Sample-Efficient Learning of Stackelberg Equilibria in General-Sum Games

Yu Bai [1] **Chi Jin** [2] **Huan Wang** [1] **Caiming Xiong** [1]

Abstract

Real world applications such as economics and policy making often involve solving multi-agent games with two unique features: (1) The agents are inherently *asymmetric* and partitioned into leaders and followers; (2) The agents have different reward functions, thus the game is *general-sum*. The majority of existing results in this field focuses on either symmetric solution concepts (e.g. Nash equilibrium) or zero-sum games. It remains vastly open how to learn the *Stackelberg equilibrium*—an asymmetric analog of the Nash equilibrium—in general-sum games efficiently from samples.

This paper initiates the theoretical study of sample-efficient learning of the Stackelberg equilibrium, in the bandit feedback setting where we only observe noisy samples of the reward. We consider three representative two-player general-sum games: bandit games, bandit-reinforcement learning (bandit-RL) games, and linear bandit games. In all these games, we identify a fundamental gap between the exact value of the Stackelberg equilibrium and its estimated version using finitely many noisy samples, which can not be closed information-theoretically regardless of the algorithm. We then establish sharp positive results on sample-efficient learning of Stackelberg equilibrium with value optimal up to the gap identified above, with matching lower bounds in the dependency on the gap, error tolerance, and the size of the action spaces. Overall, our results unveil unique challenges in learning Stackelberg equilibria under noisy bandit feedback, which we hope could shed light on future research on this topic.

[1]Salesforce Research [2]Princeton University. Correspondence to: Yu Bai <yu.bai@salesforce.com>.

Workshop on Reinforcement Learning Theory, 38^{th} International Conference on Machine Learning, 2021.

1. Introduction

Real-world problems such as economic design and policy making can often be modeled as multi-agent games that involves two levels of thinking: The policy maker—as a player in this game—needs to reason about the other player's optimal behaviors given her decision, in order to inform her own optimal decision making. Consider for example the optimal taxation problem in the AI Economist (Zheng et al., 2020), a game modeling a real-world social-economic system involving a *leader* (e.g. the government) and a group of interacting *followers* (e.g. citizens). The leader sets a tax rate which determines an economics-like game for the followers; the followers then play in this game with the objective to maximize their own reward (such as individual productivity). However, the goal of the leader is to maximize her own reward (such as overall equality) which is in general different from the followers' rewards, making these games *general-sum* (Roughgarden, 2010). Such two-level thinking appears broadly in other applications as well such as in automated mechanism design (Conitzer & Sandholm, 2002; 2004), optimal auctions (Cole & Roughgarden, 2014; Dütting et al., 2019), security games (Tambe, 2011), reward shaping (Leibo et al., 2017), and so on.

Another key feature in such games is that the players are *asymmetric*, and they act in turns: the leader first plays, then the follower sees the leader's action and then adapts to it. This makes symmetric solution concepts such as Nash equilibrium (Nash, 1951) not always appropriate. A more natural solution concept for these games is the *Stackelberg equilibrium*: the leader's optimal strategy, assuming the followers play their best response to the leader (Simaan & Cruz, 1973; Conitzer & Sandholm, 2006). The Stackelberg equlibrium is often the desired solution concept in many of the aforementioned applications. Furthermore, it is of compelling interest to understand the learning of Stackelberg equilibria from *samples*, as it is often the case that we can only learn about the game through interactively deploying policies and observing the (noisy) feedback from the game (Zheng et al., 2020).

Despite the motivations, theoretical studies of learning Stackelberg equilibria in general-sum games remain vastly open, in particular when we can only learn from random samples. A line of work provides guarantees for finding

Stackelberg equilibria in general-sum games, but restricts attention to either the full observation setting (so that the exact game is observable) or with an exact best-response oracle (Conitzer & Sandholm, 2006; Letchford et al., 2009; Von Stengel & Zamir, 2010; Peng et al., 2019). These results lay out a foundation for analyzing the Stackelberg equilibrium, but do not generalize to the bandit feedback setting in which the game can only be learned from random samples. Another line of work considers the sample complexity of learning the Nash equilibrium in Markov games (Pérolat et al., 2017; Bai & Jin, 2020; Bai et al., 2020; Liu et al., 2020; Xie et al., 2020; Zhang et al., 2020), which also do not imply algorithms for finding the Stackelberg equilibrium in these games as the Nash is in general different from the Stackelberg equilibrium in general-sum games. So far, it is unclear how to learn Stackelberg in general-sum games sample-efficiently.

In this work, we study the sample complexity of learning Stackelberg equilibrium in general-sum games. We focus on general-sum games with two players (one leader and one follower), in which we wish to learn an approximate Stackelberg equilbrium for the leader from random samples. Our contributions can be summarized as follows.

- As a warm-up, we consider *bandit games* in which the two players play an action in turns and observe their own rewards (Section 3). We identify a fundamental gap between the exact Stackelberg value and its estimated version from finite samples, which cannot be closed information-theoretically regardless of the algorithm. We then propose a rigorous definition gap_ε for this gap, and show that it is possible to sample-efficiently learn the $(\mathsf{gap}_\varepsilon + \varepsilon)$-approximate Stackelberg equilibrium with $\widetilde{O}(AB/\varepsilon^2)$ samples, where A, B are the number of actions for the two players. We further show a matching lower bound $\Omega(AB/\varepsilon^2)$. We also establish similar results for learning Stackelberg in simultaneous matrix games (Appendix D).
- We consider *bandit-RL games* in which the leader's action determines an episodic Markov Decision Process (MDP) for the follower. We show that a $(\mathsf{gap}_\varepsilon + \varepsilon)$ approximate Stackelberg equilibrium for the leader can be found in $\widetilde{O}(H^5S^2AB/\varepsilon^2)$ episodes of play, where H, S are the horizon length and number of states for the follower's MDP, and A, B are the number of actions for the two players (Section 4). Our algorithm utilizes recently developed reward-free reinforcement learning techniques to enable fast exploration for the follower within the MDPs.
- Finally, we consider *linear bandit games* in which the action spaces for the two players can be arbitrarily large, but the reward is a linear function of a d-dimensional feature representation of the actions. We design an algorithm that achieves $\widetilde{O}(d^2/\varepsilon^2)$ sample complexity upper bound for linear bandit games (Section 5). This only depends polynomially on the feature dimension instead of the size of the action spaces, and has at most an $\widetilde{O}(d)$ gap from the lower bound.

2. Preliminaries

Bandit games A general-sum two-player bandit game can be described by a tuple $M = (\mathcal{A}, \mathcal{B}, r_1, r_2)$, which defines the following game played by two players, a *leader* and a *follower*:

- The leader plays an action $a \in \mathcal{A}$, with $|\mathcal{A}| = A$.
- The follower sees the action played by the leader, and plays an action $b \in \mathcal{B}$, with $|\mathcal{B}| = B$.
- The follower observes a (potentially random) reward $r_2(a, b) \in [0, 1]$. The leader also observes her own reward $r_1(a, b) \in [0, 1]$.

Best response, Stackelberg equilibrium Let $\mu_i(a, b) := \mathbb{E}[r_i(a, b)]$ $(i = 1, 2)$ denote the mean rewards. For each leader action a, the *best response set* $\mathsf{BR}_0(a)$ is the set of follower actions that maximize $\mu_2(a, \cdot)$:

$$\mathsf{BR}_0(a) := \left\{ b : \mu_2(a, b) = \max_{b' \in \mathcal{B}} \mu_2(a, b') \right\}. \tag{1}$$

Given the best-response set $\mathsf{BR}_0(a)$, we define the function $\phi_0 : \mathcal{A} \to [0, 1]$ as the leader's value function when the follower plays the worst-case best response (henceforth the"exact ϕ-function"):

$$\phi_0(a) := \min_{b \in \mathsf{BR}_0(a)} \mu_1(a, b), \tag{2}$$

This is the value function for the leader action a, assuming the follower plays the best response to a and breaks ties in the best response set against the leader's favor. This is known as *pessimistic tie breaking* and provides a worst-case guarantee for the leader (Conitzer & Sandholm, 2006).

The Stackelberg Equilibrium (henceforth also "Stackelberg") for the leader is the "best response to the best response", i.e. any action $a_\star$ that maximizes ϕ_0 (Simaan & Cruz, 1973):

$$a_\star \in \arg\max_{a \in \mathcal{A}} \phi_0(a). \tag{3}$$

We are interested in finding approximate solutions to the Stackelberg equilibrium, that is, an action $\widehat{a}$ that approximately maximizes $\phi_0(a)$.

In this paper we consider the *bandit feedback* setting, that is, the algorithm cannot directly observe the mean rewards

$\mu_1(\cdot,\cdot)$ and $\mu_2(\cdot,\cdot)$, and can only query (a,b) and obtain random samples $(r_1(a,b), r_2(a,b))$. Our goal is to determine the number of samples in order to find an approximate maximizer of $\phi_0(a)$. Note that the bandit feedback setting assumes observation noise, which as we will see brings in a fundamental challenge that is not present in existing work on learning Stackelberg, which assumes either exact observation of the mean rewards (Conitzer & Sandholm, 2006; Letchford & Conitzer, 2010), or the best-response oracle that can query a and obtain the exact best response set $\mathsf{BR}_0(a)$ (Letchford et al., 2009; Peng et al., 2019).

3. Warm-up: bandit games

Hardness of maximizing ϕ_0 from samples Given the exact ϕ-function ϕ_0 (2), a natural notion of approximate Stackelberg equilibrium is to find an action $\widehat{a}$ that is ε near-optimal for maximizing ϕ_0:

$$\phi_0(\widehat{a}) \geq \max_{a\in\mathcal{A}} \phi_0(a) - \varepsilon. \tag{4}$$

However, the following lower bound shows that, in the worst case, it is hard to find such $\widehat{a}$ from finite samples.

Theorem 1 ($\Omega(1)$ lower bound for maximizing ϕ_0)**.** *For any sample size n and any algorithm for maximizing ϕ_0 that outputs an action $\widehat{a}\in\mathcal{A}$, there exists a bandit game with $A=B=2$ on which the algorithm must suffer from $\Omega(1)$ error with probability at least $1/3$:*

$$\phi_0(\widehat{a}) \leq \max_{a\in\mathcal{A}} \phi_0(a) - 1/2.$$

Theorem 1 stems from a *hardness of determining the best response* $\mathsf{BR}_0(a)$ *exactly* from samples. (See Table 2 for the construction of the hard instance and Appendix E.1 for the full proof of Theorem 1.) This is in stark contrast to the standard $1/\sqrt{n}$ type learning result in finding other solution concepts such as the Nash equilibrium (Bai & Jin, 2020; Liu et al., 2020), and suggests a new fundamental challenge to learning Stackelberg equilibrium from samples.

Learning Stackelberg with value optimal up to gap The lower bound in Theorem 1 shows that approximately maximizing ϕ_0 is information-theoretically hard. Motivated by this, we consider in turn a slightly relaxed notion of optimality, in which we consider maximizing ϕ_0 only up to the *gap* between ϕ_0 and its counterpart using ε-approximate best responses. More concretely, define the ε-approximate versions of the best response set and ϕ-function as

$$\phi_\varepsilon(a) := \min_{b\in\mathsf{BR}_\varepsilon(a)} \mu_1(a,b),$$
$$\mathsf{BR}_\varepsilon(a) := \left\{ b\in\mathcal{B} : \mu_2(a,b) \geq \max_{b'} \mu_2(a,b') - \varepsilon \right\}.$$

These definitions are similar to the vanilla BR_0 and ϕ_0 in (1) and (2), except that we allow any ε-approximate best response to be considered as a valid response to the leader action. Observe we always have $\mathsf{BR}_\varepsilon(a) \supseteq \mathsf{BR}_0(a)$ and $\phi_\varepsilon(a) \leq \phi_0(a)$. We then define the *gap* of the game for any $\varepsilon\in(0,1)$ as

$$\mathsf{gap}_\varepsilon := \max_{a\in\mathcal{A}} \phi_0(a) - \max_{a\in\mathcal{A}} \phi_\varepsilon(a) \geq 0. \tag{5}$$

This gap_ε is discontinuous in ε in general and can be as large as $\Theta(1)$ for any $\varepsilon>0$. With the definition of the gap, we are now ready to state our main result, which shows that it is possible to sample-efficiently learn Stackelberg Equilibria with value up to $(\mathsf{gap}_\varepsilon+\varepsilon)$. The proof can be found in Appendix E.3.

Theorem 2 (Learning Stackelberg in bandit games)**.** *For any bandit game and $\varepsilon\in(0,1)$, Algorithm 5 outputs $(\widehat{a},\widehat{b})$ such that with probability at least $1-\delta$,*

$$\phi_0(\widehat{a}) \geq \phi_{\varepsilon/2}(\widehat{a}) \geq \max_{a\in\mathcal{A}} \phi_0(a) - \mathsf{gap}_\varepsilon - \varepsilon,$$
$$\mu_2(\widehat{a},\widehat{b}) \geq \max_{b'\in\mathcal{B}} \mu_2(\widehat{a},b') - \varepsilon$$

with $n=\widetilde{O}\big(AB/\varepsilon^2\big)$ samples, where $\widetilde{O}(\cdot)$ hides log factors. Further, the algorithm runs in $O(n)=\widetilde{O}(AB/\varepsilon^2)$ time.

Theorem 2 shows that it is possible to learn $\widehat{a}$ that maximizes $\phi_0(a)$ up to $(\mathsf{gap}_\varepsilon+\varepsilon)$ accuracy, using $\widetilde{O}(AB/\varepsilon^2)$ samples. The quantity gap_ε is not bounded and can be as large as $\Theta(1)$ for any ε (see Lemma E.1 for a formal statement); however the gap is non-increasing as we decrease ε. IIn general, Theorem 2 presents a "best-effort" positive result for learning Stackelberg under this relaxed notion of optimality. To the best of our knowledge, this is the first result for sample-efficient learning of Stackelberg equilibrium in general-sum games with bandit feedbacks.

Lower bound We accompany Theorem 2 by an $\Omega(AB/\varepsilon^2)$ sample complexity lower bound, showing that Theorem 2 achieves the optimal sample complexity up to logarithmic factors, and that $(\mathsf{gap}_\varepsilon+\varepsilon)$ suboptimality is perhaps a sensible learning goal for finding approximate Stackelberg. The proof of Theorem 3 is deferred to Appendix E.4.

Theorem 3 (Lower bound for bandit games)**.** *There exists an absolute constant $c>0$ such that the following holds. For any $\varepsilon\in(0,c)$, $g\in[0,c)$, any $A,B\geq 3$, and any algorithm that queries $N\leq c\big[AB/\varepsilon^2\big]$ samples and outputs an estimate $\widehat{a}\in\mathcal{A}$, there exists a bandit game M on which $\mathsf{gap}_\varepsilon=g$ and the algorithm suffers from $(g+\varepsilon)$ error:*

$$\phi_{\varepsilon/2}(\widehat{a}) \leq \phi_0(\widehat{a}) \leq \max_{a\in\mathcal{A}} \phi_0(a) - g - \varepsilon$$

with probability at least $1/3$.

4. Bandit-RL games

Setting A bandit-RL game is described by the leader's action set $\mathcal{A}$ (with $|\mathcal{A}| = A$), and a family of MDPs $M = \{M^a : a \in \mathcal{A}\}$. Each leader action $a \in \mathcal{A}$ determines an episodic MDP $M^a = (H, \mathcal{S}, \mathcal{B}, \mathbb{P}^a, r_{1,h}(a, \cdot, \cdot), r_{2,h}(a, \cdot, \cdot))$ that contains H steps, S states, B actions, with two reward functions r_1 and r_2. In each episode of play, (1) The leader plays action $a \in \mathcal{A}$; (2) The follower sees this action and enters the MDP M^a. She observes the deterministic[1] initial state s_1, and plays in M^a with exploration feedback for one episode. While the follower plays in the MDP, she observes reward $r_{2,h}(a, s_h, b_h)$, whereas the leader also observes her own reward $r_{1,h}(a, s_h, b_h)$.

We let π^b denote a policy for the follower, and let $V_1(a, \pi^b)$ and $V_2(a, \pi^b)$ denote its value functions (in M^a) for the leader and the follower respectively. Similar as in bandit games, we define the ε-approximate best-response set $\mathsf{BR}_\varepsilon(a)$ and the ε-approximate ϕ-function $\phi_\varepsilon(a)$ for all $\varepsilon \geq 0$ as

$$\phi_\varepsilon(a) := \min_{\pi^b \in \mathsf{BR}_\varepsilon(a)} V_1(a, \pi^b),$$

$$\mathsf{BR}_\varepsilon(a) := \left\{ \pi^b : V_2(a, \pi^b) \geq \max_{\widetilde{\pi}^b} V_2(a, \widetilde{\pi}^b) - \varepsilon \right\}.$$

Define $\mathsf{gap}_\varepsilon = \max_{a \in \mathcal{A}} \phi_0(a) - \max_{a \in \mathcal{A}} \phi_\varepsilon(a)$ similarly as in (5).

We now state our theoretical guarantee for bandit-RL games. Our algorithm uses reward-free exploration to handle the two reward functions, as well as a constrained MDP solver to minimize leader's value within follower's best response set. (cf. Algorithm 7). The proof can be found in Appendix F.2.

Theorem 4 (Learning Stackelberg in bandit-RL games)**.** *For any bandit-RL game and sufficiently small $\varepsilon \leq O(1/H^2S^2)$, Algorithm 7 with $n = \widetilde{O}(H^5S^2AB/\varepsilon^2 + H^7S^4AB/\varepsilon)$ episodes of play can return $(\widehat{a}, \widehat{\pi}^b)$ such that with probability at least $1 - \delta$,*

$$\phi_0(\widehat{a}) \geq \phi_{\varepsilon/2}(\widehat{a}) \geq \max_{a \in \mathcal{A}} \phi_0(a) - \mathsf{gap}_\varepsilon - \varepsilon,$$

$$V_2(\widehat{a}, \widehat{\pi}^b) \geq \max_{\widetilde{\pi}^b} V_2(\widehat{a}, \widetilde{\pi}^b) - \varepsilon,$$

where $\widetilde{O}(\cdot)$ hides $\log(HSAB/\delta\varepsilon)$ factors. Further, the algorithm runs in $\mathrm{poly}(HSAB/\delta\varepsilon)$ time.

Sample complexity, relationship with reward-free RL Theorem 4 shows that for bandit-RL games, the approximate Stackelberg Equilibrium (with value optimal up to $\mathsf{gap}_\varepsilon + \varepsilon$) can be efficiently found with polynomial sample complexity and runtime. In particular, (for small ε) the leading term in the sample complexity scales as $\widetilde{O}(H^5S^2AB/\varepsilon^2)$. Since bandit-RL games include bandit games as a special case, the $\Omega(AB/\varepsilon^2)$ lower bound for bandit games (Theorem 3) apply here and implies that the A, B dependence in Theorem 4 is tight, while the H dependence may be suboptimal. We also remark the learning goal for the follower in our bandit-RL game is a new RL setting in between the single-reward setting and the full reward-free setting, for which the optimal S dependence is currently unclear and may be an interesting open question.

[1]The general case where s_1 is stochastic reduces to the deterministic case by adding a step $h = 0$ with a single deterministic initial state s_0, which only increases the horizon of the game by 1.

5. Linear bandit games

Setting We consider a bandit game with action space $\mathcal{A}$, $\mathcal{B}$ that are finite but potentially arbitrarily large, and assume in addition that the reward functions has a linear form

$$r_i(a, b) = \phi(a, b)^\top \theta_i^\star + z_i, \quad i = 1, 2, \tag{6}$$

where $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ is a d-dimensional feature map, $\theta_1^\star, \theta_2^\star \in \mathbb{R}^d$ are unknown ground truth parameters for the rewards, and z_1, z_2 are random noise which we assume are mean-zero and 1-sub-Gaussian. For linear bandit games, we define gap_ε same as definition (5) for bandit games.

Our algorithm for linear bandit games is based on a `CoreSet` subroutine to query representative features, and a weighted lesat square step to estimate $\theta_{1,2}^\star$ (cf. Algorithm 8). We now state its main theoretical guarantee; the proof can be found in Appendix G.2.

Theorem 5 (Learning Stackelberg in linear bandit games)**.** *For any linear bandit game, Algorithm 8 outputs a $(\mathsf{gap}_\varepsilon + \varepsilon)$-approximate Stackelberg equilibrium $(\widehat{a}, \widehat{b})$ with probability at least $1 - \delta$:*

$$\phi_0(\widehat{a}) \geq \phi_{\varepsilon/2}(\widehat{a}) \geq \max_{a \in \mathcal{A}} \phi_0(a) - \mathsf{gap}_\varepsilon - \varepsilon,$$

in at most $n = \widetilde{O}(d^2/\varepsilon^2)$ queries.

Sample complexity; computation Theorem 5 shows that Algorithm 8 achieves $\widetilde{O}(d^2/\varepsilon^2)$ sample complexity for learning Stackelberg equilibria in linear bandit games. This only depends polynomially on the feature dimension d instead of the size of the action spaces A, B, which improves over Algorithm 5 when A, B are large and is desired given the linear structure (6). This sample complexity has at most a $\widetilde{O}(d)$ gap from the lower bound: An $\Omega(d/\varepsilon^2)$ lower bound for linear bandit games can be obtained by directly adapting $\Omega(AB/\varepsilon^2)$ lower bound for bandit games in Theorem 3 (see Appendix G.3 for a formal statement and proof). We also note that, while the focus of Theorem 5 is on the sample complexity rather than the computation, Algorithm 8 is guaranteed to run in $\mathrm{poly}(A, B, d, 1/\varepsilon^2)$ time.

References

Ahmadinejad, A., Dehghani, S., Hajiaghayi, M., Lucier, B., Mahini, H., and Seddighin, S. From duels to battlefields: Computing equilibria of blotto and other games. *Mathematics of Operations Research*, 44(4):1304–1325, 2019.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Blum, A., Haghtalab, N., and Procaccia, A. D. Learning optimal commitment to overcome insecurity. 2014.

Blum, A., Haghtalab, N., Hajiaghayi, M., and Seddighin, S. Computing stackelberg equilibria of large general-sum games. In *International Symposium on Algorithmic Game Theory*, pp. 168–182. Springer, 2019.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cole, R. and Roughgarden, T. The sample complexity of revenue maximization. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 243–252, 2014.

Conitzer, V. and Sandholm, T. Complexity of mechanism design. *arXiv preprint cs/0205075*, 2002.

Conitzer, V. and Sandholm, T. Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pp. 132–141, 2004.

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

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D., and Ravindranath, S. S. Optimal auctions through deep learning. In *International Conference on Machine Learning*, pp. 1706–1715. PMLR, 2019.

Fiez, T., Chasnov, B., and Ratliff, L. J. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Foster, D. J., Li, Z., Lykouris, T., Sridharan, K., and Tardos, E. Learning in games: Robustness of fast convergence. *arXiv preprint arXiv:1606.06244*, 2016.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020a.

Jin, C., Netrapalli, P., and Jordan, M. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pp. 4880–4889. PMLR, 2020b.

Korpelevich, G. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Korzhyk, D., Yin, Z., Kiekintveld, C., Conitzer, V., and Tambe, M. Stackelberg vs. nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *Journal of Artificial Intelligence Research*, 41:297–327, 2011.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pp. 5662–5670. PMLR, 2020.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Letchford, J. and Conitzer, V. Computing optimal strategies to commit to in extensive-form games. In *Proceedings of the 11th ACM conference on Electronic commerce*, pp. 83–92, 2010.

Letchford, J., Conitzer, V., and Munagala, K. Learning and approximating the optimal strategy to commit to. In *International Symposium on Algorithmic Game Theory*, pp. 250–262. Springer, 2009.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Marchesi, A., Trovò, F., and Gatti, N. Learning probably approximately correct maximin strategies in simulation-based games with infinite strategy spaces. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 834–842, 2020.

Nash, J. Non-cooperative games. *Annals of mathematics*, pp. 286–295, 1951.

Nemirovski, A. Prox-method with rate of convergence o (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004.

Nemirovski, A. S. and Yudin, D. B. Cesari convergence of the gradient method of approximating saddle points of convex-concave functions. In *Doklady Akademii Nauk*, volume 239, pp. 1056–1059. Russian Academy of Sciences, 1978.

Nouiehed, M., Sanjabi, M., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

Peng, B., Shen, W., Tang, P., and Zuo, S. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2149–2156, 2019.

Pérolat, J., Strub, F., Piot, B., and Pietquin, O. Learning nash equilibrium for general-sum markov games from batch data. In *Artificial Intelligence and Statistics*, pp. 232–241. PMLR, 2017.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Rakhlin, A. and Sridharan, K. Optimization, learning, and games with predictable sequences. *arXiv preprint arXiv:1311.1869*, 2013.

Roughgarden, T. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Shoham, Y. and Leyton-Brown, K. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

Simaan, M. and Cruz, J. B. On the stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, 11(5):533–555, 1973.

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

Todd, M. J. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

von Neumann, J. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Von Stengel, B. and Zamir, S. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69 (2):446–457, 2010.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pp. 3674–3682. PMLR, 2020.

Zhang, K., Kakade, S. M., Başar, T., and Yang, L. F. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., and Socher, R. The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

A. Additional related work

Since the seminal paper of (von Neumann, 1928), notions of equilibria in games and their algorithmic computation have received wide attention (see, e.g., Cesa-Bianchi & Lugosi, 2006; Shoham & Leyton-Brown, 2008). For the scope of this paper, this section focuses on reviewing results that related to learning Stackelberg equilibria.

Learning Stackelberg equilibria in zero-sum games The first category of results study two-player zero-sum games, where the rewards of the two players sum to zero. Most results along this line focus on the bilinear or convex-concave setting (see, e.g., Korpelevich, 1976; Nemirovski & Yudin, 1978; Nemirovski, 2004; Rakhlin & Sridharan, 2013; Foster et al., 2016), where the Stackelberg equilibrium coincide with the Nash equilibrium due to Von Neumann's minimax theorem (von Neumann, 1928). Results for learning Stackelberg equilibria beyond convex-concave setting are much more recent, with Rafique et al. (2018); Nouiehed et al. (2019) considering the nonconvex-concave setting, and Fiez et al. (2019); Jin et al. (2020b); Marchesi et al. (2020) considering the nonconvex-nonconcave setting. Marchesi et al. (2020) provide sample complexity results for learning Stackelberg with infinite strategy spaces, using discretization techniques that may scale exponentially in the dimension without further assumptions on the problem structure.

We remark that a crucial property of zero-sum games is that any two strategies giving similar rewards for the follower will also give similar rewards for the leader. This is no longer true in general-sum games, and prevents most statistical results for learning Stackelberg equilibria in the zero-sum setting from generalizing to the general-sum setting.

Learning Stackelberg equilibria in general-sum games The computational complexity of finding Stackelberg equilibria in games with simultaneous play ("computing optimal strategy to commit to") is studied in (Conitzer & Sandholm, 2006; Letchford & Conitzer, 2010; Von Stengel & Zamir, 2010; Korzhyk et al., 2011; Ahmadinejad et al., 2019; Blum et al., 2019). These results assume full observation of the payoff function, and show that several versions of matrix games and extensive-form (multi-step) games admit polynomial time algorithms. Another line of work considers learning Stackelberg with a "best response oracle" (Letchford et al., 2009; Blum et al., 2014; Peng et al., 2019), which is in general incomparable with our bandit feedback oracle (their oracle allows querying any leader strategy and returns the *exact* best response for that strategy) and do not imply sample complexity results in our setting. Fiez et al. (2019) study the local convergence of first-order algorithms for finding Stackelberg equilibria in general-sum games. Their result also assumes exact feedback and do not allow sampling errors. Lastly, the AI Economist (Zheng et al., 2020) studied the optimal taxation problem by learning the Stackelberg equilibrium via a two-level reinforcement learning approach.

Learning equilibria in Markov games A recent line of results (Bai & Jin, 2020; Bai et al., 2020; Xie et al., 2020; Zhang et al., 2020) consider learning Markov games (Shapley, 1953)—a generalization of Markov decision process to the multi-agent setting. We remark that all three settings studied in this paper can be cast as special cases of general-sum Markov games, which is studied by (Liu et al., 2020). In particular, Liu et al. (2020) provides sample complexity guarantees for finding Nash equilibria, correlated equilibria, or coarse correlated equilibria of the general-sum Markov games. These Nash-finding algorithms are related to our setting, but do not imply results for learning Stackelberg (see Section **??** and Appendix E.5 for detailed discussions).

B. Preliminaries on Markov decision processes

We also present the basics of a Markov Decision Processes (MDPs), which will be useful for the bandit-RL games in Section 4. We consider episodic MDPs defined by a tuple $(H, \mathcal{S}, \mathcal{B}, d^1, \mathbb{P}, r)$, where H is the horizon length, $\mathcal{S}$ is the state space, $\mathcal{B}$ is the action space[2], $\mathbb{P} = \{\mathbb{P}_h(\cdot|s,b) : h \in [H], s \in \mathcal{S}, b \in \mathcal{B}\}$ is the transition probabilities, and $r = \{r_h : \mathcal{S} \times \mathcal{B} \to [0,1], h \in [H]\}$ are the (potentially random) reward functions. A policy $\pi = \{\pi_h^b(\cdot|s) \in \Delta_{\mathcal{B}} : h \in [H], s \in \mathcal{S}\}$ for the player is a set of probability distributions over actions given the state.

In this paper we consider the exploration setting as the protocol of interacting with MDPs, similar as in (Azar et al., 2017; Jin et al., 2018). The learning agent is able to play episodes repeatedly, where in each episode at step $h \in \{1, \ldots, H\}$, the agent observes state $s_h \in \mathcal{S}$, takes an action $b_h \sim \pi_h(\cdot|s_h)$, observes her reward $r_h = r_h(s_h, b_h) \in [0,1]$, and transits to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, b_h)$. The initial state is received from the MDP: $s_1 \sim d^1(\cdot)$. The overall value function (return)

[2]The notation $\mathcal{B}$ indicates that the MDP is played by the follower (cf. Section 4); we reserve $\mathcal{A}$ as the leader's action space in this paper.

Algorithm 1 Learning Stackelberg in bandit games with optimistic tie-breaking

Require: Target accuracy $\varepsilon > 0$.
set $N \leftarrow C \log(AB/\delta)/\varepsilon^2$ for some constant $C > 0$.
1: Query each $(a, b) \in \mathcal{A} \times \mathcal{B}$ for N times and obtain $\{r_1^{(j)}(a, b), r_2^{(j)}(a, b)\}_{j=1}^N$.
2: Construct empirical estimates of the means $\widehat{\mu}_i(a, b) = \frac{1}{N} \sum_{j=1}^N r_i^{(j)}(a, b)$ for $i = 1, 2$.
3: Construct approximate best response sets and values for all $a \in \mathcal{A}$:

$$\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) := \left\{ b : \widehat{\mu}_2(a, b) \geq \max_{b' \in \mathcal{B}} \widehat{\mu}_2(a, b') - 3\varepsilon/4 \right\},$$
$$\widehat{\psi}_{3\varepsilon/4}(a) := \max_{b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)} \widehat{\mu}_1(a, b).$$

4: Output $(\widehat{a}, \widehat{b})$ where $\widehat{a} = \arg\max_{a \in \mathcal{A}} \widehat{\psi}_{3\varepsilon/4}(a)$, $\widehat{b} = \arg\max_{b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \widehat{\mu}_1(\widehat{a}, b)$.

of a policy π is defined as $V(\pi) := \mathbb{E}_\pi \left[\sum_{h=1}^H r_h(s_h, b_h) \right]$.

C. Results with optimistic tie-breaking

In this section, we present alternative versions of our main results where the Stackelberg equilibrium is defined via *optimistic tie-breaking*.

C.1. Bandit games

The setting is exactly the same as in Section 3, except that now we consider optimistic versions of the ϕ-functions that take the **max** over best-response sets (henceforth the ψ-functions):

$$\psi_\varepsilon(a) := \max_{b \in \mathsf{BR}_\varepsilon(a)} \mu_1(a, b), \tag{7}$$

for all $\varepsilon \geq 0$. Notice that now $\psi_\varepsilon \geq \psi_0$, and we consider the following new definition of gap:

$$\widetilde{\mathsf{gap}}_\varepsilon := \max_{a \in \mathcal{A}_\varepsilon} [\psi_\varepsilon(a) - \psi_0(a)], \quad \text{where}$$
$$\mathcal{A}_\varepsilon := \left\{ a \in \mathcal{A} : \psi_\varepsilon(a) \geq \max_{a \in \mathcal{A}} \psi_0(a) - \varepsilon \right\}.$$

Our desired optimality guarantee is

$$\psi_0(\widehat{a}) \geq \max_{a \in \mathcal{A}} \psi_0(a) - \widetilde{\mathsf{gap}}_\varepsilon - \varepsilon.$$

We now state our sample complexity upper bound for learning Stackelberg in bandit games under optimistic tie-breaking. The proof can be found in Section H.1.

Theorem C.1 (Bandit games with optimistic tie-breaking)**.** *For the two-player bandit game and any $\varepsilon \in (0, 1)$, Algorithm 1 outputs $(\widehat{a}, \widehat{b})$ such that with probability at least $1 - \delta$,*

$$\psi_0(\widehat{a}) \geq \max_{a \in \mathcal{A}} \psi_0(a) - \widetilde{\mathsf{gap}}_\varepsilon - \varepsilon,$$
$$\mu_2(\widehat{a}, \widehat{b}) \geq \max_{b' \in \mathcal{B}} \mu_2(\widehat{a}, b') - \varepsilon$$

with $n = \widetilde{O}(AB/\varepsilon^2)$ samples, where $\widetilde{O}(\cdot)$ hides log factors. Further, the algorithm runs in $O(n) = \widetilde{O}(AB/\varepsilon^2)$ time.

Intuitions about new gap We provide some intuitions about why—in contrast to the gap_ε defined in Section 3—our sample complexity depends on this newly defined $\widetilde{\mathsf{gap}}_\varepsilon$ here. Observe that, $\widetilde{\mathsf{gap}}_\varepsilon$ measures the max gap between $\psi_\varepsilon(a) -$

Algorithm 2 Learning Stackelberg in bandit-RL games with optimistic tie-breaking

Require: Target accuracy $\varepsilon > 0$.
set $N \leftarrow \widetilde{O}(H^5 S^2 B/\varepsilon^2 + H^7 S^4 B/\varepsilon)$.
for $a \in \mathcal{A}$ **do**
  Let the leader pull arm $a \in \mathcal{A}$ and the follower run the `Reward-Free RL-Explore` algorithm for N episodes, and obtain model estimate $\widehat{M}^a$.
  Let $(\widehat{V}_1(a,\cdot), \widehat{V}_2(a,\cdot))$ denote the value functions for the model $\widehat{M}^a$.
  Compute the empirical best response value $\widehat{V}_2^\star(a) := \max_{\pi^b} \widehat{V}_2(a, \pi^b)$ by any optimal planning algorithm (e.g. value iteration) on the empirical MDP $\widehat{M}^a$.
  Solve the following program

$$\begin{aligned} &\text{maximize}_{\pi^b} \;\; \widehat{V}_1(a, \pi^b) \\ &\text{s.t.} \;\; \pi^b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a) := \Big\{ \pi^b : \widehat{V}_2(a, \pi^b) \geq \widehat{V}_2^\star(a) - 3\varepsilon/4 \Big\} \end{aligned} \tag{8}$$

  by subroutine $(\widehat{\pi}^{b,(a)}, \widehat{\psi}_{3\varepsilon/4}(a)) \leftarrow$ `BestCaseBestResponse`$(\widehat{M}^a, \widehat{V}_2^\star(a) - 3\varepsilon/4)$.
end for
output $(\widehat{a}, \widehat{\pi}^b)$, where $\widehat{a} \leftarrow \arg\max_{a\in\mathcal{A}} \widehat{\psi}_{3\varepsilon/4}(a)$ and $\widehat{\pi}^b \leftarrow \widehat{\pi}^{b,(\widehat{a})}$.

$\psi_0(a)$, over all possible a's whose ψ_ε can compete with the best ψ_0. For any of these a's, statistically (if we only have $O(AB/\varepsilon^2)$ samples), the best response set $\mathsf{BR}_\varepsilon(a)$ is indistinguishable from $\mathsf{BR}_0(a)$, and we may well pick these a's as the Stackelberg equilibrium. However, their true ψ_0 can be (much) lower than the ψ_ε, and thus picking one of these a's we may have to suffer from the so-defined gap_ε in the worst case.

C.2. Bandit-RL games

The setting is exactly the same as in Section 4, except the definition of the ψ-functions takes the max over best-response sets:

$$\psi_\varepsilon(a) := \max_{\pi^b \in \mathsf{BR}_\varepsilon(a)} V_1(a, \pi^b),$$

for all $\varepsilon \geq 0$. Similar as in Section C.1, we consider the following new definition of gap:

$$\begin{aligned} \widetilde{\mathsf{gap}}_\varepsilon &:= \max_{a\in\mathcal{A}_\varepsilon} [\psi_\varepsilon(a) - \psi_0(a)], \quad \text{where} \\ \mathcal{A}_\varepsilon &:= \Big\{ a \in \mathcal{A} : \psi_\varepsilon(a) \geq \max_{a\in\mathcal{A}} \psi_0(a) - \varepsilon \Big\}. \end{aligned}$$

We now state our sample complexity upper bound for learning Stackelberg in bandit-RL games under optimistic tie-breaking. The proof is analogous to Theorem 4, and can be found in Section H.2.

Theorem C.2 (Learning Stackelberg in bandit-RL games with optimistic tie-breaking)**.** *For any bandit-RL game and sufficiently small $\varepsilon \leq O(1/H^2S^2)$, Algorithm 2 with $n = \widetilde{O}(H^5S^2AB/\varepsilon^2 + H^7S^4AB/\varepsilon)$ episodes of play can return $(\widehat{a}, \widehat{\pi}^b)$ such that with probability at least $1-\delta$,*

$$\begin{aligned} \psi_0(\widehat{a}) &\geq \max_{a\in\mathcal{A}} \psi_0(a) - \widetilde{\mathsf{gap}}_\varepsilon - \varepsilon, \\ V_2(\widehat{a}, \widehat{\pi}^b) &\geq \max_{\widetilde{\pi}^b} V_2(\widehat{a}, \widetilde{\pi}^b) - \varepsilon, \end{aligned}$$

where $\widetilde{O}(\cdot)$ hides $\log(HSAB/\delta\varepsilon)$ factors. Further, the algorithm runs in $\mathrm{poly}(HSAB/\delta\varepsilon)$ time.

D. Matrix game with simultaneous play

In this section, we consider a variant of the two-player bandit game, in which the leader and follower instead play a matrix game simultaneously, and the follower cannot see the leader's action. The problem of finding Stackelberg in this setting is also known as learning the "optimal strategy to commit to" (Conitzer & Sandholm, 2006).

Setting A general-sum matrix game with simultaneous play can be described as $M = (\mathcal{A}, \mathcal{B}, r_1, r_2)$ with $|\mathcal{A}| = A$, $|\mathcal{B}| = B$, and $r_1, r_2 : \mathcal{A} \times \mathcal{B} \to [0, 1]$, which defines the following game:

- The leader pre-specifies a policy $\pi^a \in \Delta_{\mathcal{A}}$ and reveals this policy to the follower.
- The leader plays $a \sim \pi^a$, and the follower plays an action $b \in \mathcal{B}$ simultaneously without seeing a.
- The leader receives reward $r_1(a, b)$ and the follows receives reward $r_2(a, b)$.

(Above, $\Delta_{\mathcal{A}}$ denotes the probability simplex on $\mathcal{A}$.) Let $\mu_i(\pi^a, b) = \sum_{a' \in \mathcal{A}} \pi^a(a')\mathbb{E}[r_i(a', b)]$, $i = 1, 2$ denote the mean rewards (for mixed policies), and

$$\phi_\varepsilon(\pi^a) := \min_{b \in \mathsf{BR}_\varepsilon(\pi^a)} \mu_1(\pi^a, b),$$
$$\mathsf{BR}_\varepsilon(\pi^a) := \left\{ b \in \mathcal{B} : \mu_2(\pi^a, b) \geq \max_{b'} \mu_2(\pi^a, b') - \varepsilon \right\}$$

denote the ε-approximate best response sets and best response values for any $\varepsilon \geq 0$, similar as in bandit games. We also overload notation to let $\phi_\varepsilon(a_1) := \phi_\varepsilon(\delta_{a_1})$ to denote the ϕ_ε value at pure strategies (δ_a is the pure strategy of always taking a).

The main difference between this setting and bandit games is that now the Stackelberg equilibrium for the leader may be achieved at *mixed strategies* only, so that we can no longer restrict attention to pure strategies $a \in \mathcal{A}$ for the leader. To see why this is true, consider 2x2 game of (Conitzer & Sandholm, 2006) shown in Table 1. In this game, the two pure strategies $\{a_1, a_2\}$ achieve $\phi_0(a_1) = 2$ (since the best response is b_1) and $\phi_0(a_2) = 3$ (since the best response is b_2). However, if we take $\pi^a_p = p\delta_{a_1} + (1 - p)\delta_{a_2}$, then the follower's best response is b_2 whenever $p < 1/2$. Taking $p \to (1/2)_-$, the leader can achieve value $\phi_0(\pi^a_p) = 4p + 3(1 - p) \to 3.5$, which is higher than both pure strategies. For $p \geq 1/2$, the follower's best response is b_1, and $\phi_0(\pi^a_p) \leq 2$. Therefore the Stackelberg equilibrium for the leader is to take π^a_p with $p \to (1/2)_-$[3].

μ_1, μ_2	b_1	b_2
a_1	2, 1	4, 0
a_2	1, 0	3, 1

Table 1. Example of matrix game with simultaneous play, where the Stackelberg strategy for the leader is mixed.

D.1. Main result

Let $\mathsf{gap}_\varepsilon := \sup_{\pi^a \in \Delta_{\mathcal{A}}} \phi_0(\pi^a) - \sup_{\pi^a \in \Delta_{\mathcal{A}}} \phi_\varepsilon(\pi^a)$ denote the gap. The following result shows that $\widetilde{O}(AB/\varepsilon^2)$ samples suffice for learning the Stackelberg up to $(\mathsf{gap}_\varepsilon + \varepsilon)$ in simultaneous matrix games, similar as in bandit games. The proof can be found in Appendix I.1.

Theorem D.1 (Learning Stackelberg in simultaneous matrix games)**.** *For any matrix game with simultaneous play, Algorithm 3 queries for* $n = O(AB \log(AB/\delta)/\varepsilon^2) = \widetilde{O}(AB/\varepsilon^2)$ *samples, and outputs* $(\widehat{\pi}^a, \widehat{b})$ *such that with probability at least* $1 - \delta$*,*

$$\phi_0(\widehat{\pi}^a) \geq \phi_{\varepsilon/2}(\widehat{\pi}^a) \geq \sup_{\pi^a \in \Delta_{\mathcal{A}}} \phi_0(\pi^a) - \mathsf{gap}_\varepsilon - \varepsilon,$$
$$\mu_2(\widehat{\pi}^a, \widehat{b}) \geq \max_{b' \in \mathcal{B}} \mu_2(\widehat{\pi}^a, b') - \varepsilon.$$

Theorem D.1 implies that $\widetilde{O}(AB/\varepsilon^2)$ samples is also enough for determining the approximate (up to gap) Stackelberg equilibrium in simultaneous games. Also, as we assumed bandit feedback, Theorem D.1 extends the results of Letchford et al. (2009); Peng et al. (2019) which studied the sample complexity assuming a best response oracle (can query $\mathsf{BR}_0(\pi^a)$ for any $\pi^a \in \Delta_{\mathcal{A}}$).

[3]The reason why the optimal policy can only be approached instead of exactly achieved is because of the pessimistic tie-breaking at $p = 1/2$, and is resolved if we take optimistic tie-breaking.

Algorithm 3 Learning Stackelberg in matrix games with simultaneous play

Require: Target accuracy $\varepsilon > 0$.

set $N \leftarrow C\log(AB/\delta)/\varepsilon^2$ for some constant $C > 0$.

1: Query each $(a,b) \in \mathcal{A}\times\mathcal{B}$ for N times and obtain $\{r_1^{(j)}(a,b), r_2^{(j)}(a,b)\}_{j=1}^N$.

2: Construct empirical estimates $\widehat{\mu}_i(\pi^a, b) = \sum_{a'\in\mathcal{A}} \pi^a(a')\frac{1}{N}\sum_{j=1}^N r_i^{(j)}(a', b)$ for $i = 1, 2$.

3: Construct approximate best response sets and values for all $\pi^a \in \Delta_{\mathcal{A}}$:

$$\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a) := \left\{ b : \widehat{\mu}_2(\pi^a, b) \ge \max_{b'\in\mathcal{B}} \widehat{\mu}_2(\pi^a, b') - 3\varepsilon/4 \right\},$$
$$\widehat{\phi}_{3\varepsilon/4}(\pi^a) := \min_{b\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a)} \widehat{\mu}_1(\pi^a, b).$$

4: Output $(\widehat{\pi}^a, \widehat{b})$ such that

$$\widehat{\phi}_{3\varepsilon/4}(\widehat{\pi}^a) \ge \sup_{\pi^a\in\Delta_{\mathcal{A}}} \widehat{\phi}_{3\varepsilon/4}(\pi^a) - \varepsilon/2, \tag{9}$$
$$\widehat{b} = \arg\min_{b\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{\pi}^a)} \widehat{\mu}_1(\widehat{\pi}^a, b).$$

Comparison between learning Stackelberg and Nash We compare Theorem D.1 with existing results on learning Nash equilibria in general-sum matrix games. On the one hand, when we have $\widetilde{O}(AB/\varepsilon^2)$ samples, with only a $(\mathsf{gap}_\varepsilon + \varepsilon)$ near-optimal Stackelberg equilibrium, but we can learn an ε-approximate Nash equilibrium (Liu et al., 2020). On the other hand, the Stackelberg value is uniquely defined (as the max of ϕ_0), whereas there can be multiple Nash values (Roughgarden, 2010) for general-sum games. Additionally, at the Stackelberg equilibrium, the leader's payoff is guaranteed to be at least as good as any Nash value (the leader can pre-specify any Nash policy). This makes Stackelberg a perhaps better solution concept in asymmetric games where the learning goal focuses more on the leader.

Runtime In Algorithm 3, the step of approximately maximizing $\widehat{\phi}_{3\varepsilon/4}(\pi^a)$ in (9) requires optimizing a discontinuous function over a continuous domain. It is unclear whether this program can be reformulated to be solved efficiently in polynomial time[4][5]. However, we remark that this is special to the pessimistic tie-breaking we assumed (Letchford et al., 2009). Learning the Stackelberg equilibrium with optimistic tie-breaking has the same $\widetilde{O}(AB/\varepsilon^2)$ sample complexity while admitting an efficient polynomial-time algorithm via linear programming (see Section D.2 for the formal statement and proof).

D.2. Optimistic tie-breaking

We also study simultaneous matrix games with optimistic tie-breaking. The setting is exactly the same as above except the definition of the ψ-functions takes the max over best-response sets:

$$\psi_\varepsilon(\pi^a) := \max_{b\in\mathsf{BR}_\varepsilon(\pi^a)} \mu_1(\pi^a, b),$$

for all $\varepsilon \ge 0$. Similar as in bandit games (Section C.1), we consider the following new definition of gap:

$$\widetilde{\mathsf{gap}}_\varepsilon := \max_{\pi^a\in\mathcal{A}_\varepsilon} [\psi_\varepsilon(\pi^a) - \psi_0(\pi^a)], \quad \text{where}$$
$$\mathcal{A}_\varepsilon := \left\{ \pi^a \in \Delta_{\mathcal{A}} : \psi_\varepsilon(\pi^a) \ge \max_{\pi^a\in\Delta_{\mathcal{A}}} \psi_0(\pi^a) - \varepsilon \right\}.$$

[4]This program has a finite-time solution by the following strategy (which utilizes the specific structure of this program): First partition $\Delta_{\mathcal{A}}$ according to which subsets of $\mathcal{B}$ are $3\varepsilon/4$ best response, and then within each partition solve an linear program (over π^a) to a fixed accuracy (e.g. $\varepsilon/10$). However, the runtime is exponential because there are 2^B subsets induced by the partition.

[5]We also remark that (Von Stengel & Zamir, 2010, Theorem 9 & Proposition 10) provides an efficient reformulation of the pessimistic Stackelberg problem in simultaneous matrix games. However, their reformulation relies crucially on the best response set being *exact*, and does not generalize to our setting which requires to solve the pessimitic Stackelberg problem with *approximate* best response sets.

Algorithm 4 Learning Stackelberg in matrix games with simultaneous play (optimistic tie-breaking version)

Require: Target accuracy $\varepsilon > 0$.
set $N \leftarrow C\log(AB/\delta)/\varepsilon^2$ for some constant $C > 0$.
1: Query each $(a, b) \in \mathcal{A} \times \mathcal{B}$ for N times and obtain $\{r_1^{(j)}(a, b), r_2^{(j)}(a, b)\}_{j=1}^N$.
2: Construct empirical estimates $\widehat{\mu}_i(\pi^a, b) = \sum_{a' \in \mathcal{A}} \pi^a(a') \frac{1}{N} \sum_{j=1}^N r_i^{(j)}(a', b)$ for $i = 1, 2$.
3: Construct approximate best response sets and values for all $\pi^a \in \Delta_{\mathcal{A}}$:

$$\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a) := \left\{ b : \widehat{\mu}_2(\pi^a, b) \geq \max_{b' \in \mathcal{B}} \widehat{\mu}_2(\pi^a, b') - 3\varepsilon/4 \right\},$$
$$\widehat{\phi}_{3\varepsilon/4}(\pi^a) := \max_{b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a)} \widehat{\mu}_1(\pi^a, b).$$

4: Output

$$\widehat{\pi}^a = \arg\max_{\pi^a \in \Delta_{\mathcal{A}}} \widehat{\phi}_{3\varepsilon/4}(\pi^a), \qquad (10)$$
$$\widehat{b} = \arg\max_{b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{\pi}^a)} \widehat{\mu}_1(\widehat{\pi}^a, b).$$

By calling the subroutine $(\widehat{\pi}^a, \widehat{b}) \leftarrow$ `BestMixedLeaderStrategy`$(\widehat{\mu}_1, \widehat{\mu}_2)$.

Theorem D.2 (Learning Stackelberg in simultaneous matrix games with optimistic tie-breaking)**.** *For any matrix game with simultaneous play, Algorithm 4 queries for $n = O(AB\log(AB/\delta)/\varepsilon^2) = \widetilde{O}(AB/\varepsilon^2)$ samples, and outputs $(\widehat{\pi}^a, \widehat{b})$ such that with probability at least $1 - \delta$,*

$$\psi_0(\widehat{\pi}^a) \geq \max_{\pi^a \in \Delta_{\mathcal{A}}} \psi_0(\pi^a) - \widetilde{\mathsf{gap}}_\varepsilon - \varepsilon,$$
$$\mu_2(\widehat{\pi}^a, \widehat{b}) \geq \max_{b' \in \mathcal{B}} \mu_2(\widehat{\pi}^a, b') - \varepsilon.$$

Further, the algorithm runs in $\mathrm{poly}(n)$ *time.*

The proof can be found in Section I.2.

Efficient runtime Theorem D.2 shares the same sample complexity $\widetilde{O}(AB/\varepsilon^2)$ as its pessimistic tie-breaking counterpart (Theorem D.1), albeit with a slightly different definition of the gap. However, an additional advantage of the optimistic version is that it is guaranteed to have a polynomial runtime. The core reason behind this is that with optimistic tie-breaking now $(\widehat{\pi}^a, \widehat{b})$ solves a *max-max problem* (instead of a max-min problem), for which we can exchange the order of maximization. Concretely, we can now first maximize over π^a for each b, which admits a linear programming formulation (cf. the `BestMixedLeaderStrategy` subroutine in Algorithm 10, also in (Conitzer & Sandholm, 2006)).

E. Proofs for Section 3

E.1. Proof of Theorem 1

To prove Theorem 1, we will construct a pair of hard instances, and use Le Cam's method (Wainwright, 2019, Section 15.2) to reduce the estimation error into a testing problem between the two hard instances. Consider the following two games M_1 and M_{-1}, where the rewards follow Bernoulli distributions: $r_i(a, b) \sim \mathsf{Ber}(\mu_i(a, b))$ with means shown in Table 2, where $\delta \in (0, 1)$ is a parameter to be determined:

Based on Table 2, it is straightforward to check that $\phi_0^{M_1}(a_1) = 1$, $\phi_0^{M_{-1}}(a_1) = 0$, and $\phi_0^{M_1}(a_2) = \phi_0^{M_{-1}}(a_2) = 1/2$. Further, $a_\star^{M_1} = a_1$ and $a_\star^{M_{-1}} = a_2$.

For any algorithm that outputs a (possibly randomized) estimator $\widehat{a} \in \mathcal{A}$ of the Stackelberg equilibrium, let π denotes its querying policy, that is, given prior queries and observations $\left\{a^{(i)}, b^{(i)}, r_1^{(i)}, r_2^{(i)}\right\}_{i=1}^{k-1}$, $\pi^{(k)}(a, b|\left\{a^{(i)}, b^{(i)}, r_1^{(i)}, r_2^{(i)}\right\}_{i=1}^{k-1})$

r_1, r_2	b_1	b_2
a_1	$1, \frac{1+\delta}{2}$	$0, \frac{1-\delta}{2}$
a_2	$\frac{1}{2}, 1$	$\frac{1}{2}, 1$

r_1, r_2	b_1	b_2
a_1	$1, \frac{1-\delta}{2}$	$0, \frac{1+\delta}{2}$
a_2	$\frac{1}{2}, 1$	$\frac{1}{2}, 1$

Table 2. Pair of hard instances M_1 (left) and M_{-1} (right). Each table lists $\mu_1(a,b), \mu_2(a,b)$ for $a \in \{a_1, a_2\}$, $b \in \{b_1, b_2\}$.

denotes the distribution of the next query. Let $\mathbb{P}_{M_1,\pi}$ and $\mathbb{P}_{M_{-1},\pi}$ denote the distribution of all n observations generated by the querying policy π. For these two instances, we have

$$\begin{aligned}
&\sup_{M\in\{M_1,M_{-1}\}} \mathbb{P}_M\left(\max_{a\in\mathcal{A}} \phi_0^M(a) - \phi_0^M(\widehat{a}) \ge \frac{1}{2}\right) \\
&= \sup_{M\in\{M_1,M_{-1}\}} \mathbb{P}_M\left(\widehat{a} \neq \arg\max_{a\in\mathcal{A}} \phi_0^M(a)\right) \\
&\ge \frac{1}{2}\big(\mathbb{P}_{M_1}(\widehat{a}\neq a_1) + \mathbb{P}_{M_{-1}}(\widehat{a}\neq a_2)\big) \\
&\ge \frac{1}{2}\big(1 - \mathrm{TV}\big(\mathbb{P}_{M_1,\pi}, \mathbb{P}_{M_{-1},\pi}\big)\big) \\
&\ge \frac{1}{2}\left(1 - \sqrt{\frac{1}{2}\mathrm{KL}\big(\mathbb{P}_{M_1,\pi} \| \mathbb{P}_{M_{-1},\pi}\big)}\right),
\end{aligned}$$

where the second-to-last step used Le Cam's inequality, and the last step used Pinsker's inequality. To upper bound the KL distance between $\mathbb{P}_{M_1,\pi}$ and $\mathbb{P}_{M_{-1},\pi}$, we apply the divergence decomposition of (Lattimore & Szepesvári, 2020, Lemma 15.1) and obtain that

$$\mathrm{KL}(\mathbb{P}_{M_1,\pi}\|\mathbb{P}_{M_{-1},\pi}) \le \sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} \mathbb{E}_{M_1,\pi}[T_{a,b}(n)] \cdot \mathrm{KL}\Big(\mathbb{P}^{a,b}_{M_1} \| \mathbb{P}^{a,b}_{M_{-1}}\Big) \le n \cdot \max_{(a,b)\in\mathcal{A}\times\mathcal{B}} \mathrm{KL}\Big(\mathbb{P}^{a,b}_{M_1} \| \mathbb{P}^{a,b}_{M_{-1}}\Big),$$

where $T_{a,b}(n)$ denotes the number of queries to (a,b) among the n queries, and $\mathbb{P}^{a,b}_{M_i}$ denote the distribution of the observation $(r_1(a,b), r_2(a,b))$ in problem M_i, $i = 1, 2$. We have $\mathrm{KL}(\mathbb{P}^{a,b}_{M_1}\|\mathbb{P}^{a,b}_{M_{-1}}) = 0$ for $(a,b) = (a_2, b_1)$ and $(a,b) = (a_2, b_2)$ since these (a,b) yield exactly the same reward distributions. For $(a,b) = (a_1, b_1)$ and $(a,b) = (a_1, b_2)$, using the bound $\mathrm{KL}(\mathsf{Ber}(\frac{1+\delta}{2})\|\mathsf{Ber}(\frac{1-\delta}{2})) = \delta\log\frac{1+\delta}{1-\delta} \le 3\delta^2$ for $\delta \le 1/2$ (and the same bound for $\mathrm{KL}(\mathsf{Ber}(\frac{1-\delta}{2})\|\mathsf{Ber}(\frac{1+\delta}{2}))$). Therefore, we get

$$\mathrm{KL}\big(\mathbb{P}_{M_1,\pi}\|\mathbb{P}_{M_{-1},\pi}\big) \le 3n\delta^2.$$

Choosing $\delta = 1/\sqrt{(27/2)n}$, the above is upper bounded by $2/9$, and thus plugging back to the preceding bound yields

$$\sup_{M\in\{M_1,M_{-1}\}} \mathbb{P}\left(\widehat{a} \neq \arg\max_{a\in\mathcal{A}} \phi_0^M(a)\right) \ge \frac{1}{2}\left(1 - \sqrt{\frac{1}{2}\mathrm{KL}\big(\mathbb{P}_{M_1,\pi}\|\mathbb{P}_{M_{-1},\pi}\big)}\right) \ge \frac{1}{3}.$$

Therefore, choosing the problem class to be $\mathcal{M}_n = \{M_1, M_{-1}\}$ with $\delta = 1/\sqrt{(27/2)n}$, the above is the desired lower bound. □

E.2. A Lemma on the gap

Lemma E.1 (Gap can be $\Omega(1)$)**.** *For any $0 \le \varepsilon_1 < \varepsilon_2 < 1$, there exists a two-player bandit game $M = M_{\varepsilon_1,\varepsilon_2}$ with $A = B = 2$, such that*

$$\begin{aligned}
&\max_{a\in\mathcal{A}} \phi_{\varepsilon_1}(a) - \max_{a\in\mathcal{A}} \phi_{\varepsilon_2}(a) \ge \frac{1}{2}, \\
&\max_{a\in\mathcal{A}} \phi_{\varepsilon_1}(a) - \phi_{\varepsilon_1}\left(\arg\max_{a'\in\mathcal{A}} \phi_{\varepsilon_2}(a')\right) \ge \frac{1}{2}.
\end{aligned}$$

In particular, (taking $\varepsilon_1 = 0$), for any ε there exists a game in which $\mathsf{gap}_\varepsilon = \max_{a\in\mathcal{A}} \phi_0(a) - \max_{a\in\mathcal{A}} \phi_\varepsilon(a) \ge 1/2$.

Algorithm 5 Learning Stackelberg in bandit games

Require: Target accuracy $\varepsilon > 0$.

set $N \leftarrow C\log(AB/\delta)/\varepsilon^2$ for some constant $C > 0$.

1: Query each $(a, b) \in \mathcal{A} \times \mathcal{B}$ for N times and obtain $\{r_1^{(j)}(a, b), r_2^{(j)}(a, b)\}_{j=1}^N$.

2: Construct empirical estimates of the means $\widehat{\mu}_i(a, b) = \frac{1}{N}\sum_{j=1}^N r_i^{(j)}(a, b)$ for $i = 1, 2$.

3: Construct approximate best response sets and values for all $a \in \mathcal{A}$:

$$\widehat{\phi}_{3\varepsilon/4}(a) := \min_{b\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(a)} \widehat{\mu}_1(a, b), \quad \text{where} \quad \widehat{\mathsf{BR}}_{3\varepsilon/4}(a) := \left\{b : \widehat{\mu}_2(a, b) \geq \max_{b'\in\mathcal{B}} \widehat{\mu}_2(a, b') - 3\varepsilon/4\right\}.$$

4: Output $(\widehat{a}, \widehat{b})$, where $\widehat{a} = \arg\max_{a\in\mathcal{A}} \widehat{\phi}_{3\varepsilon/4}(a)$, $\widehat{b} = \arg\min_{b\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \widehat{\mu}_1(\widehat{a}, b)$.

Proof. Let $0 \leq \varepsilon_1 < \varepsilon_2$. We construct the problem $M = M_{\varepsilon_1,\varepsilon_2}$ as follows: $\mathcal{A} = \{a_1, a_2\}$ and $\mathcal{B} = \{b_1, b_2\}$, and the rewards $\{r_1(a, b), r_2(a, b)\}_{a\in\mathcal{A},b\in\mathcal{B}}$ are all deterministic and valued as in the following table:

r_1, r_2	b_1	b_2
a_1	$1, \frac{\varepsilon_1+\varepsilon_2}{2}$	$0, 0$
a_2	$\frac{1}{2}, 1$	$\frac{1}{2}, 1$

Table 3. Construction of $r_1(a, b), r_2(a, b)$ for $a \in \{a_1, a_2\}$, $b \in \{b_1, b_2\}$.

For the arm a_2, actions b_1 and b_2 are exactly the same, so we have $\phi_\varepsilon(a_2) = \frac{1}{2}$ for all ε. For the arm a_1, observe that $\varepsilon_1 < \frac{\varepsilon_1+\varepsilon_2}{2} < \varepsilon_2$, and thus $\mathsf{BR}_{\varepsilon_1}(a_1) = \{b_1\}$ and $\phi_{\varepsilon_1}(a_1) = 1$, but $\mathsf{BR}_{\varepsilon_2}(a_1) = \{b_1, b_2\}$ and $\phi_{\varepsilon_2}(a_1) = 0$. Therefore,

$$\max_{a\in\mathcal{A}} \phi_{\varepsilon_1}(a) = \max\left\{1, \frac{1}{2}\right\} = 1,$$
$$\max_{a\in\mathcal{A}} \phi_{\varepsilon_2}(a) = \max\left\{0, \frac{1}{2}\right\} = \frac{1}{2},$$
$$\phi_{\varepsilon_1}\left(\arg\max_{a'\in\mathcal{A}} \phi_{\varepsilon_2}(a')\right) = \phi_{\varepsilon_1}(a_2) = \frac{1}{2}.$$

This shows the desired result. $\square$

E.3. Proof of Theorem 2

Algorithm 5 pulled each arm (a, b) for $N = O(\log(AB/\delta)/\varepsilon^2)$ times, and $\widehat{\mu}_1(a, b)$, $\widehat{\mu}_2(a, b)$ are the empirical means of the observed rewards. By the Hoeffding inequality and union bound over (a, b), with probability at least $1 - \delta$, we have

$$\max_{(a,b)\in\mathcal{A}\times\mathcal{B}} |\widehat{\mu}_i(a, b) - \mu_i(a, b)| \leq \varepsilon/8 \quad \text{for } i = 1, 2. \tag{11}$$

Properties of $\widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$ On the uniform convergence event (21), we have the following: for any $b \in \mathsf{BR}_{\varepsilon/2}(a)$, we have

$$\widehat{\mu}_2(a, b) \geq \mu_2(a, b) - \varepsilon/8 \geq \max_{b'\in\mathcal{B}} \mu_2(a, b') - 5\varepsilon/8 \geq \max_{b'\in\mathcal{B}} \widehat{\mu}_2(a, b') - 3\varepsilon/4,$$

and thus $b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$. This shows that $\mathsf{BR}_{\varepsilon/2}(a) \subseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$. Similarly we can show that $\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \subseteq \mathsf{BR}_\varepsilon(a)$. In other words,

$$\mathsf{BR}_\varepsilon(a) \supseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \supseteq \mathsf{BR}_{\varepsilon/2}(a) \quad \text{for all } a \in \mathcal{A}.$$

Notably, this implies that $\widehat{b} \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a}) \in \mathsf{BR}_\varepsilon(\widehat{a})$, the desired near-optimality guarantee for $\widehat{b}$.

Near-optimality of $\widehat{a}$ On the one hand, because $\widehat{a}$ maximizes $\widehat{\mu}_1(a, \widehat{b}(a))$, we have for any $a \in \mathcal{A}$ that

$$\min_{b' \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \widehat{\mu}_1(\widehat{a}, b') \stackrel{(i)}{=} \widehat{\mu}_1(\widehat{a}, \widehat{b}) \geq \widehat{\mu}_1(a, \widehat{b}(a)) \stackrel{(ii)}{\geq} \min_{b \in \mathsf{BR}_\varepsilon(a)} \widehat{\mu}_1(a, b),$$

where (i) is because $\widehat{b}$ minimizes $\widehat{\mu}_1(\widehat{a}, \cdot)$ within $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})$, and (ii) is because $\widehat{b}(a) \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \subseteq \mathsf{BR}_\varepsilon(a)$. By the uniform convergence (21), we get that

$$\min_{b' \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \mu_1(\widehat{a}, b') \geq \min_{b \in \mathsf{BR}_\varepsilon(a)} \mu_1(a, b) - 2 \cdot \varepsilon/8 \geq \phi_\varepsilon(a) - \varepsilon.$$

Since the above holds for all $a \in \mathcal{A}$, taking the max on the right hand side gives

$$\max_{a \in \mathcal{A}} \phi_\varepsilon(a) - \varepsilon \leq \min_{b' \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \mu_1(\widehat{a}, b') \stackrel{(i)}{\leq} \min_{b' \in \mathsf{BR}_{\varepsilon/2}(\widehat{a})} \mu_1(\widehat{a}, b') = \phi_{\varepsilon/2}(\widehat{a}),$$

where (i) is because $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a}) \supseteq \mathsf{BR}_{\varepsilon/2}(a)$. This yields that

$$\phi_{\varepsilon/2}(\widehat{a}) \geq \max_{a \in \mathcal{A}} \phi_\varepsilon(a) - \varepsilon = \max_{a \in \mathcal{A}} \phi_0(a) - \mathsf{gap}_\varepsilon - \varepsilon,$$

which is the first part of the bound for $\widehat{a}$.

On the other hand, since $\phi_\varepsilon(a)$ is increasing as we decrease ε, we directly have

$$\phi_{\varepsilon/2}(\widehat{a}) \leq \phi_0(\widehat{a}) \leq \max_{a \in \mathcal{A}} \phi_0(a).$$

This is the second part of the bound for $\widehat{a}$. □

E.4. Proof of Theorem 3

Suppose $\varepsilon \in (0, c)$ and $g \in [0, c)$ where $c > 0$ is an absolute constant to be determined. For any algorithm that outputs an estimator $\widehat{a} \in \mathcal{A}$, let π denote its (sequential) querying policy, and $\mathbb{P}_{M,\pi}$ denote the joint distribution of the N observed rewards $(r_1(a^{(i)}, b^{(i)}), r_2(a^{(i)}, b^{(i)}))_{i=1}^N$ under game M. We will rely on the divergence decomposition of (Lattimore & Szepesvári, 2020, Lemma 15.1):

$$\mathrm{KL}(\mathbb{P}_{M,\pi} \| \mathbb{P}_{M',\pi}) \leq \sum_{(a,b) \in \mathcal{A} \times \mathcal{B}} \mathbb{E}_{M_1,\pi}[T_{a,b}(N)] \cdot \mathrm{KL}\Big(\mathbb{P}_M^{a,b} \| \mathbb{P}_{M'}^{a,b}\Big), \tag{12}$$

where $P_M^{a,b}$ denotes the distribution of observation $(r_1(a, b), r_2(a, b))$ under game M, and $T_{a,b}(N)$ denotes the number of queries of (a, b) using algorithm π (which is a random variable). We will also use the fact that

$$\mathrm{KL}(\mathsf{Ber}(1/2) \| \mathsf{Ber}(1/2 + \delta)) = \frac{1}{2} \log \frac{1/2}{1/2 + \delta} + \frac{1}{2} \log \frac{1/2}{1/2 - \delta} = \frac{1}{2} \log \frac{1}{1 - 4\delta^2} \leq \frac{1}{2} \cdot 8\delta^2 \leq 4\delta^2 \tag{13}$$

whenever $4\delta^2 \leq 1/2$, i.e. $|\delta| \leq 1/2\sqrt{2}$.

Construction of hard instance In our construction below, the rewards follow Bernoulli distributions: $r_i(a, b) \sim \mathsf{Ber}(\mu_i(a, b))$, so that it suffices to specify $\mu_i(a, b)$. Without loss of generality assume $B/3$ is an integer, and let $\mathcal{B} = [B] = \{1, \dots, B\}$ for notational simplicity.

We define a family of games $M_{a_\star, b_\star^1, b_\star^2}$ indexed by $a_\star \in \mathcal{A}$ and $b_\star^1, b_\star^2 \in \mathcal{B}$. Each game $M_{a_\star, b_\star^1, b_\star^2}$ is defined as follows:

- $\mu_1(a, b) = 1/2 + g + \varepsilon$ for all $a \in \mathcal{A}$ and $1 \leq b \leq B/3$;
- $\mu_1(a, b) = 1/2 + \varepsilon$ for all $a \in \mathcal{A}$ and $B/3 + 1 \leq b \leq 2B/3$;
- $\mu_1(a, b) = 1/2$ for all $a \in \mathcal{A}$ and $2B/3 + 1 \leq b \leq B$.

- $\mu_2(a_\star, b_\star^1) = 1/2 + 2\varepsilon$, where $b_\star^1 \in \{1, \dots, B/3\}$.
- $\mu_2(a_\star, b_\star^2) = 1/2 + 5\varepsilon/4$, where $b_\star^2 \in \{B/3+1, \dots, 2B/3\}$.
- $\mu_2(a_\star, b') = 1/2$ for all $b' \neq b_\star^1, b_\star^2$.
- $\mu_2(a', b) = 1/2$ for all $a' \neq a_\star$ and $b \in \mathcal{B}$.

For this game, we have $\phi_0(a_\star) = 1/2 + g + \varepsilon$, $\phi_\varepsilon(a_\star) = 1/2 + \varepsilon$, and $\phi_0(a') = \phi_\varepsilon(a') = 1/2$ for all $a' \neq a_\star$. Therefore,

$$\mathsf{gap}_\varepsilon = \max_{a \in \mathcal{A}} \phi_0(a) - \max_{a \in \mathcal{A}} \phi_\varepsilon(a) = g.$$

Further, notice that as long as $\widehat{a} \neq a_\star$, we have $\phi_0(\widehat{a}) = \max_a \phi_0(a) - (g + \varepsilon)$.

Define $\mathbb{P}_M$ as the mixture of over the prior of $M_{a_\star, b_\star^1, b_\star^2}$ where the prior samples $a_\star \sim \mathrm{Unif}(\mathcal{A})$, $b_\star^1 \sim \mathrm{Unif}(\{1, \dots, B/3\})$, and $b_\star^2 \sim \mathrm{Unif}(\{B/3+1, \dots, 2B/3\})$. Define M_0 as the "null-game" where all the r_2 are $1/2$, and r_1 has the same configuration as in the above game.

Proof of lower bound Under the mixture $\mathbb{P}_M$, we have

$$\begin{aligned}
&\mathbb{P}_M\Big(\phi_0(\widehat{a}) \le \max_{a \in \mathcal{A}} \phi_0(a) - (g+\varepsilon)\Big) = \mathbb{P}_M(\widehat{a} \neq a_\star) \\
&= \frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1}^{2B/3} \mathbb{P}_{M_{a_\star, b_\star^1, b_\star^2}}(\widehat{a} \neq a_\star) \\
&\ge \frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1}^{2B/3} \mathbb{P}_0(\widehat{a} \neq a_\star) - \frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1}^{2B/3} \mathrm{TV}\Big(\mathbb{P}_{0,\pi}, \mathbb{P}_{M_{a_\star, b_\star^1, b_\star^2}, \pi}\Big) \\
&\ge \frac{1}{A} \sum_{a_\star} \mathbb{P}_0(\widehat{a} \neq a_\star) - \frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1}^{2B/3} \sqrt{\frac{1}{2}\mathrm{KL}\Big(\mathbb{P}_{0,\pi} \| \mathbb{P}_{M_{a_\star, b_\star^1, b_\star^2}, \pi}\Big)} \\
&\ge 1 - \frac{1}{A} - \underbrace{\frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1}^{2B/3} \sqrt{\frac{1}{2}\mathrm{KL}\Big(\mathbb{P}_{0,\pi} \| \mathbb{P}_{M_{a_\star, b_\star^1, b_\star^2}, \pi}\Big)}}_{(\star)}.
\end{aligned}$$

We now show that $(\star) \le 1/3$ if $N \le c[AB/\varepsilon^2]$ for some small absolute constant $c > 0$. Using the divergence decomposition (12), we have

$$\begin{aligned}
(\star) &\le \frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1}^{2B/3} \sqrt{\frac{1}{2} \sum_{a,b} \mathbb{E}_{0,\pi}[T_{a,b}(N)] \mathrm{KL}\Big(\mathbb{P}_0^{a,b} \| \mathbb{P}^{a,b}_{M_{a_\star, b_\star^1, b_\star^2}}\Big)} \\
&\overset{(i)}{=} \frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1}^{2B/3} \Bigg(\frac{1}{2}\mathbb{E}_{0,\pi}\big[T_{a_\star, b_\star^1}(N)\big] \mathrm{KL}\Big(\mathbb{P}_0^{a_\star, b_\star^1} \| \mathbb{P}^{a_\star, b_\star^1}_{M_{a_\star, b_\star^1, b_\star^2}}\Big) + \\
&\qquad \frac{1}{2}\mathbb{E}_{0,\pi}\big[T_{a_\star, b_\star^2}(N)\big] \mathrm{KL}\Big(\mathbb{P}_0^{a_\star, b_\star^2} \| \mathbb{P}^{a_\star, b_\star^2}_{M_{a_\star, b_\star^1, b_\star^2}}\Big)\Bigg)^{1/2} \\
&\le \frac{1}{A(B^2/9)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sum_{b_\star^2=B/3+1} \sqrt{\frac{1}{2}\mathbb{E}_{0,\pi}\big[T_{a_\star, b_\star^1}(N)\big] \mathrm{KL}\Big(\mathbb{P}_0^{a_\star, b_\star^1} \| \mathbb{P}^{a_\star, b_\star^1}_{M_{a_\star, b_\star^1, b_\star^2}}\Big)} \\
&\qquad + \sqrt{\frac{1}{2}\mathbb{E}_{0,\pi}\big[T_{a_\star, b_\star^2}(N)\big] \mathrm{KL}\Big(\mathbb{P}_0^{a_\star, b_\star^2} \| \mathbb{P}^{a_\star, b_\star^2}_{M_{a_\star, b_\star^1, b_\star^2}}\Big)} \\
&\overset{(ii)}{\le} \frac{1}{A(B/3)} \sum_{a_\star} \sum_{b_\star^1=1}^{B/3} \sqrt{\frac{1}{2}\mathbb{E}_{0,\pi}\big[T_{a_\star, b_\star^1}(N)\big] \cdot 4 \cdot (2\varepsilon)^2} + \frac{1}{A(B/3)} \sum_{a_\star} \sum_{b_\star^2=1}^{B/3} \sqrt{\frac{1}{2}\mathbb{E}_{0,\pi}\big[T_{a_\star, b_\star^2}(N)\big] \cdot 4 \cdot (5\varepsilon/4)^2}
\end{aligned}$$

$$\stackrel{(iii)}{\le} \sqrt{\frac{1}{A(B/3)}\sum_{a_\star}\sum_{b_\star^1=1}^{B}\mathbb{E}_{0,\pi}\left[T_{a_\star,b_\star^1}(N)\right]\cdot 8\varepsilon^2} + \sqrt{\frac{1}{A(B/3)}\sum_{a_\star}\sum_{b_\star^2=1}^{B}\mathbb{E}_{0,\pi}\left[T_{a_\star,b_\star^2}(N)\right]\cdot 8\varepsilon^2}$$
$$= 2\sqrt{\frac{24N\varepsilon^2}{AB}}.$$

Above, (i) used the fact that for the null game M_0 and the game $M_{a_\star,b_\star^1,b_\star^2}$, only the actions $(a_\star, b_\star^1)$ and $(a_\star, b_\star^2)$ will lead to different observation distributions. (ii) used the fact that $r_2(a_\star, b_\star^1)$ has mean $1/2 + 2\varepsilon$ under $M_{a_\star,b_\star^1,b_\star^2}$ and mean $1/2$ under M_0 (and the other Bernoulli means correspondingly), and the fact that $r_1(a, b)$ are equally distributed in the two games and thus do not contribute to the KL, and finally the KL bound (13) for small enough ε such that $2\varepsilon < 1/2\sqrt{2}$. (iii) used the power mean inequality and the equality $\sum_{a_\star,b_\star} T_{a_\star,b_\star}(N) = N$ for any algorithm.

The above implies that, as long as $\varepsilon < 1/4\sqrt{2}$ and $g \le 1/4$, for $N \le AB/(864\varepsilon^2)$, we have $(\star) \le 1/3$, and thus

$$\mathbb{P}_M\left(\phi_0(\widehat{a}) \le \max_{a\in\mathcal{A}} \phi_0(a) - (g+\varepsilon)\right) = \frac{1}{A(B^2/9)}\sum_{a_\star,b_\star^1,b_\star^2}\mathbb{P}_{M_{a_\star,b_\star^1,b_\star^2}}\left(\phi_0(\widehat{a}) \le \max_{a\in\mathcal{A}} \phi_0(a) - (g+\varepsilon)\right)$$
$$\ge 1 - \frac{1}{A} - \frac{1}{3} \ge \frac{1}{3}.$$

Therefore there must exist a game $M_{a_\star,b_\star^1,b_\star^2}$ on which the error probability is at least $1/3$. This is the desired lower bound. □

E.5. Equivalence to turn-based Markov game

We consider the following general-sum turn-based Markov game[6] with two steps and state space $\mathcal{S} = \{s_a : a \in \mathcal{A}\}$:

- $(h = 1)$ Leader receives deterministic initial state s_1 and plays action $a \in \mathcal{A}$. No reward for both players.
- $(h = 2)$ The game transits deterministically to s_a. The follower plays action $b \in \mathcal{B}$ and observes reward $r_2(s_a, b) = r_2(a, b)$. The leader observes reward $r_1(s_a, b) = r_1(a, b)$.
- The game terminates.

It is straightforward to see that the bandit game $M = (\mathcal{A}, \mathcal{B}, r_1, r_2)$ is equivalent to the above turn-based Markov game. Note that the Markov game has $|\mathcal{S}| = A$ states.

Now, let $a_\star$ be the leader's exact Stackelberg equilibrium (as defined in (3)). For any a, let $b_\star(a) = \arg\min_{b\in\mathsf{BR}_0(a)} \mu_1(a, b)$ be the best response of a with the worst μ_1. Define the deterministic follower policy $\pi_\star^b$ as $\pi_\star^b(s_a) = b_\star(a)$ for all $s_a \in \mathcal{S}$.

We claim that $(a_\star, \pi_\star^b)$ is a Nash equilibrium of the above Markov game. Indeed, $\pi_\star^b$ is clearly $a_\star$'s best response on the follower's reward. Also, if we fix $\pi_\star^b$, then $a_\star$ is also the leader's best response to $\pi_\star^b$, as we have

$$\mu_1(s_a, \pi_\star^b(s_a)) = \mu_1(a, b_\star(a)) = \min_{b\in\mathsf{BR}_0(a)} \mu_1(a, b) = \phi_0(a),$$

and thus the leader's best response is exactly the argmax of $\phi_0(a)$, i.e. $a_\star$.

F. Proofs for Section 4

F.1. Subroutine `WorstCaseBestResponse`

We describe the `WorstCaseBestResponse` subroutine in Algorithm 6.

F.2. Proof of Theorem 4

[6]The formal definition of turn-based Markov games can be found in (Bai & Jin, 2020).

Algorithm 6 Subroutine `WorstCaseBestResponse`$(M, \underline{V}_2)$

Require: MDP $M = (H, \mathcal{S}, \mathcal{B}, \mathbb{P}_h(\cdot|\cdot,\cdot), r_{1,h}(\cdot,\cdot), r_{2,h}(\cdot,\cdot))$. Initial state $s_1 \in \mathcal{S}$. Target value $\underline{V}_2$.

Solve the following linear program over $\{d_h(s,b) : h \in [H], s \in \mathcal{S}, b \in \mathcal{B}\}$:

$$\begin{aligned} &\text{minimize} \ \sum_{h=1}^{H} \sum_{s\in\mathcal{S}, b\in\mathcal{B}} d_h(s,b) r_{1,h}(s,b) \\ &\text{s.t.} \ \sum_{h=1}^{H} \sum_{s\in\mathcal{S}, b\in\mathcal{B}} d_h(s,b) r_{2,h}(s,b) \ge \underline{V}_2, \\ &\qquad \sum_{s\in\mathcal{S}, b\in\mathcal{B}} d_h(s,b)\mathbb{P}_h(s'|s,b) = \sum_{b'\in\mathcal{B}} d_{h+1}(s',b') \quad \text{for all } 1 \le h \le H-1, \ s' \in \mathcal{S}, \\ &\qquad d_1(s_1,\cdot) \in \Delta_{\mathcal{B}}, \quad d_1(s_1',\cdot) = 0 \quad \text{for all } s' \ne s_1. \end{aligned} \tag{14}$$

Above, $\Delta_{\mathcal{B}}$ denotes the probability simplex on $\mathcal{B}$ (which is a set of $B+1$ linear constraints).
Let d_h denote the solution and $\underline{V}_1$ denote the value of the above program.

Set $\pi_h^b(b|s) \leftarrow d_h(s,b)/\sum_{b\in\mathcal{B}} d_h(s,b)$ for all (h,s,b) (with the convention $0/0 = 1/B$).

output $(\pi^b, \underline{V}_1)$.

Algorithm 7 Learning Stackelberg in bandit-RL games

Require: Target accuracy $\varepsilon > 0$.

for $a \in \mathcal{A}$ **do**

  Let the leader pull arm $a \in \mathcal{A}$ and the follower run the `Reward-Free RL-Explore` algorithm for $N \leftarrow \widetilde{O}(H^5S^2B/\varepsilon^2 + H^7S^4B/\varepsilon)$ episodes, and obtain model estimate $\widehat{M}^a$.

  Let $(\widehat{V}_1(a,\cdot), \widehat{V}_2(a,\cdot))$ denote the value functions for the model $\widehat{M}^a$.

  Compute the empirical best response value $\widehat{V}_2^\star(a) := \max_{\pi^b} \widehat{V}_2(a,\pi^b)$ by any optimal planning algorithm (e.g. value iteration) on the empirical MDP $\widehat{M}^a$.

  Solve the following program

$$\text{minimize}_{\pi^b} \ \ \widehat{V}_1(a,\pi^b) \ \ \text{s.t.} \ \ \pi^b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a) := \left\{\pi^b : \widehat{V}_2(a,\pi^b) \ge \widehat{V}_2^\star(a) - 3\varepsilon/4\right\} \tag{15}$$

  by subroutine $(\widehat{\pi}^{b,(a)}, \widehat{\phi}_{3\varepsilon/4}(a)) \leftarrow$ `WorstCaseBestResponse`$(\widehat{M}^a, \widehat{V}_2^\star(a) - 3\varepsilon/4)$.

end for

output $(\widehat{a}, \widehat{\pi}^b)$ where $\widehat{a} \leftarrow \arg\max_{a\in\mathcal{A}} \widehat{\phi}_{3\varepsilon/4}(a)$ and $\widehat{\pi}^b \leftarrow \widehat{\pi}^{b,(\widehat{a})}$.

Correctness of subroutine We first show that the `WorstCaseBestResponse` subroutine (Algorithm 6) with input $(\widehat{M}^a, \widehat{V}_2^\star(a) - 3\varepsilon/4)$ indeed solves the nominal problem (15). To see this, observe that in the nominal problem (15), both the objective function and the constraint are linear functions of the visitation distribution $\left\{\mathbb{P}_h^{\pi^b}(s,b)\right\}$ induced by π^b. Therefore, maximizing over all visitation distributions is equivalent to maximizing over all π^b. To ensure that a general $\{d_h(s,b)\}_{h,s,b}$ is a visitation distribution, it suffices for it to satisfy the constraints $d_1(s_1,\cdot) \in \Delta_{\mathcal{B}}$, $d_1(s_1',\cdot) = 0$ for $s_1' \ne s_1$, and at each $h \ge 2$ and each state s' the in-flow is equal to the out-flow, meaning that

$$\sum_{s\in\mathcal{S}, b\in\mathcal{B}} d_h(s,b)\mathbb{P}_h(s'|s,b) = \sum_{b\in\mathcal{B}} d_{h+1}(s',b)$$

for all $h \ge 1$, $s' \in \mathcal{S}$. These are exactly the constraints specified in the linear program (14). Finally, for a visitation distribution d_h, notice that $\pi_h^b(b|s) = d_h(s,b)/\sum_{b\in\mathcal{S}} d_h(s,b)$ (with the convention $0/0 = 1/B$) defines a policy π^b whose visitation distribution is exactly d_h. This shows that the linear program (14) is indeed a correct algorithm for solving (15).

Properties of reward-free exploration For each $a \in \mathcal{A}$, Algorithm 7 played the `Reward-Free RL-Explore` algorithm of Jin et al. (2020a) for $N = \widetilde{O}(H^5S^2B/\varepsilon^2 + H^7S^4B/\varepsilon)$ episodes and obtained an estimate of the transition dynamics $\widehat{\mathbb{P}}^a$. (More specifically, it ran $N_0 = \widetilde{O}(H^7 S^4 B/\varepsilon)$ episodes in its *exploration* phase and $N_{\rm data} = \widetilde{O}(H^5 S^2 B/\varepsilon^2)$ episodes in its *data-gathering* phase.) Further, let $\{\widehat{r}_{1,h}(a,s,b), \widehat{r}_{2,h}(a,s,b)\}$ denote the empirical mean of the observed rewards in the data-gathering phase.

Let $\widetilde{V}_1(a,\pi^b)$ and $\widetilde{V}_2(a,\pi^b)$ denote the value functions of the empirical MDPs $(\mathbb{P}^a, \widehat{r}_1)$ and $(\mathbb{P}^a, \widehat{r}_2)$ (note that these MDPs combine the true models and the *empirical* rewards). With our choice N, by (Jin et al., 2020a, Theorem 3.1), we have with probability at least $1-\delta$ that

$$\sup_{\pi^b} \left| \widehat{V}_i(a,\pi^b) - \widetilde{V}_i(a,\pi^b) \right| \le \varepsilon/16 \quad \text{for } i = 1,2. \tag{16}$$

We now argue that the `Reward-Free RL-Explore` algorithm can correctly estimate the rewards, along with estimating transitions. Indeed, we have the following

Lemma F.1. *Suppose we run the* `Reward-Free RL-Explore` *algorithm where the data gathering phase contains* $N_{\rm data} \ge \widetilde{O}(H^3 S^2 B/\varepsilon^2)$ *trajectories, and we in addition receive (stochastic) reward signals* $r_{1,h}, r_{2,h}$ *along the trajectories. Then with probability at least* $1-\delta$*, the empirical reward estimates* $\widehat{r}_{1,h}, \widehat{r}_{2,h}$ *and the associated value functions* $\widetilde{V}_1$ *and* $\widetilde{V}_2$ *satisfy that*

$$\sup_{\pi^b} \left| \widetilde{V}_i(a,\pi^b) - V_i(a,\pi^b) \right| \le \varepsilon \quad \text{for } i = 1,2.$$

We defer the proof of Lemma F.1 to Appendix F.3. As we have $N_{\rm data} = \widetilde{O}(H^5 S^2 B/\varepsilon^2)$, we can apply Lemma 6 and get (by choosing a large absolute constant in the choice of $N_{\rm data}$) that

$$\sup_{\pi^b} \left| \widetilde{V}_i(a,\pi^b) - V_i(a,\pi^b) \right| \le \varepsilon/16 \quad \text{for } i = 1,2. \tag{17}$$

Combining (16) and (17) (and noticing those are true for all $a \in \mathcal{A}$), we get

$$\sup_{a\in\mathcal{A},\pi^b} \left| \widehat{V}_i(a,\pi^b) - V_i(a,\pi^b) \right| \le \varepsilon/8 \quad \text{for } i = 1,2. \tag{18}$$

Guarantees on $\mathsf{BR}_{3\varepsilon/4}(a)$ Now, for any $a \in \mathcal{A}$, recall Algorithm 7 constructed the empirical best-response set (cf. (15))

$$\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) := \left\{ \pi^b : \widehat{V}_2(a,\pi^b) \ge \max_{\widetilde{\pi}^b} \widehat{V}_2(a,\widetilde{\pi}^b) - 3\varepsilon/4 \right\}.$$

We claim that

$$\mathsf{BR}_{\varepsilon}(a) \supseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \supseteq \mathsf{BR}_{\varepsilon/2}(a) \quad \text{for all } a \in \mathcal{A}.$$

Indeed, fixing any $a \in \mathcal{A}$, let $\pi^\star$ denote the optimal policy for $V_2(a,\cdot)$ and $\widehat{\pi}^\star$ denote the optimal policy for $\widehat{V}_2(a,\cdot)$ (dropping dependence on a for notational simplicity). Suppose $\pi_b' \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$, then we have

$$\begin{aligned}
& V_2(a,\pi^\star) - V_2(a,\pi_b') \\
& \le \underbrace{V_2(a,\pi^\star) - \widehat{V}_2(a,\pi^\star)}_{\le \varepsilon/8} + \underbrace{\widehat{V}_2(a,\pi^\star) - \widehat{V}_2(a,\widehat{\pi}^\star)}_{\le 0} + \underbrace{\widehat{V}_2(a,\widehat{\pi}^\star) - \widehat{V}_2(a,\pi_b')}_{\le 3\varepsilon/4} + \underbrace{\widehat{V}_2(a,\pi_b') - V_2(a,\pi_b')}_{\le \varepsilon/8} \\
& \le 3\varepsilon/4 + 2\cdot\varepsilon/8 \le \varepsilon.
\end{aligned}$$

This shows that $\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \subseteq \mathsf{BR}_{\varepsilon}(a)$. Notably, this implies that the output $\widehat{\pi}^b \in \mathsf{BR}_{\varepsilon}(\widehat{a})$, the desired optimality guarantee for $\widehat{\pi}^b$.

Similar as above, take any $\pi_b' \in \mathsf{BR}_{\varepsilon/2}(a)$, we have

$$\begin{aligned}
& \widehat{V}_2(a,\widehat{\pi}^\star) - \widehat{V}_2(a,\pi_b') \\
& \le \underbrace{\widehat{V}_2(a,\widehat{\pi}^\star) - V_2(a,\widehat{\pi}^\star)}_{\le \varepsilon/8} + \underbrace{V_2(a,\widehat{\pi}^\star) - V_2(a,\pi^\star)}_{\le 0} + \underbrace{V_2(a,\pi^\star) - V_2(a,\pi_b')}_{\le \varepsilon/2} + \underbrace{V_2(a,\pi_b') - \widehat{V}_2(a,\pi_b')}_{\le \varepsilon/8} \\
& \le \varepsilon/2 + 2\cdot\varepsilon/8 \le 3\varepsilon/4.
\end{aligned}$$

This shows that $\mathsf{BR}_{\varepsilon/2}(a) \subseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$, the other part of the claim.

Stackelberg guarantee for $\widehat{a}$ Finally, we show the Stackelberg guarantee for $\widehat{a}$. This part is similar as in the proof of Theorem 2. First, because $\widehat{a}$ maximizes $\widehat{\phi}_{3\varepsilon/4}(a) = \min_{\pi^b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)} \widehat{V}_1(a, \pi^b)$, we have for any $a \in \mathcal{A}$ that

$$\min_{\widetilde{\pi}^b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \widehat{V}_1(\widehat{a}, \widetilde{\pi}^b) = \widehat{\phi}_{3\varepsilon/4}(\widehat{a}) \geq \widehat{\phi}_{3\varepsilon/4}(a) = \min_{\widetilde{\pi}^b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)} \widehat{V}_1(a, \widetilde{\pi}^b) \overset{(i)}{\geq} \min_{\widetilde{\pi}^b \in \mathsf{BR}_{\varepsilon}(a)} \widehat{V}_1(a, \widetilde{\pi}^b)$$

where (i) is because $\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \subseteq \mathsf{BR}_{\varepsilon}(a)$ for all a. By the uniform convergence (18), we get

$$\min_{\widetilde{\pi}^b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} V_1(\widehat{a}, \widetilde{\pi}^b) \geq \min_{\widetilde{\pi}^b \in \mathsf{BR}_{\varepsilon}(a)} V_1(a, \widetilde{\pi}^b) - 2 \cdot \varepsilon/8 \geq \phi_{\varepsilon}(a) - \varepsilon.$$

Since the above holds for all $a \in \mathcal{A}$, taking the max on the right hand side gives

$$\max_{a \in \mathcal{A}} \phi_{\varepsilon}(a) - \varepsilon \leq \min_{\widetilde{\pi}^b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} V_1(\widehat{a}, \widetilde{\pi}^b) \overset{(i)}{\leq} \min_{\widetilde{\pi}^b \in \mathsf{BR}_{\varepsilon/2}(\widehat{a})} V_1(\widehat{a}, \widetilde{\pi}^b) = \phi_{\varepsilon/2}(\widehat{a}),$$

where (i) is because $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a}) \supseteq \mathsf{BR}_{\varepsilon/2}(a)$. In other words, we have

$$\phi_{\varepsilon/2}(\widehat{a}) \geq \max_{a \in \mathcal{A}} \phi_{\varepsilon}(a) - \varepsilon = \max_{a \in \mathcal{A}} \phi_0(a) - \mathsf{gap}_{\varepsilon} - \varepsilon.$$

This is the first part of the bound for $\widehat{a}$.

On the other hand, since $\phi_{\varepsilon}(a)$ is increasing as we decrease ε, we directly have

$$\phi_{\varepsilon/2}(\widehat{a}) \leq \phi_0(\widehat{a}) \leq \max_{a \in \mathcal{A}} \phi_0(a).$$

This is the second part of the bound for $\widehat{a}$.

□

F.3. Proof of Lemma F.1

We consider estimating a single reward $r_h = r_{1,h}$. The bound for two rewards can be obtained by setting $\delta \to \delta/2$ and applying a union bound. Consider the MDP M^a which consists of S states, B actions, and H steps. Let $V(a, \pi^b; r)$ denote the value function using the true MDP, policy π^b and reward function r, and $V(a, \pi^b; \widehat{r})$ denote the value function using the estimated reward $\widehat{r}$. Further, let

$$\mathcal{S}_h^{\delta} := \left\{ s : \max_{\pi^b} \mathbb{P}_h^{\pi^b}(s) \geq \delta \right\}.$$

denote the set of δ-significant states. By (Jin et al., 2020a), the data gathering phase of `Reward-Free RL-Explore` obtains data where the h-th step is sampled i.i.d. from some policy μ_h, such that for any $s \in \mathcal{S}_h^{\delta}$ we have

$$\max_{\pi^b} \frac{\mathbb{P}_h^{\pi^b}(s, b)}{\mu_h(s, b)} \leq 2SBH.$$

We have for any π^b that

$$\begin{aligned}
&\left| \widetilde{V}_1(a, \pi^b) - V_1(a, \pi^b) \right| = \left| V(a, \pi^b; r) - V(a, \pi^b; \widehat{r}) \right| \\
&= \left| \sum_{h=1}^{H} \sum_{s,b} \mathbb{P}_h^{\pi^b}(s, b)(\widehat{r}_h(s, b) - \mathbb{E}[r_h(s, b)]) \right| \\
&\leq \left| \sum_{h=1}^{H} \sum_{s \notin \mathcal{S}_h^{\delta}, b} \mathbb{P}_h^{\pi^b}(s, b)(\widehat{r}_h(s, b) - \mathbb{E}[r_h(s, b)]) \right| + \left| \sum_{h=1}^{H} \sum_{s \in \mathcal{S}_h^{\delta}, a} \mathbb{P}_h^{\pi^b}(s, b)(\widehat{r}_h(s, b) - \mathbb{E}[r_h(s, b)]) \right|
\end{aligned}$$

$$\leq \sum_{h=1}^{H} \sum_{s \notin \mathcal{S}_h^\delta} \mathbb{P}_h^{\pi^b}(s) + \sum_{h=1}^{H} \underbrace{\left| \sum_{s \in \mathcal{S}_h^\delta, b} \mathbb{P}_h^{\pi^b}(s,b)(\widehat{r}_h(s,b) - \mathbb{E}[r_h(s,b)]) \right|}_{:=\Delta_h}$$

$$\leq HS\delta + \sum_{h=1}^{H} \Delta_h.$$

For any h, by the Cauchy-Schwarz inequality, we have

$$\sup_{\pi^b} \Delta_h \leq \sup_{\pi^b} \left[\sum_{s \in \mathcal{S}_h^\delta, b} \underbrace{\mathbb{P}_h^{\pi^b}(s,b)}_{=\mathbb{P}_h^{\pi^b}(s)\cdot \pi_h^b(b|s)} (\widehat{r}_h(s,b) - \mathbb{E}[r_h(s,b)])^2 \right]^{1/2}$$

$$\leq \sup_{\pi^b} \max_{\nu:\mathcal{S}\to\mathcal{B}} \left[\sum_{s \in \mathcal{S}_h^\delta, b} \mathbb{P}_h^{\pi^b}(s)(\widehat{r}_h(s,b) - \mathbb{E}[r_h(s,b)])^2 \mathbf{1}\{b = \nu(s)\} \right]^{1/2}$$

$$\overset{(i)}{\leq} \max_{\nu:\mathcal{S}\to\mathcal{B}} \left[2SBH \cdot \sum_{s \in \mathcal{S}_h^\delta, b} \mu_h(s,b)(\widehat{r}_h(s,b) - \mathbb{E}[r_h(s,b)])^2 \mathbf{1}\{b = \nu(s)\} \right]^{1/2}$$

$$\overset{(ii)}{\leq} \max_{\nu:\mathcal{S}\to\mathcal{B}} \left[2SBH \cdot \sum_{s \in \mathcal{S}_h^\delta, b} \mu_h(s,b) \cdot \widetilde{O}\left(\frac{1}{N_h(s,b)} \right) \cdot \mathbf{1}\{b = \nu(s)\} \right]^{1/2}$$

$$\overset{(iii)}{\leq} \max_{\nu:\mathcal{S}\to\mathcal{B}} \left[2SBH \cdot \sum_{s \in \mathcal{S}_h^\delta, b} \mu_h(s,b) \cdot \widetilde{O}\left(\frac{1}{N\mu_h(s,b)} \right) \cdot \mathbf{1}\{b = \nu(s)\} \right]^{1/2}$$

$$= \widetilde{O}\left(\sqrt{\frac{S^2BH}{N}} \right).$$

Above, (i) used the fact that $\mathbb{P}_h^{\pi^b}(s)\mathbf{1}\{b = \nu(s)\} \leq 2SBH \cdot \mu_h(s,b)$ as $\{\pi_{h'}^b\}_{h' \leq h-1} \cup \{\nu\}$ is a valid policy. (ii) used the Hoeffding inequality (and a union bound) for the reward estimates, and the fact that the visitation of the reward-free algorithm is independent of the observed reward. (iii) used the multiplicative Chernoff bound for the visitation count $N_h(s,b) \sim \mathsf{Bin}(N, \mu_h(s,b))$ and a union bound over (s,b), which requires $N \geq O(1/\min_{s,b} \mu_h(s,b))$. Recall that `Reward-Free RL-Explore` used $\delta = \varepsilon/2H^2S$ and $\mu_h(s,b) \geq \frac{\varepsilon}{4H^3S^2B}$ for all (s,b). Thus the requirement for N is $N \geq O(H^3S^2B/\varepsilon)$ which is implied by our assumption that $N \geq \widetilde{O}(H^3S^2B/\varepsilon^2)$.

Further, plugging in the choice of N (with a sufficiently large constant) into the above bound yields

$$\sup_{\pi^b} \Delta_h \leq \widetilde{O}\left(\frac{S^2BH}{H^3S^2B/\varepsilon^2} \right) \leq \varepsilon/2H.$$

This further implies that

$$\left| \widetilde{V}_1(a,\pi^b) - V_1(a,\pi^b) \right| \leq HS \cdot \varepsilon/(2H^2S) + H \cdot \varepsilon/(2H) \leq \varepsilon,$$

the desired result. □

G. Proofs for Section 5

G.1. Algorithm description

We present our algorithm for linear bandit games in Algorithm 8. Compared with our Algorithm 5 for bandit games, Algorithm 8 takes advantage of the linear structure through the following important modifications: (1) Rather than querying

Algorithm 8 Learning Stackelberg in linear bandit games

Require: Target accuracy $\varepsilon > 0$.

1: Find $(\mathcal{K}, \rho) \leftarrow \texttt{CoreSet}(\Phi)$ (cf. (19)). Let $\mathcal{K} = \{(a_j, b_j) : 1 \le j \le K\}$ where $K = |\mathcal{K}|$.

2: Query each (a_j, b_j) for $N = O(d \log(d/\delta)/\varepsilon^2)$ times. Let $(\widehat{\mu}_{1,j}, \widehat{\mu}_{2,j})$ denote the empirical mean of the observed rewards over the N queries.

3: Estimate $(\theta_1^\star, \theta_2^\star)$ via weighted least squares

$$\widehat{\theta}_i := \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{K} \rho(a_j, b_j)\big(\phi(a_j, b_j)^\top \theta_i - \widehat{\mu}_{i,j}\big)^2, \quad i = 1, 2. \tag{20}$$

4: Construct approximate best response sets and values for all $a \in \mathcal{A}$:

$$\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) := \left\{ b : \phi(a, b)^\top \widehat{\theta}_2 \ge \max_{b' \in \mathcal{B}} \phi(a, b')^\top \widehat{\theta}_2 - 3\varepsilon/4 \right\},$$
$$\widehat{\phi}_{3\varepsilon/4}(a) := \min_{b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)} \phi(a, b)^\top \widehat{\theta}_1.$$

5: Output $(\widehat{a}, \widehat{b})$, where $\widehat{a} = \arg\max_{a \in \mathcal{A}} \widehat{\phi}_{3\varepsilon/4}(a)$, $\widehat{b} = \arg\min_{b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \phi(\widehat{a}, b)^\top \widehat{\theta}_1$.

every action pair, we only query (a, b) in a *core set* $\mathcal{K}$ found through the following subroutine

$$\begin{aligned} &\texttt{CoreSet}(\Phi) := (\mathcal{K}, \rho) \quad \text{where } \mathcal{K} \subset \mathcal{A} \times \mathcal{B},\ \rho \in \Delta_{\mathcal{K}}, \quad \text{such that} \\ &\max_{\phi \in \Phi} \phi^\top \Big(\sum_{(a,b) \in \mathcal{K}} \rho(a, b) \phi(a, b) \phi(a, b)^\top \Big)^{-1} \phi \le 2d \quad \text{and} \quad K = |\mathcal{K}| \le 4d \log\log d + 16. \end{aligned} \tag{19}$$

Such a core set is guaranteed to exist for any finite Φ (Lattimore et al., 2020, Theorem 4.4), and can be found efficiently in $O(ABd^2)$ steps of computation (Todd, 2016, Lemma 3.9). (2) Rather than estimating the reward at every (a, b) in a tabular fashion, we use a weighted least-squares (20) to obtain estimates $(\widehat{\theta}_1, \widehat{\theta}_2)$ which are then used to approximate the true reward for all (a, b).

G.2. Proof of Theorem 5

First by the guarantee (19) for the $\texttt{CoreSet}$ subroutine, we have $K = |\mathcal{K}| \le 4d \log\log d + 16$. At each $j \in [K]$ and associated $(a_j, b_j) \in \mathcal{K}$, as we queried the rewards for N times, the empirical means satisfy (letting $\phi_j := \phi(a_j, b_j)$ for shorthand)

$$\widehat{\mu}_{i,j} = \phi_j^\top \theta_i^\star + \widetilde{z}_{i,j}, \quad i = 1, 2,$$

where $\widetilde{z}_{i,j}$ is the empirical mean of N i.i.d. 1-sub-Gaussian noises, and thus is $1/N$-sub-Gaussian. Therefore, the weighted least squares estimator (20) can be expressed as (letting $\rho_j := \rho(a_j, b_j)$ for shorthand)

$$\begin{aligned} \widehat{\theta}_i &= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{K} \rho(a_j, b_j)\big(\phi(a_j, b_j)^\top \theta - \widehat{\mu}_{i,j}\big)^2 = \left(\sum_{j=1}^{K} \rho_j \phi_j \phi_j^\top \right)^{-1} \sum_{j=1}^{K} \rho_j \phi_j \cdot \widehat{\mu}_{i,j} \\ &= \left(\sum_{j=1}^{K} \rho_j \phi_j \phi_j^\top \right)^{-1} \sum_{j=1}^{K} \rho_j \phi_j \big(\phi_j^\top \theta_i^\star + \widetilde{z}_{i,j}\big) \\ &= \theta_i^\star + \left(\sum_{j=1}^{K} \rho_j \phi_j \phi_j^\top \right)^{-1} \sum_{j=1}^{K} \rho_j \phi_j \widetilde{z}_{i,j}. \end{aligned}$$

This implies the following guarantee (recall $\Phi = \{\phi(a, b) : (a, b) \in \mathcal{A} \times \mathcal{B}\}$):

$$\max_{\phi \in \Phi} \left| \phi^\top (\widehat{\theta}_i - \theta_i^\star) \right|$$

$$
\begin{aligned}
&= \max_{\phi\in\Phi} \left| \phi^\top \left(\sum_{j=1}^K \rho_j \phi_j \phi_j^\top \right)^{-1} \sum_{j=1}^K \rho_j \phi_j \widetilde{z}_{i,j} \right| \\
&\le \max_j |\widetilde{z}_{i,j}| \cdot \max_{\phi\in\Phi} \left| \sum_{j=1}^K \rho_j \phi^\top \left(\sum_{j=1}^K \rho_j \phi_j \phi_j^\top \right)^{-1} \phi_j \right| \\
&\overset{(i)}{\le} \max_j |\widetilde{z}_{i,j}| \cdot \max_{\phi\in\Phi} \left(\sum_{j=1}^K \rho_j \phi^\top \left(\sum_{j=1}^K \rho_j \phi_j \phi_j^\top \right)^{-1} \phi_j \phi_j^\top \left(\sum_{j=1}^K \rho_j \phi_j \phi_j^\top \right)^{-1} \phi \right)^{1/2} \\
&= \max_j |\widetilde{z}_{i,j}| \cdot \max_{\phi\in\Phi} \left(\phi^\top \left(\sum_{j=1}^K \rho_j \phi_j \phi_j^\top \right)^{-1} \phi \right)^{1/2} \\
&\overset{(ii)}{\le} \sqrt{2d} \cdot \max_j |\widetilde{z}_{i,j}|.
\end{aligned}
$$

Above, (i) uses Jensen's inequality (over the distribution induced by ρ_j), and (ii) used the property (19) of the core set. Now, as $\widetilde{z}_{i,j}$ is $1/N$-sub-Gaussian, with probability at least $1-\delta$, we have

$$
\max_{i=1,2} \max_{j\in[K]} |\widetilde{z}_{i,j}| \le \sqrt{\frac{\log(2K/\delta)}{N}} \le C\sqrt{\frac{\log(d/\delta)}{N}}.
$$

Substituting this into the preceding bound yields

$$
\max_{\phi\in\Phi} \left| \phi^\top (\widehat{\theta}_i - \theta_i^\star) \right| \le C\sqrt{\frac{d\log(d/\delta)}{N}}.
$$

Choosing $N = Cd\log(d/\delta)/\varepsilon^2$ guarantees that $\max_{\phi\in\Phi} \left| \phi^\top (\widehat{\theta}_i - \theta_i^\star) \right| \le \varepsilon/8$. When this happens, we have for any $(a,b) \in \mathcal{A}\times\mathcal{B}$ and any $i=1,2$ that the estimated mean reward is close to the true reward:

$$
\left| \phi(a,b)^\top \widehat{\theta}_i - \phi(a,b)^\top \theta_i^\star \right| \le \varepsilon/8.
$$

We can then proceed analogously to the proof of Theorem 2 to conclude that the output $(\widehat{a},\widehat{b})$ satisfies $\phi_0(\widehat{a}) \ge \phi_{\varepsilon/2}(\widehat{a}) \ge \max_{a\in\mathcal{A}} \phi_0(a) - \mathsf{gap}_\varepsilon - \varepsilon$ and $\widehat{b} \in \mathsf{BR}_\varepsilon(\widehat{a})$. Further, notice that the total amount of queries is

$$
NK \le Cd\log(d/\delta)/\varepsilon^2 \cdot d\log\log d = \widetilde{O}(d^2/\varepsilon^2).
$$

This proves Theorem 5. □

G.3. Lower bound

We present a $\Omega(d/\varepsilon^2)$ lower bound for linear bandit games. This shows that the sample complexity upper bound in our Theorem 5 has at most an $\widetilde{O}(d)$ factor from the lower bound.

Theorem G.1 (Lower bound for linear bandit games)**.** *There exists an absolute constant $c>0$ such that the following holds. For any $\varepsilon\in(0,c)$, $g\in[0,c)$, and any algorithm that queries $n \le c[d/\varepsilon^2]$ samples and outputs an estimate $\widehat{a}\in\mathcal{A}$, there exists a linear bandit game M with feature dimension d, on which $\mathsf{gap}_\varepsilon = g$ and the algorithm suffers from $(g+\varepsilon)$ error:*

$$
\phi_{\varepsilon/2}(\widehat{a}) \le \phi_0(\widehat{a}) \le \max_{a\in\mathcal{A}} \phi_0(a) - g - \varepsilon.
$$

Proof. This lower bound is a direct corollary of the $\Omega(AB/\varepsilon^2)$ lower bound in Theorem 3. Specifically, we can pick the size of the action spaces A', B' so that $d/2 \le A'B' \le d$, and take $\phi(a,b) = \mathbf{1}_{a,b} \in \mathbb{R}^d$ where $\mathbf{1}_{a,b}$ is the standard basis vector with one at index (a,b) (this index is understood as an index in $[d]$). This family of linear bandit games is equivalent to the family of bandit games with $A'B' \ge d/2$, for which any algorithm has to suffer from at least $(g+\varepsilon_\varepsilon)$ error if the number of queries $n \le cd/\varepsilon^2 \le cA'B'/\varepsilon^2$ by (the hard instance construction of) Theorem 3. This proves Theorem G.1. □

H. Proofs for Section C

H.1. Proof of Theorem C.1

Recall that Algorithm 1 pulled each arm (a, b) for $N = O(\log(AB/\delta)/\varepsilon^2)$ times, and $\widehat{\mu}_1(a, b)$, $\widehat{\mu}_2(a, b)$ are the empirical means of the observed rewards. By the Hoeffding inequality and union bound over (a, b), with probability at least $1 - \delta$, we have

$$\max_{(a,b)\in\mathcal{A}\times\mathcal{B}} |\widehat{\mu}_i(a, b) - \mu_i(a, b)| \le \varepsilon/8 \quad \text{for } i = 1, 2. \tag{21}$$

Properties of $\widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$ On the uniform convergence event (21), we have the following: for any $b \in \mathsf{BR}_{\varepsilon/2}(a)$, we have

$$\widehat{\mu}_2(a, b) \ge \mu_2(a, b) - \varepsilon/8 \ge \max_{b'\in\mathcal{B}} \mu_2(a, b') - 5\varepsilon/8 \ge \max_{b'\in\mathcal{B}} \widehat{\mu}_2(a, b') - 3\varepsilon/4,$$

and thus $b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$. This shows that $\mathsf{BR}_{\varepsilon/2}(a) \subseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(a)$. Similarly we can show that $\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \subseteq \mathsf{BR}_{\varepsilon}(a)$. In other words,

$$\mathsf{BR}_{\varepsilon}(a) \supseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \supseteq \mathsf{BR}_{\varepsilon/2}(a) \quad \text{for all } a \in \mathcal{A}.$$

Notably, this implies that $\widehat{b} \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a}) \in \mathsf{BR}_{\varepsilon}(\widehat{a})$, the desired near-optimality guarantee for $\widehat{b}$.

Near-optimality of $\widehat{a}$ On the one hand, because $\widehat{a}$ maximizes $\widehat{\mu}_1(a, \widehat{b}(a))$, we have for any $a \in \mathcal{A}$ that

$$\max_{b'\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \widehat{\mu}_1(\widehat{a}, b') \stackrel{(i)}{=} \widehat{\psi}_{3\varepsilon/4}(\widehat{a}) \ge \widehat{\psi}_{3\varepsilon/4}(a) \stackrel{(ii)}{\ge} \max_{b\in\mathsf{BR}_{\varepsilon/2}(a)} \widehat{\mu}_1(a, b),$$

where (i) is by definition of $\widehat{\psi}_{3\varepsilon/4}$, and (ii) is because $\widehat{\mathsf{BR}}_{3\varepsilon/4}(a) \supseteq \mathsf{BR}_{\varepsilon}(a)$. By the uniform convergence (21), we get that

$$\max_{b'\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \mu_1(\widehat{a}, b') \ge \max_{b\in\mathsf{BR}_{\varepsilon/2}(a)} \mu_1(a, b) - 2\cdot\varepsilon/8 \ge \psi_{\varepsilon/2}(a) - \varepsilon.$$

Since the above holds for all $a \in \mathcal{A}$, taking the max on the right hand side gives

$$\max_{a\in\mathcal{A}} \psi_{\varepsilon/2}(a) - \varepsilon \le \max_{b'\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a})} \mu_1(\widehat{a}, b') \stackrel{(i)}{\le} \max_{b'\in\mathsf{BR}_{\varepsilon}(\widehat{a})} \mu_1(\widehat{a}, b') = \psi_{\varepsilon}(\widehat{a}),$$

where (i) is because $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{a}) \subseteq \mathsf{BR}_{\varepsilon}(a)$. This yields that

$$\psi_{\varepsilon}(\widehat{a}) \ge \max_{a\in\mathcal{A}} \psi_{\varepsilon/2}(a) - \varepsilon \ge \max_{a\in\mathcal{A}} \psi_0(a) - \varepsilon,$$

and thus $\widehat{a} \in \mathcal{A}_{\varepsilon}$, and we can further rewrite the above as

$$\psi_0(\widehat{a}) \ge \max_{a\in\mathcal{A}} \psi_0(a) - \varepsilon - [\psi_0(\widehat{a}) - \psi_{\varepsilon}(\widehat{a})] \ge \max_{a\in\mathcal{A}} \psi_0(a) - \widetilde{\mathsf{gap}}_{\varepsilon} - \varepsilon.$$

which is the desired bound for $\widehat{a}$. □

H.2. Proof of Theorem C.2

The proof is completely analogous to that of Theorem 4 and Theorem C.1: we first establish the uniform convergence of the form (18), and then analyze the value functions similarly as in the proof of Theorem 4, except that we replace min over best response sets to max over best response sets, similar as in Theorem C.1. The guarantee we get has the same form as in Theorem 4 except that we replace ϕ-functions by ψ-functions, and replace $\mathsf{gap}_{\varepsilon}$ with $\widetilde{\mathsf{gap}}_{\varepsilon}$. □

Algorithm 9 Subroutine `BestCaseBestResponse`$(M, \underline{V}_2)$

Require: MDP $M = (H, \mathcal{S}, \mathcal{B}, \mathbb{P}_h(\cdot|\cdot,\cdot), r_{1,h}(\cdot,\cdot), r_{2,h}(\cdot,\cdot))$. Initial state $s_1 \in \mathcal{S}$. Target value $\underline{V}_2$.

1: Solve the following linear program over $\{d_h(s,b) : h \in [H], s \in \mathcal{S}, b \in \mathcal{B}\}$:

$$\begin{aligned} \text{maximize} \quad & \sum_{h=1}^{H} \sum_{s\in\mathcal{S}, b\in\mathcal{B}} d_h(s,b) r_{1,h}(s,b) \\ \text{s.t.} \quad & \sum_{h=1}^{H} \sum_{s\in\mathcal{S}, b\in\mathcal{B}} d_h(s,b) r_{2,h}(s,b) \ge \underline{V}_2, \\ & \sum_{s\in\mathcal{S}, b\in\mathcal{B}} d_h(s,b)\mathbb{P}_h(s'|s,b) = \sum_{b'\in\mathcal{B}} d_{h+1}(s',b') \quad \text{for all } 1 \le h \le H-1,\ s' \in \mathcal{S}, \\ & d_1(s_1,\cdot) \in \Delta_{\mathcal{B}}, \quad d_1(s_1',\cdot) = 0 \quad \text{for all } s' \neq s_1. \end{aligned} \tag{22}$$

Above, $\Delta_{\mathcal{B}}$ denotes the probability simplex on $\mathcal{B}$ (which is a set of $B+1$ linear constraints).
Let d_h denote the solution and $\underline{V}_1$ denote the value of the above program.

2: Set $\pi_h^b(b|s) \leftarrow d_h(s,b)/\sum_{b\in\mathcal{B}} d_h(s,b)$ for all (h,s,b) (with the convention $0/0 = 1/B$).

output $(\pi^b, \underline{V}_1)$.

I. Proofs for Section D

I.1. Proof of Theorem D.1

Recall that Algorithm 3 pulled each arm $(a,b) \in \mathcal{A} \times \mathcal{B}$ for $N = O(\log(AB/\delta)/\varepsilon^2)$ times, and $\widehat{\mu}_1(a,b)$, $\widehat{\mu}_2(a,b)$ denote the empirical means of the observed rewards. By the Hoeffding inequality and union bound over (a,b), with probability at least $1-\delta$, we have

$$\max_{\pi^a\in\Delta_{\mathcal{A}}, b\in\mathcal{B}} |\widehat{\mu}_i(\pi^a, b) - \widehat{\mu}_i(\pi^a, b)| = \max_{(a,b)\in\mathcal{A}\times\mathcal{B}} |\widehat{\mu}_i(a,b) - \mu_i(a,b)| \le \varepsilon/8 \quad \text{for } i = 1,2. \tag{23}$$

Properties of $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a)$ On the uniform convergence event (23), we have the following: for any $b \in \mathsf{BR}_{\varepsilon/2}(\pi^a)$, we have

$$\widehat{\mu}_2(\pi^a, b) \ge \mu_2(\pi^a, b) - \varepsilon/8 \ge \max_{b'\in\mathcal{B}} \mu_2(\pi^a, b') - 5\varepsilon/8 \ge \max_{b'\in\mathcal{B}} \widehat{\mu}_2(\pi^a, b') - 3\varepsilon/4,$$

and thus $b \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a)$. This shows that $\mathsf{BR}_{\varepsilon/2}(\pi^a) \subseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a)$. Similarly we can show that $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a) \subseteq \mathsf{BR}_{\varepsilon}(\pi^a)$. In other words,

$$\mathsf{BR}_{\varepsilon}(\pi^a) \supseteq \widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a) \supseteq \mathsf{BR}_{\varepsilon/2}(\pi^a) \quad \text{for all } \pi^a \in \Delta_{\mathcal{A}}.$$

Notably, this implies that $\widehat{b} \in \widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{\pi}^a) \in \mathsf{BR}_{\varepsilon}(\widehat{\pi}^a)$, the desired near-optimality guarantee for $\widehat{b}$.

Near-optimality of $\widehat{\pi}^a$ On the one hand, because $\widehat{\pi}^a$ approximately maximizes $\widehat{\mu}_1(\pi^a, \widehat{b}(\pi)^a)$ (in the sense of (9)), we have for any $\pi^a \in \Delta_{\mathcal{A}}$ that

$$\min_{b'\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{\pi}^a)} \widehat{\mu}_1(\widehat{\pi}^a, b') = \widehat{\phi}_{3\varepsilon/4}(\widehat{\pi}^a) \ge \widehat{\phi}_{3\varepsilon/4}(\pi^a) - \varepsilon/8 \overset{(i)}{\ge} \min_{b\in\mathsf{BR}_{\varepsilon}(\pi^a)} \widehat{\mu}_1(\pi^a, b) - \varepsilon/8,$$

where (i) is because $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a) \subseteq \mathsf{BR}_{\varepsilon}(\pi^a)$. By the uniform convergence (23), we get that

$$\min_{b'\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{\pi}^a)} \mu_1(\widehat{\pi}^a, b') \ge \min_{b\in\mathsf{BR}_{\varepsilon}(\pi^a)} \mu_1(\pi^a, b) - \varepsilon/4 - 2\cdot\varepsilon/8 \ge \phi_{\varepsilon}(\pi^a) - \varepsilon.$$

Since the above holds for all $\pi^a \in \Delta_{\mathcal{A}}$, taking the max on the right hand side gives

$$\sup_{\pi^a\in\Delta_{\mathcal{A}}} \phi_{\varepsilon}(\pi^a) - \varepsilon \le \min_{b'\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{\pi}^a)} \mu_1(\widehat{\pi}^a, b') \overset{(i)}{\le} \min_{b'\in\mathsf{BR}_{\varepsilon/2}(\widehat{\pi}^a)} \mu_1(\widehat{\pi}^a, b') = \phi_{\varepsilon/2}(\widehat{\pi}^a),$$

Algorithm 10 Subroutine $\texttt{BestMixedLeaderStrategy}(\widehat{\mu}_1, \widehat{\mu}_2)$

Require: Reward estimates $\widehat{\mu}_1, \widehat{\mu}_2 : \mathcal{A} \times \mathcal{B} \to [0, 1]$.

Define vectors

$$\widehat{v}_b = (\widehat{\mu}_1(a, b))_{a\in\mathcal{A}} \in [0, 1]^A, \quad \widehat{w}_b = (\widehat{\mu}_2(a, b))_{a\in\mathcal{A}} \in [0, 1]^A$$

for all $b \in \mathcal{B}$.

for $b \in \mathcal{B}$ **do**

Solve the following linear program over $\pi^a \in \Delta_{\mathcal{A}}$:

$$\begin{aligned} &\text{maximize} \;\; (\pi^a)^\top \widehat{v}_b \\ &\text{s.t.} \;\; (\pi^a)^\top (\widehat{w}_b - \widehat{w}_{b'}) \geq 0 \quad \text{for all } b' \in \mathcal{B} \setminus \{b\}. \\ &\qquad \pi^a \in \Delta_{\mathcal{A}}. \end{aligned} \tag{24}$$

Let $\widehat{\pi}^a(b), \widehat{u}(b)$ denote the solution and the value at the solution respectively.

end for

Output $\widehat{b} = \arg\max_{b\in\mathcal{B}} \widehat{u}(b)$ and $\widehat{\pi}^a = \widehat{\pi}^a(\widehat{b})$.

---

where (i) is because $\widehat{\mathsf{BR}}_{3\varepsilon/4}(\widehat{\pi}^a) \supseteq \mathsf{BR}_{\varepsilon/2}(\pi^a)$. In other words,

$$\phi_{\varepsilon/a}(\pi^a) \geq \sup_{\pi^a\in\Delta_{\mathcal{A}}} \phi_\varepsilon(\pi^a) - \varepsilon = \sup_{\pi^a\in\Delta_{\mathcal{A}}} \phi_0(\pi^a) - \mathsf{gap}_\varepsilon - \varepsilon.$$

This yields the first part of the bound for $\widehat{\pi}^a$.

On the other hand, since $\phi_\varepsilon(\pi^a)$ is increasing as we decrease ε, we directly have

$$\phi_{\varepsilon/2}(\widehat{\pi}^a) \leq \phi_0(\widehat{\pi}^a) \leq \sup_{\pi^a\in\Delta_{\mathcal{A}}} \phi_0(\pi^a).$$

This is the second part of the bound for $\widehat{\pi}^a$. □

I.2. Proof of Theorem D.2

We first check that the $\texttt{BestMixedLeaderStrategy}$ subroutine is a correct algorithm for solving (10). Let

$$\widehat{V} = (\widehat{\mu}_1(a, b))_{a,b=1}^{A,B} \quad \text{and} \quad \widehat{W} = (\widehat{\mu}_2(a, b))_{a,b=1}^{A,B}$$

denote the matrix of estimated rewards. Observe that (10) is equivalent to the following problem

$$\begin{aligned} &\max_{\pi^a\in\Delta_{\mathcal{A}}} \max_{b\in\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a)} \widehat{\mu}_1(\pi^a, b) \\ &= \max_{b\in\mathcal{B}} \max_{\pi^a:\widehat{\mathsf{BR}}_{3\varepsilon/4}(\pi^a)\ni b} (\pi^a)^\top \widehat{V} e_b \\ &= \max_{b\in\mathcal{B}} \max_{\pi^a\in\Delta_{\mathcal{A}}} (\pi^a)^\top \widehat{V} e_b \\ &\qquad \text{s.t.} \;\; (\pi^a)^\top \widehat{W} e_b \geq (\pi^a)^\top \widehat{W} e_{b'} \quad \text{for all } b' \in \mathcal{B}, \end{aligned}$$

where $e_b \in \Delta_{\mathcal{B}}$ denotes the standard basis vector in $\mathcal{B}$ (1 at b and 0 at $b' \neq b$). For each b, the above problem is exactly the same as the linear program (24). Then, the above problem requires maximizing the value over $b \in \mathcal{B}$, which is done in the output step of Algorithm 10. This shows that the $\texttt{BestMixedLeaderStrategy}$ subroutine (Algorithm 10) is correct for solving (10). Note this also proves that the $\arg\max$ in (10) is attainable (instead of the $\sup$ in Theorem D.1 which may not be attainable in general).

The rest of the proof is analogous to Theorem D.1 where we can again establish the uniform convergence (23) and obtain the suboptimality guarantee in terms of ψ and $\widetilde{\mathsf{gap}}_\varepsilon$ instead of ϕ and gap_ε, similar as in Theorem C.1. □