

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 LEARNING TO PLAY MULTI-FOLLOWER BAYESIAN STACKELBERG GAMES

Anonymous authors

Paper under double-blind review

## ABSTRACT

In a multi-follower Bayesian Stackelberg game, a leader plays a mixed strategy over  $L$  actions to which  $n \geq 1$  followers, each having one of  $K$  possible private types, best respond. The leader's optimal strategy depends on the distribution of the followers' private types. We study an online learning version of this problem: a leader interacts for  $T$  rounds with  $n$  followers with types sampled from an unknown distribution every round. The leader's goal is to minimize regret, defined as the difference between the cumulative utility of the optimal strategy and that of the actually chosen strategies. We design learning algorithms for the leader under different feedback settings. Under type feedback, where the leader observes the followers' types after each round, we design algorithms that achieve  $\mathcal{O}(\sqrt{\min\{L \log(nKAT), nK\} \cdot T})$  regret for independent type distributions and  $\mathcal{O}(\sqrt{\min\{L \log(nKAT), K^n\} \cdot T})$  regret for general type distributions. Interestingly, those bounds do not grow with  $n$  at a polynomial rate. Under action feedback, where the leader only observes the followers' actions, we design algorithms with  $\mathcal{O}(\min\{\sqrt{n^L K^L A^{2L} LT \log T}, K^n \sqrt{T} \log T\})$  regret. We also provide a lower bound of  $\Omega(\sqrt{\min\{L, nK\} T})$ , almost matching the type-feedback upper bounds.

## 1 INTRODUCTION

Stackelberg games are a fundamental model of strategic interaction in multi-agent systems. Unlike normal-form games where all agents simultaneously play their strategy, Stackelberg games model a *leader* committing to their strategy; the remaining *follower(s)* take their actions after observing the leader's commitment (Conitzer & Sandholm, 2006; Von Stackelberg, 2010). Such asymmetric interactions are ubiquitous in a wide range of setting, from a firm entering a market dominated by an established competitor (Von Stackelberg, 2010), to an online platform releasing features that influence consumers on that platform (Zhao et al., 2023; Cao et al., 2024), to security games (Balcan et al., 2015; Sinha et al., 2018) to strategic machine learning (Hardt et al., 2016; Hossain et al., 2024). They also form the foundation of seminal models in computational economics like Bayesian Persuasion (Kamenica & Gentzkow, 2011) or contract design (Düting et al., 2024) that capture more structured settings with asymmetries relating to information or payouts respectively.

In these settings and beyond, there is one key question: what is the optimal strategy for the leader to commit to? Answering this question requires knowing how the follower(s) will react to the leader's strategy, which typically boils down to knowing the followers' utilities. The Bayesian approach attempts to relax this complete information assumption. Pioneering works like Conitzer & Sandholm (2006) assume that followers' utilities are parametrized by hidden types from a known distribution. Here, the leader aims to compute the Bayesian Stackelberg Equilibrium: the strategy maximizing the leader's expected utility with the followers' types drawn from the known distribution.

In many of the settings mentioned, even the Bayesian perspective may be too strong and unrealistic: the leader (e.g., online platform, dominant firm) may only know the structure of the followers' utilities but not the distribution of their types (Cole & Roughgarden, 2014). While not much can be achieved in a single round of such a game, the leader can often interact with the followers over multiple rounds and learn about them over time. The leader must, however, balance learning with playing the optimal strategy given current information – the well-known exploration-exploitation trade-off in the online learning literature.

054 **Our Contributions:** This paper comprehensively studies the learning and computational problem  
 055 for an online Bayesian Stackelberg game (BSG). Specifically, we consider the interaction over  $T$   
 056 rounds between a leader and  $n$  followers, each realizing one of  $K$  possible private types at each  
 057 round. To our knowledge, this is the first work on online learning in BSGs with multiple followers.  
 058 We study two feedback models: observing realized types of the followers, or observing their best-  
 059 responding actions, after each round. Our core objective is exploring how these feedback models  
 060 affect the learnability of the optimal strategy, which is challenging for several reasons. First, with  
 061 multiple followers, the unknown joint type space is exponentially large. Further, followers’ taking  
 062 best-responding actions means that the leader’s utility function is discontinuous and non-convex.  
 063 Lastly, even the offline single-follower version of this problem has known computational challenges  
 064 (Conitzer & Sandholm, 2006). A key technical tool used to unravel this is a geometric characterization  
 065 of the leader’s strategy space in terms of best-response regions (presented in Section 3). Section 4  
 066 uses this and an observation about learning type distributions vis-a-vis learning utility to provide  
 067 algorithms for both general type distributions and independent ones. A matching lower bound is  
 068 also provided. Section 5 then studies algorithms for the action feedback case, where we leverage  
 069 our geometric insights along with techniques from linear bandits. Table 1 summarizes our results.  
 070 Throughout, we comment on the computational complexity of our algorithms and uncover interesting  
 071 trade-offs that situate our work with the broader literature on Stackelberg games.  
 072  
 073

Table 1: Regret bounds for learning the optimal leader strategy in Bayesian Stackelberg games with  $n$  followers under various settings. The  $\tilde{\mathcal{O}}(\cdot)$  notation omits logarithmic factors.

	Type Feedback		Action Feedback
	Independent types	General types	
Upp. Bound	$\tilde{\mathcal{O}}(\sqrt{\min\{L, nK\}T})$	$\tilde{\mathcal{O}}(\sqrt{\min\{L, K^n\}T})$	$\tilde{\mathcal{O}}(\min\{\sqrt{n^L K^L A^{2L} L}, K^n\}\sqrt{T})$
Low. Bound	$\Omega(\sqrt{\min\{L, nK\}T})$	$\Omega(\sqrt{\min\{L, nK\}T})$	$\Omega(\sqrt{\min\{L, nK\}T})$

084 **Related Works:** Our work contributes to the growing literature on the computational and learning  
 085 aspects of Stackelberg games (Conitzer & Sandholm, 2006; Conitzer & Korzhyk, 2011; Castiglioni  
 086 et al., 2020; Zhu et al., 2023). In particular, Letchford et al. (2009); Peng et al. (2019); Bacchocchi  
 087 et al. (2024) study learning in single-follower non-Bayesian Stackelberg games. Like our work,  
 088 they assume that the follower myopically best responds in each round. However, they assume the  
 089 follower’s utility function to be fixed but unknown, whereas we consider a Bayesian setting in which  
 090 followers have unknown stochastic types that parameterize a known utility function.  
 091

092 Closer to our work, Balcan et al. (2015; 2025) design online learning algorithms with  $\text{poly}(K)\sqrt{T}$   
 093 regrets for Bayesian Stackelberg games with a single follower with unknown type distribution, while  
 094 Bernasconi et al. (2023) obtain  $\tilde{\mathcal{O}}(K^{3n/2}\sqrt{T})$  regret for multi-receiver Bayesian persuasion problem  
 095 (which is similar to multi-follower Bayesian Stackelberg game) by a reduction to adversarial linear  
 096 bandit problem. Adopting previous techniques would lead to a  $\text{poly}(K^n)\sqrt{T}$  regret bound in our  
 097 multi-follower setting, which is exponential in the number of followers  $n$  (see details in Section 5)  
 098 and undesirable when followers are many. Using a different approach, we design an algorithm with  
 099  $\tilde{\mathcal{O}}(\sqrt{n^L K^L A^{2L} LT})$  regret, a better result when the number of leader’s actions  $L$  is small compared  
 100 to  $n$ . The exponential dependency on  $L$  is unavoidable from a computational perspective, as Conitzer  
 101 & Sandholm (2006) show that BSGs are NP-Hard to solve with respect to  $L$ . Our algorithm combines  
 102 the Upper Confidence Bound (UCB) principle and a partition of the leader’s strategy space into  
 103 best-response regions, which is a novel approach to our knowledge.

103 Online Bayesian Stackelberg game can be seen as a piecewise linear stochastic bandit problem.  
 104 While linear stochastic bandit problems have been well studied (Auer et al., 2002; Dani et al., 2008;  
 105 Abbasi-yadkori et al., 2011), piecewise linearity brings additional challenges. Bacchocchi et al.  
 106 (2025) study a single-dimensional piecewise linear stochastic bandit problem with unknown pieces;  
 107 in contrast, we have known pieces but a multi-dimensional space, so the techniques and results of  
 108 that work and ours are not directly comparable.

108  
109  

## 2 MODEL

110  
111 **Multi-Follower Bayesian Stackelberg Game:** We consider the interactions between a single  
112 *leader* and  $n \geq 1$  *followers*. The leader has  $L \geq 2$  actions, denoted by  $\mathcal{L} = [L] = \{1, \dots, L\}$ , and  
113 chooses a mixed strategy  $x \in \Delta(\mathcal{L})$  over them, where  $\Delta(\mathcal{L})$  is the space of probability distributions  
114 over the action set. We use  $x(\ell)$  to denote the probability of the leader playing action  $\ell \in \mathcal{L}$ .<sup>1</sup>  
115 Each follower has a finite action set  $\mathcal{A} = [A]$ . We represent the joint action of the  $n$  followers as  
116  $\mathbf{a} = (a_1, \dots, a_n)$ . Each follower  $i$  also has a private type  $\theta_i \in \Theta = [K]$ , with the vector of all follower  
117 types denoted by  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in \Theta^n$ . We consider a Bayesian setting where this type vector is  
118 drawn from a distribution  $\mathcal{D}$  (i.e.  $\boldsymbol{\theta} \sim \mathcal{D}$ ), with  $\mathcal{D}_i$  denoting the marginal distribution of  $\theta_i$ . The  
119 properties of this joint distribution play a key role in our results. We will consider two scenarios:  
120

- Independent type distributions: The followers' types are independent:  $\mathcal{D} = \mathcal{D}_1 \times \dots \times \mathcal{D}_n$ .
- General type distributions: The followers' types can be arbitrarily correlated.

121 If the leader selects action  $\ell$  and the followers select joint action  $\mathbf{a}$ , the leader receives utility  
122  $u(\ell, \mathbf{a}) \in [0, 1]$  and each follower  $i$  receives utility  $v_i(\ell, a_i, \theta_i) \in [0, 1]$ . Observe that each follower's  
123 utility depends only on their own action and type, alongside the leader's action; it does not depend  
124 on the actions of other followers.<sup>2</sup> For a mixed strategy  $x \in \Delta(\mathcal{L})$  and followers' actions  $\mathbf{a}$ , the  
125 leader's expected utility is given by  $u(x, \mathbf{a}) = \mathbb{E}_{\ell \sim x}[u(\ell, \mathbf{a})] = \sum_{\ell \in \mathcal{L}} x(\ell)u(\ell, \mathbf{a})$ . Likewise, the  $i$ th  
126 follower's expected utility under  $x$  is  $v_i(x, a_i, \theta_i) = \mathbb{E}_{\ell \sim x}[v_i(\ell, a_i, \theta_i)]$ . We assume that the leader  
127 knows each follower's utility function but not their private types.  
128

129 An instance of a multi-follower Bayesian Stackelberg game is defined by the tuple  $I =$   
130  $(n, L, A, K, u, v, \mathcal{D})$ . In this game, the leader first commits to a mixed strategy  $x$  without knowledge  
131 of the followers' types. The follower types are then jointly realized from  $\mathcal{D}$ , and each follower selects  
132 a *best-responding* action based on the leader's strategy. It is without loss of generality to consider  
133 followers choosing pure action since follower utilities are independent of one another.  
134

135 **Definition 2.1** (Followers' Best Response). *For a leader's mixed strategy  $x$ , the best response of a  
136 follower  $i$  with realized type  $\theta_i$  is given by  $\text{br}_i(\theta_i, x) \in \arg \max_{a \in \mathcal{A}} v_i(x, a, \theta_i)$ .<sup>3</sup> The vector of best  
137 responses is denoted by  $\text{br}(\boldsymbol{\theta}, x) = (\text{br}_1(\theta_1, x), \dots, \text{br}_n(\theta_n, x))$ .*  
138

139 Let  $U_{\mathcal{D}}(x) = \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[u(x, \text{br}(\boldsymbol{\theta}, x))]$  denote the leader's expected utility when the leader commits to  
140 mixed strategy  $x$ , the followers have their types drawn from  $\mathcal{D}$  and best respond.  
141

142 **Definition 2.2** (Leader's Optimal Strategy). *For a joint follower type distribution  $\mathcal{D}$ , the leader's  
143 optimal strategy, also known as the Stackelberg Equilibrium, is given as follows:*  
144

$$x^* \in \arg \max_{x \in \Delta(\mathcal{L})} U_{\mathcal{D}}(x) = \arg \max_{x \in \Delta(\mathcal{L})} \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[u(x, \text{br}(\boldsymbol{\theta}, x))].$$

145 **Online Learning Model:** When the leader knows the distribution  $\mathcal{D}$ , they can compute the optimal  
146 strategy by solving the optimization problem specified in Definition 2.2. Indeed, this is the premise  
147 of Conitzer & Sandholm (2006). Our work examines an online learning model where the leader does  
148 not know the type distribution  $\mathcal{D}$  a priori; instead, the leader must learn the optimal strategy through  
149 feedback from repeated interactions with followers over  $T$  rounds.  
150

151 We examine two feedback models. In the *type feedback* setting, the leader observes the types  $\boldsymbol{\theta}^t$  of  
152 the followers after each round  $t$ , whereas in the *action feedback* setting, the leader only observes  
153 the actions  $\mathbf{a}^t$  of the followers. Note that type feedback is strictly more informative than action  
154 feedback since the follower's actions can be inferred from their types by computing their best response  
155 (Definition 2.1). We summarize the interactions at a given round  $t$  as follows:  
156

1. The leader chooses a strategy  $x^t \in \Delta(\mathcal{L})$ .
2. Follower types for this round are realized:  $\boldsymbol{\theta}^t \sim \mathcal{D}$ .

157 <sup>1</sup>All of our results can be generalized to the setting where the leader's strategy space is an arbitrary compact  
158 convex set  $\mathcal{X} \subseteq \mathbb{R}^d$  and the leader and followers' utility functions  $u(x, \mathbf{a}), v_i(x, a_i, \theta_i)$  are linear (or affine)  
159 functions of  $x \in \mathcal{X}$ . Our presentation focuses on  $\mathcal{X} = \Delta(\mathcal{L})$  for simplicity.  
160

<sup>2</sup>This no externality assumption is common in modeling a large population of agents (Dughmi & Xu, 2017;  
Xu, 2020; Castiglioni et al., 2020).

<sup>3</sup>In case of ties, we assume that followers break ties in favor of the leader.

162 3. Followers take their best-responding actions  $\mathbf{a}^t = \mathbf{br}(\boldsymbol{\theta}, x^t)$  and the leader gets utility  $u(x^t, \mathbf{a}^t)$ .  
 163 4. Under type feedback, the leader observes the type profile  $\boldsymbol{\theta}^t$ . Under action feedback, the leader  
 164 observes only the followers' actions  $\mathbf{a}^t$ .  
 165

166 The leader deploys a learning algorithm (based on past feedback) to select strategy  $x^t$  for every round  
 167  $t$ . We study learning algorithms that minimize the cumulative *regret* with respect to the optimal  
 168 equilibrium strategy (Definition 2.2). Formally defined below, minimizing this objective requires a  
 169 careful balance between exploring the strategy space while not taking too many sub-optimal strategies.  
 170

**Definition 2.3.** *The regret of a learning algorithm that selects strategy  $x^t$  at round  $t \in [T]$  is:*

$$171 \quad \text{Reg}(T) = \sum_{t=1}^T \mathbb{E}_{\boldsymbol{\theta}^t \sim \mathcal{D}} [u(x^*, \mathbf{br}(\boldsymbol{\theta}^t, x^*)) - u(x^t, \mathbf{br}(\boldsymbol{\theta}^t, x^t))] = \sum_{t=1}^T (U_{\mathcal{D}}(x^*) - U_{\mathcal{D}}(x^t)).$$

175 Note that  $\text{Reg}(T)$  is a random variable, because the selection of  $x^t$  depends on the past type  
 176 realizations  $\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^{t-1}$ . We aim to minimize the expected regret  $\mathbb{E}[\text{Reg}(T)]$ .  
 177

178 Lastly, our model assumes followers behave *myopically*, selecting their best actions based only on  
 179 the leader's current strategy, without considering future rounds. This is consistent with the related  
 180 literature (Peng et al., 2019; Bacchocchi et al., 2024; Letchford et al., 2009) and well-motivated in  
 181 settings like online platforms or security games where followers maximize their immediate utility.  
 182

### 3 BEST RESPONSE REGIONS: A GEOMETRIC PERSPECTIVE

184 Since the followers' best-responding actions are sensitive to the leader's strategy  $x$ , the leader's  
 185 expected utility function  $U_{\mathcal{D}}(x)$  is discontinuous in  $x$ . This presents a key challenge to both learning  
 186 and optimizing over the leader's strategy space. To overcome this challenge, we first show that the  
 187 leader's strategy space  $\Delta(\mathcal{L})$  can be partitioned into a polynomial number of non-empty *best-response*  
 188 *regions* (followers have the same best-response actions within each region). While the notion of  
 189 best-response regions has been proposed by prior works (Balcan et al., 2015; Peng et al., 2019;  
 190 Bacchocchi et al., 2024; Yang & Zhang, 2024), those works consider single-follower cases. With  
 191 multiple followers, we will argue that the number of such regions does not increase exponentially in  
 192 the number of followers  $n$  (Lemma 3.2) – a key property to be used in later sections. **At a high level,**  
 193 **the best-response region approach allows us to reason about the leader strategy space in a discrete**  
 194 **sense. This is not only instructive for regret analysis (such as for Theorem 4.1) but also facilitates**  
 195 **leveraging algorithms like UCB (Algorithm 3), which are defined for discrete settings.**

#### 3.1 A SINGLE FOLLOWER

198 To build intuition, we first consider the leader playing against a single follower ( $n = 1$ ). The follower  
 199 has a utility function  $v(\ell, a, \theta)$  and a type  $\theta \in \Theta = [K]$  drawn from distribution  $\mathcal{D}$ . Next, let  
 200  $w : \Theta \rightarrow \mathcal{A}$  be a mapping from follower type to action – i.e.  $w(\theta)$  specifies an action for type  $\theta$ . For  
 201 such a mapping  $w$ , let  $R(w) \subseteq \Delta(\mathcal{L})$  be the set of leader strategies under which the follower's best  
 202 response action  $\mathbf{br}(\theta, x)$  is equal to  $w(\theta)$  for every type  $\theta \in \Theta$ . Formally:

$$203 \quad R(w) = \left\{ x \in \Delta(\mathcal{L}) \mid \mathbf{br}(\theta, x) = w(\theta), \forall \theta \in \Theta \right\} \\ 204 \quad = \left\{ x \in \Delta(\mathcal{L}) \mid v(x, w(\theta), \theta) \geq v(x, a', \theta), \forall \theta \in \Theta, \forall a' \in \mathcal{A} \right\}$$

208 where we recall that for any action  $a$ ,  $v(x, a, \theta) = \sum_{\ell \in \mathcal{L}} x(\ell) v(\ell, a, \theta)$ . The set  $R(w)$  is defined  
 209 as the *best-response region* for mapping  $w$ . This region can also equivalently be defined as the  
 210 intersection of several halfspaces (see Figure 1 in Appendix A for a visual). In particular, let

$$211 \quad d_{\theta, a, a'} = [v(1, a, \theta) - v(1, a', \theta), \dots, v(L, a, \theta) - v(L, a', \theta)]^T \in \mathbb{R}^L$$

213 denote the “advantage” of follower type  $\theta$  taking action  $a$  over  $a'$  at each of the  $L$  possible leader  
 214 actions. Then the halfspace  $H(d_{\theta, a, a'}) = \{x \in \Delta(\mathcal{L}) \mid \langle x, d_{\theta, a, a'} \rangle \geq 0\}$  contains all the leader  
 215 strategies under which the follower with type  $\theta$  prefers action  $a$  over  $a'$ . Thus, the best-response  
 region is  $R(w) = \bigcap_{\theta \in \Theta, a \in \mathcal{A}} H(d_{\theta, w(\theta), a})$ , the intersection of  $|\Theta| \cdot |\mathcal{A}| = KA$  halfspaces.

216 3.2 MULTIPLE FOLLOWERS  
217

218 We generalize the intuitions from the single-follower case to the multi-follower case. Let  $W =$   
219  $(w_1, \dots, w_n)$  denote a tuple of  $n$  mappings, where each  $w_i : \Theta \rightarrow \mathcal{A}$  is the best-response mapping  
220 for follower  $i$ . So,  $W$  is a mapping from joint type space  $\Theta^n$  to joint action space  $\mathcal{A}^n$ , where  
221  $W(\boldsymbol{\theta}) = (w_1(\theta_1), \dots, w_n(\theta_n)) \in \mathcal{A}^n$  denotes the joint action of all followers under joint type  
222  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ . Alternatively, one can think of  $W$  as a matrix,  $W \in \mathcal{A}^{n \times K}$ , where each entry  
223  $w_{ik} \in \mathcal{A}$  records the best-response action for follower  $i$  if he has type  $\theta_i = k$ . We generalize the  
224 notion of best-response region  $R(W)$  from the single-follower case to the multi-follower case:

225 **Definition 3.1** (Best-Response Region). *For a matrix  $W \in \mathcal{A}^{n \times K}$ , the best-response region for  $W$   
226 is the set of leader strategies under which the followers' best responses are given by  $W$ :*

$$227 \quad R(W) = \left\{ x \in \Delta(\mathcal{L}) \mid \text{br}(\boldsymbol{\theta}, x) = W(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta^n \right\}. \\ 228$$

229 As in the single-follower case,  $R(W)$  can be expressed as the intersection of multiple halfspaces:  
230  $R(W) = \bigcap_{i \in [n], \theta_i \in \Theta, a_i \in \mathcal{A}} H(d_{\theta_i, w_i(\theta_i), a_i})$ .

232 We make an important observation: the leader's expected utility function  $U_{\mathcal{D}}(x)$  is linear in  $x$  within  
233 each non-empty best-response region. By definition, for all  $\boldsymbol{\theta} \in \Theta^n$  and  $x \in R(W)$ , we have  
234  $\text{br}(\boldsymbol{\theta}, x) = W(\boldsymbol{\theta})$ . So,

$$235 \quad U_{\mathcal{D}}(x) = \sum_{\boldsymbol{\theta} \in \Theta^n} \mathcal{D}(\boldsymbol{\theta}) u(x, \text{br}(\boldsymbol{\theta}, x)) = \sum_{\boldsymbol{\theta} \in \Theta^n} \mathcal{D}(\boldsymbol{\theta}) \sum_{\ell \in \mathcal{L}} x(\ell) u(\ell, W(\boldsymbol{\theta})) = \sum_{\boldsymbol{\theta} \in \Theta^n} \mathcal{D}(\boldsymbol{\theta}) \langle x, z_{W, \boldsymbol{\theta}} \rangle.$$

238 where  $z_{W, \boldsymbol{\theta}}$  is the  $L$ -dimensional vector  $z_{W, \boldsymbol{\theta}} = (u(1, W(\boldsymbol{\theta})), \dots, u(L, W(\boldsymbol{\theta})))$ . So, we conclude  
239 that the leader's expected utility is linear within each region  $R(W)$ :

240 **Lemma 3.1.** *For each  $W$ , the leader's expected utility function  $U_{\mathcal{D}}(x)$  is linear in  $x \in R(W)$ .*

242 Although  $U_{\mathcal{D}}(x)$  is linear *within* each best-response region, it could be non-linear and even discontinuous  
243 across different best-response regions.

245 3.3 ENUMERATING BEST-RESPONSE REGIONS AND COMPUTING THE OFFLINE OPTIMAL  
246

247 Let  $\mathcal{W} = \{W \in \mathcal{A}^{n \times K} \mid R(W) \neq \emptyset\}$  denote the set of mappings  $W$  for which the corresponding  
248 best-response region  $R(W)$  is non-empty. Although the total number of  $W \in \mathcal{A}^{n \times K}$  is  $A^{n \times K}$ , the  
249 number of *non-empty* best-response regions is significantly smaller, especially when  $L$  (number of  
250 actions of the leader) is treated as a constant. The exact characterization is given below. The proof (in  
251 Appendix A) uses a result in computational geometry regarding the number of nonempty regions  
252 obtained by dividing  $\mathbb{R}^L$  using  $\mathcal{O}(nKA^2)$  hyperplanes.

253 **Lemma 3.2.** *The number of non-empty best-response regions,  $|\mathcal{W}|$ , is  $\mathcal{O}(n^L K^L A^{2L})$ .*

254 For any algorithm to leverage these best response regions, it is imperative that these regions can be  
255 enumerated efficiently. The following lemma shows this is always possible. Intuitively, we construct  
256 a graph where the nodes represent non-empty best-response regions and an edge exists between  
257  $W, W' \in \mathcal{W}$  if and only if  $W$  and  $W'$  differ in exactly one entry. Traversing an edge, therefore,  
258 corresponds to moving to an adjacent best-response region by crossing a single hyperplane boundary.  
259 We show that this graph is always connected and can thus be efficiently traversed using breadth-first  
260 search. The exact algorithm and proof of Lemma 3.3 are in Appendix A.

261 **Lemma 3.3.** *The set of non-empty best-response regions  $\{R(W) : W \in \mathcal{W}\}$  can be enumerated in  
262  $\text{poly}(n^L, K^L, A^L, L)$  time.*

264 We now show that the optimal strategy within each region can be efficiently computed. Recall from  
265 Definition 2.2 that, when given the followers' type distribution  $\mathcal{D}$ , computing the leader's optimal  
266 strategy requires solving  $\max_{x \in \Delta(\mathcal{L})} \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[u(x, \text{br}(\boldsymbol{\theta}, x))]$ . Since the leader's utility is linear within  
267 a region  $R(W)$ , the optimal solution within  $R(W)$  can be computed by the following linear program:

$$268 \quad \max_{x \in R(W)} \sum_{\boldsymbol{\theta} \in \Theta^n} \mathcal{D}(\boldsymbol{\theta}) u(x, \text{br}(\boldsymbol{\theta})) = \max_{x \in R(W)} \sum_{\boldsymbol{\theta} \in \Theta^n} \mathcal{D}(\boldsymbol{\theta}) u(x, W(\boldsymbol{\theta})) \quad (1)$$

270 where  $x \in R(W)$  is given by the following set of linear constraints:  
 271

$$\begin{cases} \sum_{\ell \in \mathcal{L}} x(\ell) [v_i(\ell, w_i(\theta_i), \theta_i) - v_i(\ell, a'_i, \theta_i)] \geq 0, \quad \forall i \in [n], \forall \theta_i \in \Theta, \forall a'_i \in \mathcal{A}, \\ x(\ell) \geq 0, \quad \forall \ell \in \mathcal{L}, \text{ and } \sum_{\ell \in \mathcal{L}} x(\ell) = 1. \end{cases} \quad (2)$$

272 While there are  $\mathcal{O}(nKA)$  constraints, each involving a sum over  $L$  elements, the objective involves  
 273 summing over  $K^n$  possible type profiles. While this is exponential in  $n$ , any input to the complete  
 274 information instance must provide the joint type distribution  $\mathcal{D} \in [0, 1]^{K^n}$  as input. Thus, the time to  
 275 compute the optimal solution within each region is polynomial in the input size.  
 276

277 The above results imply that, given distribution  $\mathcal{D}$ , the optimal leader strategy in BSGs can be  
 278 computed efficiently when the number  $L$  of leader's actions is small. This is because the optimal  
 279 strategy within each best-response region  $R(W)$  can be computed efficiently by the linear program  
 280 (1), the overall optimal strategy is the maximum over all non-empty best-response regions, and  
 281 there are at most  $\mathcal{O}(n^L K^L A^{2L})$  such regions by Lemma 3.2. **We thus showed above that BSGs**  
 282 **are polynomial-time solvable for a constant  $L$ . In comparison, Theorem 7 of Conitzer & Sandholm**  
 283 **(2006) proves that the optimal strategy is NP-hard to compute in BSGs when  $L$  is treated as a variable**  
 284 **and is asymptotically increasing.**  
 285

## 287 4 TYPE FEEDBACK

### 289 4.1 LEARNING ALGORITHMS AND UPPER BOUNDS

290 **General Type Distributions:** We now address the core problem of learning the optimal leader  
 291 strategy from online feedback. This section considers the type-feedback setting, where the leader  
 292 observes each follower's realized type  $\theta^t = (\theta_1^t, \dots, \theta_n^t)$  at the end of round  $t$ . We start with general  
 293 distributions – that is, the followers' types can be arbitrarily correlated. Observing types after each  
 294 round allows us to directly estimate the unknown distribution  $\mathcal{D}$  and compute an optimal strategy  
 295 accordingly. This is formalized in Algorithm 1:  
 296

---

#### 298 **ALGORITHM 1:** Type-Feedback Algorithm – General Type Distributions

---

299 At round  $t = 1$ , pick an arbitrary strategy  $x^1 \in \Delta(\mathcal{L})$ .

300 **for** round  $t \geq 2$  **do**

301    Choose  $x^t \in \arg \max_{x \in \Delta(\mathcal{L})} \sum_{s=1}^{t-1} u(x, \text{br}(\theta^s, x))$  – the empirically optimal strategy.  
 302    Observe the followers' types  $\theta^t \sim \mathcal{D}$ .

---

303  
 304 At first glance, one might think that this algorithm might suffer a large regret because the distribution  
 305  $\mathcal{D}$ , which has support size  $|\Theta^n| = K^n$ , is difficult to estimate. Indeed, the estimation error for such a  
 306 distribution using  $t$  samples is at least  $\Omega(\sqrt{\frac{K^n}{t}})$  even if  $\mathcal{D}$  is a product distribution (namely, the types  
 307 are independent) (Lin, 2022). This suggests that the empirically optimal strategy  $x^t$  might be worse  
 308 than the true optimal strategy  $x^*$  by at least  $\Omega(\sqrt{\frac{K^n}{t}})$ , which would cause an  $\Omega(\sqrt{K^n T})$  regret in  $T$   
 309 rounds in total. As we will show in Theorem 4.1, one analysis of Algorithm 1 achieves exactly this as  
 310 a regret upper bound. The proof (in Appendix B.2) upper bounds the single-round regret by the total  
 311 variation (TV) distance between the empirical distribution  $\hat{\mathcal{D}}^t$  and the true distribution  $\mathcal{D}$ .  
 312

313 While this suggests that  $\mathcal{O}(\sqrt{K^n T})$  regret might be tight, this is interestingly not true when  $n$  is  
 314 large! That is, the intuitive lower bound that arises from the estimation error for distribution  $\mathcal{D}$   
 315 is not correct. Although the empirical type distribution can differ significantly from the true type  
 316 distribution, the empirical *utility* of any strategy  $x \in \Delta(\mathcal{L})$  is actually concentrated around the true  
 317 expected utility of  $x$  with high probability. We formalize this below:  
 318

319 **Lemma 4.1.** *Given  $t$  samples  $\theta^1, \dots, \theta^t$  from distribution  $\mathcal{D}$ , let  $\hat{U}^t(x) = \frac{1}{t} \sum_{s=1}^t u(x, \text{br}(\theta^s, x))$   
 320 be the empirical expected utility of a strategy  $x \in \Delta(\mathcal{L})$  computed on the  $t$  samples. Recall that  
 321  $U_{\mathcal{D}}(x) = \mathbb{E}_{\theta \sim \mathcal{D}}[u(x, \text{br}(\theta, x))]$  denotes the true expected utility of  $x$ . With probability at least  $1 - \delta$ ,  
 322 we have: for all  $x \in \Delta(\mathcal{L})$ ,  $|U_{\mathcal{D}}(x) - \hat{U}^t(x)| \leq \mathcal{O}\left(\sqrt{\frac{L \log t}{t}} + \sqrt{\frac{L \log(nKA) + \log(1/\delta)}{t}}\right)$ .*

324 *Proof sketch.* By Lemma 3.2, the leader’s strategy space  $\Delta(\mathcal{L})$  can be divided into  $|\mathcal{W}| =$   
 325  $\mathcal{O}(n^L K^L A^{2L})$  best-response regions, and the leader’s utility function  $U_{\mathcal{D}}(x)$  is linear inside each  
 326 region (Lemma 3.1). Because the *pseudo-dimension* of linear functions in an  $L$  dimensional space  
 327 are at most  $L$ , we have with probability at least  $1 - \delta'$ , the empirical utility  $\hat{U}^t(x)$  on  $t$  samples  
 328 approximates the true expected utility  $U_{\mathcal{D}}(x)$  with accuracy  $\mathcal{O}\left(\sqrt{\frac{L \log t}{t}} + \sqrt{\frac{\log(1/\delta')}{t}}\right)$  for every  
 329 strategy  $x$  inside a best-response region. Taking a union bound over all  $\mathcal{O}(n^L K^L A^{2L})$  best-response  
 330 regions, i.e., letting  $\delta' = \delta/\mathcal{O}(n^L K^L A^{2L})$ , proves the lemma. See details in Appendix B.1.  $\square$   
 331

332

333 Note that the above concentration result holds for all strategies  $x \in \Delta(\mathcal{L})$  simultaneously, instead  
 334 of for a single fixed strategy (which easily follows from Hoeffding’s inequality). This result means  
 335 that the simple Algorithm 1 can achieve a regret that is of the order  $\sqrt{T}$ , logarithmic in  $n$ , with an  
 336 additional  $\sqrt{L}$  factor. This is better for large  $n$  and small  $L$ . This new regret bound, along with the  
 337 earlier one  $\mathcal{O}(\sqrt{K^n T})$ , is formalized in Theorem 4.1 below, with the proof given in Appendix B.2.

338

339 **Theorem 4.1.** *The type-feedback Algorithm 1 for general type distributions achieves expected regret*  
 340  $\mathcal{O}(\min\{\sqrt{LT \cdot \log(nKAT)}, \sqrt{K^n T}\})$  *and can be implemented in  $\text{poly}((nKA)^L LT)$  time.*

341

342 Theorem 4.1 also comments on the runtime of Algorithm 1, which hinges on the computability  
 343 of  $x^t \in \arg \max_{x \in \Delta(\mathcal{L})} \sum_{s=1}^{t-1} u(x, \mathbf{br}(\boldsymbol{\theta}^s, x))$ . Using the techniques developed in Section 3, this  
 344 maximization can be solved by taking the maximum over the optimal strategies from each non-empty  
 345 best-response region  $W \in \mathcal{W}$ , computed using the empirical type distribution. Using Lemmas  
 346 3.2 and 3.3 and the fact that the optimal strategy within a non-empty  $R(W)$  can be solved by the  
 347 following linear program, we obtain a runtime that is polynomial when  $L$  is constant:<sup>4</sup>

$$348 \max_{W \in \mathcal{W}} \left\{ \max_{x \in R(W)} \sum_{s=1}^t u(x, W(\boldsymbol{\theta}^s)) \text{ subject to the constraints in (2)} \right\}. \quad (3)$$

349

350 **Independent Type Distributions:** Algorithm 1 and the corresponding regret bound in Theorem  
 351 4.1 hold without any assumptions on the joint type distribution  $\mathcal{D}$ . In many settings, however, the  
 352 followers’ types may be independent of one another. Intuitively, one expects learning to be easier in  
 353 such settings since it suffices to learn the marginals as opposed to the richer joint distribution. This is  
 354 indeed correct: in Algorithm 2, we build the empirical distribution  $\hat{\mathcal{D}}_i^t$  for each marginal from samples  
 355  $\theta_i^1, \dots, \theta_i^t$  for follower  $i$  and then take the product  $\hat{\mathcal{D}}^t = \prod_{i=1}^n \hat{\mathcal{D}}_i^t$  to estimate  $\mathcal{D} = \prod_{i=1}^n \mathcal{D}_i$ .

356

---

357 **ALGORITHM 2:** Type-Feedback Algorithm - Independent Type Distributions

---

359 At  $t = 1$ , pick an arbitrary strategy  $x^1 \in \Delta(\mathcal{L})$ .

360 **for**  $round t > 1$  **do**

361   Choose  $x^t \in \arg \max_{x \in \Delta(\mathcal{L})} \mathbb{E}_{\boldsymbol{\theta} \sim \hat{\mathcal{D}}^{t-1}} [u(x, \mathbf{br}(\boldsymbol{\theta}, x))]$   
 362   Observe realized follower type  $(\theta_1^t, \dots, \theta_n^t)$   
 363   **for**  $i \in [n], k \in \Theta$  **do**  
 364      $\hat{\mathcal{D}}_i^t(k) = \frac{1}{t} \sum_{s=1}^t \mathbf{1}[\theta_i^s = k]$   
 365     $\hat{\mathcal{D}}^t(\boldsymbol{\theta}) = \prod_{i=1}^n \hat{\mathcal{D}}_i^t(\theta_i)$ ,  $\forall \boldsymbol{\theta} \in \Theta^n$

---

367

368 This algorithm achieves a much improved regret,  $\mathcal{O}(\sqrt{nKT})$ , formalized in Theorem 4.2 and  
 369 empirically verified in Appendix D. The proof (in Appendix B.3) is similar to the  $\mathcal{O}(\sqrt{K^n T})$  regret  
 370 analysis of Theorem 4.1, which upper bounds the single-round regret by the TV distance between  $\hat{\mathcal{D}}^t$   
 371 and  $\mathcal{D}$ . But for independent distributions, we can relate the TV distance with the sum of Hellinger  
 372 distances between the marginals  $\hat{\mathcal{D}}_i^t$  and  $\mathcal{D}_i$ , which is bounded by  $\mathcal{O}(\sqrt{\frac{nK}{t}})$  instead of  $\mathcal{O}(\sqrt{\frac{K^n}{t}})$ ,  
 373 so the total regret is bounded by  $\mathcal{O}(\sqrt{nKT})$ . The computational complexity, though, increases to  
 374  $\text{poly}((nKA)^L LT K^n)$  as the empirical product distribution  $\hat{\mathcal{D}}^t = \prod_i^n \hat{\mathcal{D}}_i^t$  has support size  $K^n$ .

375

<sup>4</sup>Also note that Algorithm 1 does not need as input the entire utility function of the leader  $u(\cdot, \cdot)$ , which has an exponential size  $L \cdot A^n$ . The algorithm only needs the utility function for the sampled types.

378 **Theorem 4.2.** *The type-feedback Algorithm 2 for independent type distributions achieves expected  
 379 regret  $\mathcal{O}(\sqrt{nKT})$  and can be implemented in  $\text{poly}((nKA)^L LT K^n)$  time.*  
 380

381 **Corollary 4.1.** *Taking the minimum of Theorems 4.1 and 4.2, we obtain a type-feedback algorithm  
 382 with expected regret  $\mathcal{O}(\min\{\sqrt{LT \cdot \log(nKAT)}, \sqrt{nKT}\})$  for independent type distributions.*  
 383

384 **4.2 LOWER BOUND**  
 385

386 We then provide a lower bound result: no algorithm for online Bayesian Stackelberg game has a  
 387 better regret than  $\Omega(\sqrt{\min\{L, nK\}T})$ . When the number of followers  $n$  is large, this lower bound  
 388 matches the previous upper bounds  $\tilde{\mathcal{O}}(\sqrt{LT})$ . To our knowledge, this work is the first to provide a  
 389 lower bound for the multi-follower problem and give an almost tight characterization of the factor  
 390 before the classical  $\sqrt{T}$  term. Interestingly, this  $\tilde{\mathcal{O}}(\sqrt{L})$  factor does not grow with  $n$  up to log factor.  
 391

392 **Theorem 4.3.** *The expected regret of any type-feedback algorithm is at least  $\Omega(\sqrt{\min\{L, nK\}T})$ .  
 393 This holds even if the followers' types are independent and the leader's utility does not depend  $\ell$ .*  
 394

395 The proof (given in Appendix B.5) involves two non-trivial reductions. First, we reduce the  
 396 *distribution learning* problem to a *single-follower* Bayesian Stackelberg game, obtaining an  
 397  $\Omega(\sqrt{\min\{L, K\}T})$  lower bound. Then, we reduce the single-follower game with  $nK$  types to  
 398 a game with  $n$  followers each with  $K$  types. One might wish to reduce a single-follower game with  
 399  $K^n$  types to an  $n$ -follower game to prove a lower bound of  $\Omega(\sqrt{\min\{L, K^n\}T})$  for general type  
 400 distributions, but that cannot be done easily due to no externality between the followers.  
 401

402 **5 ACTION FEEDBACK**  
 403

404 We now discuss the setting where the leader observes the followers' actions after each round. This  
 405 setting is more practical yet challenging than the type-feedback setting. We present two learning  
 406 algorithms. The first algorithm achieves  $\mathcal{O}(K^n \sqrt{T} \log T)$  regret, using a previous technique from  
 407 Bernasconi et al. (2023). The second algorithm involves a novel combination of the Upper Confidence  
 408 Bound principle and the concentration analysis of best-response regions from Lemma 4.1, achieving  
 409  $\mathcal{O}(\sqrt{n^L K^L A^{2L} LT \log T})$  regret. The latter is better when the number of followers  $n$  is large and  
 410 the number of leader actions  $L$  is small. We empirically simulate both approaches in Appendix D.  
 411

412 **Linear-bandit based approach with  $\mathcal{O}(K^n \sqrt{T} \log T)$  regret:** Bernasconi et al. (2023) developed  
 413 a technique to reduce the online learning problem of solving a linear program with unknown objective  
 414 to a linear bandit instance. A spiritually similar approach can be applied here. While the optimization  
 415 problem for our Bayesian Stackelberg games (Definition 2.2) is not a linear program, we show that  
 416 under a different formulation, this can actually be solved by a single linear program (we explain the  
 417 details in the proof of Theorem 5.1). We can thus leverage the techniques of Bernasconi et al. (2023)  
 418 to reduce this to a linear bandit problem. Since Bernasconi et al. (2023) considers an adversarial  
 419 online learning environment (ours is stochastic), directly applying their technique will lead to a  
 420 sub-optimal  $\tilde{\mathcal{O}}(K^{\frac{3n}{2}} \sqrt{T})$  regret bound. Instead, we apply the OFUL algorithm for stochastic linear  
 421 bandit (Abbasi-yadkori et al., 2011) to obtain a better regret bound of  $\tilde{\mathcal{O}}(K^n \sqrt{T})$ . Instead, we apply  
 422 the OFUL algorithm for stochastic linear bandit (Abbasi-yadkori et al., 2011) to obtain a better regret  
 423 bound of  $\tilde{\mathcal{O}}(K^n \sqrt{T})$ . See details in Appendix C.1.  
 424

425 **Theorem 5.1.** *There exists an action-feedback algorithm for online Bayesian Stackelberg game with  
 426  $\mathcal{O}(K^n \sqrt{T} \log T)$  regret.*  
 427

428 **Algorithm 3 with  $\mathcal{O}(\sqrt{n^L K^L A^{2L} LT \log T})$  regret.** We design a better algorithm for large  $n$  and  
 429 small  $L$ , not using Bernasconi et al. (2023)'s technique but using the "concentration over best-response  
 430 regions" idea we developed in the previous sections. Recall from Section 3 that the leader's strategy  
 431 space can be partitioned into best-response regions:  $\Delta(\mathcal{L}) = \bigcup_{W \in \mathcal{W}} R(W)$ . When the leader plays  
 strategy  $x$  in a region  $R(W)$ , the followers' best-response function satisfies  $\text{br}(\theta, x) = W(\theta)$ , so

432 the leader's expected utility is

$$434 U(x, R(W)) = \sum_{\theta \in \Theta^n} \mathcal{D}(\theta) u(x, W(\theta)) = \sum_{\mathbf{a} \in \mathcal{A}^n} u(x, \mathbf{a}) \sum_{\theta | W(\theta) = \mathbf{a}} \mathcal{D}(\theta) = \sum_{\mathbf{a} \in \mathcal{A}^n} u(x, \mathbf{a}) \mathcal{P}(\mathbf{a} | R(W))$$

436 where  $\mathcal{P}(\mathbf{a} | R(W)) = \sum_{\theta \in \Theta^n: W(\theta) = \mathbf{a}} \mathcal{D}(\theta)$  denotes the probability that the followers jointly take  
 437 action  $\mathbf{a} \in \mathcal{A}^n$  when the leader plays  $x \in R(W)$ . Since the distribution  $\mathcal{P}(\cdot | R(W)) \in \Delta(\mathcal{A}^n)$   
 438 does not depend on  $x$  as long as  $x \in R(W)$ , playing  $N$  strategies  $x^1, \dots, x^N$  within  $R(W)$  yields  
 439  $N$  observations  $\mathbf{a}^1, \dots, \mathbf{a}^N \sim \mathcal{P}(\cdot | R(W))$ . Using these samples, we can estimate the utility of any  
 440 other strategy  $x \in R(W)$  within the same region. We define the empirical utility estimate on  $N$   
 441 samples of joint actions as  $\hat{U}_N(x, R(W)) = \frac{1}{N} \sum_{s=1}^N u(x, \mathbf{a}^s)$ .

442 **Lemma 5.1.** Suppose  $T \geq |\mathcal{W}|$ . With probability at least  $1 - \frac{1}{T^2}$ , we have:  $\forall W \in \mathcal{W}, \forall N \in$   
 443  $\{1, \dots, T\}, \forall x \in R(W), |U(x, R(W)) - \hat{U}_N(x, R(W))| \leq \sqrt{\frac{4(L+1) \log(3T)}{N}}$ .

444 The proof of this lemma is similar to the proof of Lemma 4.1 and given in Appendix C.2.

445 For each region  $W \in \mathcal{W}$ , let  $N^t(W) = \sum_{s=1}^{t-1} \mathbb{1}[x^s \in R(W)]$  be the number of times when strategies  
 446 in region  $R(W)$  were played in the first  $t-1$  rounds. Given the result in Lemma 5.1, we define an  
 447 Upper Confidence Bound (UCB) on the expected utility of the optimal strategy in region  $R(W)$ :

$$448 \text{UCB}^t(W) = \max_{x \in R(W)} \left\{ \hat{U}_{N^t(W)}(x, R(W)) \right\} + \sqrt{\frac{4(L+1) \log(3T)}{N^t(W)}}.$$

449 We design the following algorithm: at each round  $t$ , select the region  $W \in \mathcal{W}$  with the highest  
 450  $\text{UCB}^t(W)$ , play the empirically optimal strategy in that region, and increment  $N^t(W)$  by 1. Full  
 451 description of the algorithm is given in Algorithm 3.

---

452 **ALGORITHM 3:** Upper Confidence Bound (UCB) for Best-Response Regions

---

453 Let  $\mathcal{W} = \{W \mid R(W) \neq \emptyset\}$ .

454 **for**  $W \in \mathcal{W}$  **do**

455   | Choose any strategy  $x \in R(W)$  and observe a joint action.

456 **for**  $round t > |\mathcal{W}|$  **do**

457   | **for** each  $W \in \mathcal{W}$  **do**

458   |   | Let  $N^t(W) = \sum_{s=1}^{t-1} \mathbb{1}[W^s = W]$  be the number of times region  $R(W)$  was chosen.

459   |   | Let  $\hat{\mathcal{P}}^t(\cdot | R(W))$  be the empirical distribution of joint actions in the rounds where region

460   |   |  $R(W)$  was chosen:  $\hat{\mathcal{P}}^t(\mathbf{a} | R(W)) = \frac{1}{N^t(W)} \sum_{s=1}^{t-1} \mathbb{1}[W^s = W] \cdot \mathbb{1}[\mathbf{a}^s = \mathbf{a}]$ .

461   |   | Compute the empirically optimal strategy in region  $R(W)$ :

$$462 \hat{x}_{R(W)}^* = \arg \max_{x \in R(W)} \mathbb{E}_{\mathbf{a} \sim \hat{\mathcal{P}}^t(\cdot | R(W))} [u(x, \mathbf{a})],$$

463   |   | which has empirical utility  $\hat{u}_{R(W)}^* = \mathbb{E}_{\mathbf{a} \sim \hat{\mathcal{P}}^t(\cdot | R(W))} [u(\hat{x}_{R(W)}^*, \mathbf{a})]$ .

464   |   | Let  $\text{UCB}^t(W) = \hat{u}_{R(W)}^* + \sqrt{\frac{4(L+1) \log(3T)}{N^t(W)}}$ .

465   |   | Let  $W^t \in \arg \max_{W \in \mathcal{W}} \text{UCB}^t(W)$ .

466   |   | Play strategy  $x^t = \hat{x}_{R(W^t)}^*$  and observe joint action  $\mathbf{a}^t = (a_1^t, \dots, a_n^t)$ .

---

467 **Theorem 5.2.** Algorithm 3 has expected regret  $\mathcal{O}(\sqrt{n^L K^L A^{2L} L \cdot T \log T})$ .

468 While the full proof is in Appendix C.3, we sketch the intuition. In the classical multi-armed bandit  
 469 problem, the UCB algorithm has expected regret  $\mathcal{O}(\sqrt{mT \log T})$  where  $m$  is the number of arms.  
 470 In our setting, each best-response region corresponds to an arm, and the confidence bound for each  
 471 region is  $\mathcal{O}(\sqrt{\frac{L \log T}{N^t(W)}})$ . The number of arms/regions is  $m = |\mathcal{W}| = \mathcal{O}(n^L K^L A^{2L})$  by Lemma 3.2.

472 So, the regret of Algorithm 3 is at most  $\mathcal{O}(\sqrt{|\mathcal{W}| \cdot L \cdot T \log T}) = \mathcal{O}(\sqrt{n^L K^L A^{2L} \cdot L \cdot T \log T})$ .

473 **Corollary 5.1.** By taking the better algorithm in Theorems 5.1 and 5.2, we obtain an action-feedback  
 474 algorithm with  $\tilde{\mathcal{O}}(\min \{K^n, \sqrt{n^L K^L A^{2L} L}\} \sqrt{T})$  regret.

486 **Dependencies on various parameters:** Since action-feedback is more limited than type-feedback,  
 487 the lower bound in Theorem 4.3 immediately carries over and shows that the  $\tilde{\mathcal{O}}(\sqrt{T})$  regret bounds  
 488 here are tight in  $T$ . There are several subtleties in achieving tighter bounds on the remaining  
 489 parameters. Conitzer & Sandholm (2006) show that, even with known distributions, BSG are NP-  
 490 Hard to solve with respect to  $L$ ; so an exponential computational dependence on  $L$  is unavoidable,  
 491 even if the regret could be made independent of  $L$  as shown in our  $\mathcal{O}(K^n \sqrt{T} \log T)$  result. Whether  
 492 an online learning algorithm with  $\text{poly}(n, K, L)\sqrt{T}$  regret exists is an open question, but such an  
 493 algorithm will suffer an exponential runtime in  $L$  unless  $P = NP$ .  
 494

## 495 6 DISCUSSION

496 This work designed online learning algorithms for Bayesian Stackelberg games with multiple follow-  
 497 ers with unknown type distributions. Although the joint type space of  $n$  followers has an exponentially  
 498 large size  $K^n$ , we achieved significantly smaller regrets:  $\tilde{\mathcal{O}}(\sqrt{\min\{nK, L\}T})$  when the followers’  
 499 types are independent and observable,  $\tilde{\mathcal{O}}(\sqrt{\min\{K^n, L\}T})$  when followers’ types are correlated and  
 500 observable, and  $\tilde{\mathcal{O}}(\min\{K^n, \sqrt{n^L K^L A^{2L}}\} \sqrt{T})$  when only the followers’ actions are observed.  
 501 These results exploit various geometric properties of the leader’s strategy space. The type-feedback  
 502 bounds are tight in all parameters and the action-feedback bounds are tight in  $T$ . The exponential  
 503 dependency on  $L$  is unavoidable computationally (Conitzer & Sandholm, 2006). Further closing  
 504 the gaps between upper and lower regret bounds is an open question and will likely involve tradeoff  
 505 between different parameters and tradeoff between computation and regret.  
 506

### 507 6.1 CONNECTIONS TO THE ADVERSARIAL SETTING

508 Our work considers a stochastic setting – follower types are sampled from an unknown but fixed  
 509 distribution. Alternatively, one can consider a setting where followers’ types are adversarially  
 510 generated. We conjecture that the main technique behind our results, “concentration over best  
 511 response regions”, can be generalized to adversarial settings. For the action-feedback case, our  
 512 approach is to run a UCB algorithm to pick a best-response region at every round, and then pick the  
 513 empirically optimal strategy within that region. Importantly, the leader’s optimization problem within  
 514 each best-response region is linear. With stochastic follower types, the linear optimization problem  
 515 within each best-response region can be solved by empirical optimization – this algorithm needs to  
 516 be changed in the adversarial setting. In more detail:

- 517 • For the full-feedback case (leader observes the followers’ types after each round), we can run  
 518 an adversarial full-feedback multi-armed bandit algorithm (e.g., Multiplicative Weights Update)  
 519 to pick a best-response region at each round. Within each best-response region, the problem  
 520 becomes an adversarial online linear optimization problem with full feedback, to which we can  
 521 apply algorithms such as FTRL to choose leader strategies.
- 522 • For partial-feedback (leader observes the followers’ actions after each round), we can run an  
 523 adversarial partial-feedback MAB algorithm (e.g., EXP3) to pick a best-response region at  
 524 each round. Within each region, use an adversarial partial-feedback online linear optimization  
 525 algorithm to choose leader strategies.

526 Intuitively, best-response regions allow us (1) to tame the continuous leader action space and (2) to  
 527 reduce the problem to classical adversarial problems. Formally verifying this approach is open.

### 531 6.2 INTER-FOLLOWER EXTERNALITY

532 Our work assumed no externality between the followers. While the learning algorithms for multiple  
 533 followers without externality may not differ significantly from that of a single follower, one might  
 534 expect the regret bound to grow at the exponential rate of  $O(\sqrt{K^n T})$ , because the support size of the  
 535 joint type distribution is  $K^n$ . Naive analysis does lead to that regret bound. However, we show that  
 536 with a more refined analysis, the regret bounds can be improved significantly (Theorems 4.1, 5.2).  
 537

538 The presence of inter-follower externality leads to a simultaneous game between the followers, given  
 539 a leader strategy. Putting aside the computational difficulties of inter-follower Nash Equilibrium  
 (or assuming oracle access to it), some of our results in the type-feedback setting generalize. In

540 particular, our  $\mathcal{O}(\sqrt{K^n T})$  result for general type distributions and  $\mathcal{O}(\sqrt{nKT})$  result for independent  
 541 type distributions also apply to the setting with inter-follower externality. Our other results, which  
 542 are built on best-response region characterization and depend on the followers' independent utility  
 543 functions, do not generalize to the externality case. The introduction of inter-follower externality  
 544 significantly complicates the analysis and algorithm for the action feedback case, since actions are  
 545 now equilibrium responses, and not just best responses to leader action. More information is needed  
 546 about the equilibrium to learn from this feedback.

547

548 

## REFERENCES

549

550 Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear  
 551 stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Wein-  
 552 berger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Asso-  
 553 ciates, Inc., 2011. URL [https://proceedings.neurips.cc/paper\\_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf).

554 Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Ban-  
 555 dit Problem. *Machine Learning*, 47(2/3):235–256, 2002. ISSN 08856125. doi: 10.1023/A:  
 556 1013689704352. URL <http://link.springer.com/10.1023/A:1013689704352>.

557 Francesco Bacchicocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. The  
 558 Sample Complexity of Stackelberg Games, 2024. URL <https://arxiv.org/abs/2405.06977>.

559 Francesco Bacchicocchi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Regret Minimization  
 560 for Piecewise Linear Rewards: Contracts, Auctions, and Beyond. In *Proceedings of the 26th ACM  
 561 Conference on Economics and Computation*, pp. 1020–1020, Stanford University Stanford CA  
 562 USA, July 2025. ACM.

563 Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Commitment Without  
 564 Regrets: Online Learning in Stackelberg Security Games. In *Proceedings of the Sixteenth ACM  
 565 Conference on Economics and Computation*, pp. 61–78, Portland Oregon USA, June 2015. ACM.  
 566 URL <https://dl.acm.org/doi/10.1145/2764468.2764478>.

567 Maria-Florina Balcan, Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Keegan Harris, and  
 568 Zhiwei Steven Wu. Nearly-optimal bandit learning in stackelberg games with side information,  
 569 2025. URL <https://arxiv.org/abs/2502.00204>.

570 Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and  
 571 Nicola Gatti. Optimal Rates and Efficient Algorithms for Online Bayesian Persuasion. In  
 572 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Pro-  
 573 ceedings of Machine Learning Research*, pp. 2164–2183. PMLR, July 2023. URL <https://proceedings.mlr.press/v202/bernasconi23a.html>.

574 Clément L. Canonne. A short note on learning discrete distributions, 2020. URL <https://arxiv.org/abs/2002.11457>.

575 Xiaogang Cao, Jing Yuan, Hui Wen, and Cuiwei Zhang. The pricing strategies under the online  
 576 platform selling mode with information sharing. *Kybernetes*, 53(3):1181–1207, 2024.

577 Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian per-  
 578 suasion. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Ad-  
 579 vances in Neural Information Processing Systems*, volume 33, pp. 16188–16198. Curran Asso-  
 580 ciates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/ba5451d3c91a0f982f103cdbe249bc78-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/ba5451d3c91a0f982f103cdbe249bc78-Paper.pdf).

581 Richard Cole and Tim Roughgarden. The sample complexity of revenue maximization. In *Proceedings  
 582 of the 46th Annual ACM Symposium on Theory of Computing - STOC '14*, pp. 243–252, New York,  
 583 New York, 2014. ACM Press. ISBN 978-1-4503-2710-7. doi: 10.1145/2591796.2591867. URL  
 584 <http://dl.acm.org/citation.cfm?doid=2591796.2591867>.

594 Vincent Conitzer and Dmytro Korzhik. Commitment to Correlated Strategies. *Proceedings of*  
 595 *the AAAI Conference on Artificial Intelligence*, 25(1):632–637, August 2011. ISSN 2374-3468,  
 596 2159-5399. doi: 10.1609/aaai.v25i1.7875. URL <https://ojs.aaai.org/index.php/AAAI/article/view/7875>.

598 Vincent Conitzer and Tuomas Sandholm. Computing the Optimal Strategy to Commit to. In *Proceed-  
 599 ings of the 7th ACM conference on Electronic commerce*, pp. 82–90, Ann Arbor Michigan USA,  
 600 June 2006. ACM. URL <https://dl.acm.org/doi/10.1145/1134707.1134717>.

602 Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit  
 603 Feedback. In *21st Annual Conference on Learning Theory*, pp. 355–366, 2008. Issue: 101.

604 Ilias Diakonikolas and Vasilis Kontonis. Sample complexity lower bounds. Available at: <http://www.iliasdiakonikolas.org/teaching/Fall19/scribes/lec4.pdf>, 2019.  
 605 Lecture notes for Advanced Learning Theory, [University of Wisconsin-Madison].

606 Shaddin Dughmi and Haifeng Xu. Algorithmic persuasion with no externalities. In *Proceedings of  
 607 the 2017 ACM Conference on Economics and Computation*, pp. 351–368, 2017.

608 Paul Dütting, Michal Feldman, Inbal Talgam-Cohen, et al. Algorithmic contract theory: A survey.  
 609 *Foundations and Trends® in Theoretical Computer Science*, 16(3-4):211–412, 2024.

610 Chenghao Guo, Zhiyi Huang, Zhihao Gavin Tang, and Xinzhi Zhang. Generalizing complex  
 611 hypotheses on product distributions: Auctions, prophet inequalities, and pandora’s problem, 2020.  
 612 URL <https://arxiv.org/abs/1911.11936>.

613 Dan Halperin and Micha Sharir. Arrangements. In *Handbook of discrete and computational geometry*,  
 614 pp. 723–762. Chapman and Hall/CRC, 2017.

615 Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification.  
 616 In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp.  
 617 111–122, 2016.

618 Safwan Hossain, Evi Micha, Yiling Chen, and Ariel Procaccia. Strategic classification with externali-  
 619 ties. *arXiv preprint arXiv:2410.08032*, 2024.

620 Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):  
 621 2590–2615, 2011.

622 Jasper Lee and Qianfan Chen. Learning discrete distributions. Available at: <https://cs.brown.edu/courses/csci1951-w/lec/lec%2012%20notes.pdf>, 2020. Lecture notes for  
 623 Sublinear Algorithms for Big Data, [Brown University].

624 Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the  
 625 optimal strategy to commit to. In *Proceedings of the 2nd International Symposium on Algorith-  
 626 mic Game Theory*, SAGT ’09, pp. 250–262, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN  
 627 9783642046445. doi: 10.1007/978-3-642-04645-2\_23. URL [https://doi.org/10.1007/978-3-642-04645-2\\_23](https://doi.org/10.1007/978-3-642-04645-2_23).

628 Tao Lin. How Does Independence Help Generalization? Sample Complexity of ERM on Product  
 629 Distributions, 2022. URL <https://arxiv.org/abs/2212.06422>.

630 Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*.  
 631 Adaptive computation and machine learning series. MIT Press, Cambridge, MA, 2012. ISBN  
 978-0-262-01825-8.

632 Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning Optimal Strategies to Commit  
 633 To. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2149–2156, July 2019.  
 634 URL <https://ojs.aaai.org/index.php/AAAI/article/view/4047>.

635 Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. Stackelberg security  
 636 games: Looking beyond a decade of success. *IJCAI*, 2018.

648 Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media,  
649 2010.

650

651 Haifeng Xu. On the tractability of public persuasion with no externalities. In *Proceedings of the*  
652 *Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2708–2727. SIAM, 2020.

653

654 Kunhe Yang and Hanrui Zhang. Computational aspects of bayesian persuasion under approximate  
655 best response, 2024. URL <https://arxiv.org/abs/2402.07426>.

656

657 Geng Zhao, Banghua Zhu, Jiantao Jiao, and Michael Jordan. Online learning in stackelberg games  
658 with an omniscient follower. In *International Conference on Machine Learning*, pp. 42304–42316.  
659 PMLR, 2023.

660

661 Banghua Zhu, Stephen Bates, Zhuoran Yang, Yixin Wang, Jiantao Jiao, and Michael I. Jordan. The  
662 sample complexity of online contract design, 2023. URL <https://arxiv.org/abs/2211.05732>.

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

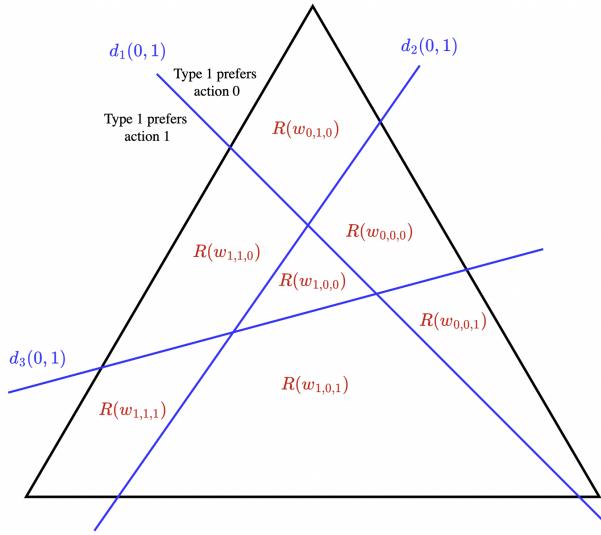
697

698

699

700

701

702 A APPENDIX FOR SECTION 3  
703  
704

705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

Figure 1: A single-follower best-response region with  $K = 3$  types and two follower actions and three leader actions –  $A = 2, L = 3$ . The triangle represents the probability simplex  $\Delta(\mathcal{L})$ . The three hyperplanes defined by  $d_1(0, 1)$ ,  $d_2(1, 0)$  and  $d_3(1, 0)$  partition the simplex into best-response regions. For example, in region  $R(w_{0,1,0})$ , the follower best-responds with action 0 for type 1, and action 1 for types 2 and 3.

## A.1 PROOF OF LEMMA 3.2

*Proof.* We have  $n$  followers each with  $K$  type and  $A$  actions. Each follower has  $K \binom{A}{2} \leq KA^2$  advantage vectors, where each advantage vector  $d_{\theta, a, a'}$  corresponds to a hyperplane in  $\mathbb{R}^L$  that separates the leader's mixed strategy space  $\Delta(\mathcal{L}) \subseteq \mathbb{R}^L$  into two halfspaces. In total,  $n$  followers have  $nKA^2$  hyperplanes. Those hyperplanes divide  $\mathbb{R}^L$  into at most  $\mathcal{O}((nKA^2)^L) = \mathcal{O}(n^L K^L A^{2L})$  regions (see, e.g., Halperin & Sharir (2017)). Each non-empty best response region is one of such regions, so the total number is  $\mathcal{O}(n^L K^L A^{2L})$ .  $\square$

## A.2 PROOF OF LEMMA 3.3

**ALGORITHM 4:** Best-Response Region Enumeration

Let  $\text{isFeasible}(x, i, \theta, a) = \mathbb{1} [\forall a' \in \mathcal{A}, \sum_{\ell \in \mathcal{L}} x(\ell)(v_i(\ell, a, \theta) - v_i(\ell, a', \theta)) \geq 0]$   
 Let  $\text{findFeasible}(W) = \{x \in \Delta(\mathcal{L}) \mid \forall i \in [n], \theta \in \Theta, \text{isFeasible}(x, i, \theta, W[i, \theta]) = 1\}$   
 Choose a random strategy  $x_{\text{init}} \in \Delta(\mathcal{L})$   
**for**  $i \in [n], \theta \in \Theta$  **do**  
**for**  $a \in \mathcal{A}$  **do**  
**if**  $\text{isFeasible}(x_{\text{init}}, i, \theta, a)$  **then**  
  $W_{\text{init}}[i, \theta] = a$   
**Let**  $queue = [W_{\text{init}}]$ , mark  $W_{\text{init}}$  as *visited*  
**while**  $queue \neq \emptyset$  **do**  
  $W = queue.pop()$   
**for**  $i \in [n], \theta \in \Theta$  **do**  
**for**  $a \in \mathcal{A}$  **and**  $a \neq W[i, \theta]$  **do**  
**Let**  $W' = W$   
**Let**  $W'[i, \theta] = a$   
**if**  $\text{findFeasible}(W') \neq \emptyset$  **and**  $W'$  **is not visited** **then**  
  $queue.append(W')$ , mark  $W'$  as *visited*

*Proof.* We construct a graph  $G = (V, E)$  where  $V$  consists of the elements  $W$ , each representing a best-response region. An edge exists between two vertices  $W, W' \in \mathcal{W}$  if and only if  $W$  and  $W'$

756 differ in exactly one entry. Traversing an edge corresponds to moving between adjacent best-response  
 757 regions by crossing a hyperplane boundary.  
 758

759 We claim that this graph is connected. Since each vertex  $W$  corresponds to a best-response region  
 760 defined by the inequalities in (2), and the leader's strategy space is the  $L$ -dimensional probability  
 761 simplex, the union of non-empty best response regions forms a partition of the strategy space. Because  
 762 these regions are convex polytopes sharing boundaries, the adjacent structure defined by differing  
 763 in one entry corresponds to crossing a shared facet. Starting from any non-empty region, we can  
 764 traverse to any other by crossing shared facets through adjacent regions, so the graph is connected.  
 765

766 Thus, to enumerate all non-empty best-response regions, we can perform a graph search (e.g.,  
 767 breadth-first search or depth-first search) starting from any initial vertex  $W$  to traverse all vertices in  
 768  $\mathcal{O}(|W|)$  steps, which is  $\mathcal{O}(n^L K^L A^{2L})$  by Lemma 3.2. Specifically, at each vertex  $W$ , we examine  
 769 all its adjacent  $nKA$  vertices. For each adjacent vertex  $W'$ , we determine whether  $R(W')$  is a  
 770 non-empty region by solving a feasibility linear program defined by the constraints in (2), which  
 771 runs in  $\text{poly}(n, K, A, L)$  time. Then, the total running time is  $\text{poly}(n^L, K^L, A^L, L)$ . We present the  
 772 algorithm formally in Algorithm 4.  $\square$   
 773

## 772 B APPENDIX FOR SECTION 4

773 The following definitions will be used in the proofs for this section:

774 **Definition B.1** (Total variation distance). *For two discrete distributions  $\mathcal{D}$  and  $\hat{\mathcal{D}}$  over support  $\Theta$ , the  
 775 total variation is half the  $L_1$  distance between the two distributions:*

$$776 \quad \delta(\mathcal{D}, \hat{\mathcal{D}}) = \frac{1}{2} \|\mathcal{D} - \hat{\mathcal{D}}\|_1 = \frac{1}{2} \sum_{\theta \in \Theta} |\mathcal{D}(\theta) - \hat{\mathcal{D}}(\theta)|.$$

777 **Definition B.2** (Hellinger distance). *For two discrete distributions  $\mathcal{D}$  and  $\hat{\mathcal{D}}$  over support  $\Theta$ , the  
 778 Hellinger distance is defined as*

$$779 \quad H(\mathcal{D}, \hat{\mathcal{D}}) = \frac{1}{\sqrt{2}} \|\sqrt{\mathcal{D}} - \sqrt{\hat{\mathcal{D}}}\|_2 = \frac{1}{\sqrt{2}} \sqrt{\sum_{\theta \in \Theta} (\mathcal{D}(\theta) - \hat{\mathcal{D}}(\theta))^2}.$$

### 780 B.1 PROOF OF LEMMA 4.1

781 We rely on several key insights about the *pseudo-dimension* of a family of functions, defined below:

782 **Definition B.3** (Definition 10.2 in Mohri et al. (2012)). *Let  $\mathcal{G}$  be a family of functions from input  
 783 space  $\mathcal{Z}$  to real numbers  $\mathbb{R}$ .*

- 784 • *A set of inputs  $\{z_1, \dots, z_m\} \subseteq \mathcal{Z}$  is shattered by  $\mathcal{G}$  if there exists thresholds  $t_1, \dots, t_m \in \mathbb{R}$  such  
 785 that for any sign vector  $\sigma = (\sigma_1, \dots, \sigma_m) \in \{-1, +1\}^m$ , there exists a function  $g \in \mathcal{G}$  satisfying  
 786  $\text{sign}(g(z_i) - t_i) = \sigma_i$  for all  $i = 1, \dots, m$ .*
- 787 • *The size of the largest set of inputs that can be shattered by  $\mathcal{G}$  is called the pseudo-dimension of  
 788  $\mathcal{G}$ , denoted by  $\text{Pdim}(\mathcal{G})$ .*

789 Given a family of functions with a finite pseudo-dimension, and samples  $z^1, \dots, z^N$  drawn from a  
 790 distribution on the input space  $\mathcal{Z}$ , the empirical mean of any function in the family will, with high  
 791 probability, be close to the true mean. Formally:

792 **Theorem B.1** (e.g., Theorem 10.6 in Mohri et al. (2012)). *Let  $\mathcal{G}$  be a family of functions from  $\mathcal{Z}$  to  
 793  $[0, 1]$  with pseudo-dimension  $\text{Pdim}(\mathcal{G}) = d$ . For any distribution  $F$  over  $\mathcal{Z}$ , with probability at least  
 794  $1 - \delta$  over the random draw of  $N$  samples  $z^1, \dots, z^N$  from  $F$ , the following holds for all  $g \in \mathcal{G}$ ,*

$$795 \quad \left| \mathbb{E}_{z \sim F}[g(z)] - \frac{1}{N} \sum_{i=1}^N g(z^i) \right| \leq \sqrt{\frac{2d \log 3N}{N}} + \sqrt{\frac{\log \frac{1}{\delta}}{2N}}.$$

800 Consider the family of linear functions over  $\mathbb{R}^L$ :  $\mathcal{G} = \{g_x : z \rightarrow \langle x, z \rangle \mid x \in \mathbb{R}^L\}$ . It is known that  
 801 the pseudo-dimension of this family is  $L$ :

810    **Lemma B.1** (e.g., Theorem 10.4 in Mohri et al. (2012)). *The family of linear functions  $\{g_x : z \rightarrow$   
 811     $\langle x, z \rangle \mid x \in \mathbb{R}^L\}$  in  $\mathbb{R}^L$  has pseudo-dimension  $L$ .*

813  
 814    *Proof of Lemma 4.1.* We now have the tools to prove Lemma 4.1. For any non-empty best-response  
 815    region defined by  $W \in \mathcal{W}$ , let  $\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^t \sim \mathcal{D}$  be  $t$  i.i.d samples. For each sample  $\boldsymbol{\theta}^i$ , we can directly  
 816    compute

$$z_{W, \boldsymbol{\theta}^i} = (u(1, W(\boldsymbol{\theta}^i)), \dots, u(L, W(\boldsymbol{\theta}^i))).$$

819    Fix any leader strategy  $x \in R(W)$ . By Lemma 3.1, the leader's expected utility by using strategy  $x$   
 820    is  $U_{\mathcal{D}}(x) = \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[\langle x, z_{W, \boldsymbol{\theta}} \rangle]$ , which is the expectation of the linear function  $g_x(z_{W, \boldsymbol{\theta}}) = \langle x, z_{W, \boldsymbol{\theta}} \rangle$ .  
 821    Therefore, by Theorem B.1 and Lemma B.1, we have

$$\Pr \left[ \forall x \in R(W), \left| \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[g_x(z_{W, \boldsymbol{\theta}})] - \frac{1}{t} \sum_{i=1}^t g_x(z_{W, \boldsymbol{\theta}^i}) \right| \leq \sqrt{\frac{2L \log 3t}{t}} + \sqrt{\frac{\log \frac{1}{\delta'}}{2t}} \right] \geq 1 - \delta'.$$

826    By definition,  $U_{\mathcal{D}}(x) = \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[g_x(z_{W, \boldsymbol{\theta}})]$  and  $\hat{U}^t(x) = \frac{1}{t} \sum_{i=1}^t g_x(z_{W, \boldsymbol{\theta}^i})$ . So,

$$\Pr \left[ \forall x \in R(W), \left| U_{\mathcal{D}}(x) - \hat{U}^t(x) \right| \leq \sqrt{\frac{2L \log 3t}{t}} + \sqrt{\frac{\log \frac{1}{\delta'}}{2t}} \right] \geq 1 - \delta'.$$

832    Let  $\delta' = \frac{\delta}{|\mathcal{W}|}$ . By the union bound, with probability at least  $1 - \delta$ , the following bound holds  
 833    simultaneously for all  $W \in \mathcal{W}$  and  $x \in R(W)$ :

$$\left| U_{\mathcal{D}}(x) - \hat{U}^t(x) \right| \leq \sqrt{\frac{2L \log(3t)}{t}} + \sqrt{\log \frac{|\mathcal{W}|}{\delta}} = \mathcal{O} \left( \sqrt{\frac{L \log t}{t}} + \sqrt{\frac{L \log(nKA) + \log(\frac{1}{\delta})}{t}} \right),$$

838    where we used the fact  $|\mathcal{W}| = O(n^L K^L A^{2L})$  from Lemma 3.2.  $\square$

## 840    B.2 PROOF OF THEOREM 4.1

842    **Analysis of  $\mathcal{O}(\sqrt{K^n T})$  Regret:** Consider a Bayesian Stackelberg game with  $n$  followers each  
 843    with  $K$  types, with joint type distribution  $\mathcal{D}$ . Let  $U(x, \mathcal{D}) = \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{D}}[u(x, \text{br}(\boldsymbol{\theta}, x))]$  be the  
 844    expected utility of the leader playing mixed strategy  $x$  when the type distribution is  $\mathcal{D}$ . Let  
 845     $x^* = \arg \max_{x \in \Delta(\mathcal{L})} U(x, \mathcal{D})$  be the optimal strategy for  $\mathcal{D}$ . At each round  $t$ , Algorithm 1 chooses  
 846    the optimal strategy  $x^t = \arg \max_{x \in \Delta(\mathcal{L})} U(x, \hat{\mathcal{D}}^{t-1})$  for the empirical distribution  $\hat{\mathcal{D}}^{t-1}$  over  $t-1$   
 847    samples. The total expected regret is equal to

$$\mathbb{E}[\text{Reg}(T)] = \sum_{t=1}^T \mathbb{E} \left[ \underbrace{U(x^*, \mathcal{D}) - U(x^t, \mathcal{D})}_{\text{single-round regret } r(t)} \right],$$

853    We upper bound the single-round regret  $r(t)$  by the total variation distance (Definition B.1) between  
 854     $\mathcal{D}$  and  $\hat{\mathcal{D}}^{t-1}$ :

856    **Claim B.1.**  $r(t) = U(x^*, \mathcal{D}) - U(x^t, \mathcal{D}) \leq 4\delta(\mathcal{D}, \hat{\mathcal{D}}^{t-1})$ .

858    *Proof.*

$$\begin{aligned} r(t) &= U(x^*, \mathcal{D}) - U(x^t, \mathcal{D}) \\ &= U(x^*, \mathcal{D}) - U(x^*, \hat{\mathcal{D}}^{t-1}) + U(x^*, \hat{\mathcal{D}}^{t-1}) - U(x^t, \hat{\mathcal{D}}^{t-1}) + U(x^t, \hat{\mathcal{D}}^{t-1}) - U(x^t, \mathcal{D}) \\ &\leq U(x^*, \mathcal{D}) - U(x^*, \hat{\mathcal{D}}^{t-1}) + 0 + U(x^t, \hat{\mathcal{D}}^{t-1}) - U(x^t, \mathcal{D}) \end{aligned} \tag{4}$$

864 where (4) follows from  $U(x^*, \hat{\mathcal{D}}^{t-1}) - U(x^t, \hat{\mathcal{D}}^{t-1}) \leq 0$  because  $x^t$  maximizes  $U(x, \hat{\mathcal{D}}^{t-1})$ . We  
 865 bound the first term in Equation (4) as follows:  
 866

$$\begin{aligned}
 867 \quad U(x^*, \mathcal{D}) - U(x^*, \hat{\mathcal{D}}^{t-1}) &= \sum_{\theta \in \Theta} \mathcal{D}(\theta) u(x, \mathbf{br}(\theta, x)) - \sum_{\theta \in \Theta} \hat{\mathcal{D}}^{t-1}(\theta) u(x, \mathbf{br}(\theta, x)) \\
 868 \\
 869 &= \sum_{\theta \in \Theta} (\mathcal{D}(\theta) - \hat{\mathcal{D}}^{t-1}(\theta)) u(x, \mathbf{br}(\theta, x)) \\
 870 \\
 871 &\leq \sum_{\theta \in \Theta} |\mathcal{D}(\theta) - \hat{\mathcal{D}}^{t-1}(\theta)| u(x, \mathbf{br}(\theta, x)) \\
 872 \\
 873 &\leq \sum_{\theta \in \Theta} |\mathcal{D}(\theta) - \hat{\mathcal{D}}^{t-1}(\theta)| \cdot 1 = 2\delta(\mathcal{D}, \hat{\mathcal{D}}^{t-1}).
 \end{aligned}$$

874 By a symmetrical argument, the second term in Equation (4) is also bounded by  $2\delta(\mathcal{D}, \hat{\mathcal{D}}^{t-1})$ .  $\square$   
 875

876 Using Claim B.1 and taking expectation, we have  
 877

$$\mathbb{E}[r(t)] \leq 4\mathbb{E}[\delta(\mathcal{D}, \hat{\mathcal{D}}^{t-1})].$$

878 According to Canonne (2020), for distributions with support size  $K^n$ ,  $\mathbb{E}[\delta(\mathcal{D}, \hat{\mathcal{D}}^{t-1})] \leq \mathcal{O}(\sqrt{\frac{K^n}{t-1}})$ .  
 879

880 Thus, we have  $\mathbb{E}[r(t)] \leq \mathcal{O}(\sqrt{\frac{K^n}{t-1}})$ . Using the inequality  $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ , we obtain  
 881

$$\mathbb{E}[\text{Reg}(T)] = \sum_{t=1}^T \mathbb{E}[r(t)] \leq \mathcal{O}\left(\sum_{t=1}^T \sqrt{\frac{K^n}{t}}\right) \leq \mathcal{O}(2\sqrt{K^n T}) = \mathcal{O}(\sqrt{K^n T}).$$

882 **Analysis of  $\mathcal{O}(\sqrt{LT \log(nKAT)})$  Regret:** Consider round  $t \geq 2$ . By Lemma 4.1, we have that  
 883 with probability at least  $1 - \delta$ :

$$\left| U_{\mathcal{D}}(x) - \hat{U}^t(x) \right| \leq \mathcal{O}\left(\sqrt{\frac{L \log t}{t}} + \sqrt{\frac{L \log(nKA) + \log(\frac{1}{\delta})}{t}}\right), \quad \forall x \in \Delta(\mathcal{L}).$$

884 Suppose this event happens. Then, the regret of the algorithm at round  $t$  is bounded as follows:  
 885

$$\begin{aligned}
 886 \quad r(t) &= \mathbb{E}_{\theta \sim \mathcal{D}}[u(x^*, \mathbf{br}(\theta, x^*)) - u(x^t, \mathbf{br}(\theta, x^t))] \\
 887 &= U_{\mathcal{D}}(x^*) - U_{\mathcal{D}}(x^t) \\
 888 &= U_{\mathcal{D}}(x^*) - \hat{U}^{t-1}(x^*) + \hat{U}^{t-1}(x^*) - \hat{U}^{t-1}(x^t) + \hat{U}^{t-1}(x^t) - U_{\mathcal{D}}(x^t) \\
 889 \\
 890 &\leq \hat{U}^{t-1}(x^*) - \hat{U}^{t-1}(x^t) + 2 \cdot \mathcal{O}\left(\sqrt{\frac{L \log(t-1)}{t-1}} + \sqrt{\frac{L \log(nKA) + \log(\frac{1}{\delta})}{t-1}}\right) \\
 891 \\
 892 &\leq 0 + 2 \cdot \mathcal{O}\left(\sqrt{\frac{L \log(t-1)}{t-1}} + \sqrt{\frac{L \log(nKA) + \log(\frac{1}{\delta})}{t-1}}\right),
 \end{aligned}$$

893 where the last inequality follows from  $\hat{U}^{t-1}(x^*) - \hat{U}^{t-1}(x^t) \leq 0$  because the algorithm selects the  
 894 strategy  $x^t$  that maximizes the empirical utility  $\hat{U}^{t-1}(x)$ . Then:  
 895

$$\begin{aligned}
 896 \quad \mathbb{E}[\text{Reg}(T)] &= \mathbb{E}\left[\sum_{t=1}^T r(t)\right] \\
 897 \\
 898 &\leq \sum_{t=1}^T (1 - \delta) \cdot \mathcal{O}\left(\sqrt{\frac{L \log t}{t}} + \sqrt{\frac{L \log(nKA) + \log(\frac{1}{\delta})}{t}}\right) + \delta T \tag{5}
 \end{aligned}$$

$$\leq \mathcal{O}\left(\sqrt{TL \log T} + \sqrt{T(L \log(nKA) + \log(\frac{1}{\delta}))}\right) + \delta T \tag{6}$$

$$\begin{aligned}
 899 \quad &\leq \mathcal{O}(\sqrt{TL(\log T + \log(nKA))}) \quad (\text{Using inequality } \sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}) \\
 900 &= \mathcal{O}(\sqrt{TL \log(nKAT)}).
 \end{aligned}$$

918 Equation (5) follows from the law of total expectation and the fact that the single-round regret is  
 919 bounded by 1. Equation (6) follows from the known inequality  $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ . We set  $\delta = \frac{1}{T}$ .  
 920

921 **Runtime Analysis:** As for the computational complexity of this algorithm, note that Lemma 3.2  
 922 states that the number of non-empty best-response regions is  $\mathcal{O}(n^L K^L A^{2L})$ . As shown by Equation  
 923 (3), we can compute the optimal strategy within each best-response region using a linear program  
 924 with  $L$  variables and at most  $\text{poly}(n^L, K^L, A^L, L, T)$  number of constraints. Further, evaluating  
 925 each constraint and the objective function can also be accomplished in this time. Since each linear  
 926 program can be solved in  $\text{poly}(n^L, K^L, A^L, L, T)$  time, and we run  $\mathcal{O}(n^L K^L A^{2L})$  linear programs  
 927 at each round, with at most  $T$  rounds, Algorithm 1 runs  $\text{poly}((nKA)^L LT)$  time.  
 928

929 **B.3 PROOF OF THEOREM 4.2**  
 930

931 *Proof.* Let  $\mathcal{D} = \prod_{i=1}^n \mathcal{D}_i$  denote the distribution over independent types. According to Claim B.1,  
 932 the single-round regret  $r(t) = U(x^*, \mathcal{D}) - U(x^t, \mathcal{D})$  satisfies  
 933

$$934 \quad r(t) \leq \mathcal{O}(\delta(\mathcal{D}, \hat{\mathcal{D}}^{t-1})),$$

936 where  $\hat{\mathcal{D}}^{t-1} = \prod_{i=1}^n \hat{\mathcal{D}}_i^{t-1}$  is the product of the empirically computed marginal type distributions.  
 937 We will use the following properties of Hellinger Distance (Definition B.2):  
 938

- 939 • (Guo et al., 2020) For any two distributions  $\mathcal{D}$  and  $\hat{\mathcal{D}}$ ,

$$940 \quad H^2(\mathcal{D}, \hat{\mathcal{D}}) \leq \delta(\mathcal{D}, \hat{\mathcal{D}}) \leq \sqrt{2}H(\mathcal{D}, \hat{\mathcal{D}}). \quad (7)$$

- 942 • (Guo et al., 2020) If both  $\mathcal{D}$  and  $\hat{\mathcal{D}}$  are product distributions, i.e.  $\mathcal{D} = \prod_{i=1}^n \mathcal{D}_i$  and  $\hat{\mathcal{D}} =$   
 943  $\prod_{i=1}^n \hat{\mathcal{D}}_i$ , then:

$$945 \quad H^2(\mathcal{D}, \hat{\mathcal{D}}) \leq \sum_{i=1}^n H^2(\mathcal{D}_i, \hat{\mathcal{D}}_i). \quad (8)$$

- 948 • (Canonne, 2020) For a distribution  $\mathcal{D}$  with support size  $K$ , the empirical distribution  $\hat{\mathcal{D}}^t$  over  $t$   
 949 samples from  $\mathcal{D}$  satisfies:

$$950 \quad \mathbb{E}[H^2(\mathcal{D}, \hat{\mathcal{D}}^t)] \leq \frac{K}{2t}. \quad (9)$$

953 We now upper bound the single-round regret  $r(t+1)$  in expectation:

$$\begin{aligned} 955 \quad \mathbb{E}[r(t+1)] &\leq \mathcal{O}(\mathbb{E}[\delta(\mathcal{D}, \hat{\mathcal{D}}^t)]) \\ 956 \quad &\leq \mathcal{O}(\mathbb{E}[H(\mathcal{D}, \hat{\mathcal{D}}^t)]) && \text{by (7)} \\ 957 \quad &\leq \mathcal{O}(\sqrt{\mathbb{E}[H^2(\mathcal{D}, \hat{\mathcal{D}}^t)]}) && \text{because } \mathbb{E}[X^2] \geq (\mathbb{E}[X])^2 \\ 958 \quad &\leq \mathcal{O}\left(\sqrt{\mathbb{E}\left[\sum_{i=1}^n H^2(\mathcal{D}_i, \hat{\mathcal{D}}_i^t)\right]}\right) && \text{by (8)} \\ 959 \quad &\leq \mathcal{O}\left(\sqrt{\frac{nK}{2t}}\right). && \text{by (9)} \end{aligned}$$

960 Using the inequality  $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ , we obtain  
 961

$$962 \quad \mathbb{E}[\text{Reg}(T)] = \sum_{t=1}^T \mathbb{E}[r(t)] \leq \sum_{t=1}^T \mathcal{O}\left(\sqrt{\frac{nK}{t}}\right) \leq \mathcal{O}(\sqrt{nKT}).$$

963  $\square$   
 964  
 965

972 B.4  $\Omega(\sqrt{\min\{L, K\}T})$  LOWER BOUND IN THE SINGLE-FOLLOWER CASE  
973974 In this section, we prove a lower bound of  $\Omega(\sqrt{\min\{L, K\}T})$  on the expected regret of any algorithm  
975 in the case of a single-follower ( $n = 1$ ), formalized in Theorem B.2.  
976977 **Theorem B.2.** *For single-follower Bayesian Stackelberg games where the follower has  $K$  types and  
978 the leader has  $L$  actions, the expected regret of any type-feedback online learning algorithm is at  
979 least  $\Omega(\sqrt{\min\{L, K\}T})$ .*  
980981 At a high level, the proof Theorem B.2 is a reduction from the *distribution learning problem*.  
982 Without loss of generality, assume that  $\min\{K, L\} = 2c$  is an even number. Further assume that  
983  $K = L = 2c$ .<sup>5</sup> The single follower has  $K = 2c$  types, with type space  $\Theta = \{\pm 1, \pm 2, \dots, \pm c\}$ .  
984 Consider a class  $\mathcal{C}$  of distributions over  $\Theta$  defined as follows:  
985986 **Definition B.4** (Class of Distributions  $\mathcal{C}$ ). *A distribution  $\mathcal{D} = \mathcal{D}_\sigma \in \mathcal{C}$  is specified by a vector  
987  $\sigma = (\sigma_1, \dots, \sigma_c) \in \{\pm 1\}^c$ . For each  $j = 1, \dots, c$ ,*  
988

989 
$$\mathcal{D}_\sigma(+j) = \frac{1}{2c}(1 + \sigma_j \epsilon), \quad \mathcal{D}_\sigma(-j) = \frac{1}{2c}(1 - \sigma_j \epsilon). \quad (10)$$
  
990

991 for some  $\epsilon > 0$ . Note that  $\mathcal{D}_\sigma(+j) > \mathcal{D}_\sigma(-j)$  if and only if  $\sigma_j = +1$ . The class  $\mathcal{C}$  consists of  $2^c$   
992 distributions.  
993994 In the distribution learning problem, given  $t$  samples  $\theta^1, \dots, \theta^t$  from an unknown distribution  $\mathcal{D} \in \mathcal{C}$ ,  
995 the goal is to construct an estimator  $\hat{\mathcal{D}}$  specified by a vector  $\hat{\sigma} \in \{\pm 1\}^c$  such that the expected total  
996 variation distance (Definition B.1) satisfies  $\mathbb{E}[\delta(\mathcal{D}, \hat{\mathcal{D}})] \leq \mathcal{O}(\epsilon)$ . It is known that solving this problem  
997 requires at least  $\Omega(\frac{2c}{\epsilon^2})$  samples.  
998999 **Theorem B.3** (e.g., (Lee & Chen, 2020; Diakonikolas & Kontonis, 2019)). *When  $\mathcal{D}$  is uniformly  
1000 sampled from the class  $\mathcal{C}$ , any algorithm that constructs estimator  $\hat{\mathcal{D}}$  using  $t$  samples from  $\mathcal{D}$  has  
1001 expected error at least  $\mathbb{E}[\delta(\hat{\mathcal{D}}, \mathcal{D})] \geq \Omega(\epsilon)$  if  $t \leq \mathcal{O}(\frac{2c}{\epsilon^2})$ .*  
1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

5If  $K > 2c$ , we can let the additional types to have probability 0. If  $L > 2c$ , we can let the additional actions  
1026 of the leader to have very low utility.  
1027

1026  
1027**Reduction from distribution learning to single-follower Bayesian Stackelberg game**

1028

**Distribution learning instance:** An unknown distribution  $\mathcal{D} \in \mathcal{C}$ .

1029

**Bayesian Stackelberg game instance:** A single follower with type space  $\Theta = \{\pm 1, \pm 2, \dots, \pm c\}$  and an unknown type distribution  $\mathcal{D}$ . The follower has binary action set  $\mathcal{A} = \{\text{Good}, \text{Bad}\}$ . The leader has action set  $\mathcal{L} = \Theta = \{\pm 1, \pm 2, \dots, \pm c\}$ . The utility functions of the two players are:

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

- Follower's utility function:

$$v(\ell, a, \theta) = \begin{cases} 1 & \text{if } \theta = +j, \ell = +j, a = \text{Good} \\ 1 & \text{if } \theta = +j, \ell = -j, a = \text{Bad} \\ 1 & \text{if } \theta = -j, \ell = -j, a = \text{Good} \\ 1 & \text{if } \theta = -j, \ell = +j, a = \text{Bad} \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

1041

1042

1043

1044

1045

1046

- Leader's utility function: For any action  $\ell \in \mathcal{L}$ ,

$$u(\ell, \text{Good}) = 1, \quad u(\ell, \text{Bad}) = 0. \quad (12)$$

1045

Note that for any mixed strategy  $x$ ,  $u(x, \text{Good}) = 1$  and  $u(x, \text{Bad}) = 0$ .

1046

**Reduction:**

Given an online learning algorithm  $\text{Alg}$  for Bayesian Stackelberg game with type feedback, we use it to construct an online learning algorithm for the distribution learning problem as follows: At each round  $t = 1, \dots, T$ ,

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

$$\sigma_j(x^t) = \begin{cases} +1, & \text{if } x^t(+j) \geq x^t(-j) \\ -1, & \text{if } x^t(+j) < x^t(-j). \end{cases} \quad (13)$$

1. Receive the leader's mixed strategy  $x^t$  from  $\text{Alg}$ .
2. Construct an estimated distribution  $\hat{\mathcal{D}}_{x^t} = \mathcal{D}_{\sigma(x^t)} \in \mathcal{C}$  based on vector  $\sigma(x^t)$  defined as follows:

$$\sigma_j(x^t) = \begin{cases} +1, & \text{if } x^t(+j) \geq x^t(-j) \\ -1, & \text{if } x^t(+j) < x^t(-j). \end{cases} \quad (13)$$

3. Observe sample  $\theta^t \sim \mathcal{D}$  and feed  $\theta^t$  to  $\text{Alg}$ .

1080 **Lemma B.2** (Follower's Best Response). *Given the leader's mixed strategy  $x \in \Delta(\Theta)$ , for each*  
 1081  *$j \in \{1, \dots, c\}$ , the best-response function of a follower with type  $+j$  or  $-j$  is:*

$$1083 \quad \text{br}(+j, x) = \begin{cases} \text{Good,} & \text{if } x(+j) \geq x(-j), \\ 1084 \quad \text{Bad,} & \text{if } x(+j) < x(-j), \end{cases}$$

$$1086 \quad \text{br}(-j, x) = \begin{cases} \text{Good,} & \text{if } x(+j) < x(-j), \\ 1087 \quad \text{Bad,} & \text{if } x(+j) \geq x(-j). \end{cases}$$

1088 *Proof.* For a follower with type  $+j$ , their utility for choosing action Good is given by

$$1090 \quad v(x, \text{Good}, +j) = \mathbb{E}_{\ell \sim x}[v(\ell, \text{Good}, +j)] = \sum_{\ell \in \mathcal{L}} x(\ell) v(\ell, \text{Good}, +j) = x(+j).$$

1093 Similarly, their utility for choosing action Bad is:

$$1094 \quad v(x, \text{Bad}, +j) = x(-j).$$

1096 Thus, by definition, the follower with type  $+j$  best responds with Good if  $x(+j) \geq x(-j)$ .

1097 Likewise, for a follower with type  $-j$ ,

$$1099 \quad v(x, \text{Good}, -j) = x(-j), \\ 1100 \quad v(x, \text{Bad}, -j) = x(+j).$$

1101 Thus, a follower with type  $-j$  best responds with Bad if  $x(+j) \geq x(-j)$ , Good otherwise.  $\square$

1103 We define  $U(x, \mathcal{D})$  as the expected utility of the leader when using mixed strategy  $x$  under the type  
 1104 distribution  $\mathcal{D}$ . By Lemma B.2, we have

$$1106 \quad U(x, \mathcal{D}) = \sum_{\theta \in \Theta} \mathcal{D}(\theta) u(x, \text{br}(\theta, x)) \\ 1107 \\ 1108 = \sum_{j=1}^c [\mathcal{D}(+j) u(x, \text{br}(+j, x)) + \mathcal{D}(-j) u(x, \text{br}(-j, x))] \\ 1109 \\ 1111 = \sum_{j=1}^c \left( \frac{1 + \sigma_j \epsilon}{2c} \mathbb{1}[x(+j) \geq x(-j)] + \frac{1 - \sigma_j \epsilon}{2c} \mathbb{1}[x(+j) < x(-j)] \right). \quad (14)$$

1114 **Definition B.5** (Disagreement Function). *The disagreement function  $\text{Disagree}(x, \mathcal{D})$  is the number*  
 1115 *of  $j \in \{1, \dots, c\}$  where the indicators  $\mathbb{1}[x(+j) \geq x(-j)]$  and  $\mathbb{1}[\mathcal{D}(+j) \geq \mathcal{D}(-j)]$  differ:*

$$1116 \quad \text{Disagree}(x, \mathcal{D}) = \sum_{j=1}^c \mathbb{1}[\mathbb{1}[x(+j) \geq x(-j)] \neq \mathbb{1}[\mathcal{D}(+j) \geq \mathcal{D}(-j)]] \\ 1117 \\ 1119 = \sum_{j=1}^c \mathbb{1}[\mathbb{1}[x(+j) \geq x(-j)] \neq \mathbb{1}[\sigma_j = +1]].$$

1122 **Lemma B.3.**  $U(\mathcal{D}, \mathcal{D}) - U(x, \mathcal{D}) = \frac{\epsilon}{c} \cdot \text{Disagree}(x, \mathcal{D})$ . In particular, the optimal strategy for the  
 1123 leader is  $x^* = \mathcal{D}$ .

1124 *Proof.*

$$1126 \quad U(\mathcal{D}, \mathcal{D}) - U(x, \mathcal{D}) = \sum_{j=1}^c \left( \frac{1 + \sigma_j \epsilon}{2c} (\mathbb{1}[\mathcal{D}(+j) \geq \mathcal{D}(-j)] - \mathbb{1}[x(+j) \geq x(-j)]) \right. \\ 1127 \\ 1129 \quad \left. + \frac{1 - \sigma_j \epsilon}{2c} (\mathbb{1}[\mathcal{D}(+j) < \mathcal{D}(-j)] - \mathbb{1}[x(+j) < x(-j)]) \right).$$

1132 For each term in the summation where  $\mathcal{D}(+j) \geq \mathcal{D}(-j)$  and  $x(+j) \geq x(-j)$  agree, the term  
 1133 evaluates to 0. When they disagree, there are two possible cases:

1134 1.  $\mathcal{D}(+j) \geq \mathcal{D}(-j)$  but  $x(+j) < x(-j)$ . Since  $\mathcal{D}(+j) \geq \mathcal{D}(-j)$  implies  $\sigma_j = +1$ , the term  
 1135 simplifies to

$$1136 \quad \frac{1+\epsilon}{2c} - \frac{1-\epsilon}{2c} = \frac{\epsilon}{c}.$$

1138 2.  $\mathcal{D}(+j) < \mathcal{D}(-j)$  but  $x(+j) \geq x(-j)$ . Here,  $\sigma_j = -1$ , so the term simplifies to

$$1140 \quad -\frac{1-\epsilon}{2c} + \frac{1+\epsilon}{2c} = \frac{\epsilon}{c}.$$

1142 Thus, we conclude that

$$1144 \quad U(\mathcal{D}, \mathcal{D}) - U(x, \mathcal{D}) = \frac{\epsilon}{c} \cdot \text{Disagree}(x, \mathcal{D}).$$

□

1147 **Lemma B.4.** *Let  $\hat{\mathcal{D}}_x$  be the estimated distribution constructed from  $x$  according to Equation (13).  
 1148 The total variation distance between  $\hat{\mathcal{D}}_x$  and  $\mathcal{D}$  is given by*

$$1150 \quad \delta(\hat{\mathcal{D}}_x, \mathcal{D}) = \frac{\epsilon}{c} \cdot \text{Disagree}(x, \mathcal{D}).$$

1152 *Proof.* By definition,

$$1154 \quad \delta(\hat{\mathcal{D}}_x, \mathcal{D}) = \frac{1}{2} \sum_{j=1}^c \left( |\hat{\mathcal{D}}_x(+j) - \mathcal{D}(+j)| + |\hat{\mathcal{D}}_x(-j) - \mathcal{D}(-j)| \right).$$

1157 If  $x$  and  $\mathcal{D}$  agree at  $(+j)$  and  $(-j)$ , the corresponding term in the total variation sum is 0. Consequently, we only need to consider the case when a disagreement occurs.

1159 1.  $x(+j) < x(-j)$  while  $\mathcal{D}(+j) \geq \mathcal{D}(-j)$ . In this case,

$$1161 \quad \hat{\mathcal{D}}_x(+j) = \frac{1-\epsilon}{2c}, \quad \hat{\mathcal{D}}_x(-j) = \frac{1+\epsilon}{2c}, \quad \mathcal{D}(+j) = \frac{1+\epsilon}{2c}, \quad \mathcal{D}(-j) = \frac{1-\epsilon}{2c}.$$

1163 Hence,

$$1165 \quad |\hat{\mathcal{D}}_x(+j) - \mathcal{D}(+j)| + |\hat{\mathcal{D}}_x(-j) - \mathcal{D}(-j)| = \frac{\epsilon}{c} + \frac{\epsilon}{c} = \frac{2\epsilon}{c}.$$

1167 2.  $x(+j) \geq x(-j)$  while  $\mathcal{D}(+j) < \mathcal{D}(-j)$ . Similarly, we have

$$1169 \quad \hat{\mathcal{D}}_x(+j) = \frac{1+\epsilon}{2c}, \quad \hat{\mathcal{D}}_x(-j) = \frac{1-\epsilon}{2c}, \quad \mathcal{D}(+j) = \frac{1-\epsilon}{2c}, \quad \mathcal{D}(-j) = \frac{1+\epsilon}{2c}.$$

1171 Again, we have

$$1173 \quad |\hat{\mathcal{D}}_x(+j) - \mathcal{D}(+j)| + |\hat{\mathcal{D}}_x(-j) - \mathcal{D}(-j)| = \frac{\epsilon}{c} + \frac{\epsilon}{c} = \frac{2\epsilon}{c}.$$

1175 Thus, it follows that

$$1177 \quad \delta(\hat{\mathcal{D}}_x, \mathcal{D}) = \frac{\epsilon}{c} \cdot \text{Disagree}(x, \mathcal{D}).$$

□

1180 From Lemma B.3 and Lemma B.4,

$$1182 \quad U(\mathcal{D}, \mathcal{D}) - U(x, \mathcal{D}) = \frac{\epsilon}{c} \cdot \text{Disagree}(x, \mathcal{D}) = \delta(\hat{\mathcal{D}}_x, \mathcal{D}). \quad (15)$$

1184 Consider the regret of the online learning algorithm Alg for the Bayesian Stackelberg game, where  
 1185 the algorithm outputs  $x^t$  at round  $t$ . By Equation (15) and Theorem B.3, the expected regret at round  
 1186  $t \leq \mathcal{O}(\frac{2c}{\epsilon^2})$  is at least

$$1187 \quad \mathbb{E}[U(\mathcal{D}, \mathcal{D}) - U(x^t, \mathcal{D})] = \mathbb{E}[\delta(\hat{\mathcal{D}}_{x^t}, \mathcal{D})] \geq \Omega(\epsilon).$$

1188 Thus, the expected regret over  $T$  rounds is at least:  
1189

$$\begin{aligned} 1190 \mathbb{E}[\text{Reg}(T)] &= \sum_{t=1}^T \mathbb{E}[U(\mathcal{D}, \mathcal{D}) - U(x^t, \mathcal{D})] \geq \min \left\{ T, \mathcal{O}\left(\frac{2c}{\epsilon^2}\right) \right\} \cdot \Omega(\epsilon) \\ 1191 &\geq \Omega(\sqrt{2cT}) = \Omega(\sqrt{\min\{K, L\}T}) \end{aligned}$$

1194 where we choose  $\epsilon = \sqrt{\frac{2c}{T}}$ .  
1195

1197 **B.5  $\Omega(\sqrt{\min\{L, nK\}T})$  LOWER BOUND FOR THE MULTI-FOLLOWER CASE: PROOF OF  
1198 THEOREM 4.3**

1200 We now prove a lower bound of  $\Omega(\sqrt{\min\{L, nK\}T})$  on the expected regret of any online learning  
1201 algorithms for Bayesian Stackelberg games with multiple followers. Without loss of generality,  
1202 assume that  $nK$  is an even integer, and assume that the number of leader actions  $L \geq nK$ . We do a  
1203 reduction from the single-follower problem to the multi-follower problem.

1204 **Single-Follower Bayesian Stackelberg Game instance:** Consider the single-follower Bayesian  
1205 Stackelberg game instance defined in Appendix B.4, but instead of a single follower with  $K$  types, we  
1206 change the instance so that the single follower has  $nK$  types, indexed by  $\Theta = \{(i, j) : i \in [n], j \in [K]\}$ . Suppose the single follower's type distribution  $\mathcal{D}$  belongs to the class  $\mathcal{C}$  in Definition B.4 with  
1207 support size  $2c = nK$  (instead of  $2c = K$ ). Note that for such a  $\mathcal{D} \in \mathcal{C}$ ,

$$\sum_{i=1}^n \sum_{j=1}^K \mathcal{D}(i, j) = 1 \quad \text{and} \quad \forall i \in [n], \quad \sum_{j=1}^K \mathcal{D}(i, j) = \frac{1}{n}.$$

1210 The follower's utility function  $v$  is given by (11), except that we now use  $\theta = (i, j)$  to represent  
1211 a type and  $\ell = (i, j)$  to represent a leader's action. The leader's action set is  $\mathcal{L} = \Theta$ , with utility  
1212 function  $u$  given by (12).

1216 **Multi-Follower Bayesian Stackelberg Game instance:** We reduce the single-follower game to an  
1217  $n$ -follower game defined below. Consider a Bayesian Stackelberg game with  $n$  followers each with  
1218  $K + 1$  types. The type distribution and the followers and leader's actions and utilities are defined  
1219 below: (To distinguish the notations from the single-follower game, we use tilde notations  $\tilde{\cdot}$ )

1221 • **Type distribution:** The followers' types are independently distributed according to distribution  
1222  $\tilde{\mathcal{D}} = \prod_{i=1}^n \tilde{\mathcal{D}}_i$  where the probability that follower  $i \in [n]$  has type  $j$  is:

$$\tilde{\mathcal{D}}_i(j) = \begin{cases} 1 - \frac{1}{100n} & \text{if } j = 0, \\ \frac{1}{100} \mathcal{D}(i, j) & \text{if } j = 1, \dots, K. \end{cases}$$

1227 • **Followers' actions and utilities:** Each follower has 3 actions  $\tilde{\mathcal{A}} = \{\text{Good, Bad, } a_0\}$ . The utility  
1228 of a follower  $i$  with type  $j \neq 0$  is equal to the utility of the single follower with type  $(i, j)$ .  
1229 Utilities for type  $j = 0$  and action  $a_0$  are specially defined:

$$\tilde{v}_i(\ell, a, \theta_i = j) = \begin{cases} v(\ell, a, (i, j)) & \text{if } \theta_i \neq 0 \text{ and } a \neq a_0 \\ -1 & \text{if } \theta_i \neq 0 \text{ and } a = a_0, \\ 1 & \text{if } \theta_i = 0 \text{ and } a = a_0, \\ -1 & \text{if } \theta_i = 0 \text{ and } a \neq a_0. \end{cases}$$

1235 Note that the best-response action of a follower with type 0 is always  $a_0$ , regardless of the  
1236 leader's strategy.

1237 • **Leader's actions and utilities:** The leader has the same action set as the single-follower game:  
1238  $\mathcal{L} = \Theta = \{(i, j) : i \in [n], j \in [K]\}$ . For any leader action  $\ell \in \mathcal{L}$ ,

$$\tilde{u}(\ell, \mathbf{a}) = \begin{cases} 1 & \text{if } n - 1 \text{ followers choose } a_0 \text{ and one plays Good,} \\ 0 & \text{otherwise.} \end{cases}$$

1242

1243

1244 Given an online learning algorithm  $\text{Alg}$  for the  $n$ -follower problem, we construct an online  
 1245 learning algorithm for the single-follower problem as follows:

1246 At each round  $t = 1, \dots, T$ :

1247

- 1248 • Obtain a strategy  $x^t \in \Delta(\mathcal{L})$  from algorithm  $\text{Alg}$ . Output  $x^t$ .
- 1249 • Receive a sample of the single follower's type  $\theta^t = (i^t, j^t) \sim \mathcal{D}$ .
- 1250 • For every follower  $i \in [n]$ , we construct their type  $\theta_i^t$  in the following way: Independently  
 1251 flip a coin that lands on head with probability  $1 - \frac{1}{100n}$ . If it lands on head, set the follower  
 1252 type  $\theta_i^t$  to 0. If it lands on tail, we select the most recent sample of the form  $(i^s = i, j^s)$   
 1253 from the history  $\{(i^s, j^s)\}_{s=1}^t$ , and set the follower's type  $\theta_i^t$  to  $j^s$ . Each sample can only  
 1254 be used once. If there are insufficient samples, we halt the algorithm.
- 1255 • Provide the constructed types  $(\theta_1^t, \dots, \theta_n^t)$  to algorithm  $\text{Alg}$ .

1256

1257 In the above reduction process, if we always have sufficient samples in the third step at each round,  
 1258 then the distribution of samples  $(\theta_1^t, \dots, \theta_n^t)$  provided to algorithm  $\text{Alg}$  is equal to the type distribution  
 1259  $\tilde{\mathcal{D}} = \prod_{i=1}^n \tilde{\mathcal{D}}_i$  of the  $n$ -follower game. Thus, from algorithm  $\text{Alg}$ 's perspective, it is solving the  
 1260  $n$ -follower game with unknown type distribution  $\tilde{\mathcal{D}}$ . We then argue that we have sufficient samples  
 1261 with high probability. Let  $H_i^t$  be the number of available samples in the history that we can use to set  
 1262 follower  $i$ 's type at round  $t$ , and let  $N_i^t$  be the number of samples that we actually need. Define

$$1 - \delta(t) = \Pr(\forall i \in [n], H_i^t \geq N_i^t),$$

1263 which is the probability that we have sufficient samples at round  $t$ .

1264 **Claim B.2.**  $\delta(t) \leq 2n \exp\left(-\frac{t^2\left(\frac{1}{100n} - \frac{1}{n}\right)^2}{2t}\right)$ .

1265

1266 *Proof.* Note that  $H_i^t$  and  $N_i^t$  are Binomial random variables:  $H_i^t \sim \text{Bin}(t, \frac{1}{n})$ ,  $N_i^t \sim \text{Bin}(t, \frac{1}{100n})$ .  
 1267 So, by union bound and Hoeffding's inequality:

$$\Pr(\exists i \in [n], H_i^t < N_i^t) \leq n \Pr(H_i^t < N_i^t) \leq 2n \exp\left(-\frac{t^2\left(\frac{1}{100n} - \frac{1}{n}\right)^2}{2t}\right).$$

□

1268

1269 Let  $\tilde{U}(x)$  be the leader's expected utility in the  $n$ -follower game (on type distribution  $\tilde{\mathcal{D}}$ ) and  $U(x)$   
 1270 be the leader's utility in the single-follower game (on type distribution  $\mathcal{D}$ ). We note that, given any  
 1271 strategy  $x \in \Delta(\mathcal{L})$  of the leader, the best-response action of follower  $i$  with type  $\theta_i = j \neq 0$  (in the  
 1272  $n$ -follower game) is equal to the best-response action of the single follower with type  $(i, j)$ , namely,  
 1273  $\text{br}_i(j, x) = \text{br}((i, j), x)$ . Thus,

$$\begin{aligned} \tilde{U}(x) &= \Pr[\text{exactly one follower has a non-0 type}] \\ &\quad \cdot \mathbb{E}[\text{leader's utility} \mid \text{exactly one follower has a non-0 type}] + 0 \\ &= \left(1 - \frac{1}{100n}\right)^{n-1} \sum_{i=1}^n \sum_{j=1}^K \frac{1}{100} \mathcal{D}(i, j) \mathbb{1}[\text{br}_i(j, x) = \text{Good}] \\ &= \left(1 - \frac{1}{100n}\right)^{n-1} \sum_{i=1}^n \sum_{j=1}^K \frac{1}{100} \mathcal{D}(i, j) \mathbb{1}[\text{br}((i, j), x) = \text{Good}] \\ &= \frac{1}{100} \left(1 - \frac{1}{100n}\right)^{n-1} U(x) \\ &\approx \frac{1}{100} e^{-\frac{1}{100}} U(x). \end{aligned}$$

1274 Define  $C = \frac{1}{100} \left(1 - \frac{1}{100n}\right)^{n-1}$ . Let  $\tilde{r}(t) = \tilde{U}(x^*) - \tilde{U}(x^t)$  denote the per-round regret of the online  
 1275 learning algorithm  $\text{Alg}$  for the  $n$ -follower game. Let  $r(t) = U(x^*) - U(x^t)$  denote the per-round

1296 regret of the single-follower algorithm constructed by the above reduction. Consider the expected  
 1297 total regret in the  $n$ -follower game:

$$\begin{aligned}
 1299 \quad \mathbb{E}[\tilde{\text{Reg}}(T)] &= \sum_{t=1}^T \mathbb{E}[\tilde{r}(t)] \\
 1300 &\geq \sum_{t=1}^T \left( (1 - \delta(t)) \cdot \mathbb{E}[\tilde{r}(t)] - \delta(t) \cdot 1 \right) \\
 1301 &= \sum_{t=1}^T (1 - \delta(t)) \cdot C \cdot \mathbb{E}[r(t)] - \sum_{t=1}^T \delta(t) \\
 1302 &\geq C \cdot \sum_{t=1}^T \mathbb{E}[r(t)] - \sum_{t=1}^T \delta(t) \cdot C \cdot 1 - \sum_{t=1}^T \delta(t) \\
 1303 &= C \cdot \mathbb{E}[\text{Reg}(T)] - (C + 1) \sum_{t=1}^T \delta(t).
 \end{aligned}$$

1313 Now, we bound  $\sum_{t=1}^T \delta(t)$ . Consider a threshold  $\tau$  such that for all  $t \geq \tau$ , we have  $\delta(t) \leq \frac{1}{T^2}$ . To  
 1314 find  $\tau$ , we solve

$$2n \exp\left(-\frac{(1/100n - 1/n)^2}{2}\tau\right) \leq \frac{1}{T^2}.$$

1315 Rearranging, we choose  $\tau$  such that

$$\tau \geq \frac{\ln(2nT^2)}{C_\tau}$$

1316 where  $C_\tau = (\frac{1}{100n} - \frac{1}{n})^2$ . If  $t \leq \tau$ , we bound  $\delta(t) \leq 1$ . For  $t > \tau$ , we use the bound  $\delta(t) \leq \frac{1}{T^2}$ .  
 1317 Now, summing over all  $t$ ,

$$\sum_{t=1}^T \delta(t) \leq \tau + (T - \tau) \frac{1}{T^2} = \frac{\ln(2nT^2)}{C_\tau} + \frac{T - \tau}{T^2} \leq \mathcal{O}\left(\frac{\ln T}{C_\tau} + \frac{1}{T}\right) = \mathcal{O}(\log T).$$

1318 Then,

$$\mathbb{E}[\tilde{\text{Reg}}(T)] \geq C \cdot \mathbb{E}[\text{Reg}(T)] - \mathcal{O}(\log T).$$

1319 The regret  $\mathbb{E}[\text{Reg}(T)]$  for a single-follower game where the follower has  $nK$  types and the leader  
 1320 has  $L = nK$  actions is at least  $\Omega(\sqrt{nKT})$  by Theorem B.2. Thus, we obtain

$$\mathbb{E}[\tilde{\text{Reg}}(T)] \geq C \cdot \Omega(\sqrt{nKT}) - \mathcal{O}(\log T) = \Omega(\sqrt{nKT}),$$

1321 which is also  $\Omega(\sqrt{\min\{L, nK\}T})$  because  $L = nK$ .

## C APPENDIX FOR SECTION 5

### C.1 $\mathcal{O}(K^n \sqrt{T} \log T)$ -REGRET ALGORITHM AND THE PROOF OF THEOREM 5.1

1322 We show that the online learning problem for a Bayesian Stackelberg game with action feedback can  
 1323 be solved with  $\mathcal{O}(K^n \sqrt{T} \log T)$  regret, by using a technique developed by Bernasconi et al. (2023).

1324 Bernasconi et al. (2023) showed that the online learning problem for a linear program with unknown  
 1325 objective parameter can be reduced to a linear bandit problem. We first show that the Bayesian  
 1326 Stackelberg game (which is not a linear program as defined in Definition 2.2) can be reformulated  
 1327 as a linear program. Then, we use Bernasconi et al. (2023)'s reduction to reduce the linear program  
 1328 formulation of online Bayesian Stackelberg game to a linear bandit problem. A difference between  
 1329 our work and Bernasconi et al. (2023) is that, while they consider an adversarial online learning  
 1330 setting, we consider a stochastic online learning setting. Directly applying Bernasconi et al. (2023)'s  
 1331 result will lead to an  $\tilde{\mathcal{O}}(K^{\frac{3n}{2}} \sqrt{T})$  regret bound. Instead, we apply the OFUL algorithm for stochastic  
 1332 linear bandit (Abbasi-yadkori et al., 2011) to obtain a better regret bound of  $\tilde{\mathcal{O}}(K^n \sqrt{T})$ .

1350  
**Step 1: Reformulate Bayesian Stackelberg game as a linear program.** First, we reformulate  
1351 the Bayesian Stackelberg game optimization problem  $\max_{x \in \Delta(\mathcal{L})} U_{\mathcal{D}}(x)$  (Definition 2.2), which is a  
1352 nonlinear program by definition, into a linear program. Let variable  $x$  represent a joint distribution  
1353 over best-response function  $W \in \mathcal{A}^{nK}$  and the leader's actions  $\mathcal{L}$ . Specifically,

$$1355 \quad x = (x(W, \ell))_{W \in \mathcal{A}^{nK}, \ell \in \mathcal{L}} \in \mathbb{R}^{A^{nK} \times L}, \quad \text{where} \quad \sum_{W \in \mathcal{A}^{nK}, \ell \in \mathcal{L}} x(W, \ell) = 1, \quad \text{and} \quad x(W, \ell) \geq 0.$$

1357 Alternatively,  $x$  can be viewed as an  $A^{nK} \times L$ -dimensional matrix, where  $W$  indexes the row and  $\ell$   
1358 indexes the columns. We maximize the following objective (which is linear in  $x$ ):

$$1360 \quad \max_x U(x) = \sum_{W \in \mathcal{A}^{nK}} \sum_{\ell \in \mathcal{L}} \sum_{\theta \in \Theta^n} \mathcal{D}(\theta) x(W, \ell) u(\ell, W(\theta)), \quad (16)$$

1362 subject to the Incentive Compatibility (IC) constraint, meaning that the followers' best-response  
1363 actions are consistent with  $W$ :  $\forall W \in \mathcal{A}^{nK}, \forall i \in [n], \forall \theta_i \in \Theta, \forall a_i \in \mathcal{A}$ ,

$$1365 \quad \sum_{\ell \in \mathcal{L}} x(W, \ell) \left( v_i(\ell, w_i(\theta_i), \theta_i) - v_i(\ell, a_i, \theta_i) \right) \geq 0. \quad (17)$$

1367 **Lemma C.1.** *With known distribution  $\mathcal{D}$ , the Bayesian Stackelberg game can be solved by the linear  
1368 program (16)(17) in the following sense: there exists a solution  $x$  to (16)(17) with only one non-zero  
1369 row  $x(W^*, \cdot)$ , and this row  $x(W^*, \cdot) \in \mathbb{R}^L$  is a solution to  $\max_{x \in \Delta(\mathcal{L})} U_{\mathcal{D}}(x)$ .*

1371 *Proof.* First, we prove that the linear program (16)(17) contains an optimal solution with only one  
1372 non-zero row. Suppose an optimal solution  $x$  has two non-zero rows  $W_1, W_2$ :

$$1374 \quad \sum_{\ell \in \mathcal{L}} x(W_1, \ell) = p_1 > 0, \quad \sum_{\ell \in \mathcal{L}} x(W_2, \ell) = p_2 > 0.$$

1376 Consider the conditional expected utility of these two rows. Because, when conditioned on row  $i$ , the  
1377 conditional probability of playing action  $\ell$  is  $\frac{x(W, \ell)}{p_1}$ , we have:

$$1379 \quad u_1 = \frac{1}{p_1} \sum_{\ell \in \mathcal{L}} \sum_{\theta \in \Theta^n} x(W_1, \ell) \mathcal{D}(\theta) u(\ell, W_1(\theta)),$$

$$1382 \quad u_2 = \frac{1}{p_2} \sum_{\ell \in \mathcal{L}} \sum_{\theta \in \Theta^n} x(W_2, \ell) \mathcal{D}(\theta) u(\ell, W_2(\theta)).$$

1384 Without loss of generality, assume  $u_1 \geq u_2$ . We construct a new solution  $x'$  by transferring probability  
1385 mass from row  $W_2$  to row  $W_1$ . Specifically,  $x'$  is defined as follows:

$$1387 \quad x'(W_1, \ell) = \frac{p_1 + p_2}{p_1} x(W_1, \ell), \quad \forall \ell \in \mathcal{L}.$$

$$1389 \quad x'(W_2, \ell) = 0, \quad \forall \ell \in \mathcal{L},$$

$$1391 \quad x'(W_j, \ell) = x(W_j, \ell), \quad \forall \text{ other } W_j, \forall \ell \in \mathcal{L}.$$

1392 It is straightforward to verify that  $x'$  satisfies the IC constraint. Now, we show that the utility of  $x'$  is  
1393 weakly greater than the utility of  $x$ .

$$1394 \quad U(x') = \sum_{\ell \in \mathcal{L}} \sum_{\theta \in \Theta^n} \frac{p_1 + p_2}{p_1} x(W_1, \ell) \mathcal{D}(\theta) u(\ell, W_1(\theta))$$

$$1397 \quad + \text{utility from rows other than } \{W_1, W_2\}$$

$$1398 \quad = (p_1 + p_2) u_1 + \text{utility from rows other than } \{W_1, W_2\}$$

$$1399 \quad \geq p_1 u_1 + p_2 u_2 + \text{utility from rows other than } \{W_1, W_2\}$$

$$1400 \quad = U(x).$$

1402 Note that the  $W_2$  row of  $x'$  has become 0. We can apply this construction iteratively until only one  
1403 row remains non-zero, without decreasing utility, thus obtaining an optimal solution with only one  
1404 non-zero row.

Let  $x^*$  be an optimal solution to the linear program (16)(17) with only one non-zero row  $W^*$ . Let  $x_{BS}^* = \max_{x \in \Delta(\mathcal{L})} U_{\mathcal{D}}(x)$  be an optimal solution for the Bayesian Stackelberg game. We prove that  $U(x^*) = U_{\mathcal{D}}(x_{BS}^*)$ .

First, we prove  $U_{\mathcal{D}}(x_{BS}^*) \leq U(x^*)$ . Let  $W_{BS}^*$  be the best-response function corresponding to  $x_{BS}^*$ , i.e.,  $x_{BS}^* \in R(W_{BS}^*)$ . We construct a feasible solution  $x$  to the linear program (16)(17) by setting the row indexed by  $W_{BS}^*$  to  $x_{BS}^*$  and assigning zero values to all other rows. By definition,  $x$  satisfies the IC constraint, so it is a feasible solution. Moreover,  $x_{BS}^*$  and  $x$  achieve the same objective value  $U_{\mathcal{D}}(x_{BS}^*) = U(x)$ . By definition,  $U(x)$  is weakly less than the optimal objective value  $U(x^*)$  of the linear program, so  $U_{\mathcal{D}}(x_{BS}^*) \leq U(x^*)$ .

Then, we prove  $U(x^*) \leq U_{\mathcal{D}}(x_{BS}^*)$ . Suppose the leader uses the strategy defined by the non-zero row of  $x^*$ , which is  $x^*(W^*, \cdot) \in \Delta(\mathcal{L})$ . By the IC constraint of the linear program, the best-response function of the followers is equal to  $W^*$ , so the expected utility of the leader is exactly equal to  $U(x^*)$ , which is  $\leq U_{\mathcal{D}}(x_{BS}^*)$  because  $x_{BS}^*$  is an optimal solution for the Bayesian Stackelberg game.  $\square$

**Step 2: Reduce online Bayesian Stackelberg game to a linear bandit problem.** Based on the linear program formulation (16)(17), we then reduce the online Bayesian Stackelberg game problem to a linear bandit problem, using the technique in Bernasconi et al. (2023). Let  $\mathcal{X} \subseteq \Delta(\mathcal{A}^{nK} \times \mathcal{L})$  be the set of feasible solutions to the linear program (16)(17). We define the loss of a strategy  $x \in \mathcal{X}$  when the follower types are  $\theta \in [K]^n$  as:

$$L_{\theta}(x) = - \sum_{W \in \mathcal{A}^{nK}} \sum_{\ell \in \mathcal{L}} x(W, \ell) u(\ell, W(\theta)).$$

We define a linear map  $\phi : \mathcal{X} \rightarrow \mathbb{R}^{K^n}$  that maps a strategy  $x \in \mathcal{X}$  to a vector in  $\mathbb{R}^{K^n}$ , representing the loss of the strategy for each type profile:

$$\phi(x) = \begin{pmatrix} L_{\theta_1}(x) \\ \vdots \\ L_{\theta_{K^n}}(x) \end{pmatrix} \in \mathbb{R}^{K^n}.$$

Its inverse,  $\phi^\dagger : \mathbb{R}^{K^n} \rightarrow \mathcal{X}$  maps a loss vector back to a strategy. Let  $\text{co } \phi(\mathcal{X})$  denote the convex hull of the image set of  $\phi$ .

Let  $\mathfrak{R}$  be a stochastic linear bandit algorithm with decision space  $\text{co } \phi(\mathcal{X}) \subseteq \mathbb{R}^{K^n}$ . In particular, we let  $\mathfrak{R}$  be the OFUL algorithm (Abbasi-yadkori et al., 2011). At each round,  $\mathfrak{R}$  outputs a strategy  $z^t \in \text{co } \phi(\mathcal{X})$ , and we invoke a Carathéodory oracle to decompose  $z^t$  into  $K^n + 1$  elements from  $\phi(\mathcal{X})$ , forming a convex combination.<sup>6</sup> We then sample one of the elements  $z_j^t$ , and apply the inverse map  $\phi^\dagger$  to obtain a strategy  $x^t \in \mathcal{X}$  for the leader. After playing strategy  $x^t$ , we observe the utility  $u^t = u(\ell^t, a^t)$  and feed the utility feedback to  $\mathfrak{R}$ .

**Theorem C.1.** *The expected regret of Algorithm 5 is  $O(K^n \sqrt{T} \log T)$ .*

*Proof.* Because  $x^t \in \mathcal{X} = \Delta(\mathcal{A}^{nK} \times \mathcal{L})$  is a feasible solutions to the linear program (16)(17), it satisfies the IC constraint. So, when the leader plays  $x^t(W, \cdot)/p(W) \in \Delta(\mathcal{L})$ , the followers (with types  $\theta$ ) will best respond according to the function  $W(\theta)$ . Thus, the leader's expected utility at round  $t$  is

$$\sum_{W \in \mathcal{A}^{nK}} p(W) \sum_{\ell \in \mathcal{L}} \frac{x^t(W, \ell)}{p(W)} \sum_{\theta \in [K]^n} \mathcal{D}(\theta) u(\ell, W(\theta)) = U(x^t) = -\mathbb{E}_{\theta \sim \mathcal{D}}[L_{\theta}(x^t)].$$

Then, the regret of Algorithm 5 in  $T$  rounds can be expressed as

$$\text{Reg}(T) = \sum_{t=1}^T \left( \mathbb{E}_{\theta \sim \mathcal{D}}[L_{\theta}(x^t)] - \mathbb{E}_{\theta \sim \mathcal{D}}[L_{\theta}(x^*)] \right),$$

<sup>6</sup>The Carathéodory oracle is based on the well-known *Carathéodory Theorem*.

---

1458   **ALGORITHM 5:** Linear Bandit Algorithm for Bayesian Stackelberg Games

1459   **Input:** A linear bandit algorithm  $\mathfrak{R}$  over decision space  $\text{co } \phi(\mathcal{X})$ , where  $\mathfrak{R}.\text{RECOMMEND}()$   
1460        returns an element in  $\text{co } \phi(\mathcal{X})$ , and  $\mathfrak{R}.\text{OBSERVELOSS}$  takes the loss feedback.

1461   **1** **for** each round  $t$  **do**

1462    **2**    Use  $\mathfrak{R}.\text{RECOMMEND}()$  to obtain  $z^t \in \text{co } \phi(\mathcal{X}) \subseteq \mathbb{R}^{K^n}$ .

1463    **3**    Call a Carathéodory oracle with input  $(z^t, \phi(\mathcal{X}))$ , which returns  $K^n + 1$  elements  
1464         $\{z_i^t, \lambda_i^t\}_{i \in [K^n + 1]}$  such that:

1465        
$$z^t = \sum_{i=1}^{K^n+1} \lambda_i^t z_i^t, \quad \text{where } \sum_{i=1}^{K^n+1} \lambda_i^t = 1.$$

1466    **4**    Draw an index  $j \in \{1, \dots, K^n + 1\}$  with probabilities  $\lambda_j^t$ .

1467    **5**    Compute  $x^t \leftarrow \phi^\dagger(z_j^t)$ . Note that  $x^t \in \mathcal{X} = \Delta(\mathcal{A}^{nK} \times \mathcal{L})$  is a matrix.

1468    **6**    Play  $x^t$  in the following sense: sample a row  $W \in \mathcal{A}^{nK}$  with probability  
1469         $p(W) = \sum_{\ell \in \mathcal{L}} x^t(W, \ell)$ , then play the mixed strategy  $x^t(W, \cdot) / p(W) \in \Delta(\mathcal{L})$ .

1470    **7**    Observe the realized utility  $u^t = u(\ell^t, \mathbf{a}^t)$ .

1471    **8**    Feed the loss to  $\mathfrak{R}$  by calling  $\mathfrak{R}.\text{OBSERVELOSS}(-u^t)$ .

1472   **9**    We let  $\mathfrak{R}$  be the OFUL algorithm (Abbasi-yadkori et al., 2011).

---

1473   where  $x^*$  is the optimal strategy in  $\mathcal{X}$ , which minimizes the expected loss (maximizes expected  
1474   utility). Let  $\text{Reg}_{\mathfrak{R}, \text{co } \phi(\mathcal{X})}(T)$  be the expected regret of the linear bandit algorithm  $\mathfrak{R}$  on decision  
1475   space  $\text{co } \phi(\mathcal{X})$  in  $T$  rounds. According to the Theorem 3.1 of Bernasconi et al. (2023),

$$\text{Reg}(T) \leq \text{Reg}_{\mathfrak{R}, \text{co } \phi(\mathcal{X})}(T).$$

1476   We let  $\mathfrak{R}$  be the OFUL algorithm (Abbasi-yadkori et al., 2011). For any  $z \in \text{co } \phi(\mathcal{X}) \subseteq \mathbb{R}^{K^n}$ , the  
1477   stochastic loss of  $z$  can be expressed as  $L^t = \langle z, \mathcal{D} \rangle + \eta^t$ , with  $|L^t| \leq 1$ ,  $\|\mathcal{D}\|_2 \leq \|\mathcal{D}\|_1 = 1$ , and  
1478    $\eta^t$  being a bounded zero-mean noise. Then, from Abbasi-yadkori et al. (2011)'s Theorem 3, we have  
1479   with probability at least  $1 - \delta$ ,

$$\text{Reg}_{\mathfrak{R}, \text{co } \phi(\mathcal{X})}(T) \leq 4 \sqrt{TK^n \log \left( \lambda + \frac{TL}{K^n} \right)} \cdot \left( \lambda^{1/2} + \sqrt{2 \log \left( \frac{1}{\delta} \right) + K^n \log \left( 1 + \frac{TL}{\lambda K^n} \right)} \right)$$

1480   where  $\lambda$  is a tunable parameter in the OFUL algorithm. By setting  $\lambda = 1$  and  $\delta = \frac{1}{T}$ , we obtain

$$\mathbb{E}[\text{Reg}(T)] \leq (1 - \delta) \cdot O(K^n \sqrt{T} \log T) + \delta \cdot T = \mathcal{O}(K^n \sqrt{T} \log T).$$

□

## 1496   C.2 PROOF OF LEMMA 5.1

1497   We can express the leader's utility function as

$$1500 \quad u(x, \mathbf{a}) = \sum_{\ell \in \mathcal{L}} x(\ell) u(\ell, \mathbf{a}) = \langle x, u_{\mathbf{a}} \rangle$$

1501   where vector  $u_{\mathbf{a}} = (u(\ell, \mathbf{a}))_{\ell \in \mathcal{L}} \in \mathbb{R}^L$ . Note that  $u(x, \mathbf{a})$  is a linear function of  $u_{\mathbf{a}}$ . Consequently,  
1502   the expected utility of a strategy  $x \in R(W)$  on the true distribution  $\mathcal{D}$  is given by

$$1503 \quad U(x, R(W)) = \mathbb{E}_{\mathbf{a} \sim \mathcal{P}(\cdot | R(W))} [\langle x, u_{\mathbf{a}} \rangle].$$

1504   Given samples  $\mathbf{a}^1, \dots, \mathbf{a}^N$ , we can compute  $u_{\mathbf{a}^1}, \dots, u_{\mathbf{a}^N}$  because we know the utility function. By  
1505   Lemma B.1, the pseudo-dimension of the family of linear functions  $\{\langle x, \cdot \rangle \mid x \in R(W) \in \mathbb{R}^L\}$  is  $L$ .  
1506   Applying Theorem B.1, with  $N$  samples, we have

$$1507 \quad \Pr \left[ \exists x \in R(W), |U(x, R(W)) - \hat{U}_N(x, R(W))| > \sqrt{\frac{2L \log 3N}{N}} + \sqrt{\frac{\log \frac{1}{\delta}}{2N}} \right] \leq \delta.$$

1512 Let  $\delta = \frac{1}{T^4}$ . Taking a union bound over all  $N \in \{1, \dots, T\}$  and all  $W \in \mathcal{W}$ , we obtain  
 1513

$$1514 \Pr \left[ \exists W \in \mathcal{W}, \exists N \in [T], \exists x \in R(W), |U(x, R(W)) - \hat{U}_N(x, R(W))| > \sqrt{\frac{2L \log(3N)}{N}} + \sqrt{\frac{\log T^4}{2N}} \right] \\ 1515 \leq |\mathcal{W}| T \delta = \frac{|\mathcal{W}| T}{T^4} \leq \frac{1}{T^2} \\ 1516$$

1519 (assuming  $T \geq |\mathcal{W}|$ ). Thus, with probability at least  $1 - \frac{1}{T^2}$ , for every  $W \in \mathcal{W}$ ,  $N \in [T]$ , and  
 1520  $x \in R(W)$ , we have

$$1521 |U(x, R(W)) - \hat{U}_N(x, R(W))| \leq \sqrt{\frac{4(L+1) \log(3T)}{N}} \\ 1522$$

1524 using the inequality  $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ .  
 1525

### 1526 C.3 PROOF OF THEOREM 5.2

1528 By Lemma 5.1, the event

$$1529 C = \left[ \forall W \in \mathcal{W}, \forall N \in [T], \forall x \in R(W), |U(x, R(W)) - \hat{U}_N(x, R(W))| \leq \sqrt{\frac{4(L+1) \log(3T)}{N}} \right] \\ 1530$$

1532 happens with probability at least  $1 - \frac{1}{T^2}$ . Suppose  $C$  happens. The regret at round  $t$  is given by  
 1533

$$1534 r(t) = U(x^*, R(W^*)) - U(x^t, R(W^t)). \\ 1535$$

1536 For any strategy  $x \in R(W)$ , we define the upper confidence bound of its utility as

$$1537 \text{UCB}^t(x) = \hat{U}_{N^t(W)}(x, R(W)) + \sqrt{\frac{4(L+1) \log(3T)}{N^t(W)}}. \\ 1538$$

1540 Since  $C$  holds, it follows that

$$1541 U(x^*, R(W^*)) \leq \text{UCB}^t(x^*).$$

1542 Because the UCB algorithm chooses the strategy with the highest upper confidence bound at round  $t$ ,  
 1543 we have  $\text{UCB}^t(x^*) \leq \text{UCB}^t(x^t)$ . Thus,  
 1544

$$1545 r(t) \leq \text{UCB}^t(x^*) - U(x^t, R(W^t)) \\ 1546 \leq \text{UCB}^t(x^t) - U(x^t, R(W^t)) \\ 1547 = \hat{U}_{N^t(W)}(x^t, R(W^t)) - U(x^t, R(W^t)) + \sqrt{\frac{4(L+1) \log(3T)}{N^t(W)}} \\ 1548 \\ 1549 \leq 2 \sqrt{\frac{4(L+1) \log(3T)}{N^t(W)}}. \\ 1550$$

1553 The total regret is at most

$$1555 \text{Reg}(T) = \sum_{t=1}^T r(t) \\ 1556 \leq 2 \sum_{t=1}^T \sqrt{\frac{4(L+1) \log(3T)}{N^t(W)}} \\ 1557 \\ 1558 = 2 \sum_{W \in \mathcal{W}} \sum_{m=1}^{N^T(W)} \sqrt{\frac{4(L+1) \log(3T)}{m}} \\ 1559 \\ 1560 \leq 8 \sum_{W \in \mathcal{W}} \sqrt{N^T(W) \cdot (L+1) \cdot \log(3T)} \\ 1561 \\ 1562 \\ 1563 \\ 1564 \\ 1565$$

1566 where we applied the inequality  $\sum_{m=1}^N \sqrt{\frac{1}{m}} \leq 2\sqrt{N}$ . By Jensen's inequality,  
 1567

$$1568 \frac{1}{|\mathcal{W}|} \sum_{W \in \mathcal{W}} \sqrt{N^T(W)} \leq \sqrt{\frac{1}{|\mathcal{W}|} \sum_{W \in \mathcal{W}} N^T(W)} = \sqrt{\frac{1}{|\mathcal{W}|} T}.$$

1571 Thus,  
 1572

$$1573 \text{Reg}(T) \leq 8 \sum_{W \in \mathcal{W}} \sqrt{|\mathcal{W}|T} \cdot \sqrt{(L+1) \log(3T)} \\ 1574 = \mathcal{O}\left(\sqrt{|\mathcal{W}|L \cdot T \log T}\right) \\ 1575 = \mathcal{O}\left(\sqrt{n^L K^L A^{2L} L \cdot T \log T}\right) \\ 1576 \\ 1577 \\ 1578$$

1579 where we used  $|\mathcal{W}| = O(n^L K^L A^{2L})$  from Lemma 3.2.  
 1580

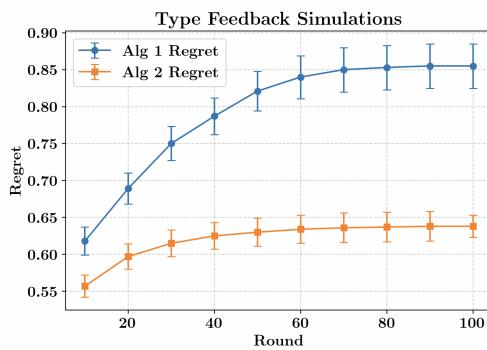
1581 Finally, considering the case where  $C$  does not happen (which has probability at most  $\frac{1}{T^2}$ ),  
 1582

$$1583 \mathbb{E}[\text{Reg}(T)] = \left(1 - \frac{1}{T^2}\right) \mathcal{O}\left(\sqrt{n^L K^L A^{2L} L \cdot T \log T}\right) + \frac{1}{T^2} \cdot T \leq \mathcal{O}\left(\sqrt{n^L K^L A^{2L} L \cdot T \log T}\right). \\ 1584$$

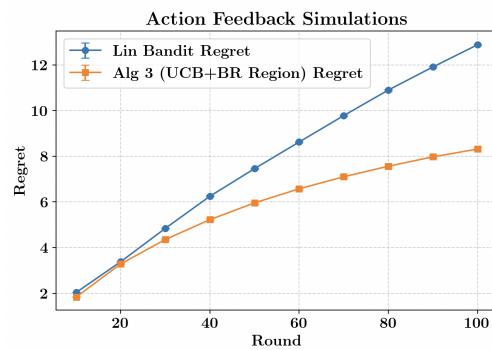
## 1585 D SIMULATIONS 1586

1587 We empirically simulate and validate the results of the studied algorithms in both the type-feedback  
 1588 setting and action feedback setting. For the former, we consider the independent type setting to  
 1589 understand how much better, in practice, is Algorithm 2 (customized for independent types) as  
 1590 opposed to the general purpose Algorithm 1 (works for general type distributions). We consider  
 1591 an  $(L = 2, K = 6, A = 2, n = 2)$  instance and simulate the results in Figure 2. As expected, the  
 1592 specialized algorithm does indeed outperform the general one.  
 1593

1594 For the action feedback case, we empirically compare our UCB-based Algorithm 3 with the linear  
 1595 bandit approach inspired by Bernasconi et al. (2023), Algorithm 5. We especially consider the small  
 1596  $n, L$  regime where our theory does not provide any concrete guidance. Shown in Figure 3, we  
 1597 consider an  $(L = 2, K = 6, A = 2, n = 2)$  instance and observe the advantage of the UCB-based  
 1598 algorithm over the linear bandit one.  
 1599



1611 Figure 2: Cumulative regret from the type-  
 1612 feedback based Algorithms 1 and 2 for an  $(L =$   
 1613  $2, K = 6, A = 2, n = 2)$  instance with inde-  
 1614 pendent types. We plot the average over 2000  
 1615 simulations with 90% confidence intervals.  
 1616  
 1617  
 1618  
 1619



1611 Figure 3: Cumulative regret from Algorithm  
 1612 5 (the Linear-Bandit approach inspired by  
 1613 Bernasconi et al. (2023)) and Algorithm 3 for  
 1614 an  $(L = 2, K = 6, A = 2, n = 2)$  instance. We  
 1615 plot the average over 2000 simulations with 90%  
 1616 confidence intervals.  
 1617  
 1618  
 1619