Online Learning in Stackelberg Games with an Omniscient Follower

Geng Zhao [1 *] **Banghua Zhu** [1 *] **Jiantao Jiao** [1] **Michael I. Jordan** [1]

Abstract

We study the problem of online learning in a two-player decentralized cooperative Stackelberg game. In each round, the leader first takes an action, followed by the follower who takes their action after observing the leader's move. The goal of the leader is to learn to minimize the cumulative regret based on the history of interactions. Differing from the traditional formulation of repeated Stackelberg games, we assume the follower is omniscient, with full knowledge of the true reward, and that they always best-respond to the leader's actions. We analyze the sample complexity of regret minimization in this repeated Stackelberg game. We show that depending on the reward structure, the existence of the omniscient follower may change the sample complexity drastically, from constant to exponential, even for linear cooperative Stackelberg games. This poses unique challenges for the learning process of the leader and the subsequent regret analysis.

1. Introduction

The multi-agent learning problem (Ferber & Weiss, 1999; Wooldridge, 2009; Filar & Vrieze, 2012; Zhang et al., 2021) has received significant attention reflecting its wide variety of real-world applications, including autonomous driving (Shalev-Shwartz et al., 2016; Sallab et al., 2017) and human-robot interaction (Kober et al., 2013; Lillicrap et al., 2015; Goodrich et al., 2008; Xie et al., 2021). In a multi-agent system, it is natural to assume that each agent possesses a different set of information due to its different viewpoint and history of actions. This phenomenon is commonly referred to as the property of *information asymmetry* (Yang et al., 2022). Such information asymmetry poses challenges to the coordination and cooperation between learning agents.

In this paper, we study how the information asymmetry affects the sample complexity of learning a two-player decentralized cooperative repeated Stackelberg game, with a focus on the setting when the follower is omniscient and myopic, and always best-responds to the leader's actions.

Consider an illustrative example in human-robot interaction where a robot is required to collaborate with a human to achieve some shared objective. This can be formulated as a repeated Stackelberg game where the interactions between human and robot happen in multiple rounds, and the human is an omniscient expert who knows the exact target and how to achieve it. In each round, the robot, as the leader who hopes to learn the world model and human behavior from scratch, first takes some action. After seeing the robot's action, the human, as an expert follower who possesses perfect information about the world, always best-responds to the robot's action to maximize their reward. The robot hopes to use as few as possible interactions to learn the world model and human behavior, and eventually find the optimal action that maximizes a shared reward.

Concretely, during each round t of the interaction, the leader first plays an action $a_t \in \mathcal{A}$, and the follower plays another action $b_t \in \mathcal{B}$ upon (perfectly) observing a_t. We assume that the two players share a reward, $r_t = h^\star(a_t, b_t) + z_t$, where $z_t \in \mathbb{R}$ is some zero-mean sub-Gaussian noise, $h^\star$ belongs to a family $\mathcal{H}$. We also assume that the follower has full knowledge of the reward and always best responds with $b_t \in \arg\max_{b\in\mathcal{B}} h^\star(a_t, b)$. However, the leader does not know $h^\star$ and can only explore via taking actions a_t and making inferences from past observations $(a_1, b_1, r_1), \cdots, (a_{t-1}, b_{t-1}, r_{t-1})$.[1] We are interested in providing tight bound for the Stackelberg regret, defined as

$$\mathcal{R}(T) = \max_{a\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{b\in\mathcal{B}} h^\star(a,b) - \max_{b_t\in\mathcal{B}} h^\star(a_t, b_t)\right)\right].$$

The Stackelberg regret characterizes the gap between the reward achieved from the optimal leader action and the reward from the actual leader action a_t.

Compared with the traditional bandit problem, the extra observation of b_t can be viewed as side information accom-

[*]Equal contribution [1]Department of EECS, University of California, Berkeley, USA. Correspondence to: Geng Zhao <gengzhao@berkeley.edu>, Banghua Zhu <banghua@berkeley.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]For simplicity, we assume in the introduction that the leader can see $b_1, \cdots, b_{t-1}$ without noise. Later we generalize to the case when the observed b_t is also noisy.

panying the usual action-reward pair. Depending on how the function family $\mathcal{H}$ and side information b are designed, the complexity of learning for the leader may vary. Here we briefly summarize several illustrative examples where the follower may help or harm the leader's learning process. We will present a general formalization that encompasses these examples in the next section.

1. **Curse of expertise.** Imagine that in a driving system, the self-driving vehicle (leader) and the human driver (follower) work together to avoid collisions. For most of the aggressive actions the leader takes, the final reward for non-collision is high since the human driver will consistently exert efforts to evade the self-driving vehicle in order to prevent collisions. From the leader's point of view, aggressive actions lead to similar outcomes as safe actions. The expertise of the human prevents the leader from learning from failure cases.

2. **Imitation Learning.** Consider an assembly robot (leader) that learns to move goods to a destination with a human expert (follower). This can be modeled by the robot choosing a drop-off location, from which the human expert continues to the correct destination. In this simple example, the robot and the human expert cooperate in a "linear" fashion—the expert can complete whatever the robot leaves undone, and upon observation of the expert's move the robot should simply imitate the behavior of the human expert in the future. This corresponds to an "imitation-based" interaction that can greatly accelerate the learning process.

3. **Expert-guided learning.** In most cases, the self-driving vehicle may have some target that is similar but not exactly the same as the human driver. For example, they both aim to avoid collision while heading to a different target. In this case, a pure imitation-based learning will fail. But the self-driving vehicle can still glean good driving standards from the human driver. With the extra observation of the behavior of human driver, the self-driving vehicle can learn much faster.

Extending beyond robotics applications, our framework is potentially applicable in various repeated cooperative game settings where direct communication is hard, unreliable, or forbidden. For instance, it captures the learning aspect of language models adjusting to human preferences, personalized digital healthcare, or other settings of AI-human interaction where explicit revelation of utility function is difficult: in such settings, the AI system (e.g., the language model, or the digital "doctor") works with a human user to achieve a common goal without direct communication of the true preferences or needs; instead, the system must learn them through repeated interactions with the users.

In this paper, we abstract and formalize these three scenarios into a simple linear Stackelberg game and analyze the sample complexity of this game. We briefly overview our main results in the next section.

1.1. Main results

Contrary to the traditional literature on linear bandits, we show that the worst-case sample complexity for achieving ϵ-Stackelberg regret is at least exponential even when $h^\star$ belongs to the linear family $\mathcal{H}_\phi = \{\theta \cdot \phi(a, b)\}$. The hard instance corresponds to the 'curse of expertise' example discussed above, where the follower's best response hurts the observation, and thus harms the whole learning process.

Theorem 1.1 (Curse of expertise, informal)**.** *There exists some ϕ such that for any algorithm, we can find some $h^\star \in \mathcal{H}_\phi$ with the regret being $\Omega(T^{(d-3)/(d-2)})$.*

This shows that the leader needs an exponential number of samples to learn a good policy even when the reward is linear. We also present an upper bound $\mathcal{O}(T^{(d+1)/(d+2)})$ for linear rewards in Theorem 3.3.

On the other hand, the side information b_t can also greatly improve the sample complexity when the linear family is structured. We provide an Upper Confidence Bound (UCB) based algorithm (Auer et al., 2002) that leads to an improved bound in this setting. In particular, we recover the rate for imitation learning when the leader can simply mimic the behavior of the follower.

Theorem 1.2 (Imitation learning, informal)**.** *There exists some ϕ such that for any $h^\star \in \mathcal{H}_\phi$, when b_t is observed, the leader can achieve regret $\mathcal{O}(\log^2(T))$ by imitating the follower behavior. However, when b_t is not observed, the regret is $\Theta(\sqrt{T})$.*

Similarly, we can also design cases where observing b_t helps reduce the problem to a traditional linear bandit, while not observing b_t suffers from exponential sample complexity.

Theorem 1.3 (Expert-guided, informal)**.** *There exists some ϕ such that for any $h^\star \in \mathcal{H}_\phi$, when b_t is observed, the leader can achieve regret $\mathcal{O}(\sqrt{T})$. However, when b_t is not observed, the regret is $\Omega(T^{(d-4)/(d-2)})$.*

In addition to these three examples, we discuss more complicated scenarios where UCB fails and we show that a careful analysis is necessary to achieve a near-optimal rate. In particular, we establish such a rate for polynomial bandits, where the best-response corresponds to a lower degree polynomial, which helps improve the rate when the noise level for reward and the observed follower behavior is similar.

Theorem 1.4 (Polynomial bandit, informal)**.** *There exists a family of $2k$-degree polynomial, such that the regret is $\Theta(\sqrt{d^{2k-1}T})$ when b_t is observed, and $\Theta(\sqrt{d^{2k}T})$ when b_t is not observed.*

1.2. Related work

Decentralized Stackelberg Games. The problem of repeated Stackelberg games has been studied extensively (von Stackelberg, 2010; Marecki et al., 2012; Lauffer et al., 2022; Kao et al., 2022), in a standard setting where the leader leads and the myopic follower follows with its best response for the current round.

Kao et al. (2022) and Lauffer et al. (2022) study a similar setting to ours, in which a leader and a follower interact through a cooperative Stackelberg game that comprises a two-stage bandit problem. However, Kao et al. (2022) restrict their focus to the tabular case where both $\mathcal{A}$ and $\mathcal{B}$ are finite and the reward $h^\star$ is uncorrelated for different actions (a, b). They also assume that both the leader and the agent are running regret-minimization algorithms independently. They show that the classic upper confidence bound (UCB) algorithm for the multi-arm bandit problem can be used for both the leader and the agent, respectively, to achieve asymptotically optimal performance (i.e., no-regret). However, it is unclear that such results can generalize to bandits with function approximation and the case of omniscient agents. Indeed, our results show that the general case (or even just the linear case) is not always statistically tractable. Note also that Lauffer et al. (2022) show that the regret can depend exponentially on the dimension of the agent's utility.

Other examples of Stackelberg games include Stackelberg security games (Conitzer & Sandholm, 2006; Tambe, 2011), strategic learning (Hardt et al., 2016; Dong et al., 2018; Liu & Chen, 2016), dynamic task pricing (Kleinberg & Leighton, 2003) and online contract design (Ho et al., 2014; Zhu et al., 2022). The problem of online learning in contract theory considers a decentralized general-sum Stackelberg game with omniscient agents. It focuses on a special case where the rewards for the leader and the agent are both linear. It is shown in Zhu et al. (2022) that one has to pay exponential sample complexity in this setting to achieve small regret in the worst case.

Centralized Stackelberg Game. Centralized Stackelberg games are also well studied in the literature (Zhong et al., 2021; Bai et al., 2021; Gerstgrasser & Parkes, 2022; Yu et al., 2022), where the machine learning algorithm has control over both the leader and the follower. Bai et al. (2021) consider the repeated Stackelberg game where both the leader and the agent learn their optimal actions (a Stackelberg equilibrium) from samples. However, they assume a central controller that can determine the actions of both the leader and the agent. Moreover, they rely on an assumption of a bounded gap between the optimal response and an ϵ-approximate best response. In contrast, in our framework, we assume that the agent's utility is unknown, and that the agent always takes the best response.

Bandit with side information. There has been significant effort in studying bandits with side information (Wang et al., 2005; Langford & Zhang, 2007; Foster et al., 2021). Such side information is generally assumed to be available before a decision. Foster et al. (2021) also consider the case when an extra observation is available after taking the actions. However, they mainly focus on the setting of reinforcement learning where the extra observation is the trajectory. Although our observation of follower behavior can also be viewed as side information, it also alters the reward in the Stackelberg game, which changes the structure of the multi-agent problem.

2. Formulation

We consider a two-player cooperative Stackelberg bandit game with an omniscient follower.

Let $\mathcal{A} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{B} \subseteq \mathbb{R}^{d_2}$ be compact sets. Up to a scaling factor, we will assume that $\mathcal{A}$ and $\mathcal{B}$ reside inside the unit ball centered at the origin. During each round $t \in [T]$ of interaction, the leader plays an action $a_t \in \mathcal{A}$, and the follower plays $b_t \in \mathcal{B}$ upon (perfectly) observing a_t. The two players both receive a reward $r_t = h^\star(a_t, b_t) + z_t$, where $z_t \in \mathbb{R}$ is zero-mean σ_r-sub-Gaussian and is independent of all past events. We will make the realizability assumption that $h^\star$ belongs to a (known) family $\mathcal{H}$ of real-valued functions on $\mathbb{B}^{d_1} \times \mathbb{B}^{d_2}$. As is common in the study of bandits, we assume that reward function is bounded, i.e., there exists $C \in (0, \infty)$ such that $0 \le h \le C$ for all $h \in \mathcal{H}$. We assume $C = 1$ throughout the paper unless stated otherwise.

We will assume that the follower, modeled after an expert human player, has full knowledge of the game and can always best respond with an optimal action $b_t \in \arg\max_{b \in \mathcal{B}} h^\star(a_t, b)$. The leader then makes a noisy observation of b_t, given by $\hat{b}_t = b_t + w_t$, where $w_t \in \mathbb{R}^{d_2}$ is zero-mean σ_b-sub-Gaussian (e.g., component-wise σ_b-sub-Gaussian with independent zero-mean coordinates) and independent of all past events.

For convenience, we denote the set of best responses to leader's action a when the ground truth reward function is h by $b_h^*(a)$. Denote $\overline{h}(a) := \max_{b \in \mathcal{B}} h(a, b)$. The optimal action, unbeknownst to the leader, is denoted $a^* := \arg\max_{a \in \mathcal{A}} \overline{h^\star}(a)$.

The leader's objective is to minimize the *regret* during T rounds of interactions, defined as

$$\mathcal{R}(T) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \overline{h}(a) - \overline{h}(a_t)\right]. \tag{1}$$

We will also focus on the sample complexity of achieving low (average) regret; that is, for some $\epsilon, \delta \in [0, 1]$, the

minimal $T \in \mathbb{N}$ such that $\mathcal{R}(T) \leq \epsilon T$.

Notations. We use calligraphic letters for sets and operators, e.g., $\mathcal{A}$. Given a set $\mathcal{A}$, we write $|\mathcal{A}|$ for the cardinality of $\mathcal{A}$. $\mathbb{B}^d$ and $\mathbb{S}^{d-1}$ denote the unit ball and the unit sphere, both centered at the origin, in d-dimensional Euclidean space. Vectors are assumed to be column vectors except for the probability and measure vectors. For a vector $v \in \mathbb{R}^d$ and an integer $i \in \mathbb{N}$, we use v_i to denote the i-th element of v, and v_{-i} to denote the vector of all elements in v except for v_i. For two n-dimensional vectors x and y, we use $x \cdot y = x^\top y$ to denote their inner product. We write $f(x) = \mathcal{O}(g(x))$ or $f(x) \lesssim g(x)$ if there exists some positive real number M and some x_0 such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$. We use $\widetilde{\mathcal{O}}(\cdot)$ to be the big-$\mathcal{O}$ notation ignoring logarithmic factors. We write $f(x) = \Omega(g(x))$ or $f(x) \gtrsim g(x)$ if there exists some positive real number M and some x_0 such that $|f(x)| \geq Mg(x)$ for all $x \geq x_0$. We write $f(x) = \Theta(g(x))$ if we have both $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$. We use $\|\cdot\|_p$ to denote the ℓ^p norm for $p \in (0, \infty]$, with $\|\cdot\|$ denoting the Euclidean (ℓ^2) norm $\|\cdot\|_2$.

Parameterized family. In subsequent discussions, we will consider the parameterized case when $\mathcal{H}$ admits a parameterization over a compact parameter space Θ. The class is denoted by $\mathcal{H}_\Theta = \{h_\theta | \theta \in \Theta\}$. When the parameterization is linear, that is,

$$h_\theta(a, b) = \theta \cdot \phi(a, b) \tag{2}$$

for some feature function $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{B}^d$, we will denote the class by $\mathcal{H}_{\Theta,\phi}$. We denote the true parameter by $\theta^\star$. For instance, when $\mathcal{A}$ and $\mathcal{B}$ are the sets of standard basis vectors in $\mathbb{R}^{|\mathcal{A}|}$ and $\mathbb{R}^{|\mathcal{B}|}$ with $\phi(a, b) = ab^\top$ and θ is bounded in $\mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$, we recover the tabular case model in Kao et al. (2022) with finite action sets. In general, however, we will focus on cases with infinite action sets.

3. Linear Stackelberg games: Curse of expertise

In this section, we study the sample complexity of learning in linear Stackelberg game, where the family of reward is restricted to $\mathcal{H}_{\Theta,\phi}$ for some given Θ and ϕ.

3.1. An exponential lower bound

It is well known that the regret for traditional linear bandits grows as $\Theta(d\sqrt{T})$ (Abbasi-Yadkori et al., 2011). In the case of a linear Stackelberg game, we present a worst-case lower bound on the regret that is exponential in dimensionality for the linear family. This suggests that the leader cannot learn the task well unless in possession of an exponential number of samples even when we restrict to linear Stackelberg games.

Assume the leader makes perfect observations of the follower's responses (i.e., $\sigma_b = 0$). We have the following lower bound.

Theorem 3.1. *For any $d \geq 4$, there exists some ϕ such that, for any algorithm that the leader runs, one can find some instance with $h_\theta \in \mathcal{H}_{\Theta,\phi}$ such that*

$$\mathcal{R}(T) \gtrsim T^{(d-4)/(d-2)}. \tag{3}$$

In other words, the sample complexity for achieving ϵ (average) regret is at least $\Omega\big((1/\epsilon)^{\frac{d-2}{2}}\big)$.

The proof is detailed in Appendix A.1. The worst-case instance presented below can be reduced to the ReLU bandit problem shown below, which is known to suffer from the exponential sample complexity (Dong et al., 2021).

Example 3.2. Let $\mathcal{A} = \mathbb{B}^{d-1}$, $\mathcal{B} = [0, 1]$ and $\Theta = \{\theta \mid \theta_{-d} \in \mathbb{S}^{d-2}, \theta_d = 1 - \Delta\}$ for some $\Delta \in (0, 1)$. Let the feature function be $\phi(a, b) = ((1 - b)a, b)$.

One can verify that in this case, one has

$$\overline{h}_\theta(a) = \max\{1 - \Delta, \theta_{-d} \cdot a\}. \tag{4}$$

Thus when a is chosen far from θ_{-d}, the reward will remain constant.

Theorem 3.1 is no mystery mathematically: the best response may destroy linearity for the leader's observations, imposing a toll. Conceptually, however, the message from the theorem is striking: it highlights a "curse of expertise"; i.e., the potential difficulty to learn with an expert on a decentralized bandit learning task with a large action space. From the classic single-agent bandit learning perspective, the task the two agents aim to solve is straightforward: a linear bandit on an action space $\phi(\mathcal{A}, \mathcal{B})$. In other words, if the expert follower lets the novice leader control the choice of b, the average regret would steadily decrease at a rate of $\widetilde{\mathcal{O}}(d\sqrt{T})$. On the other hand, with a myopic focus, the follower's expertise in best responding ironically results in a significantly higher regret, as it deprives the learner of the ability to explore.

In the context of autonomous driving, for example, this can manifest in scenarios where the autonomous vehicle takes a poor action (e.g., an aggressive lane change) yet other vehicles or pedestrian immediately respond by slowing down or steering away to avoid a possible collision, thereby hiding the potential negative consequences of the action. The lack of coordination and the constant best response from the follower, both common in practice, makes it hard for the leader to efficiently learn the reward landscape or improve their current policy.

3.2. An exponential upper bound

For any class $\mathcal{H}$ of reward functions on a pair of actions (a, b), an upper bound on the sample complexity (and regret) can be obtained using a covering argument.

Theorem 3.3. *Let $N(\epsilon) = N(\mathcal{H}, \epsilon, \|\cdot\|_\infty)$ denote the ℓ^∞ covering number of $\mathcal{H}$ with radius $\epsilon > 0$. Then we can achieve*

$$\mathcal{R}(T) \lesssim \inf_{\epsilon > 0} \epsilon T + \sqrt{N(\epsilon) T}. \tag{5}$$

To achieve this, simply compute an ϵ-covering of $\mathcal{H}$ and let the leader play no-regret algorithms on the ϵ-covering set. Note that although the covering is constructed for pair of actions $(a, b) \in \mathcal{A}_\epsilon \times \mathcal{B}_\epsilon$, it suffices for the leader to run no-regret algorithms on actions $\mathcal{A}_\epsilon$. The detailed algorithm and proof are given in Appendix A.2.

This upper bound is achieved when the leader does not even utilize the observations of the follower's responses. Indeed, in the worst case (e.g., in Example 3.2), the responses will not provide information.

As a corollary, in the linear regime with $\mathcal{H}_{\Theta,\phi}$, the covering number is $N(\epsilon) = N(\Theta, \epsilon, \|\cdot\|) \le \exp\left(O\left(d \log \frac{1}{\epsilon}\right)\right)$ (Wainwright, 2019). Choosing $\epsilon \asymp T^{-1/(d+2)}$, Theorem 3.3 reduces to the following upper bound in the linearly parameterized case.

Corollary 3.4. *In the linear case, we can achieve $\mathcal{R}(T) \lesssim T^{(d+1)/(d+2)}$.*

In other words, the sample complexity for achieving average regret equal to ϵ is upper bounded by $\mathcal{O}\left(\left(1/\epsilon\right)^{d+2}\right)$. This upper bound is agnostic to any structural property of the feature function ϕ, such as smoothness or even continuity.

4. UCB with side observations

Although the worst-case sample complexity for linear Stackelberg games is exponential, it is possible to obtain a fine-grained analysis and improved rate for the family $\mathcal{H}_{\Theta,\phi}$ when ϕ is better structured. A natural choice of algorithm for the leader is some variant of UCB that incorporates observations of the follower's actions. In this section, we will describe a general recipe for a family of UCB algorithms to incorporate the side information as well as the challenge in their design.

4.1. Algorithm description

We consider the following variant of UCB that uses the follower's responses as side information to improve the confidence set.

Remark 4.1. The regression oracles and the sequences $\{\alpha_t\}_{t\in[T]}, \{\beta_t\}_{t\in[T]}$ must be chosen appropriately so that the following condition holds: Given an error tolerance $\delta \in (0, 1)$, we require $h^\star \in \bigcap_{t=1}^T \mathcal{H}_t$ with probability at least $1 - \delta$.

Remark 4.2. A common choice for $\mathsf{Reg}^{(b)}$ and $\mathsf{Reg}^{(r)}$ is the least-squares regression oracle that computes

$$h_t^{(b)} \in \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{t-1} \|b_h^*(a_i) - \hat{b}_i\|^2 \tag{6}$$

and

$$h_t^{(r)} \in \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{t-1} (\overline{h}(a_i) - r_i)^2. \tag{7}$$

When the least-squares computation becomes infeasible under complex response-reward structures (this is common for (6)), custom oracles need to be designed. A more intricate approach may be to jointly construct the estimate using both $\{\hat{b}_\tau\}_{\tau\in[t-1]}$ and $\{r_\tau\}_{\tau\in[t-1]}$. We leave it for future research to study systematic designs of the oracles and the confidence sets.

Remark 4.3. When the responses are unobserved or ignored (e.g., by choosing $\alpha_t = \infty$), Algorithm 1 reduces to the classic Eluder UCB using the least-squares (reward) oracle with $\mathcal{H}_t = \mathcal{H}_t^{(r)}$ (Russo & Van Roy, 2013).

The choices of $\{\alpha_t\}_{t\in\mathbb{N}}$ and $\{\beta_t\}_{t\in\mathbb{N}}$ can pose another challenge. An naive attempt to get a generic upper bound on α_t is to use a covering argument as in Russo & Van Roy (2013) using the following measurement between two functions $h, h' \in \mathcal{H}$: $d^{(b)}(h, h') = \sup_a \|b_h^*(a) - b_{h'}^*(a)\|$. But note that this does not necessarily define a norm, and further the covering number of $\mathcal{H}$ in this sense can be infinite when the best response is discontinuous in the leader's action a. Thus, such an approach is often not useful and one may have to determine α_t on a per instance basis.

4.2. Examples

While Theorem 3.1 shows that the involvement of the omniscient follower can lead to "curse of expertise," a stark deterioration in the sample complexity, there are many scenarios

Algorithm 1 UCB with side information from expert

Input: Regression oracles $\mathsf{Reg}^{(b)}$ and $\mathsf{Reg}^{(r)}$ on reward and response, $\{\alpha_t\}_{t\in[T]}$, $\{\beta_t\}_{t\in[T]}$
for $t = 1$ **to** T **do**
  Compute $h_t^{(b)} = \mathsf{Reg}^{(b)}(\hat{b}_1, \ldots, \hat{b}_{t-1})$ and $h_t^{(r)} = \mathsf{Reg}^{(r)}(r_1, \ldots, r_{t-1})$
  Set $\mathcal{H}_t^{(b)} := \{h : \sum_{i=1}^{t-1} \|b_h^*(a_i) - b_{h_t^{(b)}}^*(a_i)\|^2 \le \alpha_t^2\}$
  Set $\mathcal{H}_t^{(r)} := \{h : \sum_{i=1}^{t-1} \left(\overline{h}(a_i) - \overline{h}_t^{(r)}(a_i)\right)^2 \le \beta_t^2\}$
  Construct confidence set $\mathcal{H}_t = \mathcal{H}_t^{(b)} \cap \mathcal{H}_t^{(r)}$
  Take action $a_t \in \arg\max_{a\in\mathcal{A}} \sup_{h\in\mathcal{H}_t} \overline{h}(a)$
  Observe (noisy) reward r_t and response $\hat{b}_t$
end for

where the leader's observation of the follower's responses can expedite learning significantly. In this section, we will explore a few such examples.

4.2.1. An imitation-based example

Let us consider a setting where the leader achieves efficient learning through imitation. Heuristically, imitation arises when the optimal action for the leader is equal to the best response for the omniscient follower or a function of it. This may capture, for instance, real-world robotics applications where the actions of the robot and the human expert are exchangeable and the true goal can be easily inferred from the expert's action. A simple scenario is when the robot and the human expert are supposed to carry out the same task perfectly, in which case the robot should simply treat the expert as a role model and imitate. The following is a concrete example.

Example 4.4. Let $\mathcal{A} = \mathcal{B} = \Theta = \mathbb{S}^{d-1}$ (or $\mathbb{B}^d$ equivalently)[2]. Consider the linearly parameterized function class $\mathcal{H}_{\Theta,\phi}$ with feature function

$$\phi(a, b) = a + b. \tag{8}$$

Here, the optimal response $b^*_\theta \equiv \theta$ is independent of a, and $\overline{h}_\theta(a) = \theta \cdot a + 1$.

Construction of confidence sets. The (noisy) observations of the follower's best responses simplify the problem into an imitation learning task. A simple oracle for the best-response observations is to take the $\mathcal{A}$-projected empirical average of responses, i.e., $\theta_t^{(b)} = \Pi_{\mathcal{A}}\big(\frac{1}{t-1}\sum_{i=1}^{t-1}\hat{b}_i\big)$.[3] The response-based confidence set reduces to

$$\Theta_t^{(b)} = \Big\{\theta \in \Theta \Big| \|\theta - \theta_t^{(b)}\| \le \frac{\alpha_t}{\sqrt{t-1}}\Big\}.$$

Standard sub-Gaussian concentration results suggest that the (Euclidean) radius of this confidence set shrinks at a rate of $t^{-1/2}$.

Lemma 4.5. *To ensure $\theta^\star \in \bigcap_{t\in[T]}\Theta_t$ with probability at least $1-\delta$, it suffices to choose $\alpha_t = \Theta\big(\sigma_b\sqrt{d + \log\frac{T}{\delta}}\big)$.*

UCB chooses actions on $\mathbb{S}^{d-1}$ increasingly close to the empirical estimate $\theta_t^{(b)}$.[4] The regret bound follows from these choices of confidence sets.

[2]While it is customary to consider $\Theta = \mathbb{B}^d$, we will observe below that the imitation-based algorithm does not crucially rely on $\|\theta^\star\|$ and only incurs smaller regret if $\|\theta^\star\| < 1$. This is because the algorithm asymptotically relies solely on the response observations, which are invariant under scaling of $\theta^\star$. It is also without loss of generality to restrict all actions to the sphere.

[3]Define the projection of $y \in \mathbb{R}^d$ onto a closed set $\mathcal{X} \subseteq \mathbb{R}^d$ as $\Pi_{\mathcal{X}}(y) := \arg\min_{x\in\mathcal{X}}\|y - x\|$, breaking ties arbitrarily when the minimizer is not unique.

[4]Even simpler, the leader can play the $\mathcal{A}$-projected empirical average of responses. Under our choice of constant α, the analysis will be the same, with the result differ by at most a constant factor.

Proposition 4.6. *In Example 4.4, UCB achieves a regret bound*

$$\mathcal{R}_{UCB}(T) \lesssim \sigma_b^2 \log T \cdot (d + \log T). \tag{9}$$

In other words, the average regret decays at a rate of $\widetilde{\mathcal{O}}(\sigma_b^2 d/T)$. This has also been analyzed in the setting of imitation learning (Rajaraman et al., 2021), and the results are consistent.

Remark 4.7. When the follower's responses are unobserved (still assumed to be best responses), this is simply a linear bandit, where the minimax regret is $\Omega(\sigma_b d\sqrt{T}) \gg \mathcal{O}(\sigma_b^2 d \log^2 T)$. This indicates the value of the b_t observations. When the follower's response is noiseless, one can see that a single sample suffices to find the optimal response since one always observes $b^\star_\theta = \theta$.

Remark 4.8. Note the gap in the $\Theta(\log T)$ regret when the response observations are used and the $\Theta(\sqrt{T})$ regret when they are ignored or unavailable, showing the value of those response observations. In fact, it is easy to modify this example slightly (e.g., taking $\phi(a, b) = \max\{|\theta^\top a|, \Delta\}b$ for some $\Delta \in (0, 1)$) to create an even larger gap: When the leader uses the response observations, the regret is $\widetilde{\mathcal{O}}(d \log T)$ with sample complexity $\widetilde{\mathcal{O}}\big(d \log \frac{1}{\epsilon}\big)$; When the response observations are unavailable, the sample complexity increases to $\Omega(\epsilon^{-d})$.

4.2.2. Expert-guided exploration

In many scenarios, the omniscient follower's actions may not directly reveal the exact state of the world but still provide crucial information. The next example illustrates a simple setting where the follower's response can significantly reduce the sample complexity.

Example 4.9. Let $\mathcal{A} = \mathcal{B} = \mathbb{S}^{d-1}$ and

$$\Theta = \{(\theta_a, \theta_b) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} | \theta_a \cdot \theta_b \ge \zeta\}$$

for some $\zeta \in (0, 1)$. Consider the parameterized family of functions $\mathcal{H}_\Theta = \{h_\theta | \theta \in \Theta\}$ where

$$h_\theta(a, b) = \mathrm{ReLU}(\theta_a \cdot a - \Delta) + \theta_b \cdot b,$$

for some $\Delta \in (0, 1)$. For simplicity, we will assume that the response observations are noiseless (i.e., $\sigma_b = 0$), although the noisy case can be analyzed analogously.

Confidences sets. The best response is $b^*_\theta \equiv \theta_b$, again independent of the leader's action. Upon observing $b_1 = \theta_b$, the leader should construct confidence sets $\Theta_t^{(b)} = \{\theta_a \in \mathbb{S}^{d-1} | \theta_a \cdot b_1 \ge \zeta\} \times \{b_1\}$, while $\Theta_t^{(r)}$ is chosen as in linear

UCB. As a result, all subsequent actions the leader takes must fall into

$$\mathcal{A}_1 := \{a \in \mathcal{A} | a \cdot b_1 \geq \zeta\}. \tag{10}$$

This refinement of the action set will reduce the sample complexity, and depending on the size of ζ relative to Δ, the reduction can be significant.

Strong reduction. When $1 - \zeta \leq (1 - \Delta)/4$, the leader learns that $\theta_a \cdot b_1 \geq \zeta$. In particular, any action $a \in \mathcal{A}_1$ must satisfy

$$\begin{aligned}\theta_a \cdot a = \frac{2 - \|\theta_a - a\|^2}{2} &\geq \frac{2 - (\|\theta_a - b_1\| + \|a - b_1\|)^2}{2} \\ &\geq \frac{2 - (2\sqrt{2 - 2\zeta})^2}{2} = 1 - 4(1 - \zeta) \geq \Delta, \end{aligned} \tag{11}$$

and thus $\overline{h}(a) = \theta_a \cdot a - \Delta + 1$ behaves as a linear function within $\mathcal{A}_1$. By playing UCB within $\mathcal{A}_1$, the leader reduces the problem to a linear bandit instance and thus achieves the following regret bound.

Proposition 4.10. *Assume $1 - \zeta \leq (1 - \Delta)/4$ in Example 4.9. UCB achieves*

$$\mathcal{R}_{UCB}(T) \leq \widetilde{\mathcal{O}}(d\sqrt{T}). \tag{12}$$

This leads to a sample complexity of $\widetilde{\mathcal{O}}(d^2/\epsilon^2)$, in contrast to the exponential sample complexity $\exp(\mathcal{O}(d \log \frac{1}{\epsilon}))$ if the responses were unobserved. Information from the follower's response guides the leader's exploration to the well conditioned part of the action space. Given the $\Omega(d\sqrt{T})$ sample complexity of linear bandits, the upper bound (12) is tight (up to logarithmic terms).

Weak reduction. When ζ is small relative to Δ, the problem does not immediately reduce to a linear bandit, but we have the following improved upper bound.

Proposition 4.11. *There exists an algorithm* Alg *that achieves*

$$\mathcal{R}_{\mathsf{Alg}}(T) \leq \mathcal{O}\big((C_\zeta^d T^{d+1})^{\frac{1}{d+2}}\big), \tag{13}$$

where $C_\zeta := \sqrt{1 - \zeta^2} \in (0, 1)$.

This bound improves as ζ decreases. The sample complexity is therefore $\widetilde{\mathcal{O}}(C_\zeta^d \epsilon^{-d-2})$, a C_ζ^d reduction compared with the original complexity without observing the responses in Corollary 3.4.

Since the reduced problem is still a ReLU bandit, UCB will not be suitable. Instead, (13) can be achieved through discretization of $\mathcal{A}_1$ as the upper bound in Theorem 3.3.

5. Beyond UCB

Although the UCB algorithm gives a near-optimal rate in most of the above examples. We also provide two cases where UCB fails to achieve the optimal rate. This necessitates a tailored algorithm design in specific settings.

5.1. Nonlinear (polynomial) family

UCB is known to fail to achieve the optimal rate in the case of the polynomial bandit family (Huang et al., 2021), where the reward is a polynomial activation on top of a linear family. We construct an example which utilizes the structure of the polynomial bandit, formally defined below.

Example 5.1 (Polynomial bandit)**.** Consider the convex function $f(x) = x^{2k}$ for some $k \in \mathbb{Z}_+$. Let

$$\mathcal{A} = \mathbb{B}^{d-1}, \mathcal{B} = [-1, 1], \Theta = \mathbb{B}^{d-1} \times \{1\}, \tag{14}$$

and

$$\phi(a, b) = (2kba, -f^*(2kb)), \tag{15}$$

where f^* is the convex conjugate of f. Consider the nonlinearly parameterized family

$$\mathcal{H}_\Theta := \{h_\theta(a, b) = f(\theta \cdot \phi(a, b)) \mid \theta \in \Theta\}. \tag{16}$$

By properties of the convex conjugate,

$$\overline{h}_\theta(a) = f(\theta_{-d} \cdot a) = (\theta_{-d} \cdot a)^{2k} \tag{17}$$

with the best response

$$\begin{aligned} b_\theta^*(a) &= \arg\max_{-1 \leq b \leq 1} 2kb\theta_{-d} \cdot a - f^*(2kb) \\ &= \frac{f'(\theta_{-d} \cdot a)}{2k} = (\theta_{-d} \cdot a)^{2k-1} \in [-1, 1]. \end{aligned}$$

This observation allows us to apply results on polynomial bandits (Huang et al., 2021).

Response-regret structure. Observe the following properties of the best response function in Example 5.1.

1. The expected reward is a function of the best response, independent of the true parameter. Namely,

$$\overline{h}_\theta(a) = b_\theta^*(a)^{\frac{2k}{2k-1}}. \tag{18}$$

This mapping is Lipschitz:

$$\big|\overline{h}_\theta(a) - \overline{h}_\theta(a')\big| \leq \frac{2k}{2k-1}\big|b_\theta^*(a) - b_\theta^*(a')\big|, \tag{19}$$

and further

$$\arg\max_{a \in \mathcal{A}} b_\theta^*(a) = \theta \in \arg\max_{a \in \mathcal{A}} \overline{h}_\theta(a), \tag{20}$$

with both maxima being 1.

2. The response observation, as a degree $2k-1$ polynomial, is more informative than the reward observation, a degree $2k$ polynomial, when the noise levels are the same and $\theta_{-d} \cdot a$ is small.

Based on these two observations, the leader may view the response b_t as a *proxy reward* and aim to minimize the *proxy regret*

$$\widehat{\mathcal{R}}(T) := \sum_{t=1}^{T} 1 - b^*_\theta(a_t). \tag{21}$$

This is consistent with minimizing the true regret $\mathcal{R}(T)$, which differs from the proxy regret $\widehat{\mathcal{R}}(T)$ by at most a constant factor by (19).

Regret bound. Using the response observations exclusively to minimize the proxy regret $\widehat{\mathcal{R}}(T) = \sum_{t=1}^{T} 1 - b^*_\theta(a_t)$, the leader reduces her task to a polynomial bandit problem with a degree $2k-1$ polynomial activation function. By (19), we may focus on bounding the proxy regret. Corollary 3.16 from Huang et al. (2021) suggests that

$$\widehat{\mathcal{R}}(T) \le \widetilde{\mathcal{O}}(\sqrt{d^{2k-1}T}), \tag{22}$$

or equivalently the sample complexity is $\widetilde{\mathcal{O}}(d^{2k-1}/\epsilon^2)$ for achieving ϵ average proxy regret. The following bound on the true regret follows from (19) and (22).

Proposition 5.2. *In example 5.1, there exists an algorithm* Alg*, using the response observations exclusively, that achieves*

$$\mathcal{R}_{\mathsf{Alg}}(T) \le \mathcal{O}(\sqrt{d^{2k-1}T}). \tag{23}$$

Proposition 5.2 suggests an $\widetilde{\mathcal{O}}(d^{2k-1}/\epsilon^2)$ sample complexity. For instance, the leader can achieve this regret with the zeroth-order algorithm proposed in Huang et al. (2021, Algorithm 6).

Remark 5.3 (Lower bound). Since the reward observations have a higher signal-to-noise-ratio, we should expect that the sample complexity of Example 5.1 to be the same order as the sample complexity of achieving ϵ average regret in a degree $2k-1$ polynomial bandit. Huang et al. (2021) shows that this is lower bounded by $\Omega(d^{2k-1}/\epsilon^2)$. Thus, (23) is essentially optimal.

Remark 5.4 (Benefit of observing responses). If the leader does not observe the responses, the problem is equivalent to a degree $2k$ polynomial bandit. The optimal regret without observing the experts actions will lead to an $\widetilde{\mathcal{O}}(d^{2k}/\epsilon^2)$ sample complexity. Thus, the response observations contribute to shaving of a factor of d, which can be significant when the dimensionality is high.

Remark 5.5 (Suboptimality of UCB). Using the traditional Eluder UCB algorithm leads to a suboptimal sample complexity of $\widetilde{\mathcal{O}}(d^{2k}/\epsilon^2)$ when the leader solely uses the response observations. Still, this is a factor d improvement compared to what she can achieve with UCB without the response observations.

5.2. Failure of the optimism principle

The next example is adapted from the ReLU bandit in Example 3.2, and shows that optimism-based method can have dramatic suboptimality in certain problems.

Example 5.6. Let $\mathcal{A} = \mathbb{B}^{d-1}$, $\mathcal{B} = \mathbb{B}^{d-1} \times [0,1]$, and

$$\Theta = \{(\theta_{-d}, \theta_d) \mid \theta_{-d} \in \mathbb{B}^d, \theta_d = 1 - \Delta\} \tag{24}$$

for some $\Delta \in (0,1)$. Consider the linear family $\mathcal{H}_{\Theta,\phi}$ with

$$\phi(a,b) = \|a\|((1-b_d)a, b_d - \|b_{-d}\|) + \frac{1-\|a\|}{2}(b_{-d}, 0). \tag{25}$$

For any $\theta \in \Theta$ with $\theta_{-d} \in \mathbb{S}^{d-1}$, the optimal action for the leader is θ_{-d}, with the follower best responding $(0,0)$ and achieving unit expected reward.

When $\|a\| = 1$, this function behaves exactly as in Example 3.2, where $b^*_\theta(a) = (0,1)$ whenever $\theta_{-d} \cdot a < 1 - \Delta$; When $a = 0$, the best response is $b^*_\theta(0) = (\theta_{-d}, b_d)$. Thus, if the response observations are noiseless, the leader learns the true parameter and hence the optimal action in one round by playing $a_1 = 0$.

However, any optimism-based method such as UCB will not achieve such efficient learning, even when the response are noiselessly observed. It is straightforward to verify that, for any action a with $\|a\| < 1$, the optimistic reward satisfies

$$\sup_{\theta \in \Theta} \overline{h}_\theta(a) < 1. \tag{26}$$

Thus, as long as the confidence set contains some θ with $\theta_{-d} \in \mathbb{S}^{d-1}$, which holds under our initial condition, optimism causes the leader to only take actions $a \in \mathbb{S}^{d-1}$, reducing the problem to the worst-case Example 3.2.

6. Conclusions

We have studied a model of online learning in decentralized cooperative Stackelberg games. We showed that, even with an omniscient follower who always best responds (myopically), the worst case sample complexity for a linear family can be as large as $\exp(\Theta(d \log \frac{1}{\epsilon}))$. This "curse of expertise" highlights the challenge caused by miscoordinated exploration. This also raises the question of how a non-myopic expert follower should respond to the leader's actions (without knowing the leader's exact algorithm) to expedite their learning and maximize their long-term reward.

We considered the UCB-type algorithm that incorporates response observations. A few examples of various hardness were considered, ranging from efficient learning through

imitation and guided exploration to the worst-case linear family example with an exponential sample complexity.

Besides the examples considered in the paper, there are numerous scenarios where the roles of the leader and the follower are more complex to reason about. This poses unique challenges for both the learning process of the leader and the subsequent analysis of regret, indicating a fertile ground for future research. Specifically, our current template of Algorithm 1 requires designing the confidence sets based on the specific response-reward structure of each problem. It remains open to find a general design (or prove the lack thereof) that systematically synthesizes the response and reward observations. A general framework of analysis that can provide a unified yet sharp upper bound on the examples is also valuable.

Acknowledgments

We thank William Chang and Mengxiao Zhang for pointing out a gap in the statement and the proof of Theorem 3.1 in an earlier draft, and we thank the anonymous reviewers for their valuable comments and suggestions. This work is partially supported by NSF Grants IIS-1901242, CCF-1909499, and CCF-2211210.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Bai, Y., Jin, C., Wang, H., and Xiong, C. Sample-efficient learning of Stackelberg equilibria in general-sum games. *Advances in Neural Information Processing Systems*, 34: 25799–25811, 2021.

Conitzer, V. and Sandholm, T. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pp. 82–90, 2006.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Dong, K., Yang, J., and Ma, T. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in Neural Information Processing Systems*, 34:26168–26182, 2021.

Ferber, J. and Weiss, G. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-wesley Reading, 1999.

Filar, J. and Vrieze, K. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Gerstgrasser, M. and Parkes, D. C. Oracles & followers: Stackelberg equilibria in deep multi-agent reinforcement learning. *arXiv preprint arXiv:2210.11942*, 2022.

Goodrich, M. A., Schultz, A. C., et al. Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3):203–275, 2008.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

Ho, C.-J., Slivkins, A., and Vaughan, J. W. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 359–376, 2014.

Huang, B., Huang, K., Kakade, S., Lee, J. D., Lei, Q., Wang, R., and Yang, J. Optimal gradient-based algorithms for non-concave bandit optimization. *Advances in Neural Information Processing Systems*, 34:29101–29115, 2021.

Kao, H., Wei, C.-Y., and Subramanian, V. Decentralized cooperative reinforcement learning with hierarchical information structure. In *International Conference on Algorithmic Learning Theory*, pp. 573–605. PMLR, 2022.

Kleinberg, R. and Leighton, T. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 594–605. IEEE, 2003.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Lauffer, N., Ghasemi, M., Hashemi, A., Savas, Y., and Topcu, U. No-regret learning in dynamic Stackelberg games. *arXiv preprint arXiv:2202.04786*, 2022.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Liu, Y. and Chen, Y. A bandit framework for strategic regression. *Advances in Neural Information Processing Systems*, 29, 2016.

Marecki, J., Tesauro, G., and Segal, R. Playing repeated Stackelberg games with unknown opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 821–828, 2012.

Rajaraman, N., Han, Y., Yang, L., Liu, J., Jiao, J., and Ramchandran, K. On the value of interaction and function approximation in imitation learning. *Advances in Neural Information Processing Systems*, 34:1325–1336, 2021.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Sallab, A. E., Abdou, M., Perot, E., and Yogamani, S. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Tambe, M. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press, 2011.

von Stackelberg, H. *Market Structure and Equilibrium*. Springer Science & Business Media, 2010.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, C.-C., Kulkarni, S. R., and Poor, H. V. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

Wooldridge, M. *An introduction to multiagent systems*. John wiley & sons, 2009.

Xie, A., Losey, D., Tolsma, R., Finn, C., and Sadigh, D. Learning latent representations to influence multi-agent interaction. In *Conference on robot learning*, pp. 575–588. PMLR, 2021.

Yang, B., Zheng, L., Ratliff, L. J., Boots, B., and Smith, J. R. Stackelberg maddpg: Learning emergent behaviors via information asymmetry in competitive games. 2022.

Yu, Y., Xu, H., and Chen, H. Learning correlated Stackelberg equilibrium in general-sum multi-leader-single-follower games. *arXiv preprint arXiv:2210.12470*, 2022.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

Zhong, H., Yang, Z., Wang, Z., and Jordan, M. I. Can reinforcement learning find Stackelberg-Nash equilibria in general-sum Markov games with myopic followers? *arXiv preprint arXiv:2112.13521*, 2021.

Zhu, B., Bates, S., Yang, Z., Wang, Y., Jiao, J., and Jordan, M. I. The sample complexity of online contract design. *arXiv preprint arXiv:2211.05732*, 2022.

A. Proofs in Section 3

A.1. Proof of Theorem 3.1

Proof. Consider Example 3.2. The expected reward is given by

$$h_\theta(a,b) := \theta \cdot \phi(a,b) = (1-b)\theta_{-d} \cdot a + b(1-\Delta), \tag{27}$$

Optimizing over $b \in [0,1]$ yields

$$\overline{h}_\theta(a) = \max\{1-\Delta, \theta_{-d} \cdot a\}. \tag{28}$$

Note that for any $a \in \mathcal{A}$ such that $\theta_{-d} \cdot a < 1-\Delta$, the best response of the follower is $b=1$, yielding an expected reward of $1-\Delta$; for any $a \in \mathcal{A}$ such that $\theta_{-d} \cdot a \geq 1-\Delta$, the best response of the follower is $b=0$, yielding an expected reward of $\theta_{-d} \cdot a$. The optimal joint response $a = \theta_{-d}$ and $b=0$ achieves the optimal expected reward of $\|\theta_{-d}\| = 1 > 1-\Delta$. From the leader's perspective, this now reduces to the problem of a ReLU bandit considered in Dong et al. (2021), since the response provides no information until the average regret falls below Δ.[5] Thus we have

$$\inf_{\hat{\pi}} \sup_{\theta \in \Theta} \mathcal{R}(T) \geq \Omega(T^{1-\frac{2}{d-2}}).$$

□

A.2. Proof of Theorem 3.3

Proof. Let $\mathcal{H}(\epsilon)$ be a minimal ϵ-covering of $\mathcal{H}$ under the metric $\|\cdot\|_\infty$. Let

$$\mathcal{A}(\epsilon) = \Big\{ \arg\max_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} h(a,b) \mid h \in \mathcal{H}(\epsilon) \Big\},$$

where we break ties arbitrarily when the optimal action is non-unique. Note that we have $|\mathcal{A}(\epsilon)| \leq |\mathcal{H}(\epsilon)| \leq N(\epsilon)$. Let $h^\star$ be the true reward function. By the definition of a covering, there exists some $h_\epsilon \in \mathcal{H}(\epsilon)$ such that $\|h^\star - h_\epsilon\|_\infty \leq \epsilon$. Thus we have

$$\mathcal{R}(T) = \sum_{t=1}^{T} \mathbb{E}[\overline{h}^\star(a^*) - \overline{h}^\star(a_t)] \leq \epsilon T + \sum_{t=1}^{T} \mathbb{E}[\overline{h}^\star_\epsilon(a^*) - \overline{h}^\star_\epsilon(a_t)].$$

We know that the optimal action for $\overline{h}_\epsilon$ must be inside the set $\mathcal{A}(\epsilon)$. Thus any worst-case optimal no-regret algorithm on the set $\mathcal{A}(\epsilon)$ gives a regret of $\sqrt{|\mathcal{A}(\epsilon)|T} \leq \sqrt{N(\epsilon)T}$. This gives that

$$\mathcal{R}(T) \leq \epsilon T + \sqrt{N(\epsilon)T}.$$

Taking infimum over ϵ finishes the proof. □

B. Proofs in Section 4

B.1. Proof of Lemma 4.5

Proof. Recall the notation from Example 4.4: let $\theta_t^{(b)} = \Pi_{\mathcal{A}}(\hat{\theta}_t)$ for $t \geq 2$, with $\hat{\theta}_t := \frac{1}{t-1}\sum_{i=1}^{t-1} \hat{b}_i$. The first round incurs at most a constant regret and can be ignored. It suffices to show that, with probability at least $1-\delta$,

$$\|\theta - \theta_t^{(b)}\| \leq \frac{\alpha_t}{\sqrt{t}} \tag{29}$$

for $\alpha_t = \Theta\big(\sigma_b\sqrt{d + \log\frac{T}{\delta}}\big)$.

[5] Same as in Dong et al. (2021), we allow the reward and response observations to be noiseless. We believe, however, the proof of Dong et al. (2021, Theorem 5.1) has a small gap, where the packing number should be computed for radius $\sqrt{\epsilon}$ instead of ϵ. This lower bound can be further improved if we assume noisy observations.

First, we bound the distance between $\hat{\theta}_t$ and θ. By our assumption,

$$\|\hat{\theta}_t - \theta\| = \Big\|\frac{1}{t-1}\sum_{i=1}^{t-1} w_i\Big\|,$$

where $w_1, \dots, w_t$ are i.i.d. zero-mean σ_b-sub-Gaussian. We proceed using a covering argument. Construct $U \subseteq \mathbb{S}^{d-1}$ such that

$$\inf_{v\in\mathbb{S}^{d-1}} \sup_{u\in U} u \cdot v \geq \frac{1}{2}. \tag{30}$$

Note that $\|u - v\| = \sqrt{2 - 2u \cdot v}$ for $u, v \in \mathbb{S}^{d-1}$. Hence, equivalently, we may choose U as a minimal 1-covering of $\mathbb{S}^{d-1}$ in Euclidean metric. Then

$$\log|U| \leq \log N^{\text{int}}(\mathbb{S}^{d-1}, 1, \|\cdot\|) \leq \log M(\mathbb{B}^d, 1, \|\cdot\|) = \Theta(d), \tag{31}$$

where N^{int} and M denote the internal covering number and the packing number of the space under a given metric. The choice of U ensures that

$$\|w\| \leq 2\sup_{u\in U} u \cdot w \tag{32}$$

for all $w \in \mathbb{R}^d$, and ignoring the constant factor, we may focus on upper bounding $\sup_{u\in U}\sum_{i=1}^{t-1} u \cdot w_i$.

For each choice of $u \in U$, let $Z_{u,i} = u \cdot w_i$, so that $Z_{u,1}, \dots, Z_{u,t-1}$ are i.i.d. zero-mean σ_b-sub-Gaussian by definition of sub-Gaussian random vectors. By Hoeffding's inequality for sub-Gaussian random variables, we have

$$\mathbb{P}\Big(\sum_{i=1}^{t} Z_{u,i} > x\Big) \leq \exp\Big(-\frac{x^2}{2t\sigma_b^2}\Big) \tag{33}$$

for all $x > 0$. Applying union bound over U and using (32) gives

$$\mathbb{P}\left(\Big\|\sum_{i=1}^{t} w_i\Big\| \geq 2x\right) \leq \mathbb{P}\left(\sup_{u\in U}\sum_{i=1}^{t} Z_{u,i} \geq x\right) \leq |U|\exp\Big(-\frac{x^2}{2t\sigma_b^2}\Big). \tag{34}$$

Choosing $x = \sigma_b\sqrt{2t\log(|U|T)} \lesssim \sigma_b\sqrt{t(d + \log\frac{T}{\delta})}$ ensures that, by another union bound over $t \in [T]$,

$$\|\hat{\theta}_t - \theta\| \lesssim \sigma_b\sqrt{t^{-1}\big(d + \log\frac{T}{\delta}\big)} \tag{35}$$

with probability at least $1 - \delta$. By the triangle inequality and the definition of projection,

$$\|\theta_t^{(b)} - \theta\| \leq \|\theta_t^{(b)} - \hat{\theta}_t\| + \|\hat{\theta}_t - \theta\| \leq 2\|\hat{\theta}_t - \theta\| \lesssim \sigma_b\sqrt{t^{-1}\big(d + \log\frac{T}{\delta}\big)} \tag{36}$$

with the same probability. This gives (29) and completes the proof. □

B.2. Proof of Proposition 4.6

Proof. We will condition upon the validity of the confidence sets, which happens with probability at least $1 - \delta$ per our choice of $\{\alpha_t\}_{t\in[T]}$.

UCB always chooses a_t in the confidence set Θ_t, with radius of order $\mathcal{O}\big(\sigma_b\sqrt{t^{-1}(d + \log\frac{T}{\delta})}\big)$. When $\theta^\star \in \Theta_t$, we have $\|a_t - \theta^\star\| \lesssim \sigma_b\sqrt{t^{-1}(d + \log\frac{T}{\delta})}$. Since both a_t and $\theta^\star$ are unit vectors, we have

$$\begin{aligned}\mathcal{R}_{UCB}(T) &\leq 2\delta T + \sum_{t=1}^{T}\big(1 - \theta^\star \cdot a_t\big) = 2\delta T + 2 + \frac{1}{2}\sum_{t=1}^{T}\|\theta^\star - a_t\|^2 \\ &\lesssim 2\delta T + \sum_{t=2}^{T}\frac{\sigma_b^2}{t}\Big(d + \log\frac{T}{\delta}\Big) = \mathcal{O}\left(\delta T + \sigma_b^2\log T \cdot \Big(d + \log\frac{T}{\delta}\Big)\right),\end{aligned}$$

where the term $2\delta T$ bounds the contribution of the event that the confidence sets fails to be all valid. Choosing $\delta = 1/T$ gives our desired bound. □

B.3. Proof of Proposition 4.10

Proof. After the first round, the leader's task reduces to a linear bandit with action space $\mathcal{A}_1$: only actions within $\mathcal{A}_1$ will be played, and the reward is linear in this region. As is well known for linear bandit (e.g., (Russo & Van Roy, 2013)), with probability $1-\delta$, the regret in this linear stage (i.e., excluding the first round) is upper bounded by

$$2\delta T + \mathcal{O}\big(\sqrt{d\log T\cdot(d\log T+\log\delta^{-1})\cdot T}\big).$$

The first round adds at most a constant to this and can be ignored. By choosing $\delta = T^{-1}$, we have

$$\mathcal{R}_{UCB}(T) \le \widetilde{\mathcal{O}}(d\sqrt{T}). \tag{37}$$

□

B.4. Proof of Proposition 4.11

Proof. Let $\Theta_1 = \{\theta_a \in \mathbb{S}^{d-1} | \theta_a \cdot b_1 \ge \zeta\} \times \{b_1\}$, and denote the true parameter by $\theta^\star = (\theta_a^\star, \theta_b^\star)$. By our assumption on the problem structure, we have $\theta_a^\star \in \Theta^{(b)}$.

As in the proof of Theorem 3.3, let $\Theta(\epsilon)$ be a minimal ϵ-covering of Θ_1 in Euclidean metric, with $\epsilon > 0$ to be specified later. In particular, there is some $\tilde{\theta}_a \in \Theta_1$ with $\|\tilde{\theta}_a - \theta_a^\star\| \le \epsilon$. Let $\mathcal{A}(\epsilon) = \{\arg\max_{a\in\mathcal{A}} \mathrm{ReLU}(\theta_a \cdot a - \Delta) \mid \theta_a \in \Theta(\epsilon)\}$, where we break tie arbitrarily when the optimal action is non-unique. Note that $|\mathcal{A}(\epsilon)| \le |\Theta(\epsilon)| = N(\Theta_1, \epsilon, \|\cdot\|)$.

Now, let the leader play UCB on the discrete action set $\mathcal{A}(\epsilon)$ after the first round. The regret satisfies

$$\mathcal{R}(T) \le 1 + \sum_{t=2}^{T} \mathbb{E}\big[\overline{h}^\star(a^*) - \overline{h}^\star(a_t)\big] \le 1 + T\cdot\mathbb{E}\big[\overline{h}^\star(a^*) - \overline{h}^\star(\tilde{a}^*)\big] + \sum_{t=1}^{T}\mathbb{E}\big[\overline{h}^\star(\tilde{a}^*) - \overline{h}^\star(a_t)\big], \tag{38}$$

where $a^* = \theta_a^\star$ and $\tilde{a}^* \in \arg\max_{a\in\mathcal{A}(\epsilon)} \overline{h}^\star(a)$. Note that $\overline{h}^\star(\tilde{a}^*) \ge \overline{h}^\star(\tilde{\theta}_a) \ge \overline{h}^\star(a^*) - \epsilon$ by our choice of $\tilde{\theta}_a$ and $\mathcal{A}(\epsilon)$, the second term in (38) is at most ϵT. The third term, the regret of UCB on $\mathcal{A}(\epsilon)$, is bounded by $\mathcal{O}(\sqrt{N(\Theta_1,\epsilon,\|\cdot\|)\cdot T})$ in expectation.

It remains to bound $N(\Theta_1, \epsilon, \|\cdot\|)$. Note that for any $\theta_a, \theta_a' \in \Theta_1$, we have

$$\begin{aligned}
\theta_a \cdot \theta_a' &= (\theta_a\cdot b_1)(\theta_a'\cdot b_1) + (\theta_a - (\theta_a\cdot b_1)b_1)\cdot(\theta_a' - (\theta_a'\cdot b_1)b_1)\\
&\ge \zeta^2 - \|\theta_a - (\theta_a\cdot b_1)b_1\|\|\theta_a' - (\theta_a'\cdot b_1)b_1\|\\
&\ge \zeta^2 - (1-\zeta^2) = 2\zeta^2 - 1.
\end{aligned}$$

Equivalently, $\|\theta_a - \theta_a'\| = \sqrt{2-2\theta_a\cdot\theta_a'} \le 2\sqrt{1-\zeta^2} = 2C_\zeta$. Thus, the covering number of Θ_1 is upper bounded by $\big(\frac{KC_\zeta}{\epsilon}^d\big)$ for some absolute constant K, which yields a regret bound of $1 + \epsilon T + \mathcal{O}(\sqrt{K^d C_\zeta^d T/\epsilon^d})$. Choosing $\epsilon \asymp (KC_\zeta)^{\frac{d}{d+2}} T^{-\frac{1}{d+2}}$ reduces this upper bound to $\mathcal{O}\Big(C_\zeta^{\frac{d}{d+2}} T^{\frac{d+1}{d+2}}\Big)$ as desired. □

C. Proofs in Section 5

C.1. Proof of Proposition 5.2

Proof. Let the leader run the phased elimination algorithm Huang et al. (2021, Algorithm 6) using the response $b_\theta^*(a_t)$ as the proxy reward to maximize. This proxy reward, in expectation, is a homogeneous polynomial of degree $2k-1$. By Corollary 3.16 in Huang et al. (2021), the algorithm achieves

$$\widehat{\mathcal{R}}(T) \le \widetilde{\mathcal{O}}\big(\sqrt{d^{2k-1}T}\big), \tag{39}$$

where $\widehat{\mathcal{R}}(T) = \sum_{t=1}^{T} 1 - b_\theta^*(a_t)$ is the proxy regret measured based on the the proxy reward (i.e., absolute response). Note that the reward is maximized exactly when the proxy reward is maximized. Thus, the Lipschitz property (19) suggests that

$$\mathcal{R}(T) \le \frac{2k}{2k-1}\widehat{\mathcal{R}}(T) \le \widetilde{\mathcal{O}}(\sqrt{d^{2k-1}T}). \tag{40}$$

□