CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 17: Learning in Stackelberg Games

October 28, 2025

Lecturer: Nika Haghtalab *Readings: See References*
Scribe: -

1 Lecture Overview

In the previous lecture we defined the *Stackelberg equilibrium* (SE) and showed it can be computed by solving, for each follower action $j \in [m]$, a linear program that maximizes the leader's payoff subject to j being a follower best response. This was called the *Multiple LP* (MLP) approach to solving Stackelberg equilibrium.

The MLP computation requires two types of knowledge. First, the objective of each linear program,

$$\max_{p \in \Delta([n])} \sum_{i \in [n]} p_i \, u_1(i,j),$$

requires knowing the leader's utility u_1. Second, the constraints of the linear program,

$$\forall j' \in [m] : \sum_{i \in [n]} p_i \big(u_2(i,j) - u_2(i,j')\big) \geq 0,$$

require knowledge of the follower's utility u_2. In this lecture, we consider scenarios where the leader (who is interested in computing the Stackelberg equilibrium) naturally knows her own utility function u_1 but may not know the follower's utility u_2. The high-level motivating question is:

Can we learn a Stackelberg-optimal strategy without knowing u_2 a priori, while observing only the follower's reactions (e.g., best responses) to a small number of selected leader strategies?

Notation. As in the last lecture, for each $j \in [m]$ define the polytope

$$P_j := \Big\{ p \in \Delta([n]) : \ \forall j' \in [m], \ \sum_{i \in [n]} p_i \big(u_2(i,j) - u_2(i,j')\big) \geq 0 \Big\},$$

i.e., the set of leader mixed strategies that induce follower best response j under optimistic tie-breaking. The LP for action j is

$$\text{LP}(j): \quad \max_{p \in \Delta([n])} \sum_{i \in [n]} p_i \, u_1(i,j) \quad \text{s.t.} \quad p \in P_j. \tag{1}$$

2 Revisiting: Learning in Zero-Sum Games

We start with the zero-sum game setting. This may appear odd, since zero-sum games are a special case where the Stackelberg value equals the minimax value (the value of the game). Nevertheless, it is a useful exercise.

Assume we play $p^1, \dots, p^T$ and the follower best responds to each: at time t, the follower plays $q^t = \mathrm{BR}(p^t)$. The leader observes only the realized payoff $u_1(p^t, q^t)$—not q^t itself and not u_2. Can the leader learn an *approximate* Stackelberg equilibrium? Yes. This is the same dynamics used in the standard proof of the minimax theorem: as long as the leader uses a no-external-regret algorithm to choose p^t, the time-average strategy $\bar{p} = \frac{1}{T}\sum_{t=1}^{T} p^t$ is an approximate maxmin strategy, satisfying

$$u_1\big(\bar{p}, \mathrm{BR}(\bar{p})\big) = \min_q u_1(\bar{p}, q) \;\geq\; \underbrace{\max_p \min_q u_1(p, q)}_{\text{game value}} \;-\; \frac{\text{Regret}}{T},$$

where the last step was implied by our proof of the minmax theorem.

In the bandit setting—where feedback is only $\hat{u}^t(p^t) = u_1\big(p^t, \mathrm{BR}(p^t)\big)$—standard bandit algorithms attain regret $O(\sqrt{Tn})$. Therefore, $T = O\big(\frac{n}{\varepsilon^2}\big)$ rounds suffice to learn an ε-Stackelberg equilibrium in a zero-sum game.

3 Learning Stackelberg Equilibria of General-Sum Games

For general-sum games, external regret is not the right notion, because it compares against a fixed benchmark while holding the sequence of follower responses fixed at $\mathrm{BR}(p^t)$. In contrast, Stackelberg planning compares against the *counterfactual* responses $\mathrm{BR}(p)$ that would have occurred had the leader consistently played p. Thus, we define *Stackelberg regret*.

Definition 3.1 (Stackelberg Regret)**.** Given leader utility u_1 and follower utility u_2, the Stackelberg regret of a leader strategy sequence $p^1, \dots, p^T$ is

$$\text{Stack-Regret} = \max_p \sum_{t=1}^{T} u_1\big(p, \mathrm{BR}(p)\big) \;-\; \sum_{t=1}^{T} u_1\big(p^t, \mathrm{BR}(p^t)\big),$$

where $\mathrm{BR}(p) = \arg\max_{j \in [m]} u_2(p, j)$, with optimistic tie-breaking.

A similar notion applies when the follower utility may vary over time, u_2^t. In that case,

$$\max_p \sum_{t=1}^{T} u_1\big(p, \mathrm{BR}^t(p)\big) \;-\; \sum_{t=1}^{T} u_1\big(p^t, \mathrm{BR}^t(p^t)\big),$$

where $\mathrm{BR}^t(p) = \arg\max_{j \in [m]} u_2^t(p, j)$.

This is as opposed to the External regret where:

$$\text{External-Regret} = \max_p \sum_{t=1}^{T} u_1\big(p, \mathrm{BR}(p^t)\big) \;-\; \sum_{t=1}^{T} u_1\big(p^t, \mathrm{BR}(p^t)\big).$$

These notions differ from each other precisely because the follower is assumed to best respond to the *counterfactual* leader strategy, not best-responding to historical p^ts. Indeed there is a natural incompatibility between Stackelberg Regret and External Regret: In some games, any sequence with $o(T)$ Stackelberg regret has $\Omega(T)$ external regret and any sequence with $o(T)$ external regret has $\Omega(T)$ Stackelberg regret.

3.1 Achieving No Stackelberg Regret

We do *not* know u_2. For each j, however, we *can* query membership in P_j by committing to some p and checking whether $\mathrm{BR}(p) = j$:

$$\textbf{Membership oracle for } P_j: \quad \text{Query } p \in \Delta([n]);\ \text{Answer: "Yes" iff } p \in P_j.$$

With optimistic tie-breaking this is exactly the event $j = \mathrm{BR}(p)$.

It is well known that a *separation oracle* suffices to solve linear programs over convex sets. A separation oracle is one that, in addition to stating Yes and No, whenever $p \notin P_j$ (with some margin) it produces a halfspace that separates p from P_j

In our learning setting we only have a *membership oracle*, which answers Yes/No but provides no separating hyperplane. Interestingly, one can simulate an ε-separation oracle and optimize a linear/convex function with using only polynomially many membership queries [Grötschel et al., 2012, Lee et al., 2018]. We use this to establish that finite Stackelberg games can be learned with polynomially many best-response observations.

Theorem 3.2 (Linear optimization from membership, [**?**])**.** *Let $C \subset \mathbb{R}^d$ be full-dimensional with aspect ratio κ (i.e., $rB \subseteq C \subseteq RB$ with $\kappa = R/r$ known up to polynomial factors), and suppose we have a membership oracle for C and a strictly interior point $x_0 \in C$. For any $c \in \mathbb{R}^d$ and $\varepsilon > 0$, one can compute $\hat{x} \in C$ with*

$$c^\top \hat{x} \geq \max_{x \in C} c^\top x - \varepsilon$$

using $O\big(d^2 \log(\kappa/\varepsilon)\big)$ membership queries and polynomial additional time.

Remark 3.3 (Warm start and well-roundedness). The theorem requires (i) a well-roundedness parameter κ and (ii) an interior seed x_0 for each region P_j. Condition (i) need not hold in all Stackelberg games: there may be degenerate instances where any small deviation from $p^\star$ changes the best response and the resulting value significantly. Here we restrict attention to Stackelberg games where each P_j has sufficient volume. In such cases, condition (ii) (finding $x_0 \in P_j$) can be met by random sampling in expected $O\big(\frac{\mathrm{Vol}(\Delta([n]))}{\mathrm{Vol}(P_j)}\big)$ tries.

For several structured Stackelberg games, there is no need to assume non-empty interior explicitly. For example, in Stackelberg security games, a monotonicity property can be leveraged so that warm starts are unnecessary. At a high level: if target i is attacked under coverage $p^\star$, then setting $p_i = 0$ while keeping other coordinates unchanged still induces i to be attacked. See the CLINCH algorithm of Haghtalab, Lykouris, Neitert, and Wei Haghtalab et al. [2022] for more information.

Putting it together for MLP. Assuming that a warm start can be found for all regions P_j with polynomial number of queries and that the well-roundedness condition holds. Using feasible region P_j and the linear objective in (1), by Theorem 3.2 with $d = n$ and $C = P_j$ we obtain:

Corollary 3.4. *Assuming the well-roundedness condition and that a warm start for every region P_j can be found in polynomially many queries, an ε-approximate Stackelberg-optimal strategy can be found using an additional*

$$O\big(m\, n^2 \log(1/\varepsilon)\big)$$

membership queries to the sets P_j *(suppressing dependence on well-roundedness and warm-start parameters). Each membership query is exactly one follower best-response observation to a played mixed stratety of the leader.*

4 A Bandit / Semi-Bandit Perspective

More generally, learning in Stackelberg games fits naturally into a bandit framework with additional structure. Let

$$g(p) := u_1\big(p, \mathrm{BR}(p)\big), \qquad \mathrm{BR}(p) = \arg\max_{j \in [m]} u_2(p, j).$$

Learning with unknown u_2 becomes bandit (or semi-bandit) optimization of g. Any method that provably optimizes such a g under the available feedback effectively finds a Stackelberg equilibrium.

The feedback and structure of g depend on the game class. For finite Stackelberg games, $g(p)$ is piecewise linear with m pieces in n dimensions. In addition to observing the value $g(p)$ (pure bandit feedback), we also observe the identity of the active piece via $\mathrm{BR}(p)$ (semi-bandit feedback). The approach we gave above is a bandit optimization algorithm tailored to this piecewise-linear structure, where we get to see which piece is active.

For Stackelberg security games, g has additional properties that CLINCH leverages to learn with even fewer assumptions and fewer number of observations.

Stackelberg and other bilevel optimization problems are also studied in continuous settings under Lipschitzness, where bandit and semi-bandit methods for Lipschitz functions are applicable.

When faced with a Stackelberg or bilevel optimization problem, it is often fruitful to first view the task as optimizing $g(p)$ through a bandit or semi-bandit lens, and to identify structural properties that can reduce the number of best-response observations beyond what vanilla bandit optimization would require.

References

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. Learning in stackelberg games with non-myopic agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 917–918, 2022.

Yin Tat Lee, Aaron Sidford, and Santosh S Vempala. Efficient convex optimization with membership oracles. In *Conference On Learning Theory*, pages 1292–1294. PMLR, 2018.