# Breaking the Curse of Multiagency: Provably Efficient Decentralized Multi-Agent RL with Function Approximation

**Yuanhao Wang**[⋆] YUANHAO@PRINCETON.EDU
**Qinghua Liu**[⋆] QINGHUAL@PRINCETON.EDU
*Princeton University*

**Yu Bai**[†] YU.BAI@SALESFORCE.COM
*Salesforce Research*

**Chi Jin**[†] CHIJ@PRINCETON.EDU
*Princeton University*

**Editors:** Gergely Neu and Lorenzo Rosasco

## Abstract

A unique challenge in Multi-Agent Reinforcement Learning (MARL) is the *curse of multiagency*, where the description length of the game as well as the complexity of many existing learning algorithms scale *exponentially* with the number of agents. While recent works successfully address this challenge under the model of tabular Markov Games, their mechanisms critically rely on the number of states being finite and small, and do not extend to practical scenarios with enormous state spaces where *function approximation* must be used to approximate value functions or policies.

This paper presents the first line of MARL algorithms that provably resolve the curse of multiagency under function approximation. We design a new decentralized algorithm—*V-Learning with Policy Replay*, which gives the first *polynomial* sample complexity results for learning approximate Coarse Correlated Equilibria (CCEs) of Markov Games under decentralized linear function approximation. Our algorithm always outputs *Markov* CCEs, and achieves an optimal rate of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ for finding $\varepsilon$-optimal solutions. Also, when restricted to the tabular case, our result improves over the current best decentralized result $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ for finding Markov CCEs. We further present an alternative algorithm—*Decentralized Optimistic Policy Mirror Descent*, which finds policy-class-restricted CCEs using a polynomial number of samples. In exchange for learning a weaker version of CCEs, this algorithm applies to a wider range of problems under generic function approximation, such as linear quadratic games and MARL problems with low "marginal" Eluder dimension.

## 1. Introduction

Multi-agent reinforcement learning (MARL) concerns problems in which agents learn to maximize their own utility via interacting with unknown environments as well as other agents, who may be strategic and adaptive. Modern MARL systems have achieved significant success on a wide range of challenging tasks, including the game of Go (Silver et al., 2016, 2017), Poker (Brown and Sandholm, 2018, 2019), strategic games (Vinyals et al., 2019; OpenAI, 2018; Bakhtin et al., 2022; Wurman et al., 2022), decentralized controls (Brambilla et al., 2013), autonomous driving (Shalev-Shwartz et al., 2016), as well as complex social scenarios such as hide-and-seek (Baker et al., 2020).

⋆ and † denote equal contribution.

Compared to the single-agent RL with a rich literature of theoretical understandings, MARL brings a set of new game-theoretic challenges, many of which remain open.

One unique challenge in MARL is the *curse of multiagency*, where the description length of the game (in particular, the size of the joint action space) scales exponentially with the number of agents. As a result, any learning algorithm that attempts to model the entire game (such as the transition probabilities or joint Q-values) suffers from exponentially large sample or computational complexities (Bai et al., 2020; Liu et al., 2021). These algorithms are prohibitive to run in practice even for fairly small multi-agent applications. To handle this challenge, practitioners promote the design of *decentralized* algorithms (see, e.g., Zhang et al. (2021a) for a review), where agents only aim to learn the relevant pieces of the games from their own local perspectives, such as individual policies, V-values or marginal Q-values (cf. definitions in Section 2). Decentralized algorithms further allow each agent to learn almost independently, with minimal or even no communication between the agents, which gives versatility and advantages to their implementation.

The curse of multiagency in MARL has been *provably* addressed by a recent line of theoretical works (Song et al., 2021; Jin et al., 2021b; Mao and Başar, 2022) using the V-Learning algorithm (Bai et al., 2020). However, their results only work for the basic setting of *tabular* Markov games (Shapley, 1953) where the numbers of states and actions are finite and small. Further, their mechanisms rely critically on the tabular setting that permits the synergy of (1) per-state no-regret algorithms, (2) incremental value updates, and (3) optimism; this prohibits a direct extension to practical scenarios with large state spaces. This is in contrast to modern MARL practice which commonly engages problems with an enormous number of states, where *function approximation*—typically in the form of deep neural networks—must be used to approximate either value functions or policies (Sutton and Barto, 2018). This naturally raises the following open question:

**Can we design decentralized MARL algorithms that breaks the curse of multiagency even with function approximation?**

In this paper, we answer the above question affirmatively by designing algorithms that finds approximate Coarse Correlated Equilibria (CCEs) in the presence of general function approximation, with polynomial sample complexity in problem parameters (including the number of agents). Concretely,

- We design a new decentralized meta-algorithm for MARL—*V-Learning with Policy Replay* (VLPR), and its accelerated version AVLPR (Section 3.1). Both algorithms integrate the standard V-Learning algorithm (Bai et al., 2020; Jin et al., 2021b) with new policy replay mechanisms to output Markov CCEs and facilitate learning under function approximation. VLPR is fully decentralized (assuming shared randomness among players), and AVLPR requires minimal communication (See Section 3.2). Both run in polynomial time given efficient subroutines.
- Our meta-algorithms VLPR and AVLPR calls for abstract subroutines to (1) estimate V-values for each agent (instead of joint Q-values); (2) compute stage-wise CCE policies by no-regret algorithms. We prove that under mild conditions on the subroutines, both meta-algorithms efficiently find approximate CCEs within a polynomial number of samples (Section 3). These mild conditions hold in both linear and tabular settings (Section 4).
- We instantiate AVLPR in the setting of *decentralized linear function approximation*, which gives the first decentralized MARL algorithm that provably breaks the curse of multiagency in this set-

ting (Section 4.1). Our algorithm achieves an optimal rate of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ for finding $\varepsilon$-optimal solutions. For tabular Markov Games, the current best decentralized algorithm for finding Markov CCEs requires $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ samples (Daskalakis et al., 2022). AVLPR improves over this result on the dependency of $1/\varepsilon$, the number of states, as well as the horizon (Section 4.2).

- We provide an alternative algorithm *Decentralized Optimistic Policy Mirror Descent* (DOPMD), which finds policy-class-restricted CCEs—a weaker notion of CCEs than standard definition—with sample complexity breaking the curse of multiagency (Section 5). In exchange for the weaker CCE notion, DOPMD applies to a wider range of problems with general function approximation that has bounded Bellman-Eluder dimension. These problems include linear quadratic games, and games with low "marginal" Eluder dimension or Bellman rank.

### 1.1. Related work

In this section, we review previous theoretical works on MARL under the model of Markov Games (Shapley, 1953; Littman, 1994). We acknowledge the abundant recent work on empirical MARL or under alternative mathematical models, which are beyond the scope of this paper.

**Centralized MARL** Sample-efficient learning of Markov Games has been studied extensively in a recent surge of work (Brafman and Tennenholtz, 2002; Wei et al., 2017; Jia et al., 2019; Sidford et al., 2020; Bai and Jin, 2020; Xie et al., 2020; Bai et al., 2020; Zhang et al., 2020; Tian et al., 2021; Liu et al., 2021; Bai et al., 2021; Huang et al., 2021; Jin et al., 2022; Chen et al., 2022b). Most of those approaches are centralized in nature, in that they estimate quantities (such as transition models or joint $Q$ functions) whose number of parameters scales exponentially with respect to the number of players, and thus suffer from the curse of multiagency in their sample complexities.

**Decentralized MARL** Decentralized approaches to break the curse of multiagency in Markov Games are pioneered by the V-Learning algorithm, which is initially proposed in the zero-sum setting by Bai et al. (2020), and subsequently extended to the general-sum setting (Song et al., 2021; Jin et al., 2021b; Mao and Başar, 2022; Mao et al., 2022; Cui and Du, 2022; Zhang et al., 2022), in which it can learn an approximate Correlated Equilibria (CE) or CCE of the game with sample complexity that scales polynomially with respect to the number of agents. Later, the SPoCMAR algorithm by Daskalakis et al. (2022) further learns approximate CCEs that are guaranteed to be *Markov*[1], with a slightly worse polynomial sample complexity. Both algorithms only work for tabular Markov Games and do not handle function approximation. Our algorithms VLPR and AVLPR can be seen as extensions of the V-Learning algorithm to the function approximation setting and can further output a Markov policy. Furthermore, the specialization of our algorithm to the tabular setting achieves improved sample complexity over Daskalakis et al. (2022) for learning Markov CCEs.

Decentralized algorithms for learning CE/CCEs have also been well-established in other games such as Normal-Form Games (NFGs) (Stoltz, 2005; Cesa-Bianchi and Lugosi, 2006) and Extensive-Form Games (EFGs) (Kozuno et al., 2021; Bai et al., 2022b,a; Song et al., 2022; Fiegel et al., 2022), by letting each agent run a no-regret algorithm that works even against adversarial opponents. However, this success does not extend to Markov Games due to the fundamental hardness of learning against adversarial opponents in Markov Games: there is a worst-case exponential-in-horizon regret lower

1. By contrast, the policies learned by V-Learning are non-Markov, history-dependent policies in general.

bound (Liu et al., 2022). Finally, decentralized algorithms have also been established in Markov Potential Games (Zhang et al., 2021b; Leonardos et al., 2021; Song et al., 2021; Ding et al., 2022)—a subclass of Markov Games—which however relies critically on its special potential structure.

**MARL with function approximation** A few recent works consider learning Markov Games with linear (Xie et al., 2020; Chen et al., 2022b) and general (Jin et al., 2022; Huang et al., 2021; Zhan et al., 2022; Xiong et al., 2022; Chen et al., 2022a; Ni et al., 2022) function approximation, by adapting techniques from the single-agent setting (Jiang et al., 2017; Jin et al., 2020; Zhou et al., 2021; Du et al., 2021; Jin et al., 2021a; Foster et al., 2021). All these works require *centralized* function classes and suffer from the curse of multiagency when specializing to the tabular setting. Our DOPMD algorithm differs from the related algorithms of Liu et al. (2022); Zhan et al. (2022) where our new inner loop admits decentralized function classes, which could be applied in much broader scenarios.

Technically, the policy replay mechanism used in our algorithms (in particular the one in AVLPR via doubling tricks) is similar to that of Zanette and Wainwright (2022), which is used there for designing a Q-Learning style algorithm for linear function approximation in the single-agent setting. However, our approaches are otherwise quite different, in particular in the way of updating values, where they use Q-Learning style incremental updates, whereas our algorithms use stage-wise learning with batch updates (similar in spirit to Value Iteration).

**Comparison with independent work (Cui et al., 2023)** Concurrent to this work, Cui et al. (2023) also consider the problem of breaking the curse of multiagency in the context of Markov games under *linear* function approximation, and in addition achieves the same improved sample complexity for finding Markov CCE in the basic *tabular* setting. Here we highlight a few key differences between the two works in the linear setting besides the apparent differences in the algorithm design: (1) *In terms of assumptions*, both works assume Bellman completeness with respect to certain policy classes (see, e.g., Assumption 3). We point out that it is crucial to restrict the expressiveness of the policy class, otherwise the game becomes “essentially tabular” (see Appendix H). We only require completeness with respect to linear argmax policies, while Cui et al. (2023) require completeness with respect to a policy class $\Pi^{\text{estimate}}$ that is implicitly defined by their algorithm and the no-regret learning oracle being used, which generally consists of policies that are more complex than linear argmax policies.[2] (2) *In terms of sample complexity*, this paper achieves $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ rate which has the optimal statistical dependency on error $\varepsilon$, while Cui et al. (2023) achieve a rate of $\widetilde{\mathcal{O}}(\varepsilon^{-4})$. We remark that Cui et al. (2023) have better dependency in the number of actions $A$, while our results have better dependency in dimension $d$ and horizon $H$. The differences in $A, d, H$ dependency come from the differences in both algorithmic techniques and assumptions (where the minimax-optimal rates can be potentially different).

In addition to the above differences, Cui et al. (2023) further provide results for learning under certain amount of model misspecification, learning approximate Correlated Equilibria (CEs), and learning linear Markov Potential Games, all of which have not been considered in this paper. On the

2. We remark that due to the several key differences between two papers in algorithm design and underlying mechanism, this statement about their $\Pi^{\text{estimate}}$ holds true regardless of choosing the full-information no-regret learning oracle in their algorithm as either the Exponential Weights algorithm (the choice in Cui et al. (2023)) or the Expected Follow-The-Perturbed-Leader algorithm (Hazan and Minasyan, 2020) (the choice in our paper). Please see Appendix H.1 for more details.

other hand, this paper presents results *beyond the linear function approximation setting*: both VLPR and AVLPR are generic meta-algorithms that provide guarantees for any function class as long as the required conditions for the subroutines are fulfilled. We further design a new algorithm for general function approximation that learns policy-class-restricted CCEs under weaker conditions (Section 5).

## 2. Preliminaries

**Markov Games** We consider episodic general-sum Markov Games with $m$ players, which can be described as a tuple $\mathrm{MG}(H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[m]}, \mathbb{P}, \{r_i\}_{i\in[m]})$. Here $H$ is the horizon length, $\mathcal{S}$ is the state space, $\mathcal{A}_i$ is the action space of the $i$-th player with $|\mathcal{A}_i| = A_i$[3]; we use $\mathbf{a} = (a_1, \dots, a_m) \in \prod_{i=1}^m \mathcal{A}_i =: \mathcal{A}$ to denote a joint action for all players, $\mathbb{P} = (\mathbb{P}_h)_{h\in[H]}$ are the transition probabilities, where $\mathbb{P}_h(\cdot|s, \mathbf{a}) \in \Delta(\mathcal{S})$ is the probability distribution of the next state at current state-action $(s, \mathbf{a})$ at step $h$; $r_i = (r_{i,h})_{h\in[H]}$ are the reward functions for player $i$, where each each $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a function that maps any state-action $(s, \mathbf{a})$ to a deterministic[4] reward. In each episode, the game starts at a fixed initial state $s_1$. At step $h \in [H]$ and state $s_h$, each player takes their own action $a_{i,h} \in \mathcal{A}_i$, receives their own reward $r_{i,h} = r_{i,h}(s_h, \mathbf{a}_h)$ where $\mathbf{a}_h = (a_{1,h}, \dots, a_{m,h})$, and the game transits to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, \mathbf{a}_h)$ in a Markov fashion.

A Markov policy for the $i$-th player is denoted by $\pi_i = \{\pi_{i,h}(\cdot|s) \in \Delta(\mathcal{A}_i)\}_{(s,h)\in\mathcal{S}\times[H]}$, which prescribes a distribution $\pi_{i,h}(\cdot|s) \in \Delta(\mathcal{A}_i)$ over the $i$-th player's actions at any $(s, h)$. Here, we use $\Delta(\mathcal{A}_i)$ to denote the probability simplex over the action set $\mathcal{A}_i$. A Markov *joint* policy $\pi = \{\pi_h(\cdot|s) \in \Delta(\mathcal{A})\}_{(s,h)\in\mathcal{S}\times[H]}$ is a joint policy over all players that prescribes a distribution over the joint actions, where the randomness of different players can be *correlated* in general. A special case of Markov joint policy is *product policy* $\pi = \{\pi_i\}_{i\in[m]}$ where each agent plays $\pi_i$ *independently*. For any joint policy $\pi$, we define its V-value function and (joint) Q-value function for any $(i, h) \in [m] \times [H]$ as $V^\pi_{i,h}(s) := \mathbb{E}_\pi[\sum_{h'=h}^H r_{i,h'}(s_{h'}, \mathbf{a}_{h'}) \mid s_h = s]$ and $Q^\pi_{i,h}(s, \mathbf{a}) := \mathbb{E}_\pi[\sum_{h'=h}^H r_{i,h'}(s_{h'}, \mathbf{a}_{h'}) \mid (s_h, \mathbf{a}_h) = (s, \mathbf{a})]$ respectively. Additionally, with a slight overload in notations, we define the *marginal Q-function* for player $i \in [m]$ and any $h \in [H]$ as

$$Q^\pi_{i,h}(s, a_i) := \mathbb{E}_\pi\Big[\textstyle\sum_{h'=h}^H r_{i,h'}(s_{h'}, \mathbf{a}_{h'}) \mid (s_h, a_{i,h}) = (s, a_i)\Big],$$

which measures the Q-value of player $i$ conditioned at a state and their own action, while marginalizing over the opponents' actions according policy $\pi$. For notational simplicity, we define operator $\mathbb{P}_h$ and $\mathbb{D}_\pi$ as $[\mathbb{P}_h V](s, \mathbf{a}) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,\mathbf{a})}[V(s')]$ and $\mathbb{D}_\pi[Q](s) := \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|s)}[Q(s, \mathbf{a})]$. We also use $\pi_{-i}$ to denote the joint policy of all but the $i$-th player specified by $\pi$. For any Markov product policy $\pi = \pi_i \times \pi_{-i}$, the Bellman operator $\mathcal{T}^\pi_{i,h}$ for player $i$ at step $h$ is a self-map over the $i$-th player's marginal Q-function space $(\mathcal{S} \times \mathcal{A}_i \to \mathbb{R})$, defined as

$$[\mathcal{T}^\pi_{i,h} f](s, a_i) := \mathbb{E}_{\mathbf{a}_{-i}\sim\pi_{-i,h}(\cdot|s), s'\sim\mathbb{P}_h(\cdot|s,\mathbf{a}), a_i'\sim\pi_{i,h+1}(s')}[r_{i,h}(s, \mathbf{a}) + f(s', a_i')].$$

**Coarse Correlated Equilibrium** Our goal is to find an approximate equilibrium of the Markov Game, i.e., a joint policy such that each player's own policy is near-optimal against their opponents

3. Our results in Section 3, 5 do not require $A_i$ to be finite.
4. Our results can generalize directly to the case of stochastic rewards.

**Algorithm 1** V-Learning with Policy Replay (VLPR)

**Initialize** $\pi^1$ to be the uniform policy: $\pi^1_{i,h}(\cdot|s) \leftarrow \mathrm{Unif}(\mathcal{A}_i)$ for all $(i, s, h)$.
**for** iteration $t = 1, \ldots, T$ **do**
  Set replay policy $\overline{\pi}^t \leftarrow \mathrm{Unif}(\{\pi^\tau\}_{\tau \in [t]})$ and $\overline{V}^{t+1}_{i,H+1} \leftarrow 0$.
  **for** $h = H, \ldots, 1$ **do**
    Compute $\pi^{t+1}_h \leftarrow$ CCE-APPROX$_h(\overline{\pi}^t, \{\overline{V}^{t+1}_{i,h+1}\}_{i\in[m]}, t)$.
    Compute $\{\overline{V}^{t+1}_{i,h}\}_{i\in[m]} \leftarrow$ V-APPROX$_h(\overline{\pi}^t, \pi^{t+1}_h, \{\overline{V}^{t+1}_{i,h+1}\}_{i\in[m]}, t)$.
**Output:** $\pi^{\mathrm{out}}$ sampled uniformly at random from $\{\pi^t\}_{t\in[T]}$.

in a certain sense. In our multi-player general-sum setting, the standard notion of Nash Equilibrium is both computationally PPAD-hard (Daskalakis, 2013) and statistically intractable, requiring $\exp(\Omega(m))$ samples (Rubinstein, 2017). We focus on learning Coarse Correlated Equilibrium (CCE), a common relaxed notion of equilibrium for general-sum Markov Games (Liu et al., 2021), which does not exhibit such hardness and can indeed be learned with polynomial time and samples in the basic tabular setting (Song et al., 2021; Jin et al., 2021b; Mao and Başar, 2022).

For any $\varepsilon > 0$, we say that a joint policy $\pi$ is an $\varepsilon$-approximate CCE of the game if

$$\mathrm{CCEGap}(\pi) := \max_{i\in[m]}(\max_{\pi_i^\dagger} V^{\pi_i^\dagger,\pi_{-i}}_{1,i}(s_1) - V^{\pi}_{1,i}(s_1)) \le \varepsilon,$$

Here, the maximizer $\pi_i^\dagger$ is also known as the best response. We denote $V^{\dagger,\pi_{-i}}_{i,h} := \max_{\pi_i^\dagger} V^{\pi_i^\dagger,\pi_{-i}}_{i,h}$.

We consider the standard setting of PAC learning from bandit feedback, where the agents repeatedly interact with the underlying Markov Game for many episodes, and observe the trajectory $(s_1, \mathbf{a}_1, \mathbf{r}_1, \ldots, s_H, \mathbf{a}_H, \mathbf{r}_H)$ (where $\mathbf{r}_h := (r_{i,h})_{i\in[m]}$) within each episode. The goal is to find an $\varepsilon$-approximate CCE $\widehat{\pi}$ of the game within as few episodes of play as possible.

### 2.1. Decentralized MARL with function approximation

To allow decentralized MARL with large state spaces, this paper considers *function approximation*, where each player $i \in [m]$ has her own marginal Q-value function class $\mathcal{F}_i$. Formally, we let each player $i \in [m]$ be equipped with finite[5] function class $\mathcal{F}_i = \mathcal{F}_{i,1} \times \cdots \times \mathcal{F}_{i,H}$, where each $f_{i,h} \in \mathcal{F}_{i,h} \subset (\mathcal{S} \times \mathcal{A}_i \to \mathbb{R})$ models a marginal Q-function at step $h \in [H]$.[6]

With suitable assumptions about $\mathcal{F}_i$ and the game (presented in the sequel), we are interested in finding an approximate CCE with sample complexity avoiding the *curse-of-multiagent* (Jin et al., 2021b; Song et al., 2021), i.e. scaling polynomially in $\max_{i\in[m]} \log |\mathcal{F}_i|$, the number of players $m$, as well as all other problem parameters.

## 3. Decentralized MARL via policy replay: meta-algorithms and guarantees

5. Our results extend directly to the case of infinite function classes via standard covering arguments.
6. While we focus on Q-type function approximation, our meta-algorithms can also extend to V-type function approximation, though the two types may encompass fairly different problem structures; see Appendix G for a discussion.

**Algorithm 2** CCE-APPROX$_h(\overline{\pi}, \{\overline{V}_{i,h+1}\}_{i\in[m]}, K)$

**Require:** Exploration policy mapping $\Gamma_{\text{explore}}$; subroutine NO-REGRET-ALG.

Execute $\overline{\pi}$ for $K$ episodes to collect $\{\mathcal{D}^i_{\text{init}}\}_{i\in[m]}$. Initialize $\mathcal{D}^{k,i}_{\text{sample}} \leftarrow \{\}$ for all $(i,k) \in [m] \times [K]$.
**for** $k = 1, \dots, K$ **do**
  **for** $(\widetilde{\pi}, P) \in \Gamma_{\text{explore}}(\overline{\pi}, \mu^k_h)$ **do**
    Execute $\widetilde{\pi}$ to collect a trajectory $(s_1, \mathbf{a}_1, \mathbf{r}_1, \dots, s_H, \mathbf{a}_H, \mathbf{r}_H)$.
    Update $\mathcal{D}^{k,i}_{\text{sample}} \leftarrow \mathcal{D}^{k,i}_{\text{sample}} \cup \{(s_h, a_{i,h}, r_{i,h} + \overline{V}_{i,h+1}(s_{h+1}))\}$ for all $i \in P$.
  Update $\mu^{k+1}_{i,h} \leftarrow$ NO-REGRET-ALG$(\mu^k_{i,h}, \mathcal{D}^{k,i}_{\text{sample}}, \mathcal{D}^i_{\text{init}})$ for all $i \in [m]$.

**Output:** $\pi^{\text{out}}_h := \frac{1}{K}\sum_{k=1}^K \mu^k_h$, where $\mu^k_h = \mu^k_{1,h} \times \cdots \times \mu^k_{m,h}$.

**Algorithm** Our first main algorithm, V-Learning with Policy Replay (VLPR; Algorithm 1), is a meta-algorithm for decentralized MARL with function approximation. At a high level, VLPR adopts a *policy replay* mechanism (Line 3), which in the $t$-th iteration sets the *roll-in policy* $\overline{\pi}^t = \text{Unif}(\{\pi^\tau\}_{\tau\in[t]})$ to be the uniform mixture of all previously learned policies. Using this roll-in policy, it then learns a new approximate CCE-policy $\pi^{t+1}$ by *stage-wise learning* which recursively computes the approximate CCE policies and V-values from $h = H$ to $1$ using two subroutines:

- CCE-APPROX$_h$ (Algorithm 2) takes in value estimates $\{\overline{V}^{t+1}_{i,h+1}\}_{i\in[m]}$, and computes an approximate CCE $\pi^{t+1}_h$ for the $h$-th step. It requires two ingredients: (1) An ordered set of *exploration policies* and *active players* $(\widetilde{\pi}, P) \in \Gamma_{\text{explore}}(\overline{\pi}, \mu^k_h)$ ($P \subseteq [m]$ is an index set), where each round executes each such $\widetilde{\pi}$ to observe a trajectory, and adds the observation $(s_h, a_{i,h}, r_{i,h} + \overline{V}_{i,h+1}(s_{h+1}))$ into the $i$-th player's dataset $\mathcal{D}^{k,i}_{\text{sample}}$ iff $i \in P$. (2) Each player then runs a no-regret algorithm NO-REGRET-ALG using the collected data. We require relatively strong NO-REGRET-ALG, which achieves small *per-state* regret in the face of large state spaces (in a proper sense) under bandit feedback (cf. Condition (1A)), which will be discussed momentarily.

- V-APPROX$_h$ (Algorithm 3) takes in the new policy $\pi^{t+1}_h$ and value estimates $\{\overline{V}^{t+1}_{i,h+1}\}_{i\in[m]}$, and produces estimates $\{\overline{V}^{t+1}_{i,h}\}_{i\in[m]}$ for the $h$-th step by regression algorithm OPTIMISTIC-REGRESS, which is required to achieve optimistic estimation with small errors (cf. Condition (1B)).

Notably, VLPR combines the policy replay mechanism and the V-APPROX subroutine which relearns a new value function at each iteration in a *batch* fashion. This mechanism is different from the standard V-Learning algorithm which directly plays a newly learned policy in each iteration without replay, but uses incremental value updates. That mechanism effectively learns the value of an implicit "output policy" (the "certified policy") which is different from the previously played policies (Bai et al., 2020; Jin et al., 2021b; Song et al., 2021; Mao and Başar, 2022). However, in the presence of function approximation, the batch learning in VLPR is preferred and precisely enabled by the policy replay mechanism, as it is otherwise unclear how to generalize the incremental value update approach to the case with general function classes.

**Conditions and guarantee** VLPR is a generic meta-algorithm. Once the subroutines satisfy specific requirements, the meta-algorithm will be guaranteed to learn an approximate CCE of the game.

**Condition 1 (Required conditions for VLPR)** *There exists* bonus function $G_{i,h}(s, \overline{\pi}, K, \delta)$ *for every* $(i,h) \in [m] \times [H]$ *such that the followings hold when executing Algorithm 1.*

**Algorithm 3** V-APPROX$_h(\overline{\pi}, \pi_h, \{\overline{V}_{i,h+1}\}_{i\in[m]}, K)$

**Require:** Exploration policy mapping $\Gamma_{\text{explore}}$; subroutine OPTIMISTIC-REGRESS.
1: Initialize $\mathcal{D}^i_{\text{reg}} \leftarrow \{\}$ for all $i \in [m]$.
2: **for** $k = 1, \ldots, K$ **do**
3: **for** $(\widetilde{\pi}, P) \in \Gamma_{\text{explore}}(\overline{\pi}, \pi_h)$ **do**
4: Execute $\widetilde{\pi}$ to collect a trajectory $(s_1, \mathbf{a}_1, \mathbf{r}_1, \ldots, s_H, \mathbf{a}_H, \mathbf{r}_H)$.
5: Add $(s_h, a_{i,h}, r_{i,h} + \overline{V}_{i,h+1}(s_{h+1}))$ into $\mathcal{D}^i_{\text{reg}}$ for all $i \in P$.
6: $\overline{V}_{i,h} \leftarrow$ OPTIMISTIC-REGRESS$(\pi_{i,h}, \mathcal{D}^i_{\text{reg}})$ for all $i \in [m]$.
**Output:** $\{\overline{V}_{i,h}\}_{i\in[m]}$.

*(1A)* **Per-state no-regret***: Subroutine* $\pi$ = CCE-APPROX$_h(\overline{\pi}, \{\overline{V}_{i,h+1}\}_{i\in[m]}, K)$ *(Algorithm 2) satisfies that with probability at least* $1-\delta$*, for all* $(i, s) \in [m] \times \mathcal{S}$*:*

$$\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)} \left(\mathbb{D}_{\mu_{i,h}\times\pi_{-i,h}} - \mathbb{D}_{\pi_h}\right)\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s) \le G_{i,h}(s, \overline{\pi}, K, \delta).$$

*(1B)* **Optimistic V-estimate***: Subroutine* $\overline{V}_{i,h}$ = V-APPROX$_h(\overline{\pi}, \pi_h, \{\overline{V}_{i,h+1}\}_{i\in[m]}, K)$ *(Algorithm 3) satisfies that with probability at least* $1-\delta$*, for all* $(i, s) \in [m] \times \mathcal{S}$*:*

$$\begin{cases} \overline{V}_{i,h}(s) \ge \min\left\{\mathbb{D}_{\pi_h}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s) + G_{i,h}(s, \overline{\pi}, K, \delta), H - h + 1\right\}, \\ \overline{V}_{i,h}(s) \le \qquad \mathbb{D}_{\pi_h}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s) + 2G_{i,h}(s, \overline{\pi}, K, \delta). \end{cases}$$

*(1C)* **Pigeon-hole condition***: There exists an absolute complexity measure* $L \in \mathbb{R}^+$ *such that for any* $(i, h) \in [m] \times [H]$*,* $(T, \delta) \in \mathbb{N} \times (0, 1)$*, and any policy sequence* $\{\pi^1, \ldots, \pi^T\}$*,*

$$\textstyle\sum_{t=1}^{T} \mathbb{E}_{s_h\sim\pi^{t+1}}\left[G_{i,h}(s_h, \mathrm{Unif}(\{\pi^\tau\}_{\tau\in[t]}), t, \delta)\right] \le \sqrt{LT\log^2(T/\delta)}.$$

Condition (1A) requires that the subroutine CCE-APPROX (which calls NO-REGRET-ALG) achieves *per-state* low-regret (recall in Algorithm 2 the output policy is a uniform mixture of polices that are played). This is more stringent than regret bounds w.r.t. a fixed state distribution as in standard *contextual* bandit problems (Lattimore and Szepesvári, 2020), but is crucial for learning CCEs which require the learned policies to extrapolate well to multiple roll-in distributions.

Condition (1B) requires the subroutine V-APPROX (which calls OPTIMISTIC-REGRESS) to produce *optimistic* and *accurate* value estimates for policy $\pi_h$, in a precise sense that the difference between the estimate $\overline{V}_{i,h}$ and the ground truth $\mathbb{D}_{\pi_h}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right]$ is sandwiched (modulo truncation) within $[1, 2]$ times the bonus function $G_{i,h}$.

Condition (1C) has a similar flavor to the pigeon-hole principle, and is used to ensure the expected bonuses sum up to $\widetilde{\mathcal{O}}(\sqrt{T})$ as in UCB-style algorithms, e.g., Azar et al. (2017); Jin et al. (2020).

We are now ready to state our main guarantee for VLPR.

**Theorem 1 ("Regret" guarantee for VLPR)** *Suppose Condition 1 holds for Algorithm 1. Then with probability at least* $1 - 3\delta$*, we have that*

$$\mathrm{CCEReg}(T) := \max_{i\in[m]} \textstyle\sum_{t=1}^{T} \left[V_{i,1}^{\dagger,\pi^t_{-i}}(s_1) - V_{i,1}^{\pi^t}(s_1)\right]_+ \le \widetilde{\mathcal{O}}(\sqrt{H^2LT}). \tag{1}$$

**Corollary 2 (Sample complexity)** *Choosing $T = \widetilde{\mathcal{O}}(H^2 L/\varepsilon^2)$ ensures that the output policy $\pi^{\rm out}$ of Algorithm 1 satisfies $\mathrm{CCEGap}(\pi^{\rm out}) \le \varepsilon$ further with probability at least*[7] $0.99$*, and the total number of episodes played is at most $\widetilde{\mathcal{O}}\big(H^5 L^2 \overline{\Gamma}/\varepsilon^4\big)$, where $\overline{\Gamma} := \max_{\overline{\pi},\pi'} |\Gamma_{\rm explore}(\overline{\pi},\pi')|$.*

Theorem 1 and Corollary 2 assert that an $\varepsilon$-approximate CCE can be found within $\mathrm{poly}(H, L, \overline{\Gamma}, 1/\varepsilon)$ samples, as long as all the subroutines in Algorithm 1 satisfy Condition 1. The proof (given in Appendix B.1) is relatively straightforward given the conditions, which uses performance difference arguments and combine Condition (1A) & (1B) to upper bound $\mathrm{CCEReg}(T)$ by the bonuses, and uses Condition (1C) to further bound the summation of the bonuses over $t \in [T]$.

### 3.1. Accelerated $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ algorithm via infrequent policy updates

The $\widetilde{\mathcal{O}}(1/\varepsilon^4)$ rate obtained in Theorem 1 is slower than the standard $1/\varepsilon^2$ rate. This happens as VLPR adopts the replay mechanism and updates the policy *at every iteration* $t \in [T]$, which causes the $T \times T = \widetilde{\mathcal{O}}(1/\varepsilon^4)$ rate. However, such a frequent policy update may be unnecessary if the roll-in distributions induced by the replay policies $\{\overline{\pi}^t\}_{t\ge 1}$ do not change quickly over $t$.

To address this, we design an accelerated algorithm called AVLPR (Algorithm 5) that improves this rate to $1/\varepsilon^2$ under an additional condition (Condition 2) that allows the algorithm to perform well with *infrequent policy updates*—more precisely $\mathcal{O}(\log T)$ updates—within $T$ iterations (Theorem 13). We will realize this condition by *doubling tricks*. See Appendix B.2 for details.

### 3.2. Decentralized execution

Our algorithms VLPR and AVLPR are thus far described in terms of all players jointly. Nevertheless, both algorithms can be implemented in a decentralized fashion. Rigorously, we consider the setting that each player is only able to see the shared state and their own action and reward. That is, they do not know other players' actions or rewards if without communication. We show that using certain simple protocols, VLPR can be executed in a fully decentralized fashion without any communication (assuming shared randomness among players), and AVLPR can be executed with $\mathcal{O}(\log T)$ rounds of extremely small communication only for the checking the triggering condition (Line 4 in Algorithm 5). We defer the detailed arguments to Appendix B.4.

## 4. Instantiation in linear and tabular settings

We now instantiate AVLPR concretely in two settings: decentralized linear function approximation (a new setting), and learning Markov CCEs for tabular Markov Games. We focus on the sample complexity here; both instantiations are also computationally efficient (cf. Appendix C.1 & E.1).

### 4.1. Decentralized linear function approximation

We consider Markov Games with *decentralized linear function approximation*, where each $\mathcal{F}_{i,h} = \{f_{i,h}(\cdot,\cdot) = \phi_i(\cdot,\cdot)^\top \theta_h : \|\theta_h\|_2 \le B_\theta := H\sqrt{d}\}$ is a linear function class with respect to a known

7. The success probability can be further improved to $1-\delta$ for any small $\delta > 0$ with at most an additional $\log(1/\delta)$ factor in the sample complexity, using an optimistic evaluation of the CCEGap combined with boosting.

$d$-dimensional *feature map*[8] $\phi_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}^d$. We consider the class of linear argmax policies

$$\Pi_{i,h}^{\mathrm{lin}} := \left\{ \pi_{i,h}(\cdot|s) = \arg\max_{a_i \in \mathcal{A}_i} \phi_i(s, a_i)^\top w_{i,h}, \ \forall\, s \in \mathcal{S} \mid w_{i,h} \in \mathbb{R}^d \right\}. \tag{2}$$

induced by the feature map $\phi_i$, and denote $\Pi_i^{\mathrm{lin}} = \bigtimes_{h\in[H]} \Pi_{i,h}^{\mathrm{lin}}$ and $\Pi^{\mathrm{lin}} := \bigtimes_{i\in[m]} \Pi_i^{\mathrm{lin}}$. To ensure that the feature map is informative enough, we make the following assumption.

**Assumption 3 ($\Pi^{\mathrm{lin}}$-completeness)** *For any $(i, h) \in [m] \times [H]$, any $f_{i,h+1} : \mathcal{S} \times \mathcal{A}_i \to [0, H]$, any $\pi \in \Pi^{\mathrm{lin}}$, we have $\mathcal{T}_{i,h}^{\pi} f_{i,h+1} \in \mathcal{F}_{i,h}$.*

At $m = 1$ (the single-agent setting), Assumption 3 is strictly weaker than the linear MDP assumption (Jin et al., 2020) but stronger than the linear completeness assumption (Zanette et al., 2020), both common assumptions for RL with linear function approximation. For $m \geq 2$, Assumption 3 can be seen as a decentralized multi-agent generalization of the linear MDP assumption, which requires that for every player $i \in [m]$ the Bellman backup of any $\overline{V}_{i,h+1}$ with respect to any linear argmax policy $\pi_{-i}$ is contained in $\mathcal{F}_{i,h}$ (thus is linear in $\phi_i(s, a_i)$).

We remark that in Assumption 3, requiring completeness only for the *restricted* policy class $\Pi^{\mathrm{lin}}$ is crucial: if completeness is required for all Markov policies, then the game is "essentially tabular" in the sense that the number of *non-trivial states* must be small (cf. Appendix H).

**Main result** For decentralized linear function approximation, we instantiate AVLPR to obtain the following guarantee. The algorithmic details and the proof can be found in Appendix C.

**Theorem 4 (AVLPR for decentralized linear function approximation)** *Suppose the decentralized linear function approximation satisfies Assumption 3. Then a suitable instantiation of AVLPR finds an $\varepsilon$-CCE within $\widetilde{\mathcal{O}}\left(d^4 H^6 m^2 (\max_{i\in[m]} A_i^5)/\varepsilon^2\right)$ episodes of play.*

Theorem 4 achieves a $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity with polynomial dependence on $(d, H, m, \max_{i\in[m]} A_i)$, avoiding the curse of multiagency. To our best knowledge, this is the first such result for learning Markov Games with decentralized linear function approximation.

**Overview of techniques** Establishing Theorem 4 requires instantiating the NO-REGRET-ALG and OPTIMISTIC-REGRESS subroutines in AVLPR for the linear function approximation setting such that Conditions (1A)-(1C) & 2 are satisfied. We choose OPTIMISTIC-REGRESS to be the standard ridge regression, which ensures Condition (1B) by Assumption 3.

The more challenging task is to choose NO-REGRET-ALG that satisfies Condition (1A), which, roughly speaking, requires (1) per-state regret guarantees at all $s \in \mathcal{S}$; (2) the policies $\left\{\mu_h^k\right\}_{k\in[K]}$ to lie in $\Pi^{\mathrm{lin}}$. Perhaps counter-intuitively, this rules out either running a separate linear adversarial bandit algorithm at each state, which violates (2), or adversarial *contextual* linear bandit algorithms such as LINEXP3 (Neu and Olkhovskaya, 2020), which violates (1). We resolve this by converting the problem into $\mathcal{S}$ parallel online linear optimization problems using the special structure of $\Pi^{\mathrm{lin}}$, and applying the Expected Follow-the-Perturbed-Leader algorithm (Hazan and Minasyan, 2020) to produce a single set of iterates within $\Pi^{\mathrm{lin}}$ that solves all $\mathcal{S}$ problems simultaneously (without any $|\mathcal{S}|$ dependence in rate), thereby fulfilling both requirements.

8. Without loss of generality, we assume bounded features: $\sup_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \|\phi_i(s, a_i)\|_2 \leq B_\phi := 1$ for all $i \in [m]$.

With these subroutines chosen, we show that Condition 1 is satisfied with bonus function

$$G_{i,h}(s,\overline{\pi},K,\delta) := \widetilde{\Theta}\left(\max_{a_i\in\mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h}+\lambda I)^{-1}} \times d(\max_i A_i^{1.5})H/\sqrt{K} + K^{-1}\right), \quad (3)$$
$$\text{where} \quad \Sigma^{\overline{\pi}}_{i,h} := \mathbb{E}_{s_h\sim\overline{\pi}}\,\mathbb{E}_{a_{i,h}\sim\mathrm{Unif}(\mathcal{A}_i)}\left[\phi_i(s_h,a_{i,h})\phi_i(s_h,a_{i,h})^\top\right], \quad \lambda = \widetilde{\Theta}(d(\max_i A_i)/K).$$

### 4.2. Learning Markov CCE in tabular Markov Games

We also instantiate AVLPR on tabular Markov Games (where $\mathcal{F}_i$ is the class of all possible marginal Q functions), and obtain the following result (algorithm details and proof in Appendix E).

**Theorem 5 (Tabular Markov Games)** *For tabular Markov Games with $S$ states, a suitable instantiation of AVLPR finds a* Markov *$\varepsilon$-CCE within $\widetilde{\mathcal{O}}\big(H^6S^2(\max_{i\in[m]} A_i)/\varepsilon^2\big)$ episodes of play.*

The only existing algorithm for learning Markov CCEs avoiding the curse of multiagency is the SPoCMAR algorithm of Daskalakis et al. (2022), which achieves a $\widetilde{\mathcal{O}}\left(H^{10}S^3(\max_i A_i)/\varepsilon^3\right)$ sample complexity. Theorem 5 achieves both an improved $(H,S)$ dependence and a near-optimal $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ rate. To establish Theorem 5, we instantiate No-Regret-Alg to be a separate EXP3 algorithm at every state $s\in\mathcal{S}$, and Optimistic-Regress to be simply a state-wise optimistic value estimate. We show that these ensure Conditions 1 with following bonus function:

$$G_{i,h}(s,\overline{\pi},K,\delta) := \widetilde{\Theta}\big(\eta_i^{-1}(J_h(s)+\iota)^{-1} + \eta_i H^2 A_i\big),$$

where $\eta_i$ is the learning rate for the $i$-th player's No-Regret-Alg, $J_h(s)$ is the expected visitation count of state $s$ at step $h$ when running roll-in policy $\overline{\pi}$ for $K$ episodes, and $\iota = \widetilde{\mathcal{O}}(1)$.

## 5. Learning CCE within restricted policy classes

In this section, we present an alternative approach for learning a *CCE within a restricted policy class* $\Pi$ (henceforth $\Pi$-CCE) under potentially much more relaxed assumptions on the function class.

**Restricted policy class** We let each player $i\in[m]$ be equipped with a class $\Pi_i$ of Markov policies (in addition to their marginal Q class $\mathcal{F}_i$), and let $\Pi := \prod_{i\in[m]}\Pi_i$ be the set of product policies over $\{\Pi_i\}_{i\in[m]}$. For any joint policy $\Lambda$, we say $\Lambda$ is an $\varepsilon$-approximate $\Pi$-CCE if

$$\mathrm{CCEGap}^{\Pi}(\Lambda) := \max_{i\in[m]}\left(\max_{\pi_i^\dagger\in\Pi_i} V_{1,i}^{\pi_i^\dagger\times\Lambda_{-i}}(s_1) - V_{1,i}^{\Lambda}(s_1)\right) \le \varepsilon.$$

In words, $\Lambda$ is an approximate $\Pi$-CCE as long as no player gains much by deviating to some other policy within $\Pi_i$. Note that we always have $\mathrm{CCEGap}^{\Pi}(\Lambda)\le\mathrm{CCEGap}(\Lambda)$, and the inequality is in general strict even when $\Pi_i$ is the set of all possible Markov policies for player $i$ (the largest class allowed here)[9], so that the $\Pi$-CCE is in general a more restricted notion.

**Assumptions** Our first assumption requires each function class $\mathcal{F}_i$ to be complete with respect to Bellman operators $\{\mathcal{T}^{\pi}_{i,h}\}_{\pi,h}$, a standard assumption to ensure accurate value estimation via square-loss regression (Jin et al., 2021a). This assumption relaxes Assumption 3 since this assumption only holds for $f_{i,h+1}\in\mathcal{F}_{i,h+1}$ (while Assumption 3 holds for arbitrary $f_{i,h+1}$).

9. Concretely, there exists a Markov Game in which there exists a $\Lambda\in\Delta(\Pi^{\mathrm{Mar}})$ such that $\mathrm{CCEGap}^{\Pi^{\mathrm{Mar}}}(\Lambda)=0$ but $\mathrm{CCEGap}(\Lambda)\ge H/4$ for any $H\ge 2$; see Appendix I for the construction.

**Algorithm 4** DOPMD: Decentralized Optimistic Policy Mirror Descent

**Require:** Learning rate $\{\eta_i\}_{i\in[m]}$, function class $\{\mathcal{F}_i\}_{i\in[m]}$, policy class $\{\Pi_i\}_{i\in[m]}$, $\{(K_i, \beta_i)\}_{i\in[m]}$.
1: **Initialize** $\Lambda_i^1 \leftarrow \mathrm{Unif}(\Pi_i)$ for all $i \in [m]$.
2: **for** round $t = 1, \dots, T$ **do**
3: Sample a policy $\pi_i^t \sim \Lambda_i^t$ for each $i \in [m]$, and set $\pi^t = \pi_1^t \times \dots \times \pi_m^t$.
4: **for** $i \in [m]$ **do**
5: Obtain $i$-th player's optimistic estimates $\{\overline{V}_i^{(t),\pi_i\times\pi_{-i}^t}\}_{\pi_i\in\Pi_i} \leftarrow \mathsf{APE}_i(\mathcal{F}_i, \Pi_i, \pi_{-i}^t, K_i, \beta_i)$.
6: Update $\Lambda_i^{t+1}(\pi_i) \propto_{\pi_i} \Lambda_i^t(\pi_i) \cdot \exp(\eta_i \cdot \overline{V}_i^{(t),\pi_i\times\pi_{-i}^t})$.

**Output:** Average policy $\overline{\Lambda} := \frac{1}{T}\sum_{t\in[T]} \Lambda_1^t \times \dots \times \Lambda_m^t$.

**Assumption 6 ($\Pi$-completeness)** *For every $i \in [m]$, the function class $\mathcal{F}_i$ satisfies completeness with respect to $\Pi$, that is, for any $h \in [H]$ and $(f_{i,h+1}, \pi) \in \mathcal{F}_{i,h+1} \times \Pi$, we have $\mathcal{T}_{i,h}^{\pi} f_{i,h+1} \in \mathcal{F}_{i,h}$.*

We also require each $\mathcal{F}_i \subset ((\mathcal{S} \times \mathcal{A}_i) \to [0, H])$ to have bounded Bellman-Eluder (BE) dimension (Jin et al., 2021a) to ensure sample-efficient RL. For any $i \in [m]$, we define

$$d_i(\mathcal{F}_i, \Pi, \varepsilon) := \max_{\pi_{-i}\in\Pi_{-i}} d_{\pi_{-i}}^{\mathrm{BE}}(\mathcal{F}_i, \Pi_i, \varepsilon), \tag{4}$$

where $d_{\pi_{-i}}^{\mathrm{BE}}(\mathcal{F}_i, \Pi_i, \varepsilon)$ denotes the Bellman-Eluder dimension of $\mathcal{F}_i$ with respect to the Bellman operators $\{\mathcal{T}_{i,h}^{\pi_i\times\pi_{-i}}\}_{\pi_i\in\Pi_i}$ (cf. Definition 12). The Bellman-Eluder dimension is a standard complexity measure in single-agent RL for controlling the complexity of exploration. We assume such Bellman-Eluder dimension of the marginal value functions to be bounded for all players $i \in [m]$.

**Assumption 7 (Bounded BE dimension)** *There exist scalars $\{d_i\}_{i\in[m]}$ such that for all $i \in [m]$ and $\varepsilon \in (0, 1)$, we have $d_i(\mathcal{F}_i, \Pi, \varepsilon) \le d_i \log(1/\varepsilon)$.*

Note that Assumption 6 & 7 are both *decentralized* in nature, as they only require properties about $(\mathcal{F}_i, \Pi_i)$ in the single-agent MDP induced by a fixed $\pi_{-i} \in \Pi_{-i}$. These are in contrast to previous approaches for learning Markov Games with general function approximation, which require similar structural conditions on their *centralized* function classes (Jin et al., 2022; Huang et al., 2021; Chen et al., 2022a).

### 5.1. Algorithm and guarantee

Our algorithm Decentralized Optimistic Policy Mirror Descent (DOPMD, Algorithm 4) is a double-loop algorithm. Its outer loop is similar to the policy mirror descent algorithms of (Liu et al., 2022; Zhan et al., 2022), where each player $i \in [m]$ maintains $\Lambda_i^t$—a distribution over polices in $\Pi_i$. The player then samples a policy $\pi_i^t \sim \Lambda_i^t$ (Line 3), obtains optimistic value estimates (Line 5), and performs Mirror Descent/Hedge (Line 6) in the policy space with these optimistic value estimates to obtain the update $\Lambda_i^{t+1} \in \Delta(\Pi_i)$.

The key new ingredient in our algorithm is the subroutine APE (Explorative All-Policy Evaluation; full description in Algorithm 6) for obtaining optimistic value estimates. For each player $i \in [m]$, subroutine $\mathsf{APE}_i(\mathcal{F}_i, \Pi_i, \pi_{-i}^t, K_i, \beta_i)$ plays $K_i$ episodes and obtains accurate value estimations for *all* $\pi_i \in \Pi_i$, in the MDP induced by the (fixed) opponent's policy $\pi_{-i}^t$. At a high level, APE modifies the GOLF algorithm of Jin et al. (2021a) by playing the policy that maximizes the *uncertainty*:

$$\pi_i^k := \arg\max_{\pi_i \in \Pi_i} \left\{ \max_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1)) - \min_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1)) \right\},$$

specified by the square-loss confidence set $\mathcal{B}^k$, instead of maximizing the optimistic value estimate as in GOLF.

**Theoretical guarantee** We are now ready to state the guarantee for the DOPMD algorithm. The proof can be found in Appendix F.2.

**Theorem 8 (Guarantee for DOPMD)** *Under Assumption 6 & 7, for any $\varepsilon > 0$, Algorithm 4 with $\eta_i = \sqrt{\log|\Pi_i|/(H^2T)}$, $K_i = \widetilde{\mathcal{O}}(H^4 d_i \log(\sum_{i\in[m]} |\Pi_i||\mathcal{F}_i|/\varepsilon^2)$, $\beta_i = \widetilde{\mathcal{O}}(H^2 \log(\sum_{i\in[m]} |\Pi_i||\mathcal{F}_i|))$ outputs an $\varepsilon$-approximate $\Pi$-CCE within at most $T \le \widetilde{\mathcal{O}}(H^2 \log(\sum_{i\in[m]} |\Pi_i|)/\varepsilon^2)$ rounds.*

*The total number of episodes played is at most*

$$T \times \left(\sum_{i\in[m]} K_i\right) = \widetilde{\mathcal{O}}\left(H^6 \left(\sum_{i\in[m]} d_i\right) \log^2(\textstyle\sum_i |\Pi_i||\mathcal{F}_i|)/\varepsilon^4\right).$$

*where $\widetilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors in $H, d_i, \varepsilon, \delta, \log|\mathcal{F}_i|, \log|\Pi_i|, m$.*

The sample complexity asserted in Theorem 8 for learning an $\varepsilon$-approximate $\Pi$-CCE is polynomial in the (summation of the) BE dimensions, the log-cardinality of the function classes and policy classes, as well as $1/\varepsilon$. While the $\Pi$-CCE guarantee is weaker than the VLPR or AVLPR algorithm (Theorem 1 & 13), in return, Theorem 8 only requires BE dimension and completeness assumptions, which are standard for general function approximation and potentially much more relaxed than Condition 1 required in Section 3.

**Decentralized execution** Note that the $i$-th player's APE only uses their own marginal Q class $\mathcal{F}_i$ and local observations for estimating the values for all $\pi_i \in \Pi_i$, and thus Algorithm 4 can be executed in a decentralized fashion by letting each player execute APE in lexicographic order in each round. As a result, neither communication nor shared randomness is required among players. This is different from the centralized algorithms of Liu et al. (2022); Zhan et al. (2022) that operate with joint Q classes.

### 5.2. Examples

We first show that Assumption 6 & 7 hold for learning $\Pi$-CCE in linear quadratic games (Zhang et al., 2019)—a special type of Markov Games with continuous states/actions and linear transitions—with linear policy classes and linear value classes.

**Example 1 (Linear quadratic games (LQGs))** *We consider $m$-player finite-horizon LQGs specified by a state space $\mathcal{S} \subset \mathbb{R}^{d_S}$ and action spaces $\{\mathcal{A}_i \subset \mathbb{R}^{d_{A,i}}\}_{i\in[m]}$. The initial state $s_1 \in \mathbb{R}^{d_s}$ is fixed, and the state transition at the $h$-th step is given by*

$$s_{h+1} = A_h s_h + \textstyle\sum_{i=1}^m B_{i,h} a_{i,h} + z_h, \tag{5}$$

*where $A_h \in \mathbb{R}^{d_S \times d_S}$, $B_{i,h} \in \mathbb{R}^{d_S \times d_{A,i}}$ are parameters of the game, and $z_h$ are independent mean-zero noises. The reward is given by $r_{i,h}(s, \mathbf{a}) = s_h^\top K_h^i s_h + \sum_{j=1}^m a_{j,h}^\top K_{j,h}^i a_{j,h}$ for all $(i,h) \in [m] \times [H]$, where $K_h^i \in \mathbb{R}^{d_S \times d_S}$, $K_{j,h}^i \in \mathbb{R}^{d_{A,j} \times d_{A,j}}$ are parameters of the game.*

An important policy class for LQGs is the class of *linear policies* (denoted as $\Pi$) of the form $\pi_{i,h}(s) = M_{i,h}s$, which for instance contains the CCE of the game under standard assumptions (Başar and Bernhard, 2008). In Appendix F.5, we show that such LQGs with properly chosen linear policy classes and linear value classes satisfy Assumption 6 and 7 with $d_i = \mathcal{O}((d_s + d_{A,i})^2)$, and admits sample-efficient learning of a $\Pi$-CCE with $\widetilde{\mathcal{O}}(\mathrm{poly}(H, \sum_{i\in[m]} d_i)/\varepsilon^4)$ samples by DOPMD.

By contrast, VLPR/AVLPR are unlikely to be instantiated on Example 1—Condition (1B) there typically requires $\Pi$-completeness of *optimistic values* (i.e., linear function plus bonus); a sufficient condition is $\Pi$-completeness of all values at step $h+1$ as in Assumption 3. Such optimistic values are no longer linear here and thus unlikely to be contained in our linear function class at step $h$.

Next and more generally, as Assumption 7 only requires bounded Bellman-Eluder dimension (cf. Definition 12) in a decentralized sense for each player, this contains rich subclasses such as low Eluder dimension or low Bellman rank for each player's induced *marginal* MDPs, by similar arguments as (Jin et al., 2021a, Proposition 11 & 12).

**Example 2 (Low Eluder dimension)** *Suppose that for all $i \in [m]$, $\mathcal{F}_i$ has* low Eluder dimension *(Wang et al. (2020); cf. Definition 10) in the sense that $\max_{h\in[H]} d_{\mathrm{E}}(\mathcal{F}_{i,h}, \varepsilon) \le d_i \log(1/\varepsilon)$, and satisfies $\Pi$-completeness (Assumption 6). Then, Assumption 7 also holds with the same $\{d_i\}_{i\in[m]}$.*

In particular, the class of functions with low Eluder dimension subsumes certain *non-linear* function classes such as generalized linear models (Russo and Van Roy, 2013) , which are of the form $\mathcal{F}_{i,h} = \{Q_{i,h}(\cdot,\cdot) = \sigma(\phi_i(\cdot,\cdot)^\top \theta_{i,h}) : \theta_{i,h} \in \mathbb{R}^{d_i}\}$, where $\phi_i : \mathcal{S}\times\mathcal{A}_i \to \mathbb{R}^{d_i}$ is a feature map, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a link function with $\sigma'(\cdot) \in [c_1, c_2]$ for some $0 < c_1 < c_2$.

**Example 3 (Low Bellman rank)** *Suppose for all $i \in [m]$, the single-agent MDP induced by any $\pi_{-i} \in \Pi_{-i}$ has* low Bellman rank *(Jiang et al., 2017) in the following sense: For any fixed $\pi_{-i} \in \Pi_{-i}$, there exist maps $\psi^{\pi_{-i}}_{i,h} : \Pi_i \to \mathbb{R}^{d_i}$, $\phi^{\pi_{-i}}_{i,h} : \mathcal{F}_i \times \Pi_i \to \mathbb{R}^{d_i}$ such that for any $f \in \mathcal{F}_i$, $\pi_i, \pi'_i \in \Pi_i$*

$$\mathbb{E}_{(s_h,a_{i,h})\sim\pi'_i\times\pi_{-i}}\left[(f_h - \mathcal{T}_h^{\pi_i\times\pi_{-i}} f_{h+1})(s_h, a_{i,h})\right] = \left\langle \phi^{\pi_{-i}}_{i,h}(f, \pi_i), \psi^{\pi_{-i}}_{i,h}(\pi'_i) \right\rangle.$$

*Then, Assumption 7 holds with the same $\{d_i\}_{i\in[m]}$.*

## 6. Conclusion

This paper provides the first line of results for provably efficient decentralized MARL under function approximation which avoids the curse of multiagency. We present two complementary approaches: The first one via policy replay and stage-wise no-regret learning, which we instantiate concretely in the linear and tabular setting and achieve a near-optimal $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ rate for learning an $\varepsilon$-approximate CCE in both settings; The second one via policy mirror descent with decentralized exploration, which learns a restricted version of CCE but applies to broader classes of problems. We believe our work opens up many interesting directions for future works, such as (1) deriving sharper sample complexities for both approaches, in particular improving the $(d, \max_{i\in[m]} A_i)$ dependence for AVLPR in the linear setting and the $S$ dependence in the tabular setting; (2) improving the computational efficiency for the policy mirror descent approach; and (3) identifying new problem classes amenable to the policy replay approach.

## References


Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. *Advances in Neural Information Processing Systems*, 34:25799–25811, 2021.

Yu Bai, Chi Jin, Song Mei, Ziang Song, and Tiancheng Yu. Efficient $\Phi$-regret minimization in extensive-form games via online mirror descent. *arXiv preprint arXiv:2205.15294*, 2022a.

Yu Bai, Chi Jin, Song Mei, and Tiancheng Yu. Near-optimal learning of extensive-form games with imperfect information. *arXiv preprint arXiv:2202.01752*, 2022b.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxpxJBKwS.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence*, 7(1):1–41, 2013.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

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

Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022a.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022b.

Qiwen Cui and Simon S Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *arXiv preprint arXiv:2206.00159*, 2022.

Qiwen Cui, Kaiqing Zhang, and Simon S Du. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. *arXiv preprint arXiv:2302.03673*, 2023.

Constantinos Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Rémi Munos, Vianney Perchet, and Michal Valko. Adapting to game trees in zero-sum imperfect information games. *arXiv preprint arXiv:2212.12567*, 2022.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Elad Hazan and Edgar Minasyan. Faster projection-free online learning. In *Conference on Learning Theory*, pages 1877–1893. PMLR, 2020.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

Zeyu Jia, Lin F Yang, and Mengdi Wang. Feature-based q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021b.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR, 2022.

Tadashi Kozuno, Pierre Ménard, Rémi Munos, and Michal Valko. Model-free learning for two-player zero-sum partially observable markov games with perfect recall. *arXiv preprint arXiv:2106.06279*, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. *arXiv preprint arXiv:2203.06803*, 2022.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, pages 3049–3068. PMLR, 2020.

Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. *arXiv preprint arXiv:2210.16976*, 2022.

OpenAI. Openai five. `https://blog.openai.com/openai-five/`, 2018.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. *ACM SIGecom Exchanges*, 15(2):45–49, 2017.

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

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

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

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Ziang Song, Song Mei, and Yu Bai. Sample-efficient learning of correlated equilibria in extensive-form games. *arXiv preprint arXiv:2205.07223*, 2022.

Gilles Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Université Paris Sud-Paris XI, 2005.

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

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. *arXiv preprint arXiv:2010.15020*, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michael Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896): 223–228, 2022.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang. A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning*, pages 24496–24523. PMLR, 2022.

Andrea Zanette and Martin Wainwright. Stabilizing q-learning with linear architectures for provable efficient learning. In *International Conference on Machine Learning*, pages 25920–25954. PMLR, 2022.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Wenhao Zhan, Jason D Lee, and Zhuoran Yang. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. *arXiv preprint arXiv:2206.01588*, 2022.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32, 2019.

Kaiqing Zhang, Sham M Kakade, Tamer Başar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021b.

Runyu Zhang, Qinghua Liu, Huan Wang, Caiming Xiong, Na Li, and Yu Bai. Policy optimization for markov games: Unified framework and faster convergence. *arXiv preprint arXiv:2206.02640*, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

## Contents

## Appendix A. Technical tools

### A.1. Concentration

The following Freedman's inequality can be found in (Agarwal et al., 2014, Lemma 9).

**Lemma 9 (Freedman's inequality)** *Suppose random variables $\{X_t\}_{t=1}^T$ is a martingale difference sequence, i.e. $X_t \in \mathcal{F}_t$ where $\{\mathcal{F}_t\}_{t\geq 1}$ is a filtration, and $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Suppose $X_t \leq R$ almost surely for some (non-random) $R > 0$. Then for any $\lambda \in (0, 1/R]$, we have with probability at least $1-\delta$ that*

$$\sum_{t=1}^{T} X_t \leq \lambda \cdot \sum_{t=1}^{T} \mathbb{E}\big[X_t^2|\mathcal{F}_{t-1}\big] + \frac{\log(1/\delta)}{\lambda}.$$

### A.2. Eluder & Bellman-Eluder dimension

We begin by presenting the standard definition of the Eluder dimension of a function class (Russo and Van Roy, 2013; Wang et al., 2020).

**Definition 10 (Eluder dimension)** *For any function class $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R})$, its Eluder dimension $d_{\mathrm{E}}(\mathcal{F}, \varepsilon)$ is defined as the length of the longest sequence $\{x_1, x_2, \dots, x_n\} \subset \mathcal{D}$ such that there exists $\varepsilon' \geq \varepsilon$ so that for all $i \in [n]$, $x_i$ is $\varepsilon'$-*independent *of its prefix sequence $\{x_1, \dots, x_{i-1}\}$, in the sense that there exists some $f_i, g_i \in \mathcal{F}$ such that*

$$\sqrt{\sum_{j=1}^{i-1} \left[(f_i - g_i)(x_j)\right]^2} \leq \varepsilon' \quad \text{but} \quad |(f_i - g_i)(x_i)| \geq \varepsilon'.$$

**Definition 11 (Distributional Eluder dimension)** *For any function class $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R})$, its distributional Eluder dimension $d_{\mathrm{E}}(\mathcal{F}, \mathcal{D}, \varepsilon)$ with respect to a class of distributions $\Pi \subset \Delta(\mathcal{X})$ and $\varepsilon > 0$ is defined as the length of the longest sequence $\{\mu_1, \mu_2, \ldots, \mu_n\} \subset \mathcal{D}$ such that there exists $\varepsilon' \geq \varepsilon$ so that for all $i \in [n]$, $\mu_i$ is $\varepsilon'$-*independent *of its prefix sequence $\{\mu_1, \ldots, \mu_{i-1}\}$, in the sense that there exists some $f_i \in \mathcal{F}$ such that*

$$\sqrt{\sum_{j=1}^{i-1} \left(\mathbb{E}_{X\sim\mu_j}[f_i(X)]\right)^2} \leq \varepsilon' \quad \text{but} \quad |\mathbb{E}_{X\sim\mu_i}[f_i(X)]| \geq \varepsilon'.$$

For decentralized MARL, we consider the following definition of the Bellman-Eluder dimension, which is similar to the original definition of Jin et al. (2021a) applied to the single-agent MDPs for player $i$ when facing a fixed Markov opponent $\pi_{-i}$, except that here we consider Bellman operators with respect to all policies $\pi_i \in \Pi_i$ instead of the Bellman optimality operator.

**Definition 12 (Bellman-Eluder dimension)** *For any player $i \in [m]$, any Markov policy class $\Pi_i$ for the $i$-th player, any Markov policy $\pi_{-i}$ for all but the $i$-th player, and any $\varepsilon > 0$, define*

$$d^{\mathrm{BE}}_{\pi_{-i}}(\mathcal{F}_i, \Pi_i, \varepsilon) := \min_{\mathcal{D}\in\{\mathcal{D}_{\Pi_i\times\pi_{-i}}, \mathcal{D}_\Delta\}} \max_{h\in[H]} d_{\mathrm{E}}\left(\left\{f_h - \mathcal{T}^{\pi_i\times\pi_{-i}}_{i,h} f_{h+1} : (f, \pi_i) \in \mathcal{F}\times\Pi_i\right\}, \mathcal{D}, \varepsilon\right),$$

*where $d_E(\cdot,\cdot,\varepsilon)$ denotes the distributional Eluder dimension (Definition 11), and*

$$\begin{aligned}
\mathcal{D}_{\Pi_i\times\pi_{-i}} &:= \left\{d^{\pi_i\times\pi_{-i}}_h(\cdot,\cdot) : \pi_i \in \Pi_i\right\} \subset \Delta(\mathcal{S}\times\mathcal{A}_i), \\
\mathcal{D}_\Delta &:= \left\{\delta_{(s,a_i)} : (s, a_i) \in \mathcal{S}\times\mathcal{A}_i\right\} \subset \Delta(\mathcal{S}\times\mathcal{A}_i),
\end{aligned}$$

*where $d^{\pi_i\times\pi_{-i}}_h(\cdot,\cdot) \in \Delta(\mathcal{S}\times\mathcal{A}_i)$ denotes the distribution of $(s_h, a_{i,h})$ when playing policy $\pi_i \times \pi_{-i}$ in the game, and $\delta_{(s,a_i)} \in \Delta(\mathcal{S}\times\mathcal{A}_i)$ denotes the point mass at $(s, a_i)$.*

## Appendix B. Proofs and additional details for Section 3

### B.1. Proof of Theorem 1

By the Bellman optimality equation, we have that for all $(t, i, h, s) \in [T]\times[m]\times[H]\times\mathcal{S}$

$$\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)} \mathbb{D}_{\mu_{i,h}\times\pi^t_{-i,h}}\left[r_{i,h} + \mathbb{P}_{h+1}V^{\dagger,\pi^t_{-i}}_{i,h+1}\right](s) = V^{\dagger,\pi^t_{-i}}_{i,h}(s). \tag{6}$$

On the other hand, by using Condition (1A) and the first inequality in Condition (1B), we have that with probability $1 - 2TH\delta$, for all $(t, i, h, s) \in [T]\times[m]\times[H]\times\mathcal{S}$

$$\begin{aligned}
&\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)} \mathbb{D}_{\mu_{i,h}\times\pi^t_{-i,h}}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}^t_{i,h+1}\right](s) \\
&\leq \mathbb{D}_{\pi^t_h}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}^t_{i,h+1}\right](s) + G_{i,h}(s, \overline{\pi}^{t-1}, t-1, \delta) \leq \overline{V}^t_{i,h}(s).
\end{aligned} \tag{7}$$

Therefore, by backward induction with the above two relations, we have that for all $(t, i, h, s) \in [T]\times[m]\times[H]\times\mathcal{S}$

$$\overline{V}^t_{i,h}(s) \geq V^{\dagger,\pi^t_{-i}}_{i,h}(s). \tag{8}$$

Similarly, by backward induction with the second inequality in Condition (1B), we can show that for all $(t, i, h, s) \in [T] \times [m] \times [H] \times \mathcal{S}$

$$\overline{V}_{i,h}^{t}(s) \le V_{i,h}^{\pi^t}(s) + 2 \sum_{h'=h}^{H} \mathbb{E}_{\pi^t} \left[ G_{i,h'}(s_{h'}, \overline{\pi}^{t-1}, t-1, \delta) \right]. \tag{9}$$

As a result, we can upper bound the CCE-regret by

$$\sum_{t=1}^{T} [V_{i,1}^{\dagger, \pi_{-i}^t} - V_{i,1}^{\pi^t}] \le 2 \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\pi^t} \left[ G_{i,h}(s_h, \overline{\pi}^{t-1}, t-1, \delta) \right] \le \widetilde{\mathcal{O}} \left( \sqrt{H^2 L T} \right),$$

where the final inequality follows from Condition (1C).

Finally the $\mathrm{CCEGap}$ of the output policy $\pi^{\mathrm{out}}$ can be bounded with Markov's inequality and the choice of $T = \widetilde{\mathcal{O}}\left(H^2 L / \varepsilon^2\right)$. ■

### B.2. Accelerated algorithm

**Algorithm 5** Accelerated V-Learning with Policy Replay (AVLPR)

1: **Initialize** $\pi^1$ to be the uniform policy: $\pi_{i,h}^1(\cdot|s) \leftarrow \mathrm{Unif}(\mathcal{A}_i)$ for all $(i, s, h)$, $\mathcal{B}_h^0 \leftarrow \emptyset$, $I_1 \leftarrow 0$.
2: **for** iteration $t = 1, \ldots, T$ **do**
3:     Execute $\pi^t$ to sample an episode, and update $\mathcal{B}_h^t = \mathcal{B}_h^{t-1} \bigcup \{s_h\}$.
4:     **if** $\exists (i, h) \in [m] \times [H]$ s.t. $\Psi_{i,h}(\mathcal{B}_h^t) \ge \Psi_{i,h}(\mathcal{B}_h^{I_t}) + 1$ or $t = 1$ **then**
5:         Set replay policy $\overline{\pi}^t \leftarrow \mathrm{Unif}(\{\pi^\tau\}_{\tau \in [t]})$ and $\overline{V}_{i,H+1}^{t+1} \leftarrow 0$.
6:         **for** $h = H, \ldots, 1$ **do**
7:             Compute $\pi_h^{t+1} \leftarrow \text{CCE-APPROX}(\overline{\pi}^t, \{\overline{V}_{i,h+1}^{t+1}\}_{i \in [m]}, t)$.
8:             Compute $\{\overline{V}_{i,h}^{t+1}\}_{i \in [m]} \leftarrow \text{V-APPROX}(\overline{\pi}^t, \{\overline{V}_{i,h+1}^{t+1}\}_{i \in [m]}, \pi_h^{t+1}, \{\mathcal{F}_i\}_{i \in [m]}, t)$.
9:         set $I_{t+1} \leftarrow t$
10:     **else**
11:         set $I_{t+1} \leftarrow I_t$ and $\pi^{t+1} \leftarrow \pi^t$

**Output:** $\pi^{\mathrm{out}}$ sampled uniformly at random from $\{\pi^t\}_{t \in [T]}$.

**Algorithm** We present our accelerated algorithm AVLPR in Algorithm 5. The main new ingredient in AVLPR is an *infrequent update* mechanism: The algorithm only performs the policy replay and learns a new policy $\pi^{t+1}$ if a certain *triggering condition* (Line 4) is satisfied, in which case the learning procedure is the same as in VLPR. Otherwise, it simply executes the current policy $\pi^t$ for one episode, adds the state $s_h$ into dataset $\mathcal{B}_h^t$, and sets $\pi^{t+1} \leftarrow \pi^t$ (Line 3).

Intuitively, the triggering condition requires that the dataset to have accumulated significantly since the last replay iteration $I_t < t$. This design is motivated by a *doubling-trick* type of observation: The state visitation induced by $\pi^t$ (and thus the sample complexity) does not differ significantly regardless of whether $\pi^t$ are updated or not, until some summary statistic (for example the visitation count of any state in the tabular case) is found to have increased to at least two times (or any constant factor $> 1$) since the last replay. We use $\Psi_{i,h}(\cdot)$ denote the logarithm of such a summary statistic, so that a new replay is triggered only if $\Psi_{i,h}(\mathcal{B}_h^t) \ge \Psi_{i,h}(\mathcal{B}_h^{I_t}) + 1$.

**Condition and guarantee** Concretely, AVLPR requires the following additional condition to ensure the validity of the infrequent update mechanism, which intuitively requires the bonus function can increase at most by a constant factor between consecutive policy updates.

**Condition 2 (Validity of infrequent policy update)** *The triggering criterion $\{\Psi_{i,h}\}_{(i,h)\in[m]\times[H]}$ in Algorithm 5 satisfies the following:*

*(a) With probability at least $1-\delta$, for all $(t,i,h)$, if $\Psi_{i,h}(\mathcal{B}_h^t) \le \Psi_{i,h}(\mathcal{B}_h^{I_t})+1$, then we must have $G_{i,h}(s,\overline{\pi}^{I_t},I_t,\delta) \le 8\times G_{i,h}(s,\overline{\pi}^t,t,\delta)$ for all $s\in\mathcal{S}$;*

*(b) The number of replays triggered (i.e. Line 4) in Algorithm 5 within $T$ iterations is upper bounded by $d_{\rm replay}\log T$ iterations with probability one, for some constant $d_{\rm replay}>0$.*

We now state our meta-guarantee for AVLPR; the proof can be found in Appendix B.3.

**Theorem 13 (Meta-guarantee for AVLPR)** *Suppose the subroutines in Algorithm 5 can be instantiated such that Condition (1A)-(1C) & 2 holds with the same bonus functions $\{G_{i,h}\}_{(i,h)\in[m]\times[H]}$ and the deployed triggering functions $\{\Psi_{i,h}\}_{(i,h)\in[m]\times[H]}$. Then we have with probability at least $1-\delta$ that*

$$\mathrm{CCEReg}(T) := \max_{i\in[m]}\sum_{t=1}^{T}\left[V_{i,1}^{\dagger,\pi_{-i}^t}(s_1)-V_{i,1}^{\pi^t}(s_1)\right]_+ \le \widetilde{\mathcal{O}}(\sqrt{H^2LT}).$$

*As a corollary, choosing $T=\widetilde{\mathcal{O}}(H^2L/\varepsilon^2)$ ensures that the output policy $\pi^{\rm out}$ of Algorithm 1 satisfies $\mathrm{CCEGap}(\pi^{\rm out})\le\varepsilon$ further with probability at least*[10] *0.99, and the total number of episodes played is at most (with $\overline{\Gamma}:=\max_{\overline{\pi},\pi'}|\Gamma_{\rm explore}(\overline{\pi},\pi')|$)*

$$\mathcal{O}\big(T+HT\times d_{\rm replay}\log T\times\overline{\Gamma}\big)=\widetilde{\mathcal{O}}\big(H^3L\overline{\Gamma}\cdot d_{\rm replay}/\varepsilon^2\big).$$

### B.3. Proof of Theorem 13

Let $\mathcal{I}$ denote the subset of $[T]$ where Line 4 is triggered. By Condition 2, $|\mathcal{I}|\le d_{\rm replay}\log T$.

By the Bellman optimality equation, we have that for all $(t,i,h,s)\in[T]\times[m]\times[H]\times\mathcal{S}$

$$\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)}\mathbb{D}_{\mu_{i,h}\times\pi_{-i,h}^t}\left[r_{i,h}+\mathbb{P}_{h+1}V_{i,h+1}^{\dagger,\pi_{-i}^t}\right](s)=V_{i,h}^{\dagger,\pi_{-i}^t}(s). \tag{10}$$

On the other hand, by using Condition (1A) and the first inequality in Condition (1B), we have that with probability $1-2TH\delta$, for all $(t,i,h,s)\in\mathcal{I}\times[m]\times[H]\times\mathcal{S}$

$$\begin{aligned}&\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)}\mathbb{D}_{\mu_{i,h}\times\pi_{-i,h}^{t+1}}\left[r_{i,h}+\mathbb{P}_{h+1}\overline{V}_{i,h+1}^{t+1}\right](s)\\ &\le\mathbb{D}_{\pi_h^{t+1}}\left[r_{i,h}+\mathbb{P}_{h+1}\overline{V}_{i,h+1}^{t+1}\right](s)+G_{i,h}(s,\overline{\pi}^t,t,\delta)\le\overline{V}_{i,h}^{t+1}(s).\end{aligned} \tag{11}$$

Therefore, by backward induction with the above two relations, we have that for all $(t,i,h,s)\in\mathcal{I}\times[m]\times[H]\times\mathcal{S}$

$$\overline{V}_{i,h}^{t+1}(s)\ge V_{i,h}^{\dagger,\pi_{-i}^{t+1}}(s), \tag{12}$$

10. The success probability can be further boosted to any $1-\delta$ by a similar argument as in Theorem 1.

which implies that for all for all $(t,i,h,s)\in[T]\times[m]\times[H]\times\mathcal{S}$:

$$\overline{V}_{i,h}^{I_t+1}(s)\geq V_{i,h}^{\dagger,\pi_{-i}^{I_t+1}}(s).$$

Similarly, by backward induction with the second inequality in Condition (1B), we can show that for all $(t,i,h,s)\in\mathcal{I}\times[m]\times[H]\times\mathcal{S}$

$$\overline{V}_{i,h}^{t+1}(s)\leq V_{i,h}^{\pi^{t+1}}(s)+2\sum_{h'=h}^{H}\mathbb{E}_{\pi^{t+1}}\left[G_{i,h'}(s_{h'},\overline{\pi}^{t},t,\delta)\right],$$

which implies that for all $(t,i,h,s)\in[T]\times[m]\times[H]\times\mathcal{S}$:

$$\overline{V}_{i,h}^{I_t+1}(s)\leq V_{i,h}^{\pi^{I_t+1}}(s)+2\sum_{h'=h}^{H}\mathbb{E}_{\pi^{I_t+1}}\left[G_{i,h'}(s_{h'},\overline{\pi}^{I_t},I_t,\delta)\right].$$

As a result, we can upper bound the CCE-regret by

$$\begin{aligned}
\sum_{t=1}^{T}[V_{i,1}^{\dagger,\pi_{-i}^{t}}-V_{i,1}^{\pi^t}] &\overset{(i)}{=} \sum_{t=1}^{T}[V_{i,1}^{\dagger,\pi_{-i}^{I_t+1}}-V_{i,1}^{\pi^{I_t+1}}]\\
&\leq 2\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{I_t+1}}\left[G_{i,h}(s_h,\overline{\pi}^{I_t},I_t,\delta)\right]\\
&\overset{(ii)}{\leq} 16\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{I_t+1}}\left[G_{i,h}(s_h,\overline{\pi}^{t-1},t-1,\delta)\right]\\
&\overset{(iii)}{=} 16\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{t}}\left[G_{i,h}(s_h,\overline{\pi}^{t-1},t-1,\delta)\right]\\
&\overset{(iv)}{\leq} \widetilde{\mathcal{O}}\left(\sqrt{H^2LT}\right),
\end{aligned}$$

where (i) and (iii) uses the fact that $\pi^t=\pi^{I_t+1}$, (iv) follows from Condition (1C), and (ii) follows from Lemma 14.

Finally the CCEGap of the output policy $\pi^{\rm out}$ can be bounded with Markov's inequality and the choice of $T=\widetilde{\mathcal{O}}\left(H^2L/\varepsilon^2\right)$. The total sample complexity would be bounded by

$$T+|\mathcal{I}|\times H\times\left(\underbrace{\mathcal{O}\left(\overline{\Gamma}\cdot T\right)}_{\text{cost of CCE-APPROX}}+\underbrace{\mathcal{O}\left(\overline{\Gamma}\cdot T\right)}_{\text{cost of V-APPROX}}\right)=\widetilde{\mathcal{O}}\left(H^3L\overline{\Gamma}d_{\rm replay}/\varepsilon^2\right).$$

**Lemma 14** *Suppose Condition 2 holds, then with probability at least* $1-\delta$, *for all* $(t,i,h)$, $G_{i,h}(s,\overline{\pi}^{I_t},I_t,\delta)\leq 8\times G_{i,h}(s_h,\overline{\pi}^{t-1},t-1,\delta)$ *for all* $s\in\mathcal{S}$.

**Proof** If Line 4 is triggered in the $(t-1)^{\rm th}$ iteration, then $I_t=t-1$ and the result holds. Otherwise, $I_t=I_{t-1}$ and for all $(i,h)$, $\Psi_{i,h}(\mathcal{B}_h^{t-1})\leq\Psi_{i,h}(\mathcal{B}_h^{I_{t-1}})+1$, which, by Condition 2, implies

$$G_{i,h}(s,\overline{\pi}^{I_t},I_t,\delta)=G_{i,h}(s,\overline{\pi}^{I_{t-1}},I_{t-1},\delta)\leq 8\times G_{i,h}(s_h,\overline{\pi}^{t-1},t-1,\delta)$$

for all $s\in\mathcal{S}$. ■

## B.4. Decentralized execution protocol for VLPR and AVLPR

In this section, we first describe our protocols, then argue that both VLPR and AVLPR can be made decentralized (with minimal communication for AVLPR) under these protocols.

We consider the following protocol: Before the game starts, the players sample a sequence of random bits with length polynomial in the number of episodes played, and all players can observe this (shared) sequence of random bits. Using this sequence, the players can then implement shared randomness in a decentralized fashion. For example, executing $\mathrm{Unif}(\{\pi^\tau\}_{\tau\in[T]})$ where each $\pi^\tau$ is a product policy can be done by using the shared random bits (with the same pre-determined protocol) to sample a shared $\tau \sim \mathrm{Unif}(T)$, then executing $\pi^\tau = \pi_1^\tau \times \cdots \times \pi_m^\tau$, which can be done in a fully decentralized fashion.

We further assume that exploration policy mapping $\Gamma_{\mathrm{explore}}(\overline{\pi}, \pi')$ (which we recall is an ordered set of tuples $(\widetilde{\pi}, P)$) is *marginally executable* in the following sense: The ordering is known to all the players, and for each $(\widetilde{\pi}, P) \in \Gamma_{\mathrm{explore}}(\overline{\pi}, \pi')$ in an ordered fashion, $P$ is known to all players, and the marginal policy $\widetilde{\pi}_i$ (conditioning on the shared random bits) is known to the $i$-th player as long as the marginal policies $\overline{\pi}_i$ and $\pi'_i$ (conditioning on the shared random bits) are known to the $i$-th player.

We remark that this assumption is satisfied with typical choices of $\Gamma_{\mathrm{explore}}$, such as our instantiations in both the tabular case and the linear case. In particular, our tabular setting chooses $\Gamma_{\mathrm{explore}}(\overline{\pi}, \pi') = [(\overline{\pi}_{1:h-1} \circ \pi'_h, [m])]$, which directly satisfies marginal executability. For the linear setting, recall by (14) that we have chosen

$$\Gamma_{\mathrm{explore}}(\overline{\pi}, \pi') = \left[(\widetilde{\pi}, P) = \left(\overline{\pi}_{1:h-1} \times (\mathrm{Unif}(\mathcal{A}_i) \times \pi'_{-i,h}), \{i\}\right)\right]_{i=1}^{m}.$$

It is straightforward to let all players know and abide by the schedule of the $P$ (just round-robin over $\{i\}$ for $i \in [m]$ in lexicographic order). Further the marginal policy $\widetilde{\pi}_i$ of each $\widetilde{\pi}$ in this list is fully determined by $\overline{\pi}_i$ and one of $\{\mathrm{Unif}(\mathcal{A}_i), \pi'_i\}$ (depending on whether $i \in P$), which verifies the marginal executability assumption.

**VLPR** Observe that for the VLPR algorithm described in Algorithm 1-3, most of the steps (such as NO-REGRET-ALG and OPTIMISTIC-REGRESS) are by nature decentralized and can be executed by each player independently. The only coordinations involved are executing either the replay policy $\overline{\pi}^t = \mathrm{Unif}(\{\pi^\tau\}_{\tau\in[T]})$ (Line 1 in Algorithm 2), or the exploration policies $(\widetilde{\pi}, P) \in \Gamma_{\mathrm{explore}}(\overline{\pi}^t, \mu_h^k)$ within Algorithm 2 and $(\widetilde{\pi}, P) \in \Gamma_{\mathrm{explore}}(\overline{\pi}^t, \pi_h^{t+1})$ within Algorithm 3. Executing $\overline{\pi}^t$ can be done by using the shared randomness described above. Further, as both $\mu_{i,h}^k$ and $\pi_{i,h}^{t+1}$ are known to the $i$-th player, and by the marginal executability assumption, all the exploration policies can be executed in a decentralized fashion. This verified the claim for VLPR.

**AVLPR** The only difference in AVLPR over VLPR is to check the triggering condition in Line 4 of Algorithm 5, which in each iteration $t \in [T]$ requires one communication of $m$ bits, one for each player (indicator of whether the condition holds for player $i \in [m]$). The players will enter the replay part if the triggering condition holds for at least one player, and start the next episode otherwise. Since all players know whether they have entered the replay part in each iteration, the replay index $I_t$ is a common knowledge that can be maintained by all players simultaneously. Further, we can let this communication can be triggered only when the triggering condition holds, which by Condition 2 happens for at most $\widetilde{\mathcal{O}}(d_{\mathrm{replay}} \log T)$ times within $T$ iterations of play.

# Appendix C. Proofs for Section 4.1

## C.1. Details of the linear AVLPR Algorithm

**Understanding Assumption 3** Assumption 3 has the following implication, which is used throughout the design and analysis of the linear function approximation case.

**Remark 15** *Assumption 3 implies the following statement. For any $(i,h) \in [m] \times [H]$, any function $\overline{V} = \overline{V}_{i,h+1} : \mathcal{S} \to [0,H]$ and any policy $\pi \in \Pi^{\text{lin}}$, there exists $\theta^{h,\pi_{-i},\overline{V}} \in \mathbb{R}^d$ such that*

$$\mathbb{D}_{\delta_{a_i} \times \pi_{-i,h}} \left[ r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1} \right](s) = \phi_i(s,a_i)^\top \theta^{h,\pi_{-i},\overline{V}} \quad \text{for all } (s,a_i) \in \mathcal{S} \times \mathcal{A}_i. \tag{13}$$

This can be seen by picking $f_{i,h+1}(s,a_i) = \overline{V}(s)$ and applying Assumption 3.

**Choice of $\Psi_{i,h}$** The switching condition in Algorithm 5 is chosen as

$$\Psi_{i,h}(\mathcal{B}) := \log\det\left( I + \frac{1}{A_i} \sum_{s\in\mathcal{B}} \sum_{a_i \in \mathcal{A}_i} \phi_i(s,a_i)\phi_i(s,a_i)^\top \right).$$

**Processing $\mathcal{D}_{\text{init}}$ and $\mathcal{D}_{\text{sample}}$** For linear function approximation, the dataset $\mathcal{D}^i_{\text{init}}$ will then be used to compute $H$ feature covariance matrices $\{\widehat{\Sigma}^{\overline{\pi}}_{i,h}\}_{h\in[H]}$ that measures the coverage of the exploration policy $\overline{\pi}$, defined as

$$\widehat{\Sigma}^{\overline{\pi}}_{i,h} := \frac{1}{|\mathcal{D}^i_{\text{init}}| \cdot A_i} \sum_{s_h \in \mathcal{D}^i_{\text{init}}} \sum_{a_i \in \mathcal{A}_i} \phi_i(s_h,a_i)\phi_i(s_h,a_i)^\top.$$

Additionally we define the population version

$$\Sigma^{\overline{\pi}}_{i,h} = \mathbb{E}_{s_h \sim \overline{\pi}}\, \mathbb{E}_{a \sim \text{Unif}(\mathcal{A}_i)} \left[ \phi_i(s_h,a_i)\phi_i(s_h,a_i)^\top \right].$$

For linear function approximation we choose the exploration scheme $\Gamma_{\text{explore}}(\overline{\pi}, \mu_h)$ in Algorithm 2 and 3 as the ordered set

$$[(\overline{\pi}_{1:h-1} \circ (\text{Unif}(\mathcal{A}_1) \times \mu_{-1,h}), \{1\}), \cdots, (\overline{\pi}_{1:h-1} \circ (\text{Unif}(\mathcal{A}_m) \times \mu_{-m,h}), \{m\})]. \tag{14}$$

As a result, $\mathcal{D}^{k,i}_{\text{sample}}$ will contain exactly one element, denoted as $(s^k_h, a^k_{i,h}, y^k_{i,h})$.

**NO-REGRET-ALG** Condition (1A) can be understood as a state-wise regret bound with respect to the loss function $\ell^k_{i,h}(s,a_i) := \mathbb{D}_{\delta_{a_i} \times \mu^k_{-i,h}}[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}](s)$. As per Assumption 3, $\ell^k_{i,h}$ can be written as a linear function $\ell^k_{i,h}(s,a_i) = \langle \theta^k_{i,h}, \phi_i(s,a_i) \rangle$. In order to guarantee a state-wise regret, we first construct a linear estimator of $\ell^k_{i,h}$ for all $(s,a_i) \in \mathcal{S} \times \mathcal{A}_i$:

$$\widehat{\ell}^k_{i,h}(s,a_i) = \langle \widehat{\theta}^k_{i,h}, \phi_i(s,a_i) \rangle,$$

where

$$\widehat{\theta}^k_{i,h} := \left( \widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I \right)^{-1} \phi_i(s^k_{i,h}, a^k_{i,h}) y^k_{i,h}.$$

This estimator is also used in adversarial linear bandits (Neu and Olkhovskaya, 2020). However, directly running an exponential weights algorithm with this estimator would not work in our setting because Assumption 3 requires $\mu_h^k$ to lie in (the convex hull of) $\Pi^{\text{lin}}$; otherwise under $\mu_{-i}^k$ the resulting action-value function cannot be approximated with a linear function. To that end, we first make the observation that the per-state bandit regret (with the comparator in $\Delta(\mathcal{A}_i)$) can be equivalently viewed as the regret of an online linear optimization problem (with the comparator in the convex hull of the action feature vectors)

$$
\begin{aligned}
&\max_{\mu_{i,h}\in\Delta_{\mathcal{A}_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu_{-i,h}^k} - \mathbb{D}_{\mu_h^k}\right)(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1})(s) \\
&= \max_{\mu_{i,h}\in\Delta_{\mathcal{A}_i}} \sum_{k=1}^{K} \langle \mu_{i,h} - \mu_{i,h}^k(\cdot \mid s), \ell_{i,h}^k(s,\cdot)\rangle = \max_{\phi\in CH(\Phi_i(s))} \sum_{k=1}^{K} \langle \phi - \Phi_i(s)^\top \mu_{i,h}^k, \theta_{i,h}^k\rangle.
\end{aligned}
$$

Here $\Phi_i(s) \in \mathbb{R}^{A_i\times d}$ is a matrix that stacks all feature vectors $\{\phi_i(s,\cdot)\}$, while we slightly abuse notation to use $CH(\cdot)$ to denote the convex hull of the rows of the matrix.

We will then apply the Expected Follow-the-Perturbed-Leader algorithm (Hazan and Minasyan (2020, Algorithm 3); see also Hazan et al. (2016)) to the online linear optimization problem, namely choosing

$$
\Phi_i(s)^\top \mu_{i,h}^k = \mathbb{E}_{v\sim\mathcal{V}}\left[\operatorname*{arg\,max}_{\phi\in CH(\Phi_i(s))} \langle \phi, \sum_{k'\le k} \widehat{\theta}_{i,h}^{k'} + v/\eta\rangle\right],
$$

where $\mathcal{V}$ is chosen as the uniform distribution over the ellipse $\{u \mid u^\top(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)u \le 1\}$, and $\eta$ is a parameter that plays a role similar to learning rate. This induces the following policy

$$
\mu_{i,h}^{k+1}(a_i|s) := \Pr_{v\sim\mathcal{V}}\left[a_i = \operatorname*{arg\,max}_{a_i'\in\mathcal{A}_i} \langle \phi_i(s,a_i'), \sum_{k'\le k} \widehat{\theta}_{i,h}^{k'} + v/\eta\rangle\right], \tag{15}
$$

which lies in the convex hull of $\Pi^{\text{lin}}$ and therefore satisfies the requirement of Assumption 3.

**Optimistic-Regress** The optimistic regression is implemented using ridge regression on the dataset $\mathcal{D}_{\text{reg}}^i$, which contains samples of $(s_h, a_{i,h}, y_{i,h})$ where $y_{i,h} = r_{i,h} + \overline{V}_{i,h+1}(s_{h+1})$. More specifically,

$$
\begin{aligned}
\widehat{\theta}_{i,h} &\leftarrow \operatorname*{arg\,min}_{\theta} \frac{1}{K} \sum_{(s_h,a_{i,h},y_{i,h})\in\mathcal{D}_{\text{reg}}^i} \left[\phi_i(s_h,a_{i,h})^\top\theta - y_{i,h}\right]^2 + \lambda\|\theta\|_2^2, \\
\overline{Q}_{i,h}(s,a_i) &\leftarrow \left(\phi_i(s,a_i)^\top \widehat{\theta}_{i,h} + \frac{3}{2} G_{i,h}(s,\overline{\pi},K,\delta)\right) \wedge (H-h+1), \\
\overline{V}_{i,h}(s) &\leftarrow \langle \pi_{i,h}(\cdot \mid s), \overline{Q}_{i,h}(s,\cdot)\rangle.
\end{aligned}
$$

**Computational efficiency** We remark here that $\overline{V}_{i,h}(s)$ does not need to be computed for every $s$ but only for states in the dataset, which can be done in polynomial time. Also, the policy in (15) does not need to be fully computed either, because executing the algorithm only requires an efficient sampling from the policy $\mu_{i,h}^{k+1}$, which can in turn easily achieved by sampling $v \sim \mathcal{V}$.

### C.2. Proof of Condition (1A)

As outlined in Section 4.1, we will first decompose the the per-state regret in Condition (1A) as the per-state regret measured on the loss estimator and statistical error terms:

$$\begin{aligned}
&\max_{\mu_{i,h}\in\Delta_{\mathcal{A}_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu_{-i,h}^k} - \mathbb{D}_{\mu_h^k}\right)(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1})(s) \\
&= \underbrace{\max_{\phi\in CH(\Phi_i(s))} \sum_{k=1}^{K} \langle \phi - \Phi_i(s)\mu_{i,h}^k(\cdot\mid s), \widehat{\theta}_{i,h}^k\rangle}_{(A)} + \underbrace{\max_{\phi\in CH(\Phi_i(s))} \sum_{k=1}^{K} \langle \phi, \theta_{i,h}^k - \widehat{\theta}_{i,h}^k\rangle}_{(B)} \\
&\quad + \underbrace{\sum_{k=1}^{K} \langle \Phi_i(s)\mu_{i,h}^k(\cdot\mid s), \widehat{\theta}_{i,h}^k - \theta_{i,h}^k\rangle}_{(C)}.
\end{aligned}$$

In Appendix D, we prove that under the choice of $\eta = 1/(dH\sqrt{K\log\delta^{-1}})$ and $\lambda = \widetilde{\Theta}(d\max_i A_i/K)$, the above three terms can be respectively controlled as following: with probability at least $1-\delta$, for all $s\in\mathcal{S}$

$$\begin{aligned}
\text{Term (A)} &\le \sup_{a_i\in\mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}} \cdot \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)}\right), \\
\text{Term (B)} &\le \sup_{a_i\in\mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)}\right), \\
\text{Term (C)} &\le \sup_{a_i\in\mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right) + \mathcal{O}(1).
\end{aligned}$$

As a result, we can pick

$$G_{i,h}(s,\overline{\pi},K,\delta) = \sup_{a_i\in\mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}} \times \widetilde{\Theta}\left(\frac{dHA_i^{1.5}}{\sqrt{K}}\right) + \Theta\left(\frac{1}{K}\right).$$

### C.3. Proof of Condition (1B)

Consider a fixed $(i,h)\in[m]\times[H]$. Denote $\mathcal{D}_{\text{reg}}^i$ as

$$\left\{(s_h^j, a_{i,h}^j, y_{i,h}^j)\right\}_{j\in[K]}.$$

By Assumption 3, there exists $\theta_{i,h}^*$ such that for all $j$,

$$\mathbb{E}[y_{i,h}^j | s_h^j, a_{i,h}^j] = \mathbb{D}_{\delta_{a_{i,h}^j}\times\pi_{-i,h}}[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}](s_h^j) = \langle \phi_i(s_h^j, a_{i,h}^j), \theta_{i,h}^*\rangle.$$

Define $\widehat{\Sigma}_{\text{reg},h} = \frac{1}{K}\sum_{j=1}^{K} \phi_i(s_h^j, a_{i,h}^j)\phi_i(s_h^j, a_{i,h}^j)^\top + \lambda I$ and $\zeta_j = y_{i,h}^j - \mathbb{P}_h\left[(\overline{V}_{i,h+1} + r_{i,h})\right]\left(s_h^j, a_{i,h}^j\right)$. Here $\zeta_j$ is mean-zero and $H$-bounded. It follows that $\forall (s,a_i)\in\mathcal{S}\times\mathcal{A}_i$

$$\left|\phi_i(s,a_i)^\top\widehat{\theta}_{i,h} - \mathbb{D}_{\delta_{a_i}\times\pi_{-i,h}}[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}](s)\right|$$

$$
\begin{aligned}
&= \left| \phi_i(s, a_i)^\top \widehat{\theta}_{i,h} - \phi_i(s, a_i)^\top \theta^\star_{i,h} \right| \\
&= \left| \phi_i(s, a_i)^\top \widehat{\Sigma}^{-1}_{\text{reg},h} \frac{1}{K} \sum_{j=1}^{K} \phi_i(s_h^j, a_{i,h}^j) \left( \phi_i(s_h^j, a_{i,h}^j)^\top \theta^\star_{i,h} + \zeta_j \right) - \phi_i(s, a_{i,h})^\top \theta^\star_{i,h} \right| \\
&\leq \|\phi_i(s, a_{i,h})\|_{\widehat{\Sigma}^{-1}_{\text{reg},h}} \times \left( \left\| \frac{1}{K} \sum_{j=1}^{K} \phi_i(s_h^j, a_{i,h}^j) \zeta_j \right\|_{\widehat{\Sigma}^{-1}_{\text{reg},h}} + \sqrt{\lambda} B_\theta \right).
\end{aligned}
$$

**Lemma 16** *Suppose we pick* $\lambda = \Theta(d \log(dK/\delta)/K)$*, then with probability* $1 - \delta$

$$
\left\| \sum_{j=1}^{K} \phi_i(s_h^j, a_{i,h}^j) \zeta_j \right\|_{\Sigma^{-1}_{\text{reg},h}} \leq \mathcal{O}\left( \sqrt{KdH^2} \log(KdH/\delta) \right).
$$

The proof of this lemma is identical to that of Lemma 19. Finally note that by Lemma 17, with probability $1 - \delta$,

$$
\widehat{\Sigma}_{\text{reg},h} \succcurlyeq \frac{1}{2} \Sigma^{\overline{\pi}}_{i,h} + \lambda I - \mathcal{O}\left( \frac{d \log(dK/\delta)}{K} \right) I \succcurlyeq \frac{1}{2} \left( \Sigma^{\overline{\pi}}_{i,h} + \lambda I \right).
$$

Therefore

$$
\begin{aligned}
&\left| \phi_i(s, a_i)^\top \widehat{\theta}_{i,h} - \left[ \mathbb{P}_h^{\pi^t} (\overline{V}_{i,h+1} + r_{i,h}) \right] (s, a_i) \right| \\
&\quad \leq \|\phi_i(s, a_i)\|_{\widehat{\Sigma}^{-1}_{\text{reg},h}} \cdot \mathcal{O}\left( dH \sqrt{1/K} \log(dK/\delta) \right) \\
&\quad \leq \|\phi_i(s, a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \cdot \mathcal{O}\left( dH \sqrt{1/K} \log(dK/\delta) \right) \\
&\quad \leq \frac{1}{2} G_{i,h}(s, \overline{\pi}, K, \delta).
\end{aligned}
$$

We conclude that $\forall (s, a_i)$

$$
\begin{aligned}
&\left( \left[ \mathbb{P}_h^{\pi^t} (\overline{V}_{i,h+1} + r_{i,h}) \right] (s, a) + G_{i,h}(s, \overline{\pi}, K, \delta) \right) \wedge (H - h + 1) \leq \overline{Q}_{i,h}(s, a) \\
&\leq \left( \mathbb{P}_h^{\pi^t} (\overline{V}_{i,h+1} + r_{i,h}) (s, a) + 2 G_{i,h}(s, \overline{\pi}, K, \delta) \right) \wedge (H - h + 1).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\min \left\{ \mathbb{D}_{\pi_h} \left[ r_{i,h} + \mathbb{P}_{h+1} \overline{V}_{i,h+1} \right] (s) + G_{i,h}(s, \overline{\pi}, K, \delta), H - h + 1 \right\} &\leq \overline{V}_{i,h}(s), \\
\overline{V}_{i,h}(s) \leq \mathbb{D}_{\pi_h} \left[ r_{i,h} + \mathbb{P}_{h+1} \overline{V}_{i,h+1} \right] (s) + 2 G_{i,h}(s, \overline{\pi}, K, \delta).&
\end{aligned}
$$

### C.4. Proof of Condition (1C)

Denote

$$
X_t := \mathbb{E}\left[ \phi_i(s_h, a_{i,h}) \phi_i(s_h, a_{i,h})^\top \mid s_h \sim \pi^t_{1:h-1},\ a_{i,h} \sim \text{Unif}(\mathcal{A}_i) \right], \qquad S_t := \sum_{\tau=1}^{t} X_\tau + \lambda_0 I_{d \times d},
$$

where $\lambda_0 = \widetilde{\mathcal{O}}(d)$. Then using the definition of $G_{i,h}$ in Equation (3),

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{\pi^{t+1}} \left[G_{i,h}(s, \overline{\pi}^t, t, \delta)\right] \\
\leq& \widetilde{\mathcal{O}}\left(\frac{d(\max_i A_i)^{1.5} H}{\sqrt{t}} \cdot \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{t} \max_{a_{i,h}\in\mathcal{A}_i} \|\phi(s_h, a_{i,h})\|_{S_t^{-1}} \mid s_h \sim \pi_{1:h-1}^{t+1}\right]\right) + \widetilde{\mathcal{O}}\left(1\right) \\
\leq& \widetilde{\mathcal{O}}\left(d(\max_i A_i)^{2.5} H \cdot \sum_{t=1}^{T} \mathbb{E}\left[\|\phi(s_h, a_{i,h})\|_{S_t^{-1}} \mid s_h \sim \pi_{1:h-1}^{t+1}, a_{i,h} \sim \mathrm{Unif}(\mathcal{A}_i)\right]\right) + \widetilde{\mathcal{O}}\left(1\right) \\
\leq& \widetilde{\mathcal{O}}\left(d(\max_i A_i)^{2.5} H \cdot \sqrt{T \cdot \sum_{t=1}^{T} \mathbb{E}\left[\|\phi(s_h, a_{i,h})\|^2_{S_t^{-1}} \mid s_h \sim \pi_{1:h-1}^{t+1}, a_{i,h} \sim \mathrm{Unif}(\mathcal{A}_i)\right]}\right) + \widetilde{\mathcal{O}}\left(1\right) \\
\leq& \widetilde{\mathcal{O}}\left(d(\max_i A_i)^{2.5} H \cdot \sqrt{T \cdot \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{tr}(X_{t+1} S_t^{-1})\right]}\right) + \widetilde{\mathcal{O}}\left(1\right) = \widetilde{\mathcal{O}}\left(\sqrt{d^3(\max_i A_i)^5 H^2 T}\right).
\end{aligned}$$

### C.5. Proof of Condition 2

Let us fix $(i, h, t) \in [m] \times [H] \times [T]$. Define $\widehat{S}_t := I + \frac{1}{A_i} \sum_{s\in\mathcal{B}_h^t} \sum_{a_i\in\mathcal{A}_i} \phi_i(s, a_i)\phi_i(s, a_i)^\top$. Then

$$\Psi_{i,h}(\mathcal{B}_h^t) - \Psi_{i,h}(\mathcal{B}_h^{I_t}) = \log\det\left(\widehat{S}_t \widehat{S}_{I_t}^{-1}\right).$$

Therefore that $\Psi_{i,h}(\mathcal{B}_h^t) - \Psi_{i,h}(\mathcal{B}_h^{I_t}) \leq 1$ implies

$$\|\widehat{S}_t^{\frac{1}{2}} \widehat{S}_{I_t}^{-1} \widehat{S}_t^{\frac{1}{2}}\|_2 \leq 2,$$

which further implies

$$\widehat{S}_t \preccurlyeq 2\widehat{S}_{I_t}.$$

In other words, to prove Condition 2 it suffices to show that $\widehat{S}_t \preccurlyeq 2\widehat{S}_{I_t}$ implies

$$t(\Sigma_{i,h}^{\overline{\pi}^t} + \lambda_t I) \leq 8I_t\left(\Sigma_{i,h}^{\overline{\pi}^{I_t}} + \lambda_{I_t} I\right),$$

where $\lambda_t = \widetilde{\Theta}(d \max_i A_i/t)$. This is equivalent to showing that

$$t(\Sigma_{i,h}^{\overline{\pi}^t} + \lambda_t I) \geq 8I_t\left(\Sigma_{i,h}^{\overline{\pi}^{I_t}} + \lambda_{I_t} I\right),$$

implies $\widehat{S}_t \succcurlyeq 2\widehat{S}_{I_t}$. By Lemma 17, with probability $1 - \delta$,

$$\widehat{S}_t \succcurlyeq \frac{t}{2}\Sigma_{i,h}^{\overline{\pi}^t} - \widetilde{\Theta}(d)I \succcurlyeq 4I_t\Sigma_{i,h}^{\overline{\pi}^{I_t}} + \widetilde{\Theta}(d)I \succcurlyeq 2\widehat{S}_{I_t}.$$

Finally taking a union bound w.r.t. $i$, $h$ and $t$ proves part (a) of the condition.

As for the second part, we make the observation that $\Psi_{i,h}(\emptyset) = \log\det I = 1$, and

$$\Psi_{i,h}(\mathcal{D}_h^t) \leq \log\det \widehat{S}_T \leq d\log\left(\|S_T\|_2\right) \leq d\log T.$$

Therefore the total number of switches is at most $dmH\log T$, *i.e.* part (b) is satisfied with $d_{\mathrm{replay}} = dmH$.

### C.6. Sample complexity for linear function approximation

Sections C.2 through C.3 show that Conditions (1A) through (1C) are satisfied with

$$G_{i,h}(s,\overline{\pi},K,\delta)=\sup_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}}\times\widetilde{\Theta}\left(\frac{dH(\max_i A_i)^{1.5}}{\sqrt{K}}\right)+\Theta\left(\frac{1}{K}\right).$$

and

$$L=\widetilde{\mathcal{O}}\left(\max_{i\in[m]}d^3(\max_i A_i)^5H^2\right).$$

Finally Section C.5 verified that Condition 2 is satisfied with $d_{\text{replay}}=dmH$. By (14), $\overline{\Gamma}=m$. Therefore by applying Theorem 13, we obtain following the sample complexity bound for finding an $\varepsilon$-CCE

$$\widetilde{\mathcal{O}}\left(\frac{H^3L\overline{\Gamma}d_{\text{replay}}}{\varepsilon^2}\right)=\widetilde{\mathcal{O}}\left(\frac{d^4m^2H^6\max_{i\in[m]}A_i^5}{\varepsilon^2}\right).$$

## Appendix D. Proofs for Appendix C.2

### D.1. Relative concentration

Consider the following random process: at time step $t$, we (randomly) picks a distribution $D_t$ over the $d$-dimensional unit ball based on $\{x_\tau\}_{\tau\in[t-1]}$, and then sample $x_t\sim D_t$. Denote by $\Sigma_t$ the covariance matrix of $D_t$. We have the following relative concentration lemma regarding the closeness between the empirical temporal-average covariance and the population one in the multiplicative sense.

**Lemma 17** *With probability at least $1-\delta$, for all $t\in[T]$*

$$\frac{1}{2}\sum_{\tau\in[t]}\Sigma_\tau-\beta I\preceq\sum_{\tau\in[t]}x_\tau x_\tau^\top\preceq 2\sum_{\tau\in[t]}\Sigma_\tau+\beta I$$

*where $\beta=\Theta(d\log(dT/\delta))$.*

**Proof** Let us first fix $t\in[T]$. Fix any $w\in\mathbb{R}^d$ with $\|w\|_2=1$. Define $W_\tau=\langle x_\tau,w\rangle^2$. It follows that $\mathbb{E}[W_\tau]=w^\top\Sigma_\tau w$, and $0\le W_\tau\le 1$. By Bernstein's inequality, with probability $1-\delta'$

$$\left|\sum_{\tau\in[t]}(W_\tau-\mathbb{E}[W_\tau])\right|\le\sqrt{4\log(1/\delta')\sum_{\tau\in[t]}\mathbb{E}[W_\tau^2]}+\mathcal{O}(\log 1/\delta') \tag{16}$$

$$\le\sqrt{4\log(1/\delta')\sum_{\tau\in[t]}\mathbb{E}[W_\tau]}+\mathcal{O}(\log 1/\delta') \tag{17}$$

$$\le\frac{1}{2}\sum_{\tau\in[t]}\mathbb{E}[W_\tau]+\mathcal{O}(\log 1/\delta'). \tag{18}$$

Therefore with probability $1-\delta'$

$$\sum_{\tau\le t}W_\tau\le 2\mathbb{E}\left[\sum_{\tau\le t}W_\tau\right]+O\left(\log\left(\frac{1}{\delta'}\right)\right),$$

$$\sum_{\tau \le t} W_\tau \ge \frac{1}{2}\mathbb{E}\left[\sum_{\tau \le t} W_\tau\right] - O\left(\log\left(\frac{1}{\delta'}\right)\right)$$

It remains to construct an $\varepsilon'$-cover $W_{\varepsilon'}$ of the $d$-sphere, where we choose $\varepsilon' = 0.01/T$. It follows that with probability $1-\delta$, for all $w$ in the unit sphere,

$$\sum_{\tau \le t}\langle w, x_\tau\rangle^2 \le 2\mathbb{E}[\sum_{\tau \le t}\langle w, x_\tau\rangle^2] + \mathcal{O}(\log(|W_{\varepsilon'}|/\delta),$$
$$\sum_{\tau \le t}\langle w, x_\tau\rangle^2 \ge \frac{1}{2}\mathbb{E}[\sum_{\tau \le t}\langle w, x_\tau\rangle^2] - \mathcal{O}(\log(|W_{\varepsilon'}|/\delta).$$

This implies

$$\frac{1}{2}\sum_{\tau\in[t]}\Sigma_\tau - \mathcal{O}(\log(|W_{\varepsilon'}|/\delta)I \preceq \sum_{\tau\in[t]} x_\tau x_\tau^\top \preceq 2\sum_{\tau\in[t]}\Sigma_\tau + \mathcal{O}(\log(|W_{\varepsilon'}|/\delta))I$$

Replacing $\delta$ by $\delta/T$ and plugging in $|W_{\varepsilon'}| \le \left(\frac{3}{\varepsilon'}\right)^d$ (Vershynin, 2018, Corollary 4.2.13) proves the lemma. ■

### D.2. Controlling Term (A) in Condition (1A)

In order to evoke the analysis of Expected FPL, we make the observation that

$$\begin{aligned}
\text{Term (A)} &= \max_{\mu_i\in\Delta_{\mathcal{A}_i}}\sum_{k=1}^K\langle \mu_i - \mu_{i,h}^k(\cdot\mid s), \widehat{\ell}_{i,h}^k(s,\cdot)\rangle \\
&= \max_{\mu_i\in\Delta_{\mathcal{A}_i}}\sum_{k=1}^K\langle \mu_i - \mu_{i,h}^k(\cdot\mid s), \Phi_i(s,\cdot)^\top\widehat{\theta}_{i,h}^k\rangle \\
&= \max_{x\in CH(\{\phi_i(s,\cdot)\})}\sum_{k=1}^K\langle x - \Phi_i(s,\cdot)\mu_{i,h}^k(\cdot|s), \widehat{\theta}_{i,h}^k\rangle.
\end{aligned}$$

Note that in our algorithm,

$$\mu_{i,h}^k(a_i|s) := \Pr_{v\sim\mathcal{V}}\left[a_i = \arg\max\left\langle \phi_i(s,\cdot), \sum_{k'<k}\widehat{\theta}_{i,h}^{k'} + \frac{1}{\eta}(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{-1/2}v\right\rangle\right],$$

which implies

$$\Phi_i(s,\cdot)\mu_{i,h}^k(\cdot|s) = \mathbb{E}_{v\sim\mathcal{V}}\argmax_{x\in CH(\{\phi_i(s,\cdot)\})}\left\langle(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{-1/2}x, (\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{1/2}\sum_{k'<k}\widehat{\theta}_{i,h}^{k'} + \frac{1}{\eta}v\right\rangle.$$

This is identical to the Expected Follow-the-Perturbed-Leader algorithm (see *e.g.* (Hazan et al., 2016, Algorithm 17)) on a sequence of linear loss vectors

$$(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{1/2}\widehat{\theta}_{i,h}^1, \cdots, (\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{1/2}\widehat{\theta}_{i,h}^K.$$

Therefore it follows from the regret of Expected FPL (Hazan and Minasyan, 2020, Theorem 10) that, by choosing $\mathcal{V}$ to be the uniform distribution over the $d$-dimensional unit ball,

$$\mathrm{Term(A)} \le \sup_{a_i \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \cdot \left[ \frac{1}{\eta} + \eta d \sum_{k=1}^{K} \|\widehat{\theta}^k_{i,h}\|^2_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I)} \right].$$

By Lemma 18, with probability at least $1 - \delta$

$$\sum_{k=1}^{K} \|\widehat{\theta}^k_{i,h}\|^2_{\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I} = \mathcal{O}\left( dK + \frac{\log \delta^{-1}}{\lambda} \right).$$

Therefore, by plugging in $\eta = 1/(dH\sqrt{(\max_i A_i) K \log \delta^{-1}})$ and $\lambda = \widetilde{\Theta}(d(\max_i A_i)/K)$, we have

$$\begin{aligned}
\mathrm{Term(A)} &\le \sup_{a \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \cdot \mathcal{O}\left( dH \sqrt{K(\max_i A_i) \log \delta^{-1}} \right) \\
&= \sup_{a \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \cdot \mathcal{O}\left( dH \sqrt{K(\max_i A_i) \log \delta^{-1}} \right),
\end{aligned}$$

where the equality follows from Lemma 17.

**Lemma 18** *With probability* $1 - \delta$*,*

$$\sum_{k=1}^{K} \|\widehat{\theta}^k_{i,h}\|^2_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} = \mathcal{O}\left( dKH^2 + \frac{H^2 \log \delta^{-1}}{\lambda} \right).$$

**Proof** Define

$$x_k := \phi_i(s^k_{i,h}, a^k_{i,h}), \quad z_k := x_k^\top (\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I)^{-1} x_k.$$

By definition

$$\sum_{k=1}^{K} \|\widehat{\theta}^k_{i,h}\|^2_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \le H^2 \sum_{k=1}^{K} z_k.$$

Moreover, $\{z_k\}_{k \in [K]}$ are i.i.d. samples satisfying that

$$\mathbb{E}[z_k] = \mathrm{Tr}\left( \Sigma^{\overline{\pi}}_{i,h} \left( \widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I \right)^{-1} \right) \le \mathcal{O}(d),$$

where the inequality follows from Lemma 17 and the choice of $\lambda$, and

$$\mathbb{E}[z_k^2] \le \frac{1}{\lambda} \cdot \mathbb{E}[z_k] \le \mathcal{O}\left( \frac{d}{\lambda} \right),$$

and

$$|z_k| \le \frac{1}{\lambda}.$$

Therefore by Bernstein's inequality, with high probability

$$\sum_k z_k \le \mathcal{O}\left( dK + \sqrt{\frac{dK \log \delta^{-1}}{\lambda}} + \frac{\log \delta^{-1}}{\lambda} \right) = \mathcal{O}\left( dK + \frac{\log \delta^{-1}}{\lambda} \right).$$

■

### D.3. Controlling Term (B) in Condition (1A)

Consider a fixed player $i \in [m]$ and step $h \in [H]$. To simplify notations, denote

$$x_k := \phi_i(s_{i,h}^k, a_{i,h}^k), \quad y_k := r_{i,h}^k + V_{i,h+1}(s_{h+1}^k), \quad \zeta_k := y_k - x_k^\top \theta_{i,h}^k.$$

For any $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$:

$$\begin{aligned}
&\sum_{k=1}^K \langle \phi_i(s,a_i), \widehat{\theta}_{i,h}^k - \theta_{i,h}^k \rangle \\
=& \left\langle \phi_i(s,a_i), \sum_{k=1}^K \left( \widehat{\theta}_{i,h}^k - \theta_{i,h}^k \right) \right\rangle \\
=& \left\langle \phi_i(s,a_i), \sum_{k=1}^K \left( \left( \widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1} x_k y_k - \theta_{i,h}^k \right) \right\rangle \\
=& \phi_i(s,a)^\top \left( \widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1} \left[ \sum_{k=1}^K x_k \left( x_k^\top \theta_{i,h}^k + \zeta_k \right) - \left( \widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right) \sum_{k=1}^K \theta_{i,h}^k \right] \\
\le& \|\phi_i(s,a_i)\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} \left[ \sqrt{\lambda} B_\theta K + \underbrace{\left\| \sum_{k=1}^K x_k \zeta_k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}}}_{\text{Term (B1)}} \right. \\
&\left. + \underbrace{\left\| \sum_{k=1}^K \left( x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\overline{\pi}} \right) \theta_{i,h}^k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}}}_{\text{Term (B2)}} \right].
\end{aligned}$$

By Lemma 19, 20, the choice of $\lambda = \widetilde{\mathcal{O}}(d(\max_i A_i)/K)$ and relative concentration (Lemma 17),

$$\text{Term (B)} \le \widetilde{\mathcal{O}}\left( \|\phi_i(s,a_i)\|_{\left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} \times Hd\sqrt{K(\max_i A_i)} \right).$$

**Lemma 19 (Term (B1))** *With probability at least $1-\delta$, we have*

$$\left\| \sum_{k=1}^K x_k \zeta_k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} = \mathcal{O}\left( \sqrt{KdH^2 \log(KdH/\delta)} + \frac{dH \log(KdH/\delta)}{\sqrt{\lambda}} \right).$$

**Proof** Consider a fixed $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$. Define

$$z_k := v^\top \left( \widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1/2} x_k \zeta_k.$$

Note that $\{z_k\}_{k=1}^K$ is a martingale with conditional variance and range bounded by

$$|z_k| \le H\lambda^{-1/2},$$

and

$$\begin{aligned}\mathrm{Var}(z_k \mid z_{1:k-1}) &= \mathbb{E}\left[\zeta_k^2 v^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} x_k x_k^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} v\right] \\ &\le H^2 \left\| \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} \Sigma_{i,h}^{\overline{\pi}} \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} \right\|_2 \le \mathcal{O}(H^2),\end{aligned}$$

where the second inequality uses Lemma 17, the definition of $\widehat{\Sigma}_{i,h}^{\overline{\pi}}$ and the choice of $\lambda$.

By Freedman inequality,

$$\left|\sum_{k=1}^K z_k\right| \le \mathcal{O}\left(\sqrt{KH^2 \log \delta^{-1}} + \frac{H \log \delta^{-1}}{\sqrt{\lambda}}\right).$$

Finally, by taking a union bound for all $v$ from a $(\sqrt{\lambda}/(HK))$-cover of the $d$-dimensional unit ball, we conclude that

$$\begin{aligned}\left\|\sum_{k=1}^K x_k \zeta_k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} &= \max_{v:\, \|v\|_2 = 1} \left| v^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} \sum_{k=1}^K x_k \zeta_k \right| \\ &\le \mathcal{O}\left(\sqrt{KdH^2 \log(KdH/\delta)} + \frac{dH \log(KdH/\delta)}{\sqrt{\lambda}}\right).\end{aligned}$$

■

**Lemma 20 (Term (B2))** *With probability at least* $1-\delta$, *we have*

$$\left\|\sum_{k=1}^K \left(x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right) \theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} = \mathcal{O}\left(\sqrt{KdB_\theta^2 \log(KdB_\theta/\delta)} + \frac{dB_\theta \log(KdB_\theta/\delta)}{\sqrt{\lambda}}\right).$$

**Proof** By triangle inequality and relative concentration (Lemma 17), we have

$$\begin{aligned}&\left\|\sum_{k=1}^K \left(x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right) \theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} \\ \le &\left\|\sum_{k=1}^K \left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right) \theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} + \left\|\left(\Sigma_{i,h}^{\overline{\pi}} - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right) \sum_{k=1}^K \theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} \\ \le &\mathcal{O}\left(\left\|\sum_{k=1}^K \left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right) \theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} + \left\|\left(\Sigma_{i,h}^{\overline{\pi}} - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right) \sum_{k=1}^K \theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}}\right).\end{aligned}$$

Consider an arbitrary $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$. Define

$$z_k := v^\top \left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} \left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right) \theta_{i,h}^k.$$

Notice that $\{z_k\}_{k=1}^K$ is a martingale with conditional variance and range bounded by

$$|z_k| \le B_\theta \lambda^{-1/2},$$

and

$$\begin{aligned}
\mathrm{Var}(z_k \mid z_{1:k-1}) &\le \mathbb{E}\left[\left(v^\top \left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} x_k x_k^\top \theta_{i,h}^k\right)^2\right] \\
&\le B_\theta^2 \mathbb{E}\left[\left(v^\top \left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} x_k\right)^2\right] \\
&= B_\theta^2 \mathbb{E}\left[v^\top \left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} x_k x_k^\top \left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} v\right] \\
&= B_\theta^2 v^\top \left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} (\Sigma_{i,h}^{\overline{\pi}}) \left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1/2} v \le \mathcal{O}(B_\theta^2),
\end{aligned}$$

where the second equality uses the fact that $\mathbb{E}[x_k x_k^\top] = \Sigma_{i,h}^{\overline{\pi}}$ and the last inequality uses Lemma 17.

By Freedman inequality,

$$\left|\sum_{k=1}^K z_k\right| \le \mathcal{O}\left(\sqrt{K B_\theta^2 \log \delta^{-1}} + \frac{B_\theta \log \delta^{-1}}{\sqrt{\lambda}}\right).$$

Finally, by taking a union bound for all $v$ from a $(\sqrt{\lambda}/(B_\theta K))$-cover of the $d$-dimensional unit ball, we conclude that

$$\begin{aligned}
&\left\|\sum_{k=1}^K \left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right) \theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}} \\
\le& \mathcal{O}\left(\sqrt{K d B_\theta^2 \log(K d B_\theta/\delta)} + \frac{d B_\theta \log(K d B_\theta/\delta)}{\sqrt{\lambda}}\right).
\end{aligned}$$

Now recall that $\widehat{\Sigma}_{i,h}^{\overline{\pi}}$ is estimated by using $K$ samples i.i.d. sampled from $\overline{\pi}^t$, so we can simply repeat the above concentration arguments for controlling $\left\|\sum_{k=1}^K \left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right) \theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}}$ to upper bound $\left\|\left(\Sigma_{i,h}^{\overline{\pi}} - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right) \sum_{k=1}^K \theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I\right)^{-1}}$, which results in the same bound as above. ■

### D.4. Controlling Term (C) in Condition (1A)

Consider a fixed player $i \in [m]$ and step $h \in [H]$. To simplify notations, denote

$$x_k := \phi_i(s_{i,h}^k, a_{i,h}^k), \quad y_k := r_{i,h}^k + V_{i,h+1}(s_{h+1}^k), \quad \zeta_k := y_k - x_k^\top \theta_{i,h}^k.$$

We have the following error decomposition similar to the one in controlling Term (B): for any $s \in \mathcal{S}$,

$$
\begin{aligned}
\text{Term (C)} &= \sum_{k=1}^{K} \langle \Phi_s \mu_{i,h}^k(\cdot \mid s), \theta_{i,h}^k - \widehat{\theta}_{i,h}^k \rangle \\
&= \sum_{a_i \in \mathcal{A}_i} \phi_i(s,a_i)^\top \left( \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) \theta_{i,h}^k - \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) \widehat{\theta}_{i,h}^k \right) \\
&= \sum_{a_i \in \mathcal{A}_i} \phi_i(s,a_i)^\top \left( \widehat{\Sigma}_{i,h}^{\pi} + \lambda I \right)^{-1} \Bigg[ \left( \widehat{\Sigma}_{i,h}^{\pi} + \lambda I \right) \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) \theta_{i,h}^k \\
&\qquad - \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) x_k \left( x_k^\top \theta_{i,h}^k + \zeta_k \right) \Bigg] \qquad (19) \\
&\le \sum_{a_i \in \mathcal{A}_i} \|\phi_i(s,a_i)\|_{\left(\widehat{\Sigma}_{i,h}^{\pi} + \lambda I\right)^{-1}} \Bigg[ \sqrt{\lambda} B_\theta K + \underbrace{\left\| \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) x_k \zeta_k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\pi} + \lambda I\right)^{-1}}}_{(C1)} \\
&\quad + \underbrace{\left\| \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) \left( x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\pi} \right) \theta_{i,h}^k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\pi} + \lambda I\right)^{-1}}}_{(C2)} \Bigg].
\end{aligned}
$$

It is easy to verify that the same arguments for bounding Term (B1) and (B2) can be used to bound Term (C1) and (C2), respectively. Formally, we have the following counterparts of Lemma 19 and 20 for bounding Term (C1) and (C2).

**Lemma 21 (Term (C1))** *Consider a fixed pair of state and action $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ and a unit vector $v \in \mathbb{R}^d$. With probability at least $1 - \delta$, we have*

$$
\begin{aligned}
&\left| v^\top \left( \widehat{\Sigma}_{i,h}^{\pi} + \lambda I \right)^{-1/2} \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) x_k \zeta_k \right| \\
&= \mathcal{O}\left( \sqrt{K H^2 \log(1/\delta)} + \frac{H \log(1/\delta)}{\sqrt{\lambda}} \right).
\end{aligned}
$$

**Lemma 22 (Term (C2))** *Consider a fixed pair of state and action $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ and a unit vector $v \in \mathbb{R}^d$. With probability at least $1 - \delta$, we have*

$$
\begin{aligned}
&\left| v^\top \left( \widehat{\Sigma}_{i,h}^{\pi} + \lambda I \right)^{-1/2} \sum_{k=1}^{K} \mu_{i,h}^k(a_i \mid s) \left( x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\pi} \right) \theta_{i,h}^k \right| \\
&= \mathcal{O}\left( \sqrt{K B_\theta^2 \log(1/\delta)} + \frac{B_\theta \log(1/\delta)}{\sqrt{\lambda}} \right).
\end{aligned}
$$

The proofs of Lemma 21 and 22 follow almost the same as the first half of Lemma 19 and 20 (before taking the union bound) respectively, so we omit them here.

To control Term (C) with Lemma 21 and 22, we needs to take a union bound for all state and action $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ and unit vector $v \in \mathbb{R}^d$. The following lemma essentially says that such union bound will only incur an additional factor of $\widetilde{\mathcal{O}}(dA_i)$ in the upper bound.

**Lemma 23** *Consider a policy $\pi$ defined as*

$$\pi_{i,h}(a_i|s) := \Pr_{v\sim\mathcal{V}}\left[a_i = \arg\max\langle\phi_i(s,\cdot), w + Wv\rangle\right],$$

*where $w \in \mathbb{R}^d$ s.t. $\|w\|_2 \le \gamma$, $\alpha I_{d\times d} \preceq W \preceq \beta I_{d\times d}$, $\mathcal{V}$ denotes the uniform distribution over the d-dimensional unit ball. Then for any states $s, s' \in \mathcal{S}$ satisfying $\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i) - \phi_i(s',a_i)\|_2 \le \varepsilon$, we have*

$$\left\|\mathbb{E}_{a\sim\pi_{i,h}(\cdot|s)}[\phi_i(s,a_i)] - \mathbb{E}_{a\sim\pi_{i,h}(\cdot|s')}[\phi_i(s',a_i)]\right\|_2 = \widetilde{\mathcal{O}}\left(\frac{d\beta\gamma\sqrt{\varepsilon}}{\alpha^2}\right).$$

We defer the proof of Lemma 23 to the end of this subsection.

By standard discretization argument, there exists a subset $\mathcal{S}_\varepsilon$ of $\mathcal{S}$ (i.e., a discrete cover of $\mathcal{S}$ w.r.t. metric $d(s,s') = \max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i) - \phi_i(s',a_i)\|_2$) such that

- for any $s \in \mathcal{S}$, there exists $s' \in \mathcal{S}_\varepsilon$ satisfying

$$\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i) - \phi_i(s',a_i)\|_2 \le \frac{1}{\mathrm{poly}(B_\theta, K, d, H, A_i, \lambda^{-1}, \eta^{-1}, \delta^{-1})},$$

- and

$$\log|\mathcal{S}_\varepsilon| \le \widetilde{\mathcal{O}}(dA_i).$$

For all $s \in \mathcal{S}$: denote by $s'$ the closest neighbour of $s$ in $\mathcal{S}_\varepsilon$ w.r.t. metric $d(s,s') = \max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i) - \phi_i(s',a_i)\|_2$,

$$\begin{aligned}
&\sum_{k=1}^K \langle \Phi_s \mu_{i,h}^k(\cdot \mid s), \widehat{\theta}_{i,h}^k - \theta_{i,h}^k\rangle \\
&\overset{(i)}{\le} \sum_{k=1}^K \langle \Phi_{s'} \mu_{i,h}^k(\cdot \mid s'), \widehat{\theta}_{i,h}^k - \theta_{i,h}^k\rangle + 1 \\
&\overset{(ii)}{\le} \sum_{a_i\in\mathcal{A}_i} \|\phi_i(s',a_i)\|_{(\widehat{\Sigma}_{i,h}^\pi + \lambda I)^{-1}} \Bigg[\sqrt{\lambda} B_\theta K \\
&\quad + \max_{v:\,\|v\|_2=1}\left|v^\top\left(\widehat{\Sigma}_{i,h}^\pi + \lambda I\right)^{-1/2}\sum_{k=1}^K \mu_{i,h}^k(a_i \mid s')\left(x_k x_k^\top - \widehat{\Sigma}_{i,h}^\pi\right)\theta_{i,h}^k\right| \\
&\quad + \max_{v:\,\|v\|_2=1}\left|v^\top\left(\widehat{\Sigma}_{i,h}^\pi + \lambda I\right)^{-1/2}\sum_{k=1}^K \mu_{i,h}^k(a_i \mid s')\left(x_k x_k^\top - \widehat{\Sigma}_{i,h}^\pi\right)\theta_{i,h}^k\right|\Bigg] + 1 \\
&\overset{(iii)}{\le} \max_{a_i\in\mathcal{A}_i}\|\phi_i(s',a_i)\|_{(\Sigma_{i,h}^\pi + \lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right) + 1
\end{aligned}$$

$$\overset{(iv)}{\leq} \max_{a_i \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\Sigma^{\pi}_{i,h}+\lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right) + 1,$$

where $(i)$ uses the definition of $\mu^k_{i,h}$, $\mathcal{S}_\varepsilon$ and Lemma 23, $(ii)$ uses Equation (19), $(iii)$ uses Lemma 21 and 22 along with a union bound for all $s' \in \mathcal{S}_\varepsilon$ and all $v$ from a

$$1/\text{poly}(B_\theta, K, d, H, A_i, \lambda^{-1}, \eta^{-1}, \delta^{-1})\text{-cover}$$

of the $d$-dimensional unit ball, and (iv) uses the fact that $s'$ is the closest neighbour of $s$ in $\mathcal{S}_\varepsilon$.

As a result,

$$\text{Term (C)} \leq \max_{a_i \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\Sigma^{\pi}_{i,h}+\lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right) + 1.$$

**Proof** [Proof of Lemma 23] To simplify notations, denote $x_{a_i} = \phi_i(s, a_i)$ and $\overline{x}_{a_i} = \phi_i(s', a_i)$, $a_i \in \mathcal{A}_i$. We cluster the actions in $\mathcal{A}_i$ into $\{\mathcal{C}_v\}_{v=1}^n$ according to the following rule: action $a_i$ and $a'_i$ are in the same cluster if and only if $\|x_{a_i} - x_{a'_i}\|_2 \leq \Delta$, where $\Delta > 10\varepsilon$ is a parameter to be specified later. Denote $y_v := \frac{1}{|\mathcal{C}_v|}\sum_{a_i \in \mathcal{C}_v} x_{a_i}$. We further denote by $c(a_i)$ the cluster that $a_i \in \mathcal{A}_i$ belongs to.

It is simple to verify that

$$\left\|\mathbb{E}_{a\sim\pi_{i,h}(\cdot|s)}[\phi_i(s, a_i)] - \mathbb{E}_{a\sim\pi_{i,h}(\cdot|s)}[y_{c(a_i)}]\right\|_2 \leq \Delta,$$

and

$$\left\|\mathbb{E}_{a\sim\pi_{i,h}(\cdot|s')}[\phi_i(s', a_i)] - \mathbb{E}_{a\sim\pi_{i,h}(\cdot|s')}[y_{c(a_i)}]\right\|_2 \leq 2\Delta.$$

As a result, to prove Lemma 23, it suffices to upper bound

$$\max_{v\in[n]} \left|\pi_{i,h}(\mathcal{C}_v \mid s) - \pi_{i,h}(\mathcal{C}_v \mid s')\right|.$$

For any $(a_i, a'_i, \theta) \in \mathcal{A}_i^2 \times \mathbb{R}^d$, define event

$$E_{a_i,a'_i}(v) = \{(x_{a_i} - x_{a'_i})^\top (w + Wv) > 0\}, \qquad E_{a_i}(v) = \bigcap_{a'_i \in (\mathcal{A}_i/\mathcal{C}_{c(a_i)})} E_{a_i,a'_i}(v).$$

Similarly, we define $\overline{E}_{a_i,a'_i}(v)$ and $\overline{E}_{a_i}(v)$ by replacing $x$ with $\overline{x}$ in the above definition. We have

$$\begin{aligned}
&\left|\pi_{i,h}(\mathcal{C}_v \mid s) - \pi_{i,h}(\mathcal{C}_v \mid s')\right| \\
=&\left|\mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v} E_{a_i}(v)\right) - \mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v} \overline{E}_{a_i}(v)\right)\right| \\
=&\left|\mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v} \bigcap_{a'_i\in(\mathcal{A}_i/\mathcal{C}_v)} E_{a_i,a'_i}(v)\right) - \mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v} \bigcap_{a'_i\in(\mathcal{A}_i/\mathcal{C}_v)} \overline{E}_{a_i,a'_i}(v)\right)\right|
\end{aligned}$$

$$\leq \sum_{a_i\in\mathcal{C}_v}\sum_{a_i'\in(\mathcal{A}_i/\mathcal{C}_v)}\left|\mathbb{P}(E_{a_i,a_i'}(v))-\mathbb{P}(\overline{E}_{a_i,a_i'}(v))\right|.$$

By the definition of $E_{a_i,a_i'}(v)$,

$$\begin{aligned}\mathbb{P}(E_{a_i,a_i'}(v)) &= \mathbb{P}\Big((x_{a_i}-x_{a_i'})^\top(w+Wv)>0\Big)\\ &= \mathbb{P}\Big((x_{a_i}-x_{a_i'})^\top Wv>(x_{a_i'}-x_{a_i})^\top w\Big)\\ &= \mathbb{P}\left(v_1>\frac{(x_{a_i'}-x_{a_i})^\top w}{\|(x_{a_i}-x_{a_i'})^\top W\|_2}\right),\end{aligned}$$

where the last equality uses the symmetry of distribution $\mathcal{V}$. By simple algebra, one can show the density function of $v_1$ is upper bounded by $\widetilde{\mathcal{O}}(d)$. As a result, we have

$$\begin{aligned}&\left|\mathbb{P}(E_{a_i,a_i'}(v))-\mathbb{P}(\overline{E}_{a_i,a_i'}(v))\right|\\ &\leq \widetilde{\mathcal{O}}(d)\times\left|\frac{(x_{a_i'}-x_{a_i})^\top w}{\|(x_{a_i}-x_{a_i'})^\top W\|_2}-\frac{(\overline{x}_{a_i'}-\overline{x}_{a_i})^\top w}{\|(\overline{x}_{a_i}-\overline{x}_{a_i'})^\top W\|_2}\right|\\ &\leq \widetilde{\mathcal{O}}(d)\times\left|\frac{(x_{a_i'}-x_{a_i})^\top w\times\|(\overline{x}_{a_i}-\overline{x}_{a_i'})^\top W\|_2-(\overline{x}_{a_i'}-\overline{x}_{a_i})^\top w\times\|(x_{a_i}-x_{a_i'})^\top W\|_2}{\|(x_{a_i}-x_{a_i'})^\top W\|_2\times\|(\overline{x}_{a_i}-\overline{x}_{a_i'})^\top W\|_2}\right|\\ &\leq \widetilde{\mathcal{O}}\left(\frac{d}{\alpha^2\Delta^2}\right)\times\left|(x_{a_i'}-x_{a_i})^\top w\times\|(\overline{x}_{a_i}-\overline{x}_{a_i'})^\top W\|_2-(\overline{x}_{a_i'}-\overline{x}_{a_i})^\top w\times\|(x_{a_i}-x_{a_i'})^\top W\|_2\right|\\ &\leq \widetilde{\mathcal{O}}\left(\frac{d}{\alpha^2\Delta^2}\right)\times\mathcal{O}(\beta\gamma\varepsilon)=\widetilde{\mathcal{O}}\left(\frac{d\beta\gamma\varepsilon}{\alpha^2\Delta^2}\right)\end{aligned}$$

where: (i) the third inequality uses the fact that $a_i$ and $a_{i'}$ are from different clusters, $W\succeq\alpha I$, and $\|\overline{x}_{a_i}-x_{a_i}\|_2\leq\varepsilon\leq 0.1\Delta$; (ii) the last inequality uses triangle inequality, $\|w\|_2\leq\gamma$, $W\preceq\beta I$ and $\|\overline{x}_{a_i}-x_{a_i}\|_2\leq\varepsilon$. We complete the proof by choosing $\Delta=20\varepsilon^{1/4}$. ∎

## Appendix E. Proofs for Section 4.2

### E.1. Details of the tabular AVLPR algorithm

The tabular MG case is a special case of the linear function approximation setting with finite number of states, *i.e.* $|\mathcal{S}|\leq S$. For the tabular setting, we choose the switching criterion function $\Psi_h$ as

$$\Psi_h(\mathcal{B}_h):=\ln\prod_{s\in\mathcal{S}}\max\left\{\sum_{s_h\in\mathcal{B}_h}\mathbb{1}(s_h=s),1\right\},$$

while the exploration scheme is chosen as $\Gamma_{\text{explore}}(\overline{\pi},\mu_h):=\{(\overline{\pi}_{1:h-1}\odot\mu_h,[m])\}$. In other words, in Line 4 of Algorithm 2 and Line 3 of Algorithm 3, all players jointly play $\mu_h^k$ (or $\pi_h$) once.

**NO-REGRET-ALG** Notice that $\mathcal{D}_{\text{sample}}^{k,i} = \{(s_h^k, a_{i,h}^k, r_{i,h}^k + V_{i,h+1}(s_{h+1}^k))\}$ always consists of a single sample. We will use it to perform an EXP3-IX style update (Neu, 2015), that is

$$\widehat{\ell}_{i,h}^k(s, a_i) = \frac{H - r_{i,h}^k - V_{i,h+1}(s_{h+1}^k)}{\mu_{i,h}^k(a_i \mid s) + \gamma_i} \times \mathbb{1}((s, a_i) = (s_h^k, a_{i,h}^k)),$$

$$\mu_{i,h}^{k+1}(\cdot \mid s) \propto \exp\left(-\eta_i \sum_{k' \le k} \widehat{\ell}_{i,h}^{k'}(s, \cdot)\right),$$

where $\eta_i = \sqrt{\frac{S \log T}{H^2 A_i T}}$ and $\gamma_i = \frac{\eta}{2}$.

**OPTIMISTIC-REGRESS** Denote the data tuple in $\mathcal{D}_{\text{reg}}^i$ by $\{(s_h^k, a_{i,h}^k, r_{i,h}^k + V_{i,h+1}(s_{h+1}^k))\}_{k \in [K]}$. Define $N_h(s) := \sum_{k=1}^K \mathbb{1}(s_h^k = s)$ and

$$\beta_i(n) := \Theta\left(\frac{\iota}{\eta_i(n + \iota)} + \eta_i H^2 A_i\right),$$

where $\iota = \log(KSA_iHm/\delta)$. The optimistic regression is performed by an empirical averaging step with bonus: if $N_h(s) > 0$, set $V_{i,h}(s) = H - h + 1$, otherwise,

$$V_{i,h}(s) = \min\left\{\frac{1}{N_h(s)} \sum_{k=1}^K (r_{i,h}^k + V_{i,h+1}(s_{h+1}^k)) \times \mathbb{1}(s_h^k = s) + \beta_i(N_h(s)), H - h + 1\right\}.$$

**Computational efficiency** It is straightforward to see that, as our instantiation only involves standard EXP3 algorithm with exponential weights updates, bouns computations, and simple averaging, the entire algorithm runs in polynomial time in $(T, H, S, \{A_i\}_{i \in [m]})$.

The rest of this section is devoted to proving Theorem 5, by checking Conditions (1A) through 2 and then applying Theorem 13.

### E.2. Proof of Condition (1A)

Denote by $N_h(s)$ the number of times state $s$ is visited at step $h$ during the $K$ episodes of executing $\overline{\pi}$ in CCE-APPROX. Let $\iota = \log(mSK \max_i A_i/\delta)$ and $p_s := \mathbb{P}^{\overline{\pi}}(s_h = s)$. By invoking the theoretical guarantee of Exp3-IX (e.g., Theorem 12.1 in Lattimore and Szepesvári (2020)) and taking a union bound for all $(i, s) \in [m] \times \mathcal{S}$, we have that with probability at least $1 - \delta$: for all $(i, s) \in [m] \times \mathcal{S}$:

$$\max_{\mu_{i,h} \in \Delta_{A_i}} \sum_{k=1}^K \left(\mathbb{D}_{\mu_{i,h} \times \mu_{-i,h}^k} - \mathbb{D}_{\mu_h^k}\right) [r_{i,h} + \mathbb{P}_{h+1} V_{i,h+1}] (s) \times \mathbb{1}(s_h^k = s)$$
$$\le \mathcal{O}\left(\frac{\iota}{\eta_i} + \eta H^2 A_i N_h(s)\right).$$

By Freedman's inequality and taking a union bound for all $(i, s, \mu_{i,h}) \in [m] \times \mathcal{S} \times \{e_i\}_{i \in [A_i]}$, we have that with probability at least $1 - \delta$: for all $(i, s) \in [m] \times \mathcal{S}$:

$$\max_{\mu_{i,h} \in \Delta_{A_i}} \sum_{k=1}^K \left(\mathbb{D}_{\mu_{i,h} \times \mu_{-i,h}^k} - \mathbb{D}_{\mu_h^k}\right) [r_{i,h} + \mathbb{P}_{h+1} V_{i,h+1}] (s) \times \mathbb{1}(s_h^k = s)$$

$$\geq p_s \times \max_{\mu_{i,h}\in\Delta_{A_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu_{-i,h}^k} - \mathbb{D}_{\mu_h^k}\right)\left[r_{i,h} + \mathbb{P}_{h+1}V_{i,h+1}\right](s) - \mathcal{O}(H\sqrt{p_s K\iota} + H\iota),$$

and

$$N_h(s) \leq \mathcal{O}(p_s K + \iota).$$

Combining all above relations gives that

$$\begin{aligned}
&\frac{1}{K}\max_{\mu_{i,h}\in\Delta_{A_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu_{-i,h}^k} - \mathbb{D}_{\mu_h^k}\right)\left[r_{i,h} + \mathbb{P}_{h+1}V_{i,h+1}\right](s)\\
&\quad\leq \min\left\{\mathcal{O}\left(\frac{\iota}{\eta_i p_s K} + \eta_i H^2 A_i\left(1+\frac{\iota}{p_s K}\right)\right), H\right\}\\
&\quad\leq \mathcal{O}\left(\frac{\iota}{\eta_i(p_s K + \iota)} + \eta_i H^2 A_i\right),
\end{aligned}$$

where the last inequality uses the fact that $\eta_i^{-2} \geq A_i$. As a result, we can pick

$$G_{i,h}(s,\overline{\pi},K,\delta) = \mathcal{O}\bigg(\frac{\iota}{\eta_i(K\mathbb{P}^{\overline{\pi}}(s_h = s) + \iota)} + \eta_i H^2 A_i\bigg).$$

### E.3. Proof of Condition (1B)

Denote by $N_h(s)$ the number of times state $s$ is visited at step $h$ during the $K$ episodes of executing $\overline{\pi}$ in V-APPROX. Let $\iota = \log(mSK\max_i A_i/\delta)$ and $p_s := \mathbb{P}^{\overline{\pi}}(s_h = s)$. Since the case of $N_h(s) = 0$ is trivial, below we only consider those state $s$ such that $N_h(s) > 0$.

By Azuma-Hoeffding inequality and taking a union bound for all $(i, s, \mu_{i,h}) \in [m]\times\mathcal{S}\times\{e_i\}_{i\in[A_i]}$, we have that with probability at least $1-\delta$: for all $(i,s)\in[m]\times\mathcal{S}$:

$$\begin{aligned}
&\left|\frac{1}{N_h(s)}\sum_{k=1}^{K}\left(r_{i,h}^k + V_{i,h+1}(s_h^{k+1})\right)\times \mathbb{1}(s_h^k = s) - \mathbb{D}_{\pi_h}\left[r_{i,h} + \mathbb{P}_{h+1}V_{i,h+1}\right](s)\right|\\
&\qquad\qquad \leq \mathcal{O}\left(H\sqrt{\frac{\iota}{N_h(s)}}\right) \leq \mathcal{O}\left(\frac{\iota}{\eta_i(N_h(s)+\iota)} + \eta_i H^2 A_i\right),
\end{aligned}$$

where the second inequality uses the fact that $\eta_i^{-1} \geq \iota$. As a result, to prove both relations in Condition (1B), it suffices to show for all $(i,s)\in[m]\times\mathcal{S}$:

$$G_{i,h}(s,\overline{\pi},K,\delta) = \Theta\left(\frac{\iota}{\eta_i(N_h(s)+\iota)} + \eta_i H^2 A_i\right).$$

By Bernstein inequality and taking a union bound for all $s \in \mathcal{S}$, we have that with probability at least $1-\delta$: for all $s\in\mathcal{S}$:

$$\frac{1}{2}p_s K - \frac{1}{2}\iota \leq N_h(s) \leq 2p_s K + \frac{1}{2}\iota.$$

We complete the proof by plugging the above sandwich relation back into the definition of $G_{i,h}(s,\overline{\pi},K,\delta)$.

## E.4. Proof of Condition (1C)

Let $\iota = \log(mSK \max_i A_i/\delta)$, $w_s^t := \mathbb{P}^{\pi^t}(s_h = s)$ and $W_s^t := \sum_{\tau \le t} \mathbb{P}^{\pi^\tau}(s_h = s)$. By plugging in the definition of $G_{i,h}$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{\pi^{t+1}} \left[ G_{i,h}(s, \overline{\pi}^t, t, \delta) \right] \\
=&\mathcal{O}\left( \sum_{t=1}^{T} \mathbb{E}_{\pi^{t+1}} \left[ \frac{\iota}{\eta_i (W_s^t + \iota)} + \eta_i H^2 A_i \right] \right) \\
=&\mathcal{O}\left( \sum_{s \in \mathcal{S}} \sum_{t=1}^{T} w_s^{t+1} \left[ \frac{\iota}{\eta_i (W_s^t + \iota)} + \eta_i H^2 A_i \right] \right) = \mathcal{O}\left( \frac{S\iota \log(T)}{\eta_i} + \eta_i H^2 A_i T \right).
\end{aligned}
$$

## E.5. Proof of Condition 2

Denote by $N_h(s, \mathcal{B}_h)$ the number of times state $s$ occurs in dataset $\mathcal{B}_h$. By Bernstein inequality and taking a union bound for all $(t, s) \in [T] \times \mathcal{S}$, we have that with probability at least $1 - \delta$: for all $(t, s) \in [T] \times \mathcal{S}$:

$$
\frac{1}{2} \sum_{\tau=1}^{t} \mathbb{P}^{\pi^\tau}(s_h = s) - \frac{1}{2}\iota \le N_h(s, \mathcal{B}_h^t) \le 2 \sum_{\tau=1}^{t} \mathbb{P}^{\pi^\tau}(s_h = s) + \frac{1}{2}\iota.
$$

Since $\Psi_{i,h}(\mathcal{B}_h^t) \le \Psi_{i,h}(\mathcal{B}_h^{I_t}) + 1$, we have $N_h(s, \mathcal{B}_h^t) \le 2N_h(s, \mathcal{B}_h^{I_t})$. Using the above relative concentration result, we obtain

$$
\sum_{\tau=1}^{t} \mathbb{P}^{\pi^\tau}(s_h = s) + \iota \le 2(N_h(s, \mathcal{B}_h^t) + \iota) \le 4N_h(s, \mathcal{B}_h^{I_t}) + 2\iota \le 8 \sum_{\tau=1}^{I_t} \mathbb{P}^{\pi^\tau}(s_h = s) + 4\iota.
$$

Finally, we complete the proof of Condition 2(a) by recalling

$$
G_{i,h}(s, \overline{\pi}^t, t, \delta) = \frac{\iota}{\eta_i (\sum_{\tau=1}^{t} \mathbb{P}^{\pi^\tau}(s_h = s) + \iota)} + \eta_i H^2 A_i.
$$

As for Condition 2(b), simply observe that: (1) $\Psi_{i,h}$ does not depend on $i$; (2) $\Psi_{i,h}(\mathcal{D}_h^t) \le S \log t$. Therefore the total number of switches up to iteration $T$ is bounded by $SH \log T$. In other words Condition 2(b) is satisfied with $d_{\text{replay}} = SH$.

## E.6. Sample complexity for tabular MG

Sections E.2 and E.3 shows that Condition (1A) and (1B) are satisfied with

$$
G_{i,h}(s, \overline{\pi}, t, \delta) = \mathcal{O}\left( \frac{\iota}{\eta_i (K\mathbb{P}^{\overline{\pi}}(s_h = s) + \iota)} + \eta_i H^2 A_i \right).
$$

Meanwhile Section E.4 shows that this choice of $G$ satisfies Condition (1C) with

$$
L = \widetilde{\mathcal{O}}\left( SH^2 \max_i A_i \right).
$$

Finally Section E.5 shows that Condition 2 is satisfied with $d_{\text{replay}} = SH$. It remains to apply Theorem 13, which gives the sample complexity bound of

$$\widetilde{\mathcal{O}}\left(\frac{H^3 L d_{\text{replay}}}{\varepsilon^2}\right) = \widetilde{\mathcal{O}}\left(\frac{S^2 H^6 \max_i A_i}{\varepsilon^2}\right).$$

Note that in the tabular algorithm, $\Gamma_{\text{explore}}$ contains a single element, so $\overline{\Gamma} = 1$.

## Appendix F. Proofs for Section 5

### F.1. Explorative All-Policy Evaluation (APE)

We provide the full description of the APE algorithm in Algorithm 6.

**Algorithm 6** $\mathsf{APE}_i(\mathcal{F}_i, \Pi_i, \pi_{-i}, K, \beta)$: Explorative All-Policy Evaluation ($i$-th player)

1: **Initialize** confidence set $\mathcal{B}^1 \leftarrow \mathcal{F}_i \times \Pi_i$, $\mathcal{D}_h \leftarrow \{\}$.
2: **for** $k = 1, \ldots, K$ **do**
3: Compute upper and lower value estimates for all $\pi_i \in \Pi_i$:

$$(\overline{V}^{k,\pi_i\times\pi_{-i}}, \underline{V}^{k,\pi_i\times\pi_{-i}}) \leftarrow \left(\max_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1)), \min_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1))\right).$$

4: Choose $\pi_i^k \leftarrow \arg\max_{\pi_i\in\Pi_i}(\overline{V}^{k,\pi_i\times\pi_{-i}} - \underline{V}^{k,\pi_i\times\pi_{-i}})$.
5: Execute $\pi_i^k \times \pi_{-i}$, and collect the trajectory $(s_1^k, a_{i,1}^k, r_{i,1}^k, \ldots, s_H^k, a_{i,H}^k, r_{i,H}^k)$ for the $i$-th player.
6: Update $\mathcal{D}_h \leftarrow \mathcal{D}_h \cup \{(s_h^k, a_{i,h}^k, r_{i,h}^k, s_{h+1}^k)\}$ for all $h \in [H]$.
7: Update confidence set

$$\mathcal{B}^{k+1} = \left\{(f, \pi_i) \in \mathcal{F}_i \times \Pi_i : \mathcal{L}_h^{\mathcal{D}_h}(f_h, f_{h+1}, \pi_i) \le \min_{f'_h\in\mathcal{F}_{i,h}} \mathcal{L}_h^{\mathcal{D}_h}(f'_h, f_{h+1}, \pi_i) + \beta, \forall\, h \in [H]\right\} \bigcap \mathcal{B}^k,$$

where $\mathcal{L}_h^{\mathcal{D}_h}(f_h, f_{h+1}, \pi_i) := \sum_{(s,a_i,r,s')\in\mathcal{D}_h} [f_h(s, a_i) - r - f_{h+1}(s', \pi_{i,h}(s'))]^2.$

**Output:** Optimistic value estimates $\{\overline{V}^{K,\pi_i\times\pi_{-i}}\}_{\pi_i\in\Pi_i}$.

### F.2. Proof of Theorem 8

In this section we prove Theorem 8. We first present the guarantee for the APE subroutine in the following proposition, whose proof can be found in Appendix F.4.

**Proposition 24 (Learning accurate Q-functions for all policies by APE)** *Under Assumption 6 & 7, there exists an absolute constant $c > 0$ so that for any player $i \in [m]$, if we choose $\beta = cH^2\log(|\Pi_i||\mathcal{F}_i|KH/\delta)$ in Algorithm 6, then with probability at least $1-\delta$ we have*

*(a)* $\underline{V}^{K,\pi_i\times\pi_{-i}} \le V_{i,1}^{\pi_i,\pi_{-i}}(s_1) \le \overline{V}^{K,\pi_i\times\pi_{-i}}$ *for all* $\pi_i \in \Pi_i$.

*(b)* $\max_{\pi_i\in\Pi_i}\left(\overline{V}^{K,\pi_i\times\pi_{-i}} - \underline{V}^{K,\pi_i\times\pi_{-i}}\right) \le \mathcal{O}\left(H\sqrt{\frac{d_i \log K\cdot\beta}{K}}\right).$

Since Algorithm 4 calls the APE subroutine for $T$ round with $m$ players per round with parameters $(\beta, K) \leftarrow (\beta_i, K_i)$, applying Proposition 24 with a union bound yields that, with probability at least $1-\delta/2$, the optimistic value estimates $\{\overline{V}_i^{(t),\pi_i\times\pi_{-i}^t}\}_{\pi_i\in\Pi_i}$ satisfy that

$$\begin{aligned} V_{i,1}^{\pi_i,\pi_{-i}}(s_1) &\overset{(i)}{\le} V_{i,1}^{\pi_i,\pi_{-i}}(s_1) + \mathcal{O}\left(H^2\sqrt{\frac{d_i\log(K_i)\cdot\log(\sum_i|\Pi_i||\mathcal{F}_i|TK_iH/\delta)}{K_i}}\right) \\ &\overset{(ii)}{\le} V_{i,1}^{\pi_i,\pi_{-i}}(s_1) + \varepsilon/2 \end{aligned} \tag{20}$$

for all $i\in[m]$, $\pi_i\in\Pi_i$, and $t\in[T]$ simultaneously. Above, (i) used our choice of $\beta_i$, and (ii) can be satisfied by choosing

$$K_i = \widetilde{\mathcal{O}}\left(\frac{H^4 d_i\cdot\log(\sum_i|\Pi_i||\mathcal{F}_i|)}{\varepsilon^2}\right). \tag{21}$$

We next show that DOPMD achieves small regret for any optimistic value estimate satisfying (20). The proof can be found in Appendix F.3.

**Proposition 25 (Regret guarantee of DOPMD)** *Suppose the optimistic value estimates in Algorithm 4 achieve valid optimism and uniformly small error, i.e.*

$$V_{i,1}^{\pi_i\times\pi_{-i}^t}(s_1) \le \overline{V}_i^{(t),\pi_i\times\pi_{-i}^t} \le V_{i,1}^{\pi_i\times\pi_{-i}^t}(s_1) + \varepsilon \tag{22}$$

*for all $t\in[T]$, $i\in[m]$, and $\pi_i\in\Pi_i$. Then, Algorithm 4 with $\eta_i = \sqrt{\log|\Pi_i|/(H^2T)}$ achieves with probability at least $1-\delta$ that*

$$\max_{i\in[m]}\max_{\pi_i\in\Pi_i}\sum_{t=1}^T\left[V_{i,1}^{\pi_i\times\Lambda_{-i}^t}(s_1) - V_{i,1}^{\Lambda^t}(s_1)\right] \le \varepsilon T + \mathcal{O}\left(H\sqrt{T\log\Big(\sum_{i\in[m]}|\Pi_i|/\delta\Big)}\right). \tag{23}$$

By (20) and Proposition 25, we have that with probability at least $1-\delta$, the output policy $\overline{\Lambda}$ of Algorithm 4 achieves

$$\mathrm{CCEGap}^{\Pi}(\overline{\Lambda}) = \max_{i\in[m]}\max_{\pi_i\in\Pi_i}\left(V_{i,1}^{\pi_i\times\overline{\Lambda}_{-i}} - V_{i,1}^{\overline{\Lambda}}\right) = \frac{1}{T}\max_{i\in[m]}\max_{\pi_i\in\Pi_i}\sum_{t=1}^T\left[V_{i,1}^{\pi_i\times\Lambda_{-i}^t}(s_1) - V_{i,1}^{\Lambda^t}(s_1)\right]$$

$$\le \varepsilon/2 + \mathcal{O}\left(H\sqrt{\log\Big(\sum_{i\in[m]}|\Pi_i|/\delta\Big)/T}\right) \le \varepsilon,$$

where the last inequality requires choosing

$$T = \widetilde{\mathcal{O}}\left(\frac{H^2\log(\sum_{i\in[m]}|\Pi_i|)}{\varepsilon^2}\right). \tag{24}$$

Combining (21) with (24), the total number of episodes played is at most

$$T\times\left(\sum_{i\in[m]}K_i\right) = \widetilde{\mathcal{O}}\left(\frac{H^6\left(\sum_{i\in[m]}d_i\right)\cdot\log^2(\sum_i|\Pi_i||\mathcal{F}_i|)}{\varepsilon^4}\right).$$

This completes the proof of Theorem 8. ■

### F.3. Proof of Proposition 25

Fix any player $i \in [m]$. We have

$$\begin{aligned}
\mathrm{Reg}_T^i &:= \max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[ V_{i,1}^{\pi_i \times \Lambda_{-i}^t}(s_1) - V_{i,1}^{\Lambda^t}(s_1) \right] \\
&\le \underbrace{\max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[ V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) - V_{i,1}^{\Lambda_i^t \times \pi_{-i}^t}(s_1) \right]}_{\mathrm{I}} + \mathcal{O}\left( H \sqrt{T \log \Big( \sum_{i \in [m]} |\Pi_i| / \delta \Big)} \right)
\end{aligned}$$

with probability at least $1 - \delta$, where the inequality uses the fact that

$$\begin{aligned}
&\max_{\Lambda_i \in \Delta(\Pi_i)} \left| \sum_{t=1}^{T} \left[ V_{i,1}^{\Lambda_i \times \pi_{-i}^t}(s_1) - V_{i,1}^{\Lambda_i \times \Lambda_{-i}^t}(s_1) \right] \right| \\
&= \max_{\Lambda_i \in \Delta(\Pi_i)} \left| \sum_{\pi_i \in \Pi_i} \Lambda_i(\pi_i) \sum_{t=1}^{T} \left[ V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) - V_{i,1}^{\pi_i \times \Lambda_{-i}^t}(s_1) \right] \right| \\
&= \max_{\pi_i \in \Pi_i} \left| \sum_{t=1}^{T} \left[ V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) - V_{i,1}^{\pi_i \times \Lambda_{-i}^t}(s_1) \right] \right| \le \mathcal{O}(H \sqrt{T \log \Big( \sum_{i \in [m]} |\Pi_i| / \delta \Big)}),
\end{aligned}$$

following by applying Azuma-Hoeffding's inequality for all $i \in [m]$ and all $\pi_i \in \Pi_i$ simultaneously.

Next, to bound term I, we have

$$\begin{aligned}
\mathrm{I} &= \max_{\pi_i \in \Pi_i} \left( \sum_{t=1}^{T} \left[ V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) - V_{i,1}^{\Lambda_i^t \times \pi_{-i}^t}(s_1) \right] \right) \\
&= \underbrace{\max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[ \overline{V}_i^{(t), \pi_i \times \pi_{-i}^t} - \overline{V}_i^{(t), \Lambda_i^t \times \pi_{-i}^t} \right]}_{(a)} + \underbrace{\max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[ V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) - \overline{V}_i^{(t), \pi_i \times \pi_{-i}^t} \right]}_{(b)} \\
&\quad + \underbrace{\sum_{t=1}^{T} \left[ \overline{V}_i^{(t), \Lambda_i^t \times \pi_{-i}^t} - V_{i,1}^{\Lambda_i^t \times \pi_{-i}^t}(s_1) \right]}_{(c)}.
\end{aligned}$$

By (22), we have $(b) \le 0$ and $(c) \le \varepsilon_1 \cdot T$. To bound $(a)$, note that by Algorithm 4, $\Lambda_i^t$ has the following equivalent Follow-The-Regularized-Leader (FTRL) form: $\Lambda_i^t(\pi_i) \propto_{\pi_i} \exp\Big( \eta_i \sum_{\tau=1}^{t-1} \overline{V}_i^{(t), \pi_i \times \pi_{-i}^t} \Big)$, where each $\overline{V}_i^{(t), \pi_i \times \pi_{-i}^t} \in [0, H]$. Therefore, by standard FTRL analysis (Orabona, 2019, Section 6.6),

$$\max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[ \overline{V}_i^{(t), \pi_i \times \pi_{-i}^t} - \overline{V}_i^{(t), \Lambda_i^t \times \pi_{-i}^t} \right] \le \frac{\log |\Pi_i|}{\eta_i} + \frac{\eta_i}{2} H^2 T \le \mathcal{O}\Big( H \sqrt{\log |\Pi_i| \cdot T} \Big),$$

where in the last inequality we have picked $\eta_i = \sqrt{\log |\Pi_i|/(H^2 T)}$. This gives that $\mathrm{I} \le \varepsilon T + \mathcal{O}\Big(H\sqrt{\log |\Pi_i| \cdot T}\Big)$, which when plugged back into the regret bound yields that, with probability at least $1-\delta$, we have for all $i \in [m]$ simultaneously

$$\begin{aligned} \mathrm{Reg}_T^i &\le \varepsilon T + \mathcal{O}\Big(H\sqrt{\log |\Pi_i| \cdot T}\Big) + \mathcal{O}\left(H\sqrt{T\log\Big(\sum_{i\in[m]} |\Pi_i|/\delta\Big)}\right) \\ &\le \varepsilon T + \mathcal{O}\left(H\sqrt{T\log\Big(\sum_{i\in[m]} |\Pi_i|/\delta\Big)}\right). \end{aligned}$$

This proves the desired result. ■

### F.4. Proof of Proposition 24

We begin by providing the following lemma, which shows that the confidence sets at every iteration contain the true value function of any policy $\pi_i$, and achieves small estimation errors with respect to the visited state-actions. The proof relies on the $\Pi$-completeness assumption (Assumption 6) and standard fast-rate concentration arguments for the square loss, and can be found in Appendix F.4.1.

**Lemma 26 (Properties of $\mathcal{B}^k$)** *Under Assumption 6, there exists an absolute constant $c > 0$ so that if we choose $\beta = cH^2 \log(|\Pi_i||\mathcal{F}_i|KH/\delta)$ in Algorithm 6, then with probability at least $1-\delta$,*

*(a)* $(Q_i^{\pi_i,\pi_{-i}}, \pi_i) \in \mathcal{B}^k$ *for all* $(\pi_i, k) \in \Pi_i \times [K]$,

*(b)* $\sum_{t=1}^{k-1} \Big[(f_h - \mathcal{T}_{i,h}^{\pi_i\times\pi_{-i}} f_{h+1})(s_h^t, a_{i,h}^t)\Big]^2 \le \mathcal{O}(\beta)$ *for all* $(k,h) \in [K]\times[H]$ *and* $(f,\pi_i) \in \mathcal{B}^k$,

*(c)* $\sum_{t=1}^{k-1} \mathbb{E}_{(s_h,a_{i,h})\sim\pi_i^t\times\pi_{-i}} \Big[(f_h - \mathcal{T}_{i,h}^{\pi_i\times\pi_{-i}} f_{h+1})(s_h, a_{i,h})^2\Big] \le \mathcal{O}(\beta)$ *for all* $(k,h) \in [K]\times[H]$ *and* $(f,\pi_i) \in \mathcal{B}^k$.

By Lemma 26(a), on the good event it ensures (with probability at least $1-\delta/2$) and by the definition of $\overline{V}^{K,\pi_i\times\pi_{-i}}$ and $\underline{V}^{K,\pi_i\times\pi_{-i}}$ in Algorithm 6, we immediately have $\underline{V}^{K,\pi_i\times\pi_{-i}} \le V_{i,1}^{\pi_i,\pi_{-i}}(s_1) \le \overline{V}^{K,\pi_i\times\pi_{-i}}$ for all $(\pi_i, k) \in \Pi_i \times [K]$, which proves part (a).

To prove part (b), for any $k \in [K]$, denote the optimistic and pessimistic Q estimates of the "exploration policy" $\pi_i^k$ by

$$\overline{f}^k = \arg\max_{f:(f,\pi_i^k)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}^k(s_1)) \quad \text{and} \quad \underline{f}^k = \arg\min_{f:(f,\pi_i^k)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}^k(s_1)),$$

where we recall that $\pi_i^k$ is chosen to maximize the difference between the above two values over all $\pi_i \in \Pi_i$. This combined with the monotonicity of $\mathcal{B}^k$ gives that, for any fixed $\pi_i \in \Pi_i$,

$$\begin{aligned} &K \times \Big(\overline{V}^{K,\pi_i\times\pi_{-i}} - \underline{V}^{K,\pi_i\times\pi_{-i}}\Big) \\ \le &\sum_{k=1}^K \left(\max_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1)) - \min_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1))\right) \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K} \left( \overline{f}_1^k(s_1, \pi_{i,1}^k(s_1)) - \underline{f}_1^k(s_1, \pi_{i,1}^k(s_1)) \right) \\
&= \sum_{k=1}^{K} \left( \overline{f}_1^k(s_1, \pi_{i,1}^k(s_1)) - V_{i,1}^{\pi_i^k \times \pi_{-i}}(s_1, \pi_{i,1}^k(s_1)) \right) + \sum_{k=1}^{K} \left( V_{i,1}^{\pi_i^k \times \pi_{-i}}(s_1, \pi_{i,1}^k(s_1)) - \underline{f}_1^k(s_1, \pi_{i,1}^k(s_1)) \right).
\end{aligned}
$$

The above two terms can be bounded by the same arguments. WLOG, below we focus on the first term.

Recall that the BE dimension assumption (Assumption 7) asserts that either the $\mathcal{D}_{\Pi_i \times \pi_{-i}}$-type or the $\mathcal{D}_\Delta$-type distributional Eluder dimension is bounded (cf. Definition 12). We first consider the case for the $\mathcal{D}_\Delta$-type distributional Eluder dimension, where we have for any $\varepsilon > 0$,

$$
d_i(\varepsilon) := \max_{h \in [H]} d_{\mathrm{E}}\left( \left\{ f_h - \mathcal{T}_{i,h}^{\pi_i \times \pi_{-i}} f_{h+1} : (f, \pi_i) \in \mathcal{F} \times \Pi_i \right\}, \mathcal{D}, \varepsilon \right) \leq d_i \log(1/\varepsilon).
$$

In this case, we have

$$
\begin{aligned}
&\sum_{k=1}^{K} \left( \overline{f}_1^k(s_1, \pi_{i,1}^k(s_1)) - V_1^{\pi_i^k \times \pi_{-i}}(s_1, \pi_{i,1}^k(s_1)) \right) \\
&= \sum_{h=1}^{H} \sum_{k=1}^{K} \mathbb{E}_{\pi_i^k \times \pi_{-i}} \left[ \overline{f}_h^k(s_h, \pi_{i,h}^k(s_h)) - r_h - \overline{f}_{h+1}^k(s_{h+1}, \pi_{i,h+1}^k(s_{h+1})) \right] \\
&\overset{(i)}{\leq} \sum_{h=1}^{H} \sum_{k=1}^{K} \left[ \left( \overline{f}_h^k - \mathcal{T}_{i,h}^{\pi_i^k \times \pi_{-i}} \overline{f}_{h+1}^k)(s_h^k, a_{i,h}^k) \right) \right] + \mathcal{O}\left( H \sqrt{K \log(H/\delta)} \right) \\
&\overset{(ii)}{\leq} \mathcal{O}\left( H \sqrt{d_i(K^{-1/2}) K \beta} \right) + \mathcal{O}\left( H \sqrt{K \log(H/\delta)} \right) \leq \mathcal{O}\left( H \sqrt{d_i K \log K \cdot \beta} \right).
\end{aligned}
\tag{25}
$$

Above, (i) follows by Azuma-Hoeffding's inequality; (ii) follows by combining Lemma 26(b) applied on $(\overline{f}^k, \pi_i^k)$ with an Eluder dimension argument (Jin et al., 2021a, Lemma 41), which gives that for all $h \in [H]$,

$$
\begin{aligned}
&\sum_{t=1}^{k-1} \left[ (\overline{f}_h^k - \mathcal{T}_{i,h}^{\pi_i^k \times \pi_{-i}} \overline{f}_{h+1}^k)(s_h^t, a_{i,h}^t) \right]^2 \leq \mathcal{O}(\beta) \quad \text{for all } k \in [K] \\
\Longrightarrow &\sum_{k=1}^{K} \left[ \left( \overline{f}_h^k - \mathcal{T}_{i,h}^{\pi_i^k \times \pi_{-i}} \overline{f}_{h+1}^k)(s_h^k, a_{i,h}^k) \right) \right] \leq \mathcal{O}\Big( \sqrt{d_i(K^{-1/2}) K \beta} \Big) \leq \mathcal{O}\Big( \sqrt{d_i \log K \cdot K \beta} \Big).
\end{aligned}
$$

For the other case of the $\mathcal{D}_{\Pi_i \times \pi_{-i}}$-type distributional-Eluder dimension, we conduct the same arguments up to the point before inequality (i) in (25), and apply the same Eluder dimension argument with respect to roll-in distributions $\{d_h^{\pi_i^k \times \pi_{-i}}\}_{k \geq 1}$ combined with Lemma 26(c) to obtain the same bound as the $\mathcal{D}_\Delta$ case.

Together with the same bound for the second term, we obtain

$$
K \times \left( \overline{V}^{K, \pi_i \times \pi_{-i}} - \underline{V}^{K, \pi_i \times \pi_{-i}} \right) \leq \mathcal{O}\left( H \sqrt{d_i K \log K \cdot \beta} \right).
$$

Dividing by $K$ on both sides proves the desired result. ■

### F.4.1. Proof of Lemma 26

The proof is similar to that of Jin et al. (2021a, Lemma 39(b) & 40). Recall that we consider a fixed $\pi_{-i}$, and let us use $\pi = \pi_i \times \pi_{-i}$ for shorthand. Define random variable

$$X_h^t(f, \pi_i) := 2(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t) \times \left[r_{i,h}^t + f_{h+1}(s_{h+1}^t, \pi_{i,h+1}(s_{h+1}^t)) - (\mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]$$

for all $(f, \pi_i, t, h) \in \mathcal{F}_i \times \Pi_i \times [K] \times [H]$.

Consider the filtration $\{\mathcal{G}_h^t\}_{t \geq 1}$ that includes all historical observations up to $(s_h^t, a_{i,h}^t)$ within iteration $t$, but not $(r_{i,h}^t, s_{h+1}^t)$. Note that $X_h^t(f, \pi_i)$ is a martingale difference sequence with respect to $\{\mathcal{G}_h^t\}_{t \in [K]}$ (as the second term is mean-zero on $\mathcal{G}_h^t$). Further, we have $X_h^t(f, \pi_i) \leq 2H^2$ almost surely as $f_h(\cdot, \cdot) \in [0, H - h + 1]$ for all $h \in [H]$. Therefore, by Freedman's inequality (Lemma 9) and a union bound, for any fixed $\lambda \leq 1/(2H^2)$, we have with probability at least $1 - \delta$ that

$$\begin{aligned} &\sum_{t=1}^{k} X_h^t(f, \pi_i) \leq 4\lambda H^2 \sum_{t=1}^{k} \left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2 + \frac{\log(|\mathcal{F}_i||\Pi_i|KH/\delta)}{\lambda} \\ &= \frac{1}{2} \sum_{t=1}^{k} \left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2 + 8H^2 \log(|\mathcal{F}_i||\Pi_i|KH/\delta). \end{aligned} \tag{26}$$

for all $(f, \pi_i, k, h)$ simultaneously, where in the second line we have picked $\lambda = 1/(8H^2)$.

Let $\mathcal{D}_h^k$ denote the dataset $\mathcal{D}_h$ maintained in Algorithm 6 before the start of the $k$-th iteration (i.e. used in forming $\mathcal{B}^k$). To prove part (b), take any $(k, h) \in [K] \times [H]$ and $(f, \pi_i) \in \mathcal{B}^k$. We have by definition of $\mathcal{B}^k$ that

$$\begin{aligned} \beta &\geq \mathcal{L}_h^{\mathcal{D}_h^k}(f_h, f_{h+1}, \pi_i) - \min_{f_h' \in \mathcal{F}_{i,h}} \mathcal{L}_h^{\mathcal{D}_h^k}(f_h', f_{h+1}, \pi_i) \\ &\overset{(i)}{\geq} \mathcal{L}_h^{\mathcal{D}_h^k}(f_h, f_{h+1}, \pi_i) - \mathcal{L}_h^{\mathcal{D}_h^k}(\mathcal{T}_{i,h}^{\pi} f_{h+1}, f_{h+1}, \pi_i) \\ &= \sum_{t=1}^{k-1} \left[f_h(s_h^t, a_{i,h}^t) - r_{i,h}^t - f_{h+1}(s_{h+1}^t, \pi_{i,h+1}(s_{h+1}^t))\right]^2 \\ &\qquad - \sum_{t=1}^{k-1} \left[(\mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t) - r_{i,h}^t - f_{h+1}(s_{h+1}^t, \pi_{i,h+1}(s_{h+1}^t))\right]^2 \\ &= -\sum_{t=1}^{k-1} X_h^t(f, \pi_i) + \sum_{t=1}^{k-1} \left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2 \\ &\overset{(ii)}{\geq} -8H^2 \log(|\mathcal{F}_i||\Pi_i|KH/\delta) + \frac{1}{2} \sum_{t=1}^{k-1} \left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2. \end{aligned}$$

Above, (i) follows by $\Pi$-completeness (Assumption 6), and (ii) follows by (26). Therefore, choosing $\beta = 8H^2 \log(|\mathcal{F}_i||\Pi_i|KH/\delta)$ ensures that

$$\sum_{t=1}^{k-1} \left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2 \leq 4\beta,$$

which proves part (b).

To prove part (a), first note that $Q^\pi_i = Q^{\pi_i\times\pi_{-i}}_i \in \mathcal{F}_i$, as we have $Q^\pi_{i,h} \in \mathcal{F}_{i,h}$ for $h = H, \dots, 1$ by Assumption 6 repeatedly. Therefore, fix any $(k, h) \in [K] \times [H]$ and $f'_h \in \mathcal{F}_{i,h}$, and let $\widetilde{Q} \in \mathcal{F}$ be defined as $\widetilde{Q}_h = f'_h$ and $\widetilde{Q}_{h'} = Q^\pi_{i,h'}$ for all $h' \neq h$. Similar as above, we have

$$\begin{aligned}
&\mathcal{L}^{\mathcal{D}^k_h}_h(Q^\pi_{i,h}, Q^\pi_{i,h+1}, \pi_i) - \mathcal{L}^{\mathcal{D}^k_h}_h(f'_h, Q^\pi_{i,h+1}, \pi_i) \\
&= \mathcal{L}^{\mathcal{D}^k_h}_h(\mathcal{T}^\pi_{i,h}\widetilde{Q}_h, \widetilde{Q}_{h+1}, \pi_i) - \mathcal{L}^{\mathcal{D}^k_h}_h(\widetilde{Q}_h, \widetilde{Q}_{h+1}, \pi_i) \\
&= \sum_{t=1}^{k-1} X^t_h(\widetilde{Q}, \pi_i) - \sum_{t=1}^{k-1} \left[(f'_h - Q^\pi_{i,h})(s^t_h, a^t_{i,h})\right]^2 \\
&\overset{(i)}{\leq} 8H^2\log(|\mathcal{F}_i||\Pi_i|KH/\delta) - \frac{1}{2}\sum_{t=1}^{k-1}\left[(f'_h - Q^\pi_{i,h})(s^t_h, a^t_{i,h})\right]^2 \overset{(ii)}{\leq} \beta,
\end{aligned}$$

where (i) follows by (26) and (ii) follows by our choice of $\beta = 8H^2\log(|\mathcal{F}_i||\Pi_i|KH/\delta)$. As this holds for any $f'_h \in \mathcal{F}_{i,h}$, taking supremum over the left-hand side above gives that

$$\mathcal{L}^{\mathcal{D}^k_h}_h(Q^\pi_{i,h}, Q^\pi_{i,h+1}, \pi_i) - \inf_{f'_h\in\mathcal{F}_{i,h}} \mathcal{L}^{\mathcal{D}^k_h}_h(f'_h, Q^\pi_{i,h+1}, \pi_i) \leq \beta.$$

As this holds for all $h \in [H]$, by definition we have $(Q^\pi_{i,h}, \pi) \in \mathcal{B}^k$ for all $k \in [K]$. This proves part (a).

Finally, part (c) can be proved by exactly the same arguments as part (b), except for redefining the filtration $\{\mathcal{G}^t_h\}_{t\geq1}$ to include all historical observations *before* episode $t$ starts, so that $(s^t_h, a^t_{i,h}) \sim d^{\pi^k_i\times\pi_{-i}}_h$ conditioned on $\mathcal{G}^t_h$, and rescaling the tail probability $\delta \to \delta/2$ in both (26) and its analog with respect to the new filtration here. ∎

### F.5. Details for Linear Quadratic Games

Here we provide the details for the LQG example (Example 1). Define the following feature map for all $i \in [m]$ (with $d_{\phi,i} := d_S + d_{A,i} + 1$):

$$\phi_i(s, a_i) = \begin{bmatrix} s \\ a_i \\ 1 \end{bmatrix} \begin{bmatrix} s^\top & a_i^\top & 1 \end{bmatrix} \in \mathbb{R}^{d_{\phi,i}\times d_{\phi,i}}.$$

We consider the following linear value class and linear policy class for all $i \in [m]$:

- $\mathcal{F}_{i,h} := \{f_{i,h}(s, a_i) = \langle\phi_i(s, a_i), \theta_h\rangle : \theta_h \in \mathbb{R}^{d_{\phi,i}\times d_{\phi,i}}, \|\theta_h\|_{\mathsf{Fr}} \leq B_{\theta,h}\}$.
- $\Pi_i := \{\pi_i = \{\pi_{i,h}(s) = M_{i,h}s\}_{h\in[H]} : M_{i,h} \in \mathbb{R}^{d_{A,i}\times d_S}, \|M_{i,h}\|_{\mathsf{Fr}} \leq B_{M,h}\}$.

Fixing any linear policy $\pi_{-i} \in \Pi_{-i}$ for the opponents, by the structure of the transition (5) and the reward, the MDP faced by player $i$ reduces to a Linear Quadratic Regulator (LQR), which we denote for simplicity of notation as

$$\begin{cases} s_{h+1} = C_h s_h + D_h a_{i,h} + z_h, \\ r_{i,h}(s, a_i) = \langle J_{i,h}, \phi_i(s, a_i)\rangle. \end{cases}$$

The above $C_h, D_h, J_{i,h}$ can be computed from $A_h, \{B_{i,h}\}_i, \{K_h^i\}_i, \{K_{j,h}^i\}_{i,j}$, and $\pi_{-i}$. It is straightforward to see that, with proper choice of $B_{\theta,h} = \mathcal{O}(\text{poly}(d_{\phi,i}, B_M)^{H-h+1})$ (the final sample complexity will only depend on its logarithm, by covering arguments), we have $\mathcal{T}_{i,h}^{\pi_i \times \pi_{-i}} f_{h+1} \in \mathcal{F}_{i,h}$ for any $f_{h+1} \in \mathcal{F}_{i,h+1}$. This verifies Assumption 6.

Further, observe that the function class

$$\left\{ f_h - \mathcal{T}_h^{\pi_i \times \pi^{-i}} f_{h+1} \mid (f, \pi_i) \in \mathcal{F}_i \times \Pi_i \right\}$$

is a linear function class with a $d_{\phi,i}^2$-dimensional feature map $\phi_i(\cdot,\cdot)$. By standard Eluder dimension bounds for linear function classes, the $\mathcal{D}_\Delta$-type BE dimension (Definition 12) is bounded by $\widetilde{\mathcal{O}}(d_{\phi,i}^2)$, thus verifying Assumption 7 with $d_i := \mathcal{O}(d_{\phi,i}^2) = \mathcal{O}((d_S + d_{A,i})^2)$. Further by standard covering arguments, we can construct finite coverings of $\mathcal{F}_i, \Pi_i$ both with log-cardinality $\widetilde{\mathcal{O}}(\text{poly}(H) \cdot d_{\phi,i}^2)$. Plugging these into Theorem 8, we obtain that DOPMD learns a $\Pi$-CCE for LQGs within

$$\widetilde{\mathcal{O}}\Big(\text{poly}(H, \textstyle\sum_{i \in [m]} d_{\phi,i})/\varepsilon^4\Big)$$

episodes of play.

## Appendix G. Discussions about V-type function approximation

Our meta-algorithms VLPR and AVLPR and their guarantees can extend directly to V-type function approximation. Indeed, at their meta-algorithm level (Algorithm 1-3), VLPR and AVLPR do not strictly speaking require $\mathcal{F}_i$ to be marginal Q classes—They directly apply as-is if $\{\mathcal{F}_i\}_{i \in [m]}$ are instead V classes, so long as the subroutines NO-REGRET-ALG and OPTIMISTIC-REGRESS can be designed Conditions (1A)-(1C) (and Condition 2) can still be satisfied with some bonus functions $\{G_{i,h}\}_{(i,h) \in [m] \times [H]}$.

However, we remark that when instantiated concretely, V-type function approximation may encompass problems with fairly different structures from Q-type function approximation. For instance, imagine adapting the linear function approximation results in Section 4.1 to linear V classes. A sensible choice of the V class would be $\mathcal{F}_{i,h} \subset \left\{ f_{i,h}(\cdot) = \phi_i(\cdot)^\top \theta_{i,h} : \theta_{i,h} \in \mathbb{R}^{d_i} \right\}$, where $\phi_i : \mathcal{S} \to \mathbb{R}^{d_i}$ are feature maps for the state. In this case, a suitable choice of the policy class is linear policies of the form $\pi_{i,h}(\cdot|s) = \arg\max_{a_i \in \mathcal{A}_i} \phi_{i,h}(s)^\top \theta_{i,h}^{a_i}$ where $\{\theta_{i,h}^{a_i}\}_{a_i \in \mathcal{A}_i} \subset \mathbb{R}^{d_i}$ is a collection of vectors. However, such a policy class can be interpreted as requiring any action $a_i \in \mathcal{A}_i$ to "have the same meaning" across all states, which could be rather unnatural compared with the Q-type feature map $\phi_i(s, a_i)$ which allows $a_i \in \mathcal{A}_i$ to be a general action *index* that could mean different things at different states.

## Appendix H. All-policy completeness implies "essentially tabular" games

Here we argue that the restriction to linear argmax policies in Assumption 3 (or some other kind of restriction) is necessary, by showing that the unrestricted *all-policy completeness* assumption places a strong implicit requirement on the game.

Consider the following all-policy completeness assumption for decentralized linear function approximation, which strengthens Assumption 3 by removing the $\Pi^{\text{lin}}$ restriction.

**Assumption 27 (All-policy completeness)** *For any $(i,h) \in [m] \times [H]$, any function $\overline{V} = \overline{V}_{i,h+1} : \mathcal{S} \to [0,H]$ and any policy $\pi$, there exists $\theta^{h,\pi_{-i},\overline{V}} \in \mathbb{R}^d$ with $\|\theta^{h,\pi_{-i},\overline{V}}\|_2 \le B_\theta$ such that*

$$\mathbb{D}_{\delta_{a_i} \times \pi_{-i}}\big[r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}\big](s) = \phi_i(s,a_i)^\top \theta^{h,\pi_{-i},\overline{V}} \quad \textit{for all } (s,a_i) \in \mathcal{S} \times \mathcal{A}_i. \tag{27}$$

Fix any $(h,s^\star) \in [H] \times \mathcal{S}$, fix any player $i \in [m]$ and $s' \in \mathcal{S}$. Let $\pi^1$ and $\pi^2$ be any two joint policies that are different only at $(h,s^\star)$. By applying Assumption 27 with zero reward (*i.e.* $r_{i,h} = 0$) and function $\overline{V}(\cdot) = 1\{s' = \cdot\}$, there exists $\theta^{\pi^{\{1,2\}}}$ such that for all $(s,a_i) \in \mathcal{S} \times \mathcal{A}_i$,

$$\phi_i(s,a_i)^\top \theta^{\pi^1} = \mathbb{E}_{a_{-i} \sim \pi^1_{-i}(\cdot|a)} \Pr[s'|s,a_i,a_{-i}],$$
$$\phi_i(s,a_i)^\top \theta^{\pi^2} = \mathbb{E}_{a_{-i} \sim \pi^2_{-i}(\cdot|a)} \Pr[s'|s,a_i,a_{-i}].$$

As $\pi^1 = \pi^2$ at $(h,s)$ with any $s \neq s^\star$, we have for every $s \neq s^\star \in \mathcal{S}$ and $a_i \in \mathcal{A}_i$ that

$$\phi_i(s,a_i)^\top \left(\theta^{\pi^1} - \theta^{\pi^2}\right) = 0,$$

and for $s = s^\star$ that

$$\phi_i(s^\star,a_i)^\top \left(\theta^{\pi^1} - \theta^{\pi^2}\right) = \mathbb{E}_{a_{-i} \sim \pi^1_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}] - \mathbb{E}_{a_{-i} \sim \pi^2_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}].$$

We say that a state $(s,h) \in \mathcal{S} \times [H]$ is *irrelevant* if the transition of this state can be affected by the action of some players. If a state is $s^\star$ *relevant*, by definition $\exists i \in [m]$, $s' \in \mathcal{S}$, $a_i \in \mathcal{A}_i$, and $\pi^1, \pi^2$ such that

$$\phi_i(s^\star,a_i)^\top \left(\theta^{\pi^1} - \theta^{\pi^2}\right) = \mathbb{E}_{a_{-i} \sim \pi^1_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}] - \mathbb{E}_{a_{-i} \sim \pi^2_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}] \neq 0.$$

It follows that (1) $v := \theta^{\pi^1} - \theta^{\pi^2} \neq 0$; (2) $v$ is orthogonal to $\phi_i(s,a_i')$ for all other $s \neq s^\star$ and $a_i' \in \mathcal{A}_i$; (3) $\phi_i(s^\star,a_i)$ is not orthogonal to $v$, and thus linearly independent from $\{\phi_i(s,a_i')\}_{s \neq s^\star, a_i' \in \mathcal{A}_i}$. Since the features $\phi_i(s,a_i) \in \mathbb{R}^d$, there could be at most $d$ such feature vectors that are linearly independent from everyone else, and therefore there are at most $d$ relevant states for player $i$.

It follows that except for at most $dm$ states, all other states are irrelevant: the transition probabilities at such states are not a function of the players' joint action. If we simply omit such states (and play an arbitrary policy when visiting such states) from the trajectory, the resulting dynamics would be a Markov game dynamics over a small (at most $dm$) number of states. In this sense such a Markov game would be "essentially tabular".

### H.1. Explicit forms of the policy class in Cui et al. (2023)

If the no-regret-learning oracle in Cui et al. (2023) is chosen as the Exponential Weights algorithm, then it will induce a policy class of the following form: $\Pi^{\text{estimate}} = \Pi_1^{\text{estimate}} \times \cdots \times \Pi_m^{\text{estimate}}$ with

$$\Pi_i^{\text{estimate}} := \Bigg\{ \pi_i(\cdot \mid s) \propto \exp\left(\eta \sum_{i=1}^{K} \left[\phi_i(s,\cdot)^\top \theta^k + \beta \|\phi_i(s,\cdot)\|_{\Sigma^{-1}}\right]_{[0,H]}\right) : \\ \theta^k \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d},\ \Sigma \succeq \lambda I \Bigg\},$$

where $[\cdot]_{[0,H]}$ denotes a truncation operator s.t. $[x]_{[0,H]} = \min\{\max\{x, 0\}, H\}$ and $\eta, \lambda, \beta$ are some tunable parameters in their algorithm. Note that linear argmax policies can be parameterized by a single $d$-dimension vector, while policies in above class are specified by a much larger number of parameters ($K$ different $d$-dimension vectors, a $d \times d$ matrix, and a few additional scalars) and involve $K$ truncations that make the exponents potentially highly nonlinear. In this sense, the above policy class is more complex than the linear argmax policy class $\Pi^{\text{lin}}$ considered in this paper. We further note that $\Pi_i^{\text{estimate}}$ reduces to $\Pi_i^{\text{lin}}$ if we remove the truncation operator, choose $\beta = 0$ and let $\eta$ go to infinity in the above definition.

If the no-regret-learning oracle is instead chosen as Expected Follow the Perturbed Leader, then we will have $\Pi^{\text{estimate}} = \Pi_1^{\text{estimate}} \times \cdots \times \Pi_m^{\text{estimate}}$ with

$$\Pi_i^{\text{estimate}} := \Bigg\{ \pi_i(a_i \mid s) = \mathbb{P}_{v\sim D_i}\left[ a_i \in \arg\max_{\widehat{a}_i} \left( \sum_{i=1}^{K} \left[ \phi_i(s, \widehat{a}_i)^\top \theta^k + \beta \|\phi_i(s, \widehat{a}_i)\|_{\Sigma^{-1}} \right]_{[0,H]} + \eta^{-1} v_{\widehat{a}_i} \right) \right] \\ : \theta^k \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d},\ \Sigma \succeq \lambda I \Bigg\},$$

where vector $v \in \mathbb{R}^{A_i}$ is sampled from some distribution $D_i$ over $\mathbb{R}^{A_i}$ and $v_{\widehat{a}_i}$ denotes the $\widehat{a}_i$-th coordinate of $v$. Similar to the argument above, this $\Pi_i^{\text{estimate}}$ is still more involved than $\Pi_i^{\text{lin}}$. It can again be reduced to $\Pi_i^{\text{lin}}$ by removing the truncation operator, choosing $\beta = 0$ and picking $D = \delta_{\vec{0}}$: the Dirac distribution at point $\vec{0}$.

## Appendix I. Difference between $\Pi^{\mathsf{Mar}}$-CCE and CCE

Here we provide an example of a toy Markov Game in which there exists a correlated policy $\Lambda \in \Delta(\Pi^{\mathsf{Mar}})$, where $\Pi^{\mathsf{Mar}}$ is the set of all Markov product policies, such that $\mathrm{CCEGap}^{\Pi^{\mathsf{Mar}}}(\Lambda) = 0$ but $\mathrm{CCEGap}(\Lambda) \geq H/4$ for any $H \geq 2$.

Consider the following "sequential rock-paper-scissors" game with horizon $H \geq 2$. The game is two-player zero-sum (with $m = 2$ and $r_2 \equiv 1 - r_1$). The state space is a singleton ($\mathcal{S} = \{s_0\}$ and $S = 1$), and each player has three actions corresponding to rock, paper, and scissors ($A_1 = A_2 = 3$). The instantaneous reward $r_1(a_1, a_2) \in \{0, 1/2, 1\}$ for player 1 is determined by the standard rock-paper-scissors rule (for example, $r_1(\text{rock}, \text{scissors}) = 1$ and $r_1(\text{rock}, \text{rock}) = 1/2$). Let $\Pi_1^{\mathsf{Mar}}$, $\Pi_2^{\mathsf{Mar}}$ denote the set of all Markov policies for each player, and $\Pi^{\mathsf{Mar}} = \Pi_1^{\mathsf{Mar}} \times \Pi_2^{\mathsf{Mar}}$. A Markov policy in this game corresponds to running a memoryless (non history-dependent) policy at each stage $h \in [H]$.

Let $\Lambda = \mathrm{Unif}(\{\pi^{\text{rock}}, \pi^{\text{paper}}, \pi^{\text{scissors}}\})$, where for each $\mathsf{a} \in \{\text{rock}, \text{paper}, \text{scissors}\}$,

$$\pi^{\mathsf{a}} := \pi_1^{\mathsf{a}} \times \pi_2^{\mathsf{a}}, \quad \text{where} \quad \pi_{i,h}^{\mathsf{a}}(\cdot|s_0) = \delta_{\mathsf{a}} \ \text{ for all } (i, h) \in [2] \times [H]$$

specifies the policy where both players play action $\mathsf{a}$ deterministically within all $H$ steps. Note that $\pi^{\mathsf{a}} \in \Pi^{\mathsf{Mar}}$ and thus $\Lambda \in \Delta(\Pi^{\mathsf{Mar}})$.

By definition of $\Lambda$, we have $V_1^{\Lambda} = H/2$. Further, it is straightforward to see that $\max_{\pi_1^\dagger \in \Pi_1^{\mathsf{Mar}}} V_1^{\pi_1^\dagger, \Lambda_{-1}} = H/2$, as this is achievable by picking $\pi_1^\dagger = \pi_1^{\text{rock}}$, and no other Markov policy $\pi_1^\dagger \in \Pi_1^{\mathsf{Mar}}$ (which is memoryless) can achieve a reward greater than $1/2$ at any step against $\Lambda_{-1}$, which plays uniformly within $\{\mathrm{rock}, \mathrm{paper}, \mathrm{scissors}\}$ at every step. This shows that $\mathrm{CCEGap}^{\Pi^{\mathsf{Mar}}}(\Lambda) = 0$.

However, consider the non-Markov policy $\widetilde{\pi}_1$ that plays uniformly at random at $h = 1$, observes the action played by the opponent (or infers the opponent's played action from the received reward), and henceforth plays the winning action against that action at step $h \in \{2, \ldots, H\}$. By definition of $\Lambda$, such a non-Markov policy will deterministically achieve reward $1$ at all steps $h \geq 2$, and thus

$$V_1^{\widetilde{\pi}_1, \Lambda_{-1}} = \frac{1}{2} + H - 1 = H - \frac{1}{2},$$

which gives

$$\mathrm{CCEGap}(\Lambda) \geq V_1^{\widetilde{\pi}_1, \Lambda_{-1}} - V_1^{\Lambda} = H - \frac{1}{2} - \frac{H}{2} \geq \frac{H}{4}$$

for any $H \geq 2$.