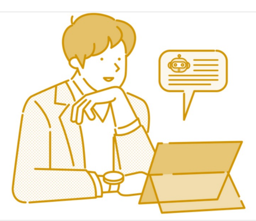
Impact of Decentralized Learning on Player Utilities in Stackelberg Games

Kate Donahue (Cornell), Nicole Immorlica (MSR), Meena Jagadeesan (UC Berkeley), Brendan Lucier (MSR), Alex Slivkins (MSR)

(Authors in alphabetical order)

Motivating examples



User chooses prompts

Chatbot chooses outputs

Rec sys selects slate

User selects content from slate

- **Sequential:** one player goes first
- **Misaligned:** players have different utilities
- **Decentralized learning:** players learn best action while only observing their own utility.

Main questions: How quickly do these two agent systems learn over time? What are the implications of algorithm design on each player's utility?

Model: decentralized Stackelberg

The static environment is a **Stackelberg game**.

- Action spaces: A = leader, B = follower
- Utility: u_1 = leader, u_2 = follower

Best response:

- Follower: $b^*(a) = argmax_{b \in B}(u_2(a,b))$
- Leader: $a^* = argmax_{a \in A}(u_1(a, b^*(a))$

Our setup:

Leader	**Follower**
ALG_1 = **bandit algorithm**	ALG_2 = **bandit algorithm**

At each time step t:

Chooses a_t using ALG_1	Observes a_t & chooses b_t using ALG_2
Stochastic reward $u_1(a_t, b_t)$ + noise	**Stochastic reward** $u_2(a_t, b_t)$ + noise

Cumulative reward: $\sum_{t=1}^{T} u_1(a_t, b_t)$

Cumulative reward: $\sum_{t=1}^{T} u_2(a_t, b_t)$

Our goal: low regret for **both** leader and follower.

Impossibility of Stackelberg benchmarks

Original Stackelberg benchmarks: utility at Stackelberg equilibrium

- $\alpha_1^{orig} := u_1(a^*, b^*(a^*))$ and $\alpha_2^{orig} := u_2(a^*, b^*(a^*))$

Theorem (Informal): For any pair of algorithms, **at least one player incurs linear regret** w.r.t. their original Stackelberg benchmark on one of the following two instances.

	b_1	b_2
a_1	$(0.6, \delta)$	$(0.2, *)$
a_2	$(0.5, 0.6)$	$(0.4, 0.4)$

Two instances:
$* = 0$ (SV = 0.6, δ)
versus
$* = 2\delta$ (SV = 0.5, 0.6)

Our error-tolerant benchmarks

Tolerant to the other player's errors due to learning.

Definition (benchmarks):

$$\alpha_1^{tol} := \inf_{\epsilon \le \gamma} \left(\max_{a \in A} \min_{b \in B_\epsilon(a)} u_1(a,b) + \epsilon \right)$$

$$\alpha_2^{tol} := \inf_{\epsilon \le \gamma} \left(\min_{a \in A_\epsilon} \max_{b \in B} u_2(a,b) + \epsilon \right)$$

γ = tolerance parameter

worst-case error level; ϵ-relaxed Stackelberg utility; ϵ-regularizer

ϵ-tolerant response sets:

$$B_\epsilon(a) := \left\{ b \in B \mid u_2(a,b) \ge \max_{b' \in B} u_2(a,b') - \epsilon \right\}$$

$$A_\epsilon := \left\{ a \in A \mid \max_{b \in B_\epsilon(a)} u_1(a,b) \ge \max_{a' \in A} \min_{b' \in B_\epsilon(a')} u_1(a',b') - \epsilon \right\}$$

Regret bounds w.r.t. new benchmarks

Negative: Both players running ExploreThenCommit leads to linear regret for both.

Positive: algorithms where both players achieve sublinear regret:

Theorem (Informal): When the leader runs **ExploreThenUCB** and the follower has low high-probability instantaneous regret, then **both players achieve $\tilde{O}(T^{\frac{2}{3}})$ regret** w.r.t. their error-tolerant benchmark.

Key algorithmic idea: the leader waits for the follower to sufficiently converge ("Explore") before starting to learn ("then UCB").

Permits flexibility in the follower's choice of algorithm

Faster learning? Not in general

Theorem (Informal): For any pair of algorithms **at least one player incurs $\Omega(T^{\frac{2}{3}})$ regret** w.r.t. their error tolerant benchmark on one of the following two instances.

	b_1	b_2
a_1	$(0.5+\delta, \delta)$	$(0, *)$
a_2	$(0.5, 3\delta)$	$(0.5, 3\delta)$

Two instances:
$* = 0$ ($\alpha_1^{tol} = 0.5 + \delta, \alpha_2^{tol} = \delta$) versus
$* = 2\delta$ ($\alpha_1^{tol} = 0.5$, $\alpha_2^{tol} = 3\delta$)

Faster learning in relaxed settings

Setting 1: Continuity condition

Result (Informal): If players agree on which actions are similar in reward (Lipschitz condition), then both players can achieve $O(\sqrt{T})$ regret w.r.t. their original Stackelberg benchmark.

Setting 2: Weaker benchmarks

Result (Informal): Consider **self-tolerant benchmarks** where players are also tolerant to their own errors. Then, both players can achieve $O(\sqrt{T})$ regret with respect to their self-tolerant benchmark.

Summary and Discussion

We proposed a model for two-agent sequential, misaligned environments with decentralized learning.

- **Our focus: how learning affects both player's utilities.**
- We showed the impossibility of Stackelberg benchmarks.
- We proposed **error-tolerant benchmarks** and constructed algorithms achieving $T^{2/3}$ regret.
- We showed scenarios which permit faster learning.

Selected related works:

Bai, Jin, Wang, Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games . NeurIPS 2021.
Camara, Hartline, Johnsen. "Mechanisms for a no-regret agent: Beyond the common prior". FOCS 2020.
Haghtalab, Podimata, Yang. "Calibrated Stackelberg Games: Learning optimal commitment against calibrated agents." NeurIPS 2023.