Can SGD Select Good Fishermen? Local Convergence under Self-Selection Biases (Extended Abstract)

Alkis Kalavasis ALKIS.KALAVASIS@YALE.EDU
Yale University

Anay Mehrotra ANAYMEHROTRA1@GMAIL.COM
Stanford University

Felix Zhou FELIX.ZHOU@YALE.EDU
Yale University

Editors: Steve Hanneke and Tor Lattimore

Abstract

We revisit the problem of estimating k linear regressors in d dimensions from samples affected by self-selection bias under the maximum selection rule, introduced by Cherapanamjeri et al. (2023). Our main result is an algorithm with sample complexity $O(d)\cdot\mathrm{poly}(k, 1/\varepsilon)$ and running time $\mathrm{poly}(d, k, 1/\varepsilon) + (k \log k)^{O(k)}$ for recovering the regressors up to joint squared error ε^2, improving the dependence on the accuracy parameter in prior algorithms of Cherapanamjeri et al. (2023) and Gaitonde and Mossel (2024).

The key ingredient is the first local-convergence algorithm for the maximum self-selection model, resolving an open problem of Cherapanamjeri et al. (2023). Our approach reduces self-selection to estimation from coarse observations, a setting studied by Fotakis et al. (2021), where the learner observes only the cell of a partition containing the latent sample. The self-selection reduction induces a structured non-convex partition, outside the scope of existing convex-partition algorithms. We prove that this partition preserves enough information locally and that the resulting negative log-likelihood is locally convex around the true parameters. These two geometric properties allow projected stochastic gradient descent, initialized from the warm start of Gaitonde and Mossel (2024), to obtain the stated end-to-end guarantee.

Keywords: high-dimensional statistics, coarse observations, self-selection bias

Acknowledgments

We thank the anonymous reviewers for comments and suggestions. We thank Manolis Zampetakis for feedback on an initial draft of this work, discussions about the literature on estimation from auction data, and for sharing the reference to Meilijson (1981). Alkis Kalavasis was supported by the Institute for Foundations of Data Science at Yale. Felix Zhou acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Extended abstract. Full version appears as arXiv:2504.07133.

References

Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Andrew Ilyas, and Manolis Zampetakis. What Makes a Good Fisherman? Linear Regression Under Self-Selection Bias. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1699–1712, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585177. URL https://doi.org/10.1145/3564246.3585177.

Dimitris Fotakis, Alkis Kalavasis, Vasilis Kontonis, and Christos Tzamos. Efficient Algorithms for Learning from Coarse Labels. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2060–2079. PMLR, 2021. URL https://proceedings.mlr.press/v134/fotakis21a.html.

Jason Gaitonde and Elchanan Mossel. Sample-Efficient Linear Regression with Self-Selection Bias, 2024. URL https://arxiv.org/abs/2402.14229.

Isaac Meilijson. Estimation of the Lifetime Distribution of the Parts from the Autopsy Statistics of the Machine. *Journal of Applied Probability*, 18(4):829–838, 1981. ISSN 00219002. URL http://www.jstor.org/stable/3213058.