Responding to Promises: No-regret learning against followers with memory

Vijeth Hebbar VHEBBAR2@ILLINOIS.EDU and **Cédric Langbort** LANGBORT@ILLINOIS.EDU
Coordinated Science Laboratory, University of Illinois Urbana–Champaign, Urbana, IL 61801, USA.

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

We consider a repeated Stackelberg game setup where the leader faces a sequence of followers of unknown types and must learn what commitments to make. While previous works have considered followers that best respond to the commitment announced by the leader in every round, we relax this setup in two ways. Motivated by natural scenarios where the leader's *reputation* factors into how the followers choose their response, we consider followers with memory. Specifically, we model followers that base their response on not just the leader's current commitment but on an aggregate of their past commitments. In developing learning strategies that the leader can employ against such followers, we make the second relaxation and assume boundedly rational followers – in particular – *quantal responding* followers. Interestingly, we observe that the smoothness property offered by the quantal response (QR) model helps in addressing the challenge posed by learning against followers with memory. Utilizing techniques from online learning, we develop algorithms that guarantee $O(\sqrt{T})$ regret for quantal responding memory-less followers and $O(\sqrt{BT})$ regret for followers with bounded memory of length B with both scaling polynomially in game parameters.

Keywords: Online Learning with Memory, Learning in Games, Stackelberg Games.

1. Introduction

Stackelberg games (SGs) offer a natural framework to capture strategic interactions with hierarchical play between agents. They have been used to model scenarios in a wide range of domains like defence (Tambe, 2011), market behavior (Anderson and Engers, 1992), persuasive signaling (Kamenica and Gentzkow, 2011; Hebbar and Langbort, 2020; Massicot and Langbort, 2019) and more (Farokhi et al., 2016; Zrnic et al., 2021). This hierarchy is typically delineated by naming one agent as the *leader* (she), who acts first and commits to a strategy, and the other as a *follower* (he), who then best responds to the leader's committed strategy. A key feature of SGs is the leader's ability to estimate the follower's best response, without which she cannot compute the optimal strategy to commit to. However, in the event that the follower's payoff function is unknown, a possible (and well-studied) work-around for the leader is to learn her optimal strategy via repeated interaction with the follower, resulting in a so-called repeated SG (Letchford et al., 2009; Blum et al., 2014; Balcan et al., 2015). In this work, we consider one such setup where a leader is facing a sequence of followers whose type (and hence, payoff function) is unknown to her and she is tasked with finding her optimal strategy.

When learning in such repeated games, the leader sequentially updates her strategy in every round of the repeated game based on feedback from past interactions with the followers. In the event that the leader's strategy in revealed to the follower in each round,[1] a prevalent assumption in the literature is that the follower selects his best response to this strategy (Balcan et al., 2015; Castiglioni

[1] Such revelation can take various forms, e.g. through attacker surveillance in security games (Tambe, 2011) or leader announcements in persuasive signaling (Kamenica and Gentzkow, 2011)

et al., 2020). While this assumption is well justified in single-shot SGs or in repeated SGs where the leader consistently plays a single strategy, it may not always hold when she continually changes her strategy. Indeed, one potential reason for this deviation – one that becomes the focus of our study – could be that the follower possesses memory and bases decisions not only on the leader's strategy in the current round, but also on her past plays. This could occur in scenarios where the leader merely '*announces*' her committed strategy[2] (e.g. persuasive signaling (Kamenica and Gentzkow, 2011)) and does not 'act' it out (like, e.g., actually deploying patrols in security games (Tambe, 2011)) before the follower picks his response. In such cases, the follower de facto responds to a promise from the leader (hence the title of this paper) and, if the veracity of this promise can only be verified ex-post, the reputation or credibility of the leader become paramount. Using an average of past promises (or any other aggregation of past strategies) is a way for a follower endowed with memory to generate and evaluate such a reputation.

Note that we are not claiming that it is necessarily beneficial for the follower to act in this way. We are, however, saying that such reliance on past promises should be considered to be within the range of possible (or even expected) behaviors of a boundedly rational follower. Consider, e.g., a politician interacting with a diverse set of voters over the course of a public campaign. Even if the politician commits to honoring the promises she makes, a skeptical voter will base their voting decisions on her past promises as well and not solely the one she made to him. In the presence of such followers, a very natural question arises for the leader, namely:

Q: How can a leader learn her optimal strategy when facing a sequence of followers that make decisions that depend on a history of her strategies?

In this work, we focus solely on a process whereby the followers respond to a weighted average of past leader strategies (which we denote as "reputation", in accordance with the discussion above).

1.1. Road-map and Contributions

In Section 2, we present some background on online optimization that is relevant to our work. We then formalize our problem framework in Section 3, focusing on a *matrix Stackelberg game* played between a leader and a sequence of followers of varying type. First, we consider memory-less followers that base their decisions only on the leader's current strategy in every round. However, we introduce a generalized response model for the followers, representing a departure from previous approaches. This model enables us to accommodate quantal and best-responding followers, among others. Subsequently, we extend our framework by endowing followers with memory, marking the second point of deviation from existing literature.

We present our results in Section 4 in the form of two algorithmic approaches that sequentially pick leader strategies. Our first approach, presented in Section 4.1, tackles learning against *memory-less* follower and guarantees $\mathcal{O}(\sqrt{T})$ regret (Theorem 4). Unlike prior works (Balcan et al., 2015; Castiglioni et al., 2020) that rely crucially on followers playing best responses, this novel framework can accommodates varied follower response models. While this ability to learn in SGs with general response models may be of independent interest, we see that is becomes particularly important when learning against followers with memory. Our second approach, presented in Section 4.2, targets the problem of learning against followers with memory and once again guarantees $\mathcal{O}(\sqrt{T})$ regret (Theorem 6). This alternate approach focuses on quantal responding followers and relies crucially on the smoothness properties of such a response model. Thus, in addition to the Quantal Response

[2] For a detailed discussion on leader commitment in such settings, see Kamenica and Gentzkow (2011).

model's ability to capture more realistic follower behavior, our result showcases its usefulness in enabling leader's learning. Notably, the presence of memory requires regret analysis techniques that differ vastly from those used in establishing Theorem 4.

1.2. Related Work

Learning in Stackelberg games: Research on learning optimal Stackelberg strategies from repeated SGs has explored various scenarios where the leader lacks information about the follower's decision-making process. Several works consider a leader repeatedly interacting with a single follower and use best response queries to learn the optimal Stackelberg strategy (Letchford et al., 2009; Marecki et al., 2012; Blum et al., 2014). In contrast, our work considers a leader that faces a (possibly, adversarially generated) sequence of followers of unknown *types*. Such a setup was first considered by Balcan et al. (2015), who frame the problem as an instance of online learning and develop no-regret learning algorithms for it. However, a key assumption in theirs and similar studies (Castiglioni et al., 2020; Sessa et al., 2020; Velicheti et al., 2024) is that followers always best respond to the leader's commitment. In a bid to capture reputation effects, we relax this assumption, and allow the followers' responses to depend on the leader's past strategies (albeit in a structured way). Notably, Xu et al. (2016) consider a Stackelberg security game (SSG) framework (Tambe, 2011) and develop an online learning algorithm that is agnostic to the nature of attackers' (followers') response. However, their approach exploits the structural properties of SSGs and it is unclear how their approach could be extended to general Stackelberg matrix games.

Non-best responding followers: A widely studied setup considers followers to also be learning agents that rely on their past payoffs to make decisions rather than the leader's strategy (Zrnic et al., 2021; Lin and Chen, 2024; Deng et al., 2019; Fiez et al., 2020). Haghtalab et al. (2024) consider learning from a single follower that, in the absence of access to leader's strategy, responds to a calibrated forecasts of it. In contrast to these works, we assume that followers have access to the leader's commitments. Another work considers a single *non-myopic* follower who deviates from best-responding to mislead the leader's learning (Haghtalab et al., 2022). In contrast, followers in our work deviate because they also consider the leader's past strategies when making decisions.

A popular choice for modeling agents that deviate from best response due to bounded rationality is the Quantal Response (QR) model. It has been used to capture realistic human behavior in various game theoretic settings (McKelvey and Palfrey, 1995; Feng et al., 2024; Yang et al., 2012). Wu et al. (2022) demonstrate that the smoothness property of QR models can help in improving the query efficiency when learning a single follower's payoff functions in Quantal Stackelberg Games. In a similar spirit, we show that when facing a sequence of memory-endowed followers, this smoothness allows us to develop no-regret algorithms.

Online Optimization with Memory (OOM) Our work is also closely related to the problem of online learning against adversaries with memory (Arora et al., 2012; Anava et al., 2015; Hebbar and Langbort, 2023) i.e. online optimization setups where the stage costs are allowed to depend on past learner decisions. Indeed, we observe that when followers have memory our problem can be cast as one instance of such a problem. However, non-convexity of stage costs in our framework poses an additional challenge. To address this, we extend approaches developed for online non-convex learning with *memory-less* costs (Agarwal et al., 2019; Suggala and Netrapalli, 2020) to our problem.

2. Preliminaries on online learning

In this section, we present a quick summary of concepts in online optimization that are relevant to our work. Let $\{f^t\}_{t=1}^H$ denote the sequence of cost functions – mapping decision set $\mathcal{X} \subset \mathbb{R}^n$ to $\mathbb{R}$ – faced by the learner over a horizon of length H. In line with the standard online optimization framework, we assume that when making decision x^t the learner only has access to $\{f^\tau\}_{\tau=1}^{t-1}$. The learner wishes to pick decisions $\{x^t\}_{t=1}^H$ such that their regret, defined as

$$\mathcal{R}(H) = \sum_{t=1}^{H} f^t(x^t) - \min_{x\in\mathcal{X}} \sum_{t=1}^{H} f^t(x), \tag{1}$$

grows sub-linearly with H. Let us now look more closely at one class of algorithms that are routinely applied to solve online optimization problems: Follow-The-Perturbed-Leader (FTPL) (Kalai and Vempala, 2005) (or equivalently, Follow-The-Regularized-Leader (FTRL) with randomized regularizers (Hazan, 2022)). Let us suppose that we have access to an approximate optimization oracle $\mathbb{O}_\epsilon$ which, given f and set $\mathcal{X}$, outputs

$$x^* = \mathbb{O}_\epsilon\big(f, \mathcal{X}\big) \qquad \text{such that} \qquad f(x^*) \le \min_{x\in\mathcal{X}} f(x) + \epsilon.$$

Now, given a random vector $\sigma \in \mathbb{R}^m$ picked from distribution $\mathcal{D}$ and a regularizing function $\rho : \mathbb{R}^m \times \mathcal{X} \to \mathbb{R}$, suppose that we generate decisions $\{x^t\}_{t=1}^H$ according to

$$x^1 = \mathbb{O}_\epsilon\big(\rho(\sigma,\cdot), \mathcal{X}\big) \text{ and } x^t = \mathbb{O}_\epsilon\Big(\textstyle\sum_{\tau=1}^{t-1} f^\tau + \rho(\sigma,\cdot), \mathcal{X}\Big),\ \forall t > 1. \tag{2}$$

Note that the decisions $\{x^t\}_t$ generated above are themselves random variables due to their dependence on σ and so, we are interested in bounding the expected value of the regret defined in (1). We can now state the following refinement of the well known bound from (Kalai and Vempala, 2005; Cesa-Bianchi and Lugosi, 2006) on the regret of FTRL algorithms.

Theorem 1 *When $\{x^t\}_t$ is generated according to (2) we can bound the expected regret as*

$$\mathbb{E}[\mathcal{R}(H)] \le \textstyle\sum_{t=1}^H \mathbb{E}[f^t(x^t) - f^t(x^{t+1})] + \mathbb{E}[\rho(\sigma, x') - \rho(\sigma, x^1)] + \epsilon(H+1)$$

The terms of the form $\mathbb{E}[f^t(x^t) - f^t(x^{t+1})]$ are typically called *stability* error terms and bounding them is a standard step in the regret analysis of many online learning algorithms. As we see ahead, that will be the case for our approach as well. Finally, we state a proposition that will aid us with bounding these stability error terms in our setup.

Proposition 2 *(Kalai and Vempala (2005))*
Let f be a bounded integrable function mapping $\mathbb{R}^n \to \mathbb{R}$. Let $\sigma \sim U[0, \frac{2}{\nu}]^n$ and $|f| \le B$ for some constants B, ν. Then for any $G, c \in \mathbb{R}^n$

$$\big|\mathbb{E}[f(G + c + \sigma) - f(G+\sigma)]\big| \le \nu B \|c\|_1$$

3. Problem Setup

Before we begin describing our framework, let us define some notation we will use throughout this paper. For any integer $S \in \mathbb{Z}_{\ge 0}$, we define the set $[S] \triangleq \{1, \dots, S\}$. We denote the j^{th} column of a matrix A as A_j and the element corresponding to the i^{th} row and j^{th} column as A_{ij}. We denote the maximal absolute element of a matrix A as $\overline{A} \triangleq \max_{i,j} |A_{ij}|$ and the vector induced p-norm as $\|A\|_p$. We denote the standard n-dimensional simplex as $\mathbf{\Delta}_n$ i.e. $\mathbf{\Delta}_n = \{s \in \mathbb{R}^n_{\ge 0} | \sum_i s_i = 1\}$.

3.1. Single-shot Stackelberg game

Consider a Stackelberg matrix game between a single leader (she) and a follower (he) that has one of K types. The leader's utility matrix is $U \in \mathbb{R}_{\geq 0}^{N \times M}$, while a follower of type $k \in \{1, \dots, K\}$ has utility matrix $V^k \in \mathbb{R}^{N \times M}$. Here, N and M are the number of actions for the leader and follower respectively. We can represent any strategy played by the leader (resp. follower) as an element x (resp. y) of $\mathbf{\Delta}_N$ (resp. $\mathbf{\Delta}_M$) where x_i (resp. y_i) denotes the probability of them playing action i. Then, the expected payoff of the leader and follower (of type k) can be captured simply as $x^\intercal U y$ and $x^\intercal V^k y$ respectively.

Let us now consider a generalized model of a *memoryless* follower's response, focusing for now on type k . Given a leader strategy $x \in \mathbf{\Delta}_N$, we define the follower's *perturbed* response function as

$$y^k(x) \in \arg\max_{y \in \mathbf{\Delta}_M} x^\intercal V^k y + \tfrac{1}{\eta} h(y). \tag{3}$$

where $h : \mathbf{\Delta}_M \to \mathbb{R}$ is a perturbation function. Various choices of perturbation functions can be made to model a variety of decision-making behavior. However, for the purpose of our work, we are interested in two primary choices: one when $h = 0$, modelling a perfectly rational follower and the other, when it is the Gibbs entropy function $h(y) = -\sum_{i=1}^m y_i \ln y_i$, modelling a quantal responding follower (Mertikopoulos and Sandholm, 2016). To highlight their nature, we will denote the follower's response in the absence of perturbation (i.e. $h = 0$) as $y^k_{BR}(x)$ and under the entropy perturbation function as $y^k_{QR}(x)$.

Note that $y^k_{BR}(x)$ may not be uniquely defined for every value of x and so, we assume that all players break ties according to a deterministic rule that is common knowledge. On the other hand, $y^k_{QR}(x)$ is uniquely defined as it is the solution to a convex optimization problem. In-fact, it is well known (Sutton and Barto, 2018) that it takes the closed form given by the logit choice model

$$y^k_{i,QR} = \frac{\exp\left(\eta \langle V^k_i, x\rangle\right)}{\sum_{j=1}^m \exp\left(\eta \langle V^k_j, x\rangle\right)} \quad \forall i \in [M].$$

In the context of the QR model, η acts as the bounded rationality constant and for the purpose of this work, we assume that it is known and constant across all follower types[3]. The QR model also offers some nice smoothness properties for the follower's response, namely we can make the following claim:

Claim 3 ***(Lipschitz Quantal Response)*** *Setting $L \triangleq 2\eta \max_k \|V^k\|_1$, we have*

$$\|y^k_{QR}(x) - y^k_{QR}(z)\|_1 \leq L\|x - z\|_\infty \quad \forall x, y \in \mathbf{\Delta}_N, k \in [K].$$

To end our discussion on the follower's response model, we define the matrix valued function $\mathcal{Y} : \mathbf{\Delta}_N \to \mathbb{R}^{M \times K}$ by stacking the response functions from (3) for each follower type as

$$\mathcal{Y}(x) \triangleq [y^1(x), \; \dots, \; y^k(x), \; \dots, \; y^K(x)].$$

When we wish to highlight the nature of the response function, we will use $\mathcal{Y}_{BR}(x)$ and $\mathcal{Y}_{QR}(x)$ for best responding and quantal responding follower's respectively. Let e^k denote the k^{th} basis vector in $\mathbb{R}^K$. We can then recover the response of a follower of type k simple as $y^k(x) = \mathcal{Y}(x)e_k$.

[3] However, our results are not limited by this assumption. Indeed, we could treat this constant as an unknown (but, parametrized by types) in a follower's payoff model and modify the approach we develop to learn this constant.

Finally, let us characterize the leader's utility when faced with a single follower of type k. For now, we do not impose any specific form on the follower's response function $y^k(\cdot)$, so long as it follows the general structure outlined in (3) and is uniquely defined. However, we assume that the leader is informed about the follower's response mechanism, allowing her to evaluate her payoff function as $x^\intercal U \mathcal{Y}(x) e_k$ for every strategy $x \in \mathbf{\Delta}_N$. This assumption implicitly means that the leader knows the set of follower payoff matrices $\{V^k | k \in [K]\}$ without which she cannot evaluate $\mathcal{Y}(\cdot)$.

3.2. Repeated SG against a sequence of followers

Consider now a repeated SG being played between the leader and a sequence of followers over a horizon of length H. Let $\{x^t\}_{t=1}^H$ be the sequence of the strategies that the leader commits to over the horizon H. We represent the types of followers encountered by the leader through the sequence $\{g^t\}_{t=1}^H$, where $g^t \in \{e_k | k \in [K]\}$ at any time t, and $g^t = e_k$ signifies that the leader is facing a follower of type k at t.

Crucially, we will now assume that the leader is initially unaware of the sequence of follower types she will encounter and must learn what strategies to pick at every round based on past outcomes. While we will allow the sequence of types to be generated adversarially, we will consider the adversary to be *oblivious* (Cesa-Bianchi and Lugosi, 2006; Hazan, 2022), i.e., the sequence of followers is determined without seeing any learner strategies.

For the leader to have any meaningful chance to "learn" what strategies to play, it is important for her to receive feedback. Two types of feedback are typically considered in such scenarios (Balcan et al., 2015; Castiglioni et al., 2020): *full information feedback*, where the follower's type g^t is revealed after the leader picks x^t and *partial information feedback*, where only the follower's response $\mathcal{Y}(x^t)g^t$ is revealed. For the purpose of this work, we will assume that leader's have full information feedback and leave the extension to other forms of feedback as future work. We can now consider followers of two types.

3.2.1. MEMORY-LESS FOLLOWERS

We say a follower is "memory-less" when, at time t, his best response is $\mathcal{Y}(x^t)$, thus being independent of all past play by the leader. In this case, the cumulative payoff gained by the leader at the end of H time steps is $\sum_{t=1}^H \langle \mathcal{Y}(x^t)^\intercal U^\intercal x^t, g^t \rangle$. Naturally, the leader aims to achieve a high cumulative payoff, and if she had access to g^t when picking her strategy x^t, she could simply pick x^t as a maximizer of $xU\mathcal{Y}(x)g^t$ at time t. However, under our assumption that the type is revealed to her only after she picks x^t, the leader's problem turns into an instance of online learning with non-convex cost functions. Motivated by this observation, a natural measure of performance for the leader's decision-making approach is *regret*, defined for our problem as

$$Regret(H) \triangleq \max_{x \in \mathbf{\Delta}_N} \langle \mathcal{Y}(x)^\intercal U^\intercal x, G^H \rangle - \sum_{t=1}^{H} \langle \mathcal{Y}(x^t)^\intercal U^\intercal x^t, g^t \rangle \quad \text{where } G^t \triangleq \sum_{\tau=1}^{t} g^\tau. \tag{4}$$

In other words, we compare the cumulative payoff accumulated by the leader at the end of H rounds to the payoff of her *best-in-hindsight* strategy i.e. the static strategy that would have obtained her the highest total payoff. Then, our first goal is

P1: *Design algorithms that sequentially pick leader strategies $\{x^t\}_{t=1}^H$ such that she has* no-regret *i.e $\frac{Regret(H)}{H}$ vanishes as H goes to infinity.*

Notably, the presented setup with memory-less followers is closely related to the one considered by Balcan et al. (2015). Indeed, if we further restrict the followers to be perfectly rational and replace the general response function $\mathcal{Y}$ with $\mathcal{Y}_{BR}$ it is identical to their setup. We will now introduce the primary point of deviation of our work and consider followers with memory.

3.2.2. Followers with memory

We will now consider the situation where the followers base their action not only on the leader's current commitment, but also on her past commitments. While such dependence on the past can occur in many different ways, with the motivation of capturing the effect of leader's reputation on a follower's response, we assume that he bases his decision *solely* on a weighted average of her past play. Formally, we define the *time-averaged* leader strategy as

$$z^t \triangleq \tfrac{1}{b_t} \textstyle\sum_{\tau=1}^t a_{t-\tau} x^\tau \text{ where } b_t \triangleq \textstyle\sum_{\tau=1}^t a_{t-\tau} \text{ and } a_s \in \mathbb{R}_{\geq 0} \; \forall \; s \geq 0. \tag{5}$$

Viewing z^t as the *reputation* established by the leader in the eyes of each follower through her past play, a_s captures the weight of her strategy from s rounds ago on her current reputation. Such weighted averaging approach can capture a wide range of models of the follower's memory. For example, setting $a_s = 1$ when $s < B$ and $a_s = 0$ otherwise, models a follower with bounded memory of length B who views reputation as a simple unweighted average of leader's past play. Setting $a_s = \gamma^s$ for some $0 < \gamma < 1$ instead, captures a follower that models reputation as a discounted average of leader strategies. Regardless of the weights in (5), we will assume that a follower's response at time t is captured in $\mathcal{Y}(z^t)$. In other words, we model each follower as responding to an average of past play by the leader. Under this modelling assumption, the expected cumulative payoff of the leader at the end of H rounds is simply $\sum_{t=1}^H \langle \mathcal{Y}(z^t)^\intercal U^\intercal x^t, g^t \rangle$, which now depends explicitly on the time-averaged leader strategy z^t.

Apart from capturing reputation effects, considering that each follower responds to such a weighted average has another important consequence. To see it, consider a leader that adheres to a single strategy x over all rounds of the repeated game. Owing to the averaging process, we then have $z^t = x$ for all $t \in [H]$ and consequently, the leader's total payoff is then $\langle \mathcal{Y}(x)^\intercal U^\intercal x, G^H \rangle$. Motivated by this, we can once again compare the leader's cumulative payoff to the payoff of the best *static* action in hindsight and define regret as

$$Regret_M(H) \triangleq \max_{x \in \mathbf{\Delta}_N} \langle \mathcal{Y}^\intercal(x) U^\intercal x, G^H \rangle - \sum_{t=1}^H \langle \mathcal{Y}(z^t)^\intercal U^\intercal x^t, g^t \rangle \tag{6}$$

where the subscript M highlights that we are considering followers with memory. We can now state our second goal as

P2: *Design an algorithm that picks leader strategies* $\{x^t\}_{t=1}^H$ *such that she has* no-regret *when facing followers with memory i.e* $\frac{Regret_M(H)}{H}$ *vanishes as* H *goes to infinity.*

In comparing with the regret definition in (4), we see that the benchmark to which we compare the leader's cumulative payoffs is identical in both cases (with and without memory). As a result, developing algorithms that are no-regret w.r.t (6), means that despite the followers possessing memory, in the long run we are able achieve the performance of the best static action when followers are memory-less.

Lastly, note that since the leader payoffs now depend on her past strategies as well, the learning problem faced by her is an instance of Online Optimization with Memory (OOM) and consequently, the standard notion of regret to consider should be *policy regret* (Arora et al., 2012). Indeed, it can be easily seen that the regret defined in (6) is identical to the *policy regret* under the constant action policy class i.e. setting $x^t = x$ for all t (Hebbar and Langbort, 2023).

4. Main Results

4.1. Learning from memory-less followers

Let us first consider the problem **P1** posed in Section 3.2.1. We can state the following

Theorem 4 *Let $\sigma = [\sigma_1, \dots, \sigma_K]$ denote a uniform random vector such that $\sigma_k \overset{i.i.d}{\sim} U[0, {}^2\!/\!_\nu]$ for all $k \in [K]$. Then, setting $\nu = \sqrt{K/H}$ and picking strategies according to (2) with $\mathcal{X} = \mathbf{\Delta}_N$, $\rho(\sigma, x) \triangleq -\langle \mathcal{Y}(x)^\intercal U^\intercal x, \sigma \rangle$ and $f^t = -\langle \mathcal{Y}(x)^\intercal U^\intercal x, g^t \rangle$ gives us the following bound:*

$$\mathbb{E}[Regret(H)] \leq 2\overline{U}\sqrt{KH} + \epsilon(H+1)$$

where the expectation is taken over the distribution of σ.

Proof Sketch: We begin by invoking Theorem 1 which gives us

$$\mathbb{E}\big[Regret(H)\big] \leq \textstyle\sum_{t=1}^{H} \mathbb{E}\big[\langle \mathcal{Y}^\intercal(x^{t+1})U^\intercal x^{t+1} - \mathcal{Y}^\intercal(x^t)U^\intercal x^t, g^t \rangle\big] + \text{ other terms}$$

In this sketch, we will focus solely on bounding the stability error terms delineated above. For complete proofs of this and all other results, readers are referred to the extended version of this paper (Hebbar and Langbort, 2024). Defining the function $\mathcal{W} : \mathbb{R}^K \to \mathbf{\Delta}^N$ as $\mathcal{W}(G) = \mathbb{O}_\epsilon\big(-\langle \mathcal{Y}(\cdot)^\intercal U^\intercal \cdot, G\rangle, \mathbf{\Delta}_N\big)$ allows us to rewrite these stability error term as

$$\mathbb{E}\big[\langle g^t, \mathcal{W}(G^t + \sigma)\rangle\big] - \mathbb{E}\big[\langle g^t, \mathcal{W}(G^{t-1} + \sigma)\rangle\big].$$

Note that this expression is a difference in expected values of functions that differ in their argument by a translation. This allows us to invoke Proposition 2 and show that each stability error term scales linearly with ν. While there are H such terms, picking ν that scales inversely with $\sqrt{H}$ allows us to ensure that the total stability error grows sublinearly in H. □

This result suggests that it suffices to consider an oracle $\mathbb{O}_\epsilon$ with $\epsilon \in \mathcal{O}(H^{-1/2})$ to ensure $\mathcal{O}(\sqrt{H})$ regret. Crucially, the regret bound we obtained does not depend on the nature of the followers' response function in any way. However, it is still important for the leader to precisely know $\mathcal{Y}(\cdot)$ (regardless of its nature), otherwise she cannot employ the oracle $\mathbb{O}_\epsilon$ as described in Theorem 4.

Note that the stage payoff $\langle \mathcal{Y}^\intercal(x^t)U^\intercal x^t, g^t \rangle$ satisfies two properties; (1) it depends *linearly* on the unknown quantity g^t and (2) it depends only on the current strategy x^t and not on past strategies. These two properties are crucial for rewriting the stability error terms in a manner that allows us to invoke Proposition 2 to bound them. Consequently, despite having non-linear stage costs, we are able to employ regret analysis techniques similar to those developed for online linear optimization (Kalai and Vempala, 2005). This is similar in spirit to the approach taken by Balcan et al. (2015), who convert problem **P1** into an instance of prediction-from-expert-advice by observing that it suffices to consider *finitely* many (albeit, exponential in game parameters) leader strategies when facing best-responding followers. However, this observation no longer holds when facing quantal responding followers, making it unclear how their approach can be extended to this case.

Finally, we present a negative result on the computational hardness of implementing the oracle O_ϵ as outlined in Theorem 4.

Claim 5 *When followers are picking best responses (i.e. $\mathcal{Y} = \mathcal{Y}_{BR}$), the optimization problem solved by the oracle O_ϵ in Theorem 4 is NP-hard.*

Note that this computational hardness is also encountered in other studies on learning in SGs against best-responding followers (Balcan et al., 2015; Castiglioni et al., 2020; Sessa et al., 2020). Much like these studies, our work focuses on developing approaches that allow the leader to quickly learn what strategy to commit to, rather than on addressing the computational bottlenecks involved. Similar computational challenges arise when followers employ quantal responses (i.e. $\mathcal{Y} = \mathcal{Y}_{QR}$) and a more detailed discussion on the computational aspect of the various oracles employed in this work can be found in the extended version (Hebbar and Langbort, 2024). This version also presents empirical results where we showcase the regret incurred by the various methods proposed in this work.

4.2. Learning from followers with memory

Now, we tackle the main problem we set out to solve and address **P2** as outlined in Section 3.2.2. In doing so, we first introduce one additional notation and define $\Theta_H = \sum_{t=1}^{H} \frac{1}{b_t} \sum_{\tau=1}^{t} a_{t-\tau}(t-\tau)$. For the remainder of this paper we will assume that followers are picking quantal responses i.e. $\mathcal{Y} = \mathcal{Y}_{QR}$. We can state the following

Theorem 6 *Let $\sigma = [\sigma_1, \dots, \sigma_N]$ denote an exponential random vector such that $\sigma_n \overset{i.i.d}{\sim} \exp(\nu)$ for all $n \in [N]$. Then, setting $\nu = \big(\|U\|_1(1+L)\sqrt{50N(\Theta_H + H)}\big)^{-1}$ and picking leader strategies according to (2) with $\mathcal{X} = \mathbf{\Delta}_N$, $\rho(\sigma, x) \triangleq -\langle x, \sigma\rangle$ and $f^t = -\langle \mathcal{Y}_{QR}(x)^\intercal U^\intercal x, g^t\rangle$ results in*

$$\mathbb{E}[Regret_M(H)] \in \mathcal{O}\Big(N^{3/2}\|U\|_1 L\sqrt{(H+\Theta_H)} + \epsilon H\Big)$$

where the expectation is taken over the distribution of σ.

Proof Sketch: By adding and subtracting terms and by invoking Theorem 1 we can show that

$$\begin{aligned}\mathbb{E}[Regret_M(H)] = {} & \textstyle\sum_{t=1}^{H} \overbrace{\mathbb{E}\big[\langle \mathcal{Y}_{QR}^\intercal(x^{t+1})U^\intercal x^{t+1}, g^t\rangle\big] - \mathbb{E}\big[\langle \mathcal{Y}_{QR}^\intercal(x^{t})U^\intercal x^{t}, g^t\rangle\big]}^{(A)} \\ & + \textstyle\sum_{t=1}^{H} \underbrace{\mathbb{E}\big[\langle \mathcal{Y}_{QR}^\intercal(x^{t})U^\intercal x^{t}, g^t\rangle\big] - \mathbb{E}\big[\langle \mathcal{Y}_{QR}^\intercal(z^{t})U^\intercal x^{t}, g^t\rangle\big]}_{(B)} + \text{other terms}\end{aligned}$$

In this sketch, we focus solely on bounding the stability error terms (A) and (B). Term (A) is identical to the stability term in the proof of Theorem 4 and as shown there, it could be written as a difference of functions that differ in their argument by an offset. This then allowed us to bound term (A) using Proposition 2. However, a similar approach cannot be employed to bound term (B) as the stage payoff $\langle \mathcal{Y}_{QR}^\intercal(z^t)U^\intercal x^t, g^t\rangle$ also depends on past strategies. Prompted by this, we employ a regret analysis methodology in this proof that differs vastly from the one used in proving Theorem 4.

Invoking the smoothness property of the quantal response model highlighted in Claim 3, we can bound the terms (A) and (B) as

$$(A) \leq (1+L)\|U\|_1\mathbb{E}[\|x^{t+1} - x^t\|_1] \quad \text{and} \quad (B) \leq \tfrac{L\|U\|_1}{b_t}\textstyle\sum_{\tau=1}^{t-1} a_{t-\tau}\mathbb{E}\big[\|x^t - x^\tau\|_1\big].$$

Our approach then involves showing that $\mathbb{E}\big[\|x^t - x^\tau\|_1\big]$ scales linearly with $\nu|t-\tau|$ and draws heavily from the analysis techniques developed by Suggala and Netrapalli (2020). Consequently, picking a small ν allows us to ensure that the contribution of each term of the form in (A) and (B) is small. Indeed, picking ν that scales inversely with $\sqrt{H + \Theta_H}$ gives us the presented bound. $\square$

Our result suggests that as long as (i) $\Theta_H \in o(H^2)$ and (ii) the oracle $\mathbb{O}_\epsilon$ is such that $\epsilon \in \mathcal{O}(H^{-1/2})$, we can guarantee a sub-linear expected regret. Also, note that the only property of the QR model we employ in proving this result is the Lipschitz continuity of the response. Thus, our result could be easily extended to consider other forms of such *smooth* response models.

Revisiting our discussion in Section 1, if we view the strategies $\{x^t\}_t$ as 'promises' being made by the leader, our result leads to another interesting observation. After announcing x^t, a leader could choose to deviate and play a different strategy. However, Theorem 6 suggests that it is advantageous for the leader to honor her promise and play x^t (and earn the expected payoff of $\langle \mathcal{Y}(z^t)^\intercal_{QR} U^\intercal x^t, g^t \rangle$) as it allows her to learn with no-regret.

We can now consider two special models of followers' memory and state the following

Corollary 7 ***(Bounded Memory)***

1. *When the followers have a finite memory of length B and they weigh all past leader strategies equally (i.e. $a_s = 1$ if $s < B$ and $a_s = 0$ otherwise), then $\mathbb{E}[Regret_M(H)] \in \mathcal{O}(\sqrt{BH})$.*
2. *When followers weigh past actions by a discount factor ($a_s = \gamma^s$ for some $0 < \gamma < 1$), $\mathbb{E}[Regret_M(H)] \in \mathcal{O}(\sqrt{H(1-\gamma)^{-1}})$.*

Qualitatively, Corollary 7 argues that for our approach to have sublinear regret we need followers to be 'forgetful enough' i.e. the utility of the leader at any stage must not depend strongly on strategies played far in the past. Indeed this is not a surprising result and agrees with similar requirements on the memory of an adversary in online learning (Arora et al., 2012; Hebbar and Langbort, 2023). Also note that while ν depends on Θ_H in Theorem 6, our approach does not depend on the weights in (5) in any other way. However, if the leader lacks precise knowledge of these weights, estimating Θ_H is difficult. Fortunately, it suffices to know a suitable upper bound $\bar{\Theta}$ on Θ_H to make the following

Corollary 8 *Picking $\nu = \left(\|U\|_1(1+L)\sqrt{50N(\overline{\Theta}+H)}\right)^{-1}$ in Theorem 6 gives the alternate bound*

$$\mathbb{E}[Regret_M(H)] \in \mathcal{O}\Big(N^{3/2}\|U\|_1 L\sqrt{H+\overline{\Theta}} + \epsilon H\Big)$$

5. Conclusion and Future work

In this work, we sought to design learning algorithms that allow a leader to learn her optimal strategy when faced with followers that deviate from perfect best response. However, our approach assumes the availability of specific optimization oracles which, in practice, are computationally expensive to implement. Further research is needed to develop more computationally efficient oracles, potentially by constraining the class of games under consideration.

Another limitation of our approach is its reliance on setups with full information feedback, where the type of the follower is revealed at the end of each round. An important direction for future work is to extend our framework to more realistic settings with partial feedback, where the leader only has access to the followers' responses. Addressing this typically involves constructing an unbiased estimator of a follower's type from best responses (Balcan et al., 2015). However, it remains to be seen if such an approach extends to the case when followers are quantal responding.

Acknowledgments

This work was supported by the ARO MURI grant W911NF-20-0252 (76582 NSMUR).

References

Naman Agarwal, Alon Gonen, and Elad Hazan. Learning in non-convex games with an optimization oracle. In *Conference on Learning Theory*, pages 18–29. PMLR, 2019.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. *Advances in Neural Information Processing Systems*, 28, 2015.

Simon P Anderson and Maxim Engers. Stackelberg versus cournot oligopoly equilibrium. *International Journal of Industrial Organization*, 10(1):127–135, 1992.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the sixteenth ACM conference on economics and computation*, pages 61–78, 2015.

Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. *Advances in Neural Information Processing Systems*, 27, 2014.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. *Advances in Neural Information Processing Systems*, 33:16188–16198, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

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

Farhad Farokhi, André MH Teixeira, and Cédric Langbort. Estimation with strategic sensors. *IEEE Transactions on Automatic Control*, 62(2):724–739, 2016.

Yiding Feng, Chien-Ju Ho, and Wei Tang. Rationality-robust information design: Bayesian persuasion under quantal response. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 501–546. SIAM, 2024.

Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International Conference on Machine Learning*, pages 3133–3144. PMLR, 2020.

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.

Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36, 2024.

Elad Hazan. *Introduction to online convex optimization*, chapter 5. The MIT Press, 2 edition, 2022.

Vijeth Hebbar and Cedric Langbort. A stackelberg signaling game for human-uav collaboration in a search-and-rescue context. *IFAC-PapersOnLine*, 53(5):297–302, 2020.

Vijeth Hebbar and Cedric Langbort. Online decision making with history-average dependent costs (extended). *arXiv preprint arXiv:2312.06641*, 2023.

Vijeth Hebbar and Cédric Langbort. Responding to promises: No-regret learning against followers with memory. *arXiv preprint arXiv:2410.07457*, 2024.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *Algorithmic Game Theory: Second International Symposium, SAGT 2009, Paphos, Cyprus, October 18-20, 2009. Proceedings 2*, pages 250–262. Springer, 2009.

Tao Lin and Yiling Chen. Generalized Principal-Agent Problem with a Learning Agent, February 2024. URL https://arxiv.org/abs/2402.09721v5.

Janusz Marecki, Gerry Tesauro, and Richard Segal. Playing repeated stackelberg games with unknown opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 821–828, 2012.

Olivier Massicot and Cedric Langbort. Public signals and persuasion for road network congestion games under vagaries. *IFAC-PapersOnLine*, 51(34):124–130, 2019.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Panayotis Mertikopoulos and William H Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. Learning to play sequential games versus unknown opponents. *Advances in neural information processing systems*, 33:8971–8981, 2020.

Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pages 845–861. PMLR, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Milind Tambe. *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge university press, 2011.

Raj Kiriti Velicheti, Melih Bastopcu, S Rasoul Etesami, and Tamer Başar. Learning how to strategically disclose information. *arXiv preprint arXiv:2403.08741*, 2024.

Jibang Wu, Weiran Shen, Fei Fang, and Haifeng Xu. Inverse game theory for stackelberg games: the blessing of bounded rationality. *Advances in Neural Information Processing Systems*, 35: 32186–32198, 2022.

Haifeng Xu, Long Tran-Thanh, and Nick Jennings. Playing repeated security games with no prior knowledge. In *AAMAS'16: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 104–112. ACM Press, 2016.

Rong Yang, Fernando Ordonez, and Milind Tambe. Computing optimal strategy against quantal response in security games. In *AAMAS*, pages 847–854. Citeseer, 2012.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems*, 34:15257–15269, 2021.