CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 18: Learning in Stackelberg Games II

October 30, 2025

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

1 Lecture Overview

In the previous lecture, we studied how a leader who initially does not know u_2 can still learn to find an approximate Stackelberg equilibrium by experimenting, playing some p^ts, and observing the responses $BR(p^t)$. While this is quite a powerful method and message when the strategies of the leader are fully observable to the followers at every time step (e.g., published, announced, or otherwise verifiably committed), learning in Stackelberg games occupies an interesting space juxtaposing several different forces at play.

On one hand, the point of Stackelberg games is that a leader publicly and verifiably *commits* to a mixed strategy, thereby inducing the actions of a follower that are best responses. On the other hand, a leader who is learning to optimize her strategy requires exploration and hedging, which make her sequence of actions involve changing strategies, which seems antithetical to *commitment*! This means that a follower may not have caught on to what the leader's strategy is at any given time step and therefore may not be able to perfectly best respond to the leader's strategy.

In this lecture, we ask: what happens in a Stackelberg game when the *follower is not a perfect best responder* each round, but instead is learning how to respond to this ever-changing sequence of leader actions using an online learning algorithm? Can the leader exploit the learning dynamics? Is there a learning algorithm that ensures the follower won't be exploited?

As we will see in today's lecture:

1. If the follower guarantees only **no *external* regret**, then there are games in which the leader can exploit the follower. That is, there exists a sequence of leader strategies in which the leader does *strictly better* than her Stackelberg value while the follower does strictly *worse* than the value guaranteed to them in the Stackelberg equilibrium.

2. If the follower guarantees **no *swap* regret** (a stronger requirement), then *no matter what the leader does*, the leader cannot exceed her Stackelberg value by more than $o(T)$.

Both of these results are established in the paper by Deng et al. [2019]. While we won't show this here, it is also good to know that a third message holds:

3. If the follower guarantees **no *swap* regret**, then a leader who does not know u_2 *a priori* can also guarantee herself utility that is at least her Stackelberg value minus $o(T)$.

The last message is from Haghtalab et al. [2023].

2 Setup and Notation

We consider a two-player Stackelberg game repeated over T rounds.

- Leader/principal action set A_1; follower/agent action set A_2.
- Utilities $u_1(a_1, a_2)$ for the leader and $u_2(a_1, a_2)$ for the follower. We overload these terms to also denote the expected utility of mixed strategy profiles.
- The leader can commit to a mixed strategy $p \in \Delta(A_1)$. The follower's best response is $\mathrm{BR}(p) \in A_2$, where ties are broken *in favor of the leader*.

The Stackelberg value for the leader is defined as

$$V := \max_{p \in \Delta(A_1)} u_1(p, \mathrm{BR}(p)),$$

where the best-response set $\mathrm{BR}(p)$ and the tie-breaking rule are as above. If $p^\star$ is a leader strategy achieving the maximum, the follower's Stackelberg payoff is

$$V' := u_2(p^\star, \mathrm{BR}(p^\star)).$$

3 Exploiting No-External-Regret Followers

Consider the following payoff matrix, written as (u_1, u_2). Columns are follower actions; rows are leader actions:

	Quit	Left	Right
Up	$(0, 0)$	$\left(\frac{1}{2}, -\frac{1}{6}\right)$	$(1, -\frac{1}{2})$
Down	$(0, 0)$	$(0, \frac{1}{3})$	$(0, \frac{1}{2})$

3.1 The Stackelberg Equilibrium and Payoffs

Let $p \in [0, 1]$ denote the leader's probability of playing Up (so $1 - p$ is Down). Using the matrix:

$$\begin{aligned} u_2(p, \textsf{Left}) &= p\cdot(-\tfrac{1}{6}) + (1-p)\cdot(\tfrac{1}{3}) \;=\; \tfrac{1}{3} - \tfrac{p}{2}, \\ u_2(p, \textsf{Right}) &= p\cdot(-\tfrac{1}{2}) + (1-p)\cdot(\tfrac{1}{2}) \;=\; \tfrac{1}{2} - p, \\ u_2(p, \textsf{Quit}) &= 0. \end{aligned}$$

Therefore the best-response map is:

$$\mathrm{BR}(p) = \begin{cases} \textsf{Right}, & p \le \frac{1}{3}, \\ \textsf{Left}, & \frac{1}{3} < p \le \frac{2}{3}, \\ \textsf{Quit}, & p > \frac{2}{3}. \end{cases}$$

Leader's Stackelberg value in this game. When the follower plays Right, the leader's expected utility is

$$u_1(p, \mathsf{Right}) = p \cdot 1 + (1-p) \cdot 0 = p.$$

Hence, the leader chooses the largest p that still induces Right, namely $p^\star = \frac{1}{3}$, so

$$V \;=\; u_1\big(\tfrac{1}{3}, \mathsf{Right}\big) \;=\; \tfrac{1}{3}.$$

The follower's Stackelberg payoff at $p^\star$ is

$$V' \;=\; u_2\big(\tfrac{1}{3}, \mathsf{Right}\big) \;=\; \tfrac{1}{6}.$$

3.2 Exploiting the Follower

We next specify a sequence of leader mixed strategies p^ts and follower best response b^ts. We establish that 1) b^t have no-external regret for the follower, 2) follower's accumulated utility is significantly less than $V'T$ and the leader's is significantly more, 3) and b^t are not arbitrary, rather they are the result of playing a natural class of no-regret algorithms, called mean-based algorithms – which do not play actions that have historically been dominated with non-negligible probability.

The sequence: For the leader, $p^t = (\frac{1}{3}, \frac{2}{3}) = p^\star$ for $t = 1, \ldots, T/2$, and $p^t = (1, 0)$ for $t = T/2 + 1, \ldots, T$. For the follower, $b^t = \mathsf{right}$ for $t = 1, \ldots, T/2$ and then $b^t = \mathsf{left}$ for $t = T/2+1, \ldots, T$.

Leader's utility: In this sequence, the leader's utility is $\sum_t u_1(p^t, b^t) = \frac{T}{2}\frac{1}{3} + \frac{T}{2}\frac{1}{2} = \frac{5}{12}T > VT + \Omega(T)$.

Follower has no external regret: Note that

$$\sum_{t=1}^{T} u_2(p^t, b^t) = \frac{T}{2} \times \frac{1}{6} + \frac{T}{2} \times (-\frac{1}{6}) = 0 < V'T - \Omega(T).$$

Moreover, for any fixed $b \in A_2$, the utility of the follower who plays b on every time step is at most 0. That is,

$$\sum_{t=1}^{T} u_2(p^t, \mathsf{quit}) = 0.$$

$$\sum_{t=1}^{T} u_2(p^t, \mathsf{left}) = \frac{T}{2} \times (\frac{1}{3} - \frac{p^\star}{2}) + \frac{T}{2} \times (\frac{1}{3} - \frac{1}{2}) = 0$$

$$\sum_{t=1}^{T} u_2(p^t, \mathsf{right}) = \frac{T}{2}(\frac{1}{2} - \frac{1}{3}) - \frac{T}{2}(\frac{1}{2} - 1) < 0.$$

The follower's response is consistent with a mean based algorithm. The figure below displays the follower's utility for any fixed action over time. In the first $T/2$ steps, both right and left are optimal for the follower. A follower who takes the historically optimal action up to that point will choose right, breaking ties in favor of the leader. In the last $T/2$ steps, the action left becomes historically optimal; therefore, the follower's choice in every step remains historically optimal.

Mean-based algorithms are precisely those algorithms that avoid playing historically suboptimal actions with non-negligible probability. Thus, this sequence of follower actions is consistent with the behavior of mean-based algorithms.

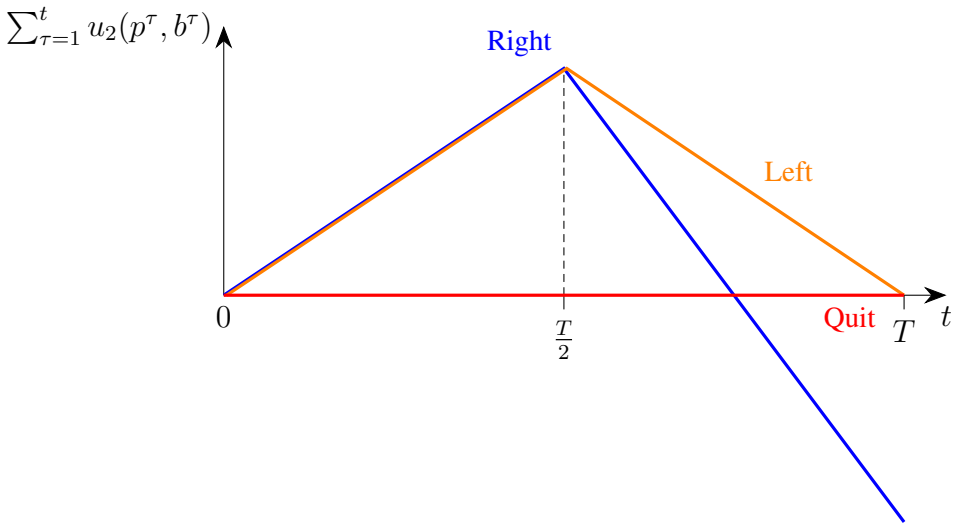

Figure 1: Cumulative utility of follower action for any fixed action.

3.3 Mean-Based Algorithms.

Definition 3.1. Any algorithm with property that if an action i is such that some actions j,

$$\sum_{\tau=1}^{t-1} \left(u_2^\tau(j) - u_2^\tau(i)\right) \geq \gamma T,$$

then the algorithm plays i with probability $p^\tau(i) \leq \gamma$, where $\gamma \in o(1)$.

Note that Hedge, MWU, and Follow the Perturbed Leader are all is mean-based algorithms. Indeed most natural no-regret algorithms that one thinks about are mean-based.

4 No-swap regret followers cannot be exploited

First note that the sequence in the previous section, the follower has swap regret. In particular, she regret not swapping to quit — which gives her 0 utility per round — every-time she played

left which gave her negative utility per round. We show below that this is the only reason she was vulnerable to exploitation.

Theorem 4.1. *If the follower guarantees* no swap regret*, then for* any *sequence of leader strategies* $p^1, \ldots, p^t$*,*

$$\sum_{t=1}^{T} u_1(p^t, b^t) \leq V \cdot T + o(T).$$

Proof. Let $a^t \in \Delta(A_1)$ denote the leader's action at time t, and let $b^t \in A_2$ be the follower's realized action. For each follower action b, define the *conditional average leader strategy*

$$\alpha^b(i) := \frac{\sum_{t=1}^{T} \Pr[b^t = b] 1(a^t = i)}{\sum_{t=1}^{T} \Pr[b^t = b]} \qquad \text{(whenever the denominator is nonzero).}$$

That is $\alpha^b \in \Delta(A_1)$.

Define the *bad* set $B \subseteq A_2$ of follower actions that are *not* best responses to these conditional mixtures:

$$B := \{\, b \in A_2 : \; b \notin \mathrm{BR}(\alpha^b) \,\}.$$

For each $b \in B$, let $\mathrm{Swap}(b) \in \arg\max_{b' \in A_2} u_2(\alpha^b, b')$ be a best response to α^b, and define the suboptimality margin

$$\delta := \min_{b \in B} u_2(\alpha^b, \mathrm{Swap}(b)) - u_2(\alpha^b, b) > 0.$$

Using swap regret. Consider the particular *swap mapping* that sends each $b \in B$ to $\mathrm{Swap}(b)$ defined above and leaves $b \notin B$ unchanged. The follower's swap regret against this mapping equals

$$\begin{aligned} \mathrm{SwapRegret}_{\mathrm{swap}} &= \sum_{b \in B} \big(u_2(\alpha^b, \mathrm{Swap}(b)) - u_2(\alpha^b, b)\big) \left(\sum_t \Pr[b^t = b]\right) \\ &\geq \delta \sum_{b \in B} \sum_t \Pr[b^t = b]. \end{aligned}$$

This gives

$$\sum_{b \in B} \sum_{t=1}^{T} \Pr[b^t = b] \in o(T), \tag{1}$$

as the follower has no swap regret (and δ is a constant).

Bounding the leader's payoff. Split the rounds into "good" ($b^t \notin B$) and "bad" ($b^t \in B$). On a good round $b \notin B$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} u_1(a^t, b) 1(b^t = b)\right] = \underbrace{u_1(\alpha^b, b)}_{= u_1(\alpha^b, \mathrm{BR}(\alpha^b)) \leq V} \cdot \sum_{t=1}^{T} \Pr[b^t = b]. \tag{2}$$

So on the whole

$$\sum_{t=1}^{T} u_1(p^t, b^t) = \mathbb{E}\left[\sum_{b\in B} u_1(a^t, b)\left(\sum_{t=1}^{T} 1(b^t = b)\right)\right] + \mathbb{E}\left[\sum_{b\notin B} u_1(a^t, b)\left(\sum_{t=1}^{T} 1(b^t = b)\right)\right] \leq o(T) + V \cdot T.$$

where the first term is bounded by Equation 1 and the right term is bounded by 2. □

References

Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. *Advances in neural information processing systems*, 32, 2019.

Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36:61645–61677, 2023.