Learning in Stackelberg Games with Non-myopic Agents

NIKA HAGHTALAB, University of California, Berkeley
THODORIS LYKOURIS, Massachusetts Institute of Technology
SLOAN NIETERT, Cornell University
ALEXANDER WEI, University of California, Berkeley

Stackelberg games are a canonical model for strategic principal-agent interactions. Consider, for instance, a defense system that distributes its security resources across high-risk targets prior to attacks being executed; or a tax policymaker who sets rules on when audits are triggered prior to seeing filed tax reports; or a seller who chooses a price prior to knowing a customer's proclivity to buy. In each of these scenarios, a *principal* first selects an action $x \in \mathcal{X}$ and then an *agent* reacts with an action $y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the principal's and agent's action spaces, respectively. In the examples above, agent actions correspond to which target to attack, how much tax to pay to evade an audit, and how much to purchase, respectively. Typically, the principal wants an x that maximizes their payoff when the agent plays a best response $y = \mathrm{br}(x)$; such a pair (x, y) is a *Stackelberg equilibrium*. By *committing* to a strategy, the principal can guarantee they achieve a higher payoff than in the fixed point equilibrium of the corresponding simultaneous-play game. However, finding such a strategy requires knowledge of the agent's payoff function.

When faced with unknown agent payoffs, the principal can attempt to learn a best response via repeated interactions with the agent. If a (naïve) agent is unaware that such learning occurs and always plays a best response, the principal can use classical online learning approaches to optimize their own payoff in the stage game. Learning from *myopic* agents has been extensively studied in multiple Stackelberg games, including security games [2, 6, 7], demand learning [1, 5], and strategic classification [3, 4].

However, long-lived agents will generally not volunteer information that can be used against them in the future. This is especially the case in online environments where a learner seeks to exploit recently learned patterns of behavior as soon as possible, and the agent can see a tangible advantage for deviating from its instantaneous best response and leading the learner astray. This trade-off between the (statistical) efficiency of learning algorithms and the perverse incentives they may create over the long-term brings us to the main questions of this work:

What are principled approaches to learning against non-myopic agents in general Stackelberg games?
How can insights from learning against myopic agents be applied to learning in the non-myopic case?

Non-myopic long-lived agents are typically modeled as receiving γ-discounted utility in the future; the discount factor can be interpreted as capturing the agent's uncertainty on whether she will participate in future rounds or the present bias she may experience. This modeling choice works particularly well for designing algorithms that are *slow* to implement lessons learned from each individual round of feedback. These algorithms make it less appealing for the agent to sacrifice payoff in the present for the effect her actions will have only far into the future and lead to agents that ε-approximately best respond each round.

High-dimensional Stackelberg games pose a unique hurdle because the set of actions that can be rationalized by an ε-approximate best responding agent can be complex. A key challenge is to design *robust learning algorithms* that can learn from erroneous best-responses. This is made worse by discontinuity in the principal's payoff function, the existence of large regions where the agent does not always best respond, and the difficulty of identifying well-behaved optimization subproblems in large Stackelberg games. Furthermore, the statistical efficiency of the principal's learning algorithm must be traded off against its effectiveness to lead an agent to

EC '22, July 11–15, 2022, Boulder, CO, USA

ACM ISBN 978-1-4503-9150-4/22/07.
https://doi.org/10.1145/3490486.3538308

ε-best respond. Therefore, another challenge is to devise principled approaches to *optimally slow down* robust learning algorithms in order to support and encourage ε-best responding behavior.

We seek principal learning algorithms that achieve vanishing regret with respect to the Stackelberg equilibrium, in the presence of non-myopic agents. We provide a framework to address both of the aforementioned challenges, namely designing robust online algorithms that learn effectively from erroneous best response queries in high dimensions, and identifying mechanisms that incentivize agents to approximately best respond. Our framework reduces learning in presence of non-myopic agents to *robust bandit optimization*. We envision principal-agent interactions as taking place over an information channel that is regulated by a third party who acts as an information screen. For example, the third party can *delay the revelation* of the agent's actions or release the principal's actions *in delayed batches*. Combined with the agent's discounted future utility, these induce the agent to ε-approximately best-respond. This information screen allows us to discuss the trade-off between the learning algorithm's statistical efficiency and agent's incentives to deviate. In each of the applications of interest, approximate best-responses lead to different error types that we must be robust to.

- For Stackelberg security games, we devise an algorithm, Clinch, for the myopic setting achieving near-optimal query complexity $\widetilde{O}(n)$, which improves upon the state-of-the-art $O(n^3)$ complexity and is of independent interest even beyond non-myopia. Our algorithm seamlessly extends to the non-myopic setting via our aforementioned framework with a bounded-region adversarial error type.
- For demand learning, we observe that we need robustness to pointwise adversarial perturbations. We show that ActiveArmElimination can be robustified in this manner and adapted to a batched bandits setting.
- For strategic classification, we prove that an existing non-myopic algorithm translates to our setting using the robustness of "gradient descent without a gradient" to pointwise adversarial error.
- For general finite Stackelberg games, we extend a multiple LP approach to our setting using robust algorithms for convex optimization with membership oracles, under bounded-region adversarial error.

The full version of this paper is available at https://eecs.berkeley.edu/~nika/pubs/nonmyopic.

CCS Concepts: • **Theory of computation** → **Convergence and learning in games**.

Additional Key Words and Phrases: Stackelberg games, security games, bandit optimization, non-myopic learning

ACM Reference Format:
Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. 2022. Learning in Stackelberg Games with Non-myopic Agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation (EC '22), July 11–15, 2022, Boulder, CO, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3490486.3538308

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation under grant CCF-2145898 and Graduate Research Fellowships DGE-1650441 and DGE-2146752.

REFERENCES

[1] Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.
[2] Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. In *Advances in Neural Information Processing Systems*, volume 27, pages 1826–1834, 2014.
[3] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pages 15265–15276, 2020.
[4] 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*, page 55–70, 2018.
[5] Robert D. Kleinberg and Frank Thomson Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Symposium on Foundations of Computer Science*, pages 594–605, 2003.
[6] Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International Symposium on Algorithmic Game Theory*, pages 250–262. Springer, 2009.
[7] Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2149–2156, 2019.