Strategyproof Learning with Advice

Eric Balkanski EB3224@COLUMBIA.EDU
Columbia University

Cherlin Zhu CZ2740@COLUMBIA.EDU
Columbia University

Editors: Gautam Kamath and Po-Ling Loh

Abstract

An important challenge in robust machine learning is when training data is provided by strategic sources who may intentionally report erroneous data for their own benefit (Caro and Gallien, 2010; Caro et al., 2010). A line of work at the intersection of machine learning and mechanism design aims to deter strategic agents from reporting erroneous training data by designing learning algorithms that are strategyproof (Dekel et al., 2010; Perote and Perote-Pena, 2004; Chen et al., 2018; Cummings et al., 2015; Meir and Rosenschein, 2011; Meir et al., 2008, 2010, 2012; Hardt et al., 2016; Ghalme et al., 2021; Dong et al., 2018; Ahmadi et al., 2021). Strategyproofness is a strong and desirable property, but it comes at a cost in the approximation ratio of even simple risk minimization problems.

A recent line of work on mechanism design with advice has shown that side information can be leveraged to overcome worst-case bounds in mechanism design. Strategyproof mechanisms that achieve an improved approximation ratio when the advice is accurate (consistency) and an acceptable approximation ratio when the advice is inaccurate (robustness) have been designed for problems such as strategic facility location (Agrawal et al., 2022; Xu and Lu, 2022; Balkanski et al., 2024a), auction design (Lu et al., 2024; Caragiannis and Kalantzis, 2024; Balkanski et al., 2024b), strategic scheduling (Balkanski et al., 2023), strategic assignment (Colini-Baldeschi et al., 2024), and metric distortion (Berger et al., 2024). In this paper, we study strategyproof learning in two settings: regression and classification; we provide the first non-trivial consistency-robustness tradeoffs for both.

In strategic learning, each agent $i \in \{1, \dots, n\}$ reports a set $S_i = \{(x_{i,j}, y_{i,j})\}_j$ of labeled data points to the learner. The points $x_{i,j}$ are public information, but the labels $y_{i,j}$ are private information that agent i can potentially misreport. The goal of the mechanism is to learn a function f from a function class $\mathcal{F}$ that minimizes the global risk $R(f, S) = \frac{1}{|S|}\sum_{i=1}^{n}\sum_{j=1}^{|S_i|}\ell(f(x_{i,j}), y_{i,j})$ for some loss function ℓ. The agents are strategic and aim to minimize their personal risk $R_i(f, S) = \frac{1}{|S_i|}\sum_{j=1}^{|S_i|}\ell(f(x_{i,j}), y_{i,j})$. We augment the problem of strategic learning with a potentially erroneous advice $\tilde{f} \in \mathcal{F}$ about the global risk minimizer that is given as input to the mechanism, in addition to the labeled data reported by the agents.

We first consider regression problems with the absolute loss function $\ell(x, y) = |x - y|$. For the classes of constant and homogeneous linear functions, Dekel et al. (2010) gave deterministic and group-strategyproof mechanisms that achieved 3 approximation to the minimum global risk and showed that these guarantees were tight. We provide the following results for regression over constant functions:

- Introduce a deterministic and strategyproof mechanism that is, for any $\gamma \in (0, 2]$, $1+\gamma$ consistent and $1 + 4/\gamma$ robust,
- Show that this mechanism is group strategyproof when agents have unique personal risk minimizers, and

- Prove that the consistency-robustness tradeoff is tight under mild assumptions.

This mechanism and its guarantees also extend to homogeneous linear functions. We then consider binary classification problems over the 0-1 loss function in the shared input setting, where agents have identical points but may disagree on their labels, and the function class is a set of specific labelings for the points. Meir et al. (2012) give a deterministic mechanism that achieves a $2n-1$ approximation and show that no deterministic mechanism can achieve a sublinear approximation. They also give a 3 approximate randomized mechanism, which they also show to be tight. We provide the following results:

- Any deterministic and $o(n)$ consistent mechanism has unbounded robustness, and
- Any random mechanism with consistency better than 3 is $\Omega(n)$ robust.

For the special case of function classes with only two labelings, we extend the deterministic mechanism for regression and its guarantees, as well as provide a randomized mechanism parametrized by $\gamma \in (0,1]$ that is $1+\gamma$ consistent and $1+1/\gamma$ robust.[1]

Keywords: Mechanism design, algorithms with predictions, strategyproof learning

Acknowledgments

We thank the anonymous reviewers for their constructive comments. The work of E. Balkanski and C. Zhu was supported in part by the National Science Foundation [Grants CCF-2210501 and IIS-2147361].

References

Priyank Agrawal, Eric Balkanski, Vasilis Gkatzelis, Tingting Ou, and Xizhi Tan. Learning-augmented mechanism design: Leveraging predictions for facility location. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 497–528, 2022.

Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 6–25, 2021.

Eric Balkanski and Cherlin Zhu. Strategyproof learning with advice. *arXiv preprint arXiv:2411.07354*, 2024.

Eric Balkanski, Vasilis Gkatzelis, and Xizhi Tan. Strategyproof scheduling with predictions. In *14th Innovations in Theoretical Computer Science Conference*, 2023.

Eric Balkanski, Vasilis Gkatzelis, and Golnoosh Shahkarami. Randomized strategic facility location with predictions. *arXiv preprint arXiv:2409.07142*, 2024a.

Eric Balkanski, Vasilis Gkatzelis, Xizhi Tan, and Cherlin Zhu. Online mechanism design with predictions. In *25th ACM Conference on Economics and Computation*, 2024b.

Ben Berger, Michal Feldman, Vasilis Gkatzelis, and Xizhi Tan. Optimal metric distortion with predictions. In *25th ACM Conference on Economics and Computation*, 2024.

[1] Extended abstract. Full version appears as (Balkanski and Zhu, 2024).

Ioannis Caragiannis and Georgios Kalantzis. Randomized learning-augmented auctions with revenue guarantees. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

Felipe Caro and Jérémie Gallien. Inventory management of a fast-fashion retail network. *Operations research*, 58(2):257–273, 2010.

Felipe Caro, Jérémie Gallien, Miguel Díaz, Javier García, José Manuel Corredoira, Marcos Montes, José Antonio Ramos, and Juan Correa. Zara uses operations research to reengineer its global distribution process. *Interfaces*, 40(1):71–84, 2010.

Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26, 2018.

Riccardo Colini-Baldeschi, Sophie Klumper, Guido Schäfer, and Artem Tsikiridis. To trust or not to trust: Assignment mechanisms with predictions in the private graph model. *arXiv preprint arXiv:2403.03725*, 2024.

Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. Truthful linear regression. In *Conference on Learning Theory*, pages 448–483. PMLR, 2015.

Ofer Dekel, Felix Fischer, and Ariel D Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016.

Pinyan Lu, Zongqi Wan, and Jialin Zhang. Competitive auctions with imperfect predictions. In *25th ACM Conference on Economics and Computation*, 2024.

Reshef Meir and Jeffrey S Rosenschein. Strategyproof classification. *ACM SIGecom Exchanges*, 10(3):21–25, 2011.

Reshef Meir, Ariel D Procaccia, and Jeffrey S Rosenschein. Strategyproof classification under constant hypotheses: A tale of two functions. In *AAAI*, volume 8, pages 126–131, 2008.

Reshef Meir, Ariel D Procaccia, and Jeffrey S Rosenschein. On the limits of dictatorial classification. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 609–616. Citeseer, 2010.

Reshef Meir, Ariel D Procaccia, and Jeffrey S Rosenschein. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012.

Javier Perote and Juan Perote-Pena. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47(2):153–176, 2004.

Chenyang Xu and Pinyan Lu. Mechanism design with predictions. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 571–577, 2022.