions, the theory of empirical processes was developed, with seminal papers establishing non-asymptotic rates of convergence in terms of Rademacher complexity, covering numbers, chaining, VC dimension, and scale-sensitive combinatorial parameters. More recently, there has been interest in extending the classical theory to the world of online learning, where there can be dependencies among the data; as in the classical case, the online regime requires guarantees of uniform convergence, with sequential analogues of the aforementioned complexity measures. Despite the recent development of such sequential counterparts to classical complexities, there have remained challenging open questions in the relations between these measures. One of the problems in extending the classical results to the sequential setting is the lack of covering-packing duality, a technical tool used in many classical methods to bound covering numbers by other quantities, such as the VC dimension using the technique of Dudley extraction. In this work, we draw inspiration from yet another classical notion of complexity, majorizing measures, and prove the exact analogues of many classical bounds in the online setting, bypassing the need for Dudley extraction. Furthermore, we relate the majorizing measures approach to the recently introduced notion of fractional covering numbers. More precisely, using this relation, we show that sequential Rademacher complexity is bounded by majorizing measures, which are dominated by chaining with respect to fractional covering numbers, which we in turn control by sequential, scale-sensitive, combinatorial parameters. Finally, we establish a tight contraction inequality for worst-case sequential Rademacher complexity. The above constitutes the resolution of a number of outstanding open problems in extending the classical theory of empirical processes to the sequential case, and, in turn, establishes sharp results for online learning. Furthermore, the introduction of majorizing measures to the online regime provides an additional tool in theorists' approaches to proving uniform convergence bounds in the sequential setting.

1. Extended abstract. Full version appears on arXiv as 2102.01729

References

Jacob Abernethy, Alekh Agarwal, Peter L Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. *arXiv preprint arXiv:0903.5328*, 2009.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019*, pages 852–860. Association for Computing Machinery, 2019.

Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification, 2021.

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

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. *journal of computer and system sciences*, 52(3):434–452, 1996.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5):1848–1869, 2008.

Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*. Number 27. Cambridge university press, 1989.

Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *61st Annual IEEE Symposium on Foundations of Computer Science*, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Victor H. de la Peña. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564, 1999.

R. M. Dudley. Sample functions of the gaussian process. *Ann. Probab.*, 1(1):66–103, 02 1973. doi: 10.1214/aop/1176997026. URL `https://doi.org/10.1214/aop/1176997026`.

Xavier Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. In *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*, pages 1–96. Springer, 1975.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper pac learning with approximate differential privacy. *arXiv preprint arXiv:2012.03893*, 2020.

Evarist Gine and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Number 40. Cambridge University Press, 2016.

Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

Young Jung, Baekjin Kim, and Ambuj Tewari. On the equivalence between online and private learnability beyond binary classification. *Advances in Neural Information Processing Systems*, 33, 2020.

Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1):37–55, 2003.

Alexander Rakhlin and Karthik Sridharan. On martingale extensions of vapnik–chervonenkis theory with applications to online learning. In *Measures of Complexity*, pages 197–215. Springer, 2015.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems*, 23:1984–1992, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

Mark Rudelson and Roman Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3): 1049–1103, 1996.

Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

VN Vapnik and A Ya Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

VN Vapnik and A Ya Chervonenkis. The necessary and sufficient conditions for the uniform convergence of averages to their expected values. *Teoriya Veroyatnostei i Ee Primeneniya*, 26(3): 543–564, 1981.