TOWARDS GENERAL FUNCTION APPROXIMATION IN ZERO-SUM MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper considers two-player zero-sum finite-horizon Markov games with simultaneous moves. The study focuses on the challenging settings where the value function or the model is parameterized by general function classes. Provably efficient algorithms for both decoupled and coordinated settings are developed. In the decoupled setting where the agent controls a single player and plays against an arbitrary opponent, we propose a new model-free algorithm. The sample complexity is governed by the Minimax Eluder dimension—a new dimension of the function class in Markov games. As a special case, this method improves the state-of-the-art algorithm by a $\sqrt{d}$ factor in the regret when the reward function and transition kernel are parameterized with d-dimensional linear features. In the coordinated setting where both players are controlled by the agent, we propose a model-based algorithm and a model-free algorithm. In the model-based algorithm, we prove that sample complexity can be bounded by a generalization of Witness rank to Markov games. The model-free algorithm enjoys a $\sqrt{K}$-regret upper bound where K is the number of episodes.

1 INTRODUCTION

In competitive reinforcement learning, there are two agents competing against each other by taking actions. Their actions together determine the state evolutions and rewards. Function approximation, especially deep neural networks (LeCun et al., 2015), contributes to the success of RL in many real world applications, such as Atari (Mnih et al., 2013), Go (Silver et al., 2015), autonomous driving (Shalev-Shwartz et al., 2016), Texas holdem poker (Sandholm, 2017), and Dota (Berner et al., 2019).

The goal of competitive reinforcement learning aims to learn the Nash Equilibrium (NE) policy in a trial-and-error fashion. In a NE, no agent can be better off by unilaterally deviating from her policy. Most of the existing sample efficient competitive RL algorithms focus on the tabular or linear function approximation cases (Pérolat et al., 2017; Bai & Jin., 2020; Bai et al., 2021; 2020; Xie et al., 2020; Zhang et al., 2020). The huge empirical success of neural network based RL methods has remained largely unexplained. Thus, there exists a wide gap between theory and practice in competitive RL with general function approximation. One demanding question to ask is:

Can we establish provably efficient competitive RL algorithms with general function approximation?

In this paper, we make a step towards answering this question by providing structural conditions and complexity measures of Markov Games and function classes that allow for efficient learning. We focus on episodic zero-sum Markov Game (MG) with simultaneous move, where each episode consists of H time steps and two players act simultaneously at each time step. The state space is arbitrarily large and can be infinite. We consider two function approximation settings: 1) use a general function class F to approximate the action-value function (Q function); 2) use a general function class $\mathcal{M}$ to approximate the environment.

To this end, we introduce algorithms based on optimistic principles. Due to subtle game-theoretic issues, naive optimism is no longer working in Markov games where minimax optimization is performed. To deal with this issue, we use the algorithmic idea called 'alternate optimism'. Specifically, we develop algorithms for both coordinated and decoupled setting. In the decoupled setting, the agent controls one player and plays against an arbitrary and potentially adversarial opponent. The sample

efficiency is measured by the gap between the learned value and value of the Nash Equilibrium. In the coordinated setting, the agent controls both players and the goal is to find the approximate Nash Equilibrium, i.e. with small duality gap. We identify key complexity measures of function classes to examine the effectiveness of elimination. By doing so, we prove upper bounds on sample complexity and regrets of the presented procedures that are independent of the number of states.

Our contributions are summarized into the following two folds:

- In the decoupled setting, we introduce Minimax Eluder dimension–a new complexity measure for competitive RL problems. We propose an algorithm that incurs at most $\widetilde{O}(H\sqrt{d_E K \log \mathcal{N}_{\mathcal{F}}})$ [1] regret in K episodes where d_E denotes the Minimax Eluder dimension $\mathcal{N}_{\mathcal{F}}$ denotes the covering number of function class, with probability at least $1-p$. As a special case, this result improves Xie et al. (2020) by a $\sqrt{d}$ multiplicative factor in their setting when the reward function and transition kernel are linearly parameterized and d is the dimension of feature mapping.
- In coordinated settings, we propose both model-based and model free algorithms. In the model-based setting, we generalize the witness rank (Sun et al., 2019a) to competitive RL. We prove that $\widetilde{O}(H^3W^2/\epsilon^2)$ samples are enough to learn a policy ϵ-close to the Nash Equilibrium, where W is witness rank. In the model-free setting, we develop algorithm for agnostic learning with a candidate policy class Π. The algorithm incurs at most $\widetilde{O}(H\sqrt{dK\log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\Pi})})$ regret. Here d is a variant of Minimax Eluder dimension and $\mathcal{N}_{\Pi}$ denotes the covering number of policy class.

1.1 Related works

There is a rich literature studying the learning and decision-making of Markov Games (Littman & Szepesvari, 1996; Greenwald et al., 2003; Grau-Moya et al., 2018; Pérolat et al., 2018; Srinivasan et al., 2018; Sidford et al., 2020; Wei et al., 2017; Pérolat et al., 2017; Bai & Jin., 2020; Bai et al., 2021; 2020; Zhang et al., 2020; Zhao et al., 2021). The most related to us are perhaps (Xie et al., 2020; Chen et al., 2021; Jin et al., 2021b), where the authors address the challenge of exploration-exploitation tradeoff in large state spaces. Due to space constraint, a detailed literature discussion is deferred to Appendix A.

1.2 Technical challenges

Previous work (Xie et al., 2020) imposes optimistic bonus on the action-value functions in every state-action pairs and performs planning by the Coarse Correlated Equilibrium (CCE) on the optimistic value functions. To achieve improved rates, we leverage the idea of 'global optimism' (Zanette et al., 2020; Jin et al., 2021a; Du et al., 2021), which maintains a constraint set of candidate functions that do not deviate much from the empirical estimates and performs optimistic planning on the initial state. However, going beyond MDPs towards MGs, two problems arise. First, the concentration property of functions in constraint set is hard to characterize due to multi-agent interplay. For this, we use the concentration methods in Jin et al. (2021a) and extend it from MDPs to MGs. The second and more prominent issue is the exploration and exploitation tradeoff. Since 'global optimism' only obtains optimism along the trajectories of behaviour policies, it may not guarantee optimism on the trajectories of target policies (i.e. NE). As a result, directly using CCE to plan will cause the duality gaps to diverge. To deal with this problem, we apply 'alternate optimism' to guide explorations, which was previously used in Wei et al. (2017) for model-based methods. The 'alternate optimism' used in this work is slightly different for value-based methods. We prove two regret decomposition lemmata to support this optimism principle.

2 Preliminaries

We consider two-player zero-sum simultaneous-moves episodic Markov game, defined by the tuple

$$(\mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, r, \mathbb{P}, H)$$

[1] We use $\widetilde{O}$ to hide logarithmic terms in H, $1/p$ and K.

where $\mathcal{S}$ is the state space, $\mathcal{A}_i$ is a finite set of actions that player $i \in \{1, 2\}$ can take, r is the reward function, $\mathbb{P}$ is the transition kernel and H is the number of time steps. At each time step $h \in [H]$, player P1 and P2 take actions $a \in \mathcal{A}_1$ and $b \in \mathcal{A}_2$ respectively upon observing the state $x \in \mathcal{S}$, and then both receive the reward $r_h(x, a, b)$. The system then transitions to a new state $x' \sim \mathbb{P}_h(\cdot|x, a, b)$ according to the transition kernel $\mathbb{P}$. Throughout this paper, we assume for simplicity that $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{A}$ and that the rewards $r_h(x, a, b)$ are deterministic functions of the tuple (x, a, b) taking values in $[-1, 1]$. Turn-based games are special cases of simultaneous games in the sense that at each state the reward and transition are independent of one player's action (Xie et al., 2020). Generalizations to $\mathcal{A}_1 \neq \mathcal{A}_2$ and stochastic reward are also straightforward.

Denote by $\Delta := \Delta(\mathcal{A})$ the probability simplex over the action space $\mathcal{A}$. A stochastic policy of P1 is a length-H sequence of functions $\pi := \{\pi_h : \mathcal{S} \mapsto \Delta\}_{h\in[H]}$. At each step $h \in [H]$ and state $x \in \mathcal{S}$, P1 takes an action sampled from the distribution $\pi_h(x)$ over $\mathcal{A}$. Similarly, a stochastic policy of P2 is given by the sequence $\nu := \{\nu_h : \mathcal{S} \mapsto \Delta\}_{h\in[H]}$. The learning happens in K episodes. In each episode k, each player P1 and P2 proposes a policy π^k and ν^k respectively based on history up to the end of episode $k - 1$ and then executes the policy to observe the trajectories $\{x_h^k, a_h^k, b_h^k\}_{h=1}^H$.

For a fixed policy pair (π, ν), the value function and Q functions for zero-sum game is defined by

$$V_{h_0}^{\pi,\nu}(x) := \mathbb{E}\left[\sum_{h=h_0}^{H} r_h(x_h, a_h, b_h)|x_{h_0} = x\right],$$

$$Q_{h_0}^{\pi,\nu}(x, a, b) := \mathbb{E}\left[\sum_{h=h_0}^{H} r_h(x_h, a_h, b_h)|x_{h_0} = x, a_{h_0} = a, b_{h_0} = b\right],$$

where the expectation is over $a_h \sim \pi_h(x_h), b_h \sim \nu_h(x_h)$ and $x_{h+1} \sim \mathbb{P}_h(\cdot|x_h, a_h, b_h)$. $V_1^{\pi,\nu}$ and $Q_1^{\pi,\nu}$ are often abbreviated to $V^{\pi,\nu}$ and $Q^{\pi,\nu}$. In zero-sum games, PI aims to maximize $V^{\pi,\nu}(x_1)$ and P2 aims to minimize $V^{\pi,\nu}(x_1)$. Given policy π of P1, the best response policy of P2 is defined by $\nu_\pi^* := \text{argmin}_\nu V^{\pi,\nu}(x_1)$. Similarly given policy ν of P2, the best response policy of P1 is defined by $\pi_\nu^* := \text{argmax}_\pi V^{\pi,\nu}(x_1)$. Then from definitions,

$$V^{\pi,\nu_\pi^*}(x_1) \leq V^{\pi^*,\nu^*}(x_1) \leq V^{\pi_\nu^*,\nu}(x_1)$$

and the equality holds for the Nash Equilibrium (NE) of the game (π^*, ν^*). We abbreviate $V^{\pi^*,\nu^*}(x_1)$ as V^* and $Q^{\pi^*,\nu^*}(x_1)$ as Q^*. V^* is often referred to as the value of the game. As common in Markov games literature (Pérolat et al., 2015), we will use the following Bellman operator

$$\mathcal{T}_h Q_{h+1}(x, a, b) := r(x, a, b) + \mathop{\mathbb{E}}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)}[\max_{\pi'} \min_{\nu'} V_{h+1}^{\pi',\nu'}(x')], \tag{1}$$

and the following Bellman operator for fixed P1's policy

$$\mathcal{T}_h^\pi Q_{h+1}(x, a, b) := r(x, a, b) + \mathop{\mathbb{E}}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)}[\min_{\nu'} V_{h+1}^{\pi,\nu'}(x')]. \tag{2}$$

Notations For a integer H, $[H]$ denotes the set $\{1, 2, \ldots, H\}$. For a finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. For a matrix $A \in \mathbb{R}^d$, $A_{i,*}$ and $A_{*,j}$ denote the i-th row and j-th column of A respectively. For a function $f : \mathcal{S} \mapsto \mathbb{R}$, $\|f\|_\infty$ denotes $\sup_{s\in\mathcal{S}} |f(s)|$. We use $\mathcal{N}(\mathcal{F}, \epsilon)$ to denote the ϵ-covering number of $\mathcal{F}$ under metric $d(f, g) = \max_h \|f_h - g_h\|_\infty$.

3 Decoupled setting

In decoupled setting, the algorithm controls P1 to play against P2. In k-th episode, P1 chooses a policy π^k and P2 chooses a policy ν^k based on π^k. Then P1 and P2 interact in the Markov Game by playing π^k and ν^k respectively, and observe the rewards and states in this episode. Notice that P2 can be adversarial and its policy is never revealed, meaning that P1 can only maximize its reward by playing π^*. The goal is to thus minimize the following 'value regret'

$$\text{Reg}(K) = \sum_{k=1}^{K} [V_1^*(x_1) - V_1^{\pi^k,\nu^k}(x_1)]. \tag{3}$$

A small value regret indicates that we learn the value of MG. If an algorithm obtains an upper bound on value regret that scales sublinearly with K for all ν^k, then it can not be exploited by opponents.

3.1 FUNCTION APPROXIMATION

We consider function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h \subset \{f : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \mapsto [0, 1]\}$ to provide function approximations for Q^* — the Q-function of Nash Equilibrium. For a policy pair (π, ν) and function f we abuse the notations and use $f(x, \pi, \nu) = \mathbb{E}_{a\sim\pi, b\sim\nu}[f(x, a, b)]$ to denote the expected reward of executing (π, ν) in function f. We make the following assumptions about $\mathcal{F}$.

Assumption 3.1 (Realizability)**.** *We assume $Q^*_h \in \mathcal{F}_h$ for all $h \in [H]$.*

Assumption 3.2 (Completeness)**.** *We assume for all $h \in [H]$ and $f \in \mathcal{F}_{h+1}$, $\mathcal{T}_h f \in \mathcal{F}_h$.*

Realizability means the function class contains the target Q function. Completeness says the function class is closed under the Bellman operator. These assumptions are common in value function approximation literature (e.g. Jin et al. (2020); Wang et al. (2020b); Jin et al. (2021a)). Note that they are weaker than Xie et al. (2020), as their setting satisfies $\mathcal{T}_h f \in \mathcal{F}_h$ for any action-value function f.

For a function $f \in \mathcal{F}$, we denote the max-min policy of P1 by π_f. Specifically,

$$(\pi_f)_h(x) := \text{argmax}_\pi \min_\nu f_h(x, \pi, \nu), \forall h \in [H].$$

Similarly we denote the min-max policy of P2 by ν_f. Given policy π of P1, the best response policy of P2 in terms of function f is defined by

$$(\nu^f_\pi)_h(x) := \text{argmin}_\nu f_h(x, \pi_h, \nu), \forall h \in [H].$$

Similarly given policy ν of P2, the best response policy of P1 in terms of function f is denoted by π^f_ν. When there is no confusion, we will drop the subscript h for simplicity.

Learnability requires the function class to have bounded complexity. We introduce Minimax Eluder dimension to capture the structural complexity of function classes in zero-sum games. To define, we use ϵ-independence of distributions and Distributional Eluder dimension (DE) from Jin et al. (2021a).

Definition 3.3 (ϵ-independence of distributions)**.** *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and $\nu, \mu_1, \ldots, \mu_n$ be probability measures over $\mathcal{X}$. We say that ν is ϵ-independent of $\{\mu_1, \ldots, \mu_n\}$ with respect to $\mathcal{G}$ if there exists $g \in \mathcal{G}$ such that $\sqrt{\sum_{i=1}^n (\mathbb{E}_{\mu_i}[g])^2} \leq \epsilon$ but $|\mathbb{E}_\nu[g]| \geq \epsilon$.*

Definition 3.4 (Distributional Eluder dimension (DE))**.** *Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and $\mathcal{D}$ be a family of probability measures over $\mathcal{X}$. The distributional Eluder dimension $\dim_{\text{DE}}(\mathcal{G}, \mathcal{D}, \epsilon)$ is the length of the longest sequence $\{\rho_1, \ldots, \rho_n\} \subset \mathcal{D}$ such that there exists $\epsilon' \geq \epsilon$ where ρ_i is ϵ'-independent of $\{\rho_1, \ldots, \rho_{i-1}\}$ for all $i \in [n]$.*

Now we introduce Minimax Eluder dimension. Recall the minimax Bellman operator from Eq (1):

$$\mathcal{T}_h f_{h+1}(x, a, b) := r(x, a, b) + \mathop{\mathbb{E}}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)}[\max_{\pi'} \min_{\nu'} f_{h+1}(x', \pi', \nu')]$$

thus Minimax Eluder dimension is the Eluder dimension on the Bellman residues with respect to the above minimax Bellman operator.

Definition 3.5 (Minimax Eluder dimension)**.** *Let $(I - \mathcal{T}_h)\mathcal{F} := \{f_h - \mathcal{T}_h f_{h+1} : f \in \mathcal{F}\}$ be the class of minimax residuals induced by function class $\mathcal{F}$ at level h and $\mathcal{D}_\Delta = \{\mathcal{D}_{\Delta,h}\}_{h\in[H]}$ where $\mathcal{D}_{\Delta,h} := \{\delta(x, a, b) : x \in \mathcal{S}, a \in \mathcal{A}_1, b \in \mathcal{A}_2\}$ is the set of Dirac measures on state-action pair (x, a, b). The ϵ-Minimax Eluder dimension of $\mathcal{F}$ is defined by $\dim_{\text{ME}}(\mathcal{F}, \epsilon) := \max_{h\in[H]} \dim_{\text{DE}}((I - \mathcal{T}_h)\mathcal{F}, \mathcal{D}_\Delta, \epsilon)$.*

3.2 OPTIMISTIC NASH ELIMINATION FOR MARKOV GAMES

Now we introduce Optimistic Nash Elimination for Markov Games (ONEMG), presented in Algorithm 1. This algorithm maintains a confidence set $\mathcal{V}^k$ that always contains the Q^*, and sequentially eliminates inaccurate hypothesis from it. In k-th episode, P1 first optimally chooses a value function f^k in $\mathcal{V}^{k-1}$ that maximizes the value of the game. Intuitively, this step performs optimistic planning in the pessimistic scenarios, i.e. assuming P2 plays the best response. Next, it plays against P2 and augments the data from this episode into replay buffer $\mathcal{B}$. The confidence set $\mathcal{V}^k$ is then updated by keeping the functions f with small minimax Bellman errors $f_h - \mathcal{T}_h f_{h+1}$, in Line 12. To estimate $f_h - \mathcal{T}_h f_{h+1}$, we use $\mathcal{E}_{\mathcal{B}_h}(f_h, f_{h+1}) - \inf_{g\in\mathcal{F}_h} \mathcal{E}_{\mathcal{B}_h}(g, f_{h+1})$, a standard variance reduction trick to avoid the double-sample issue. It can be shown that when the value function class is complete, $\mathcal{E}_{\mathcal{B}_h}(f_h, f_{h+1}) - \inf_{g\in\mathcal{F}_h} \mathcal{E}_{\mathcal{B}_h}(g, f_{h+1})$ is an unbiased estimator of $f_h - \mathcal{T}_h f_{h+1}$ (Lemma B.1).

Notice that unlike many previous works that add optimistic bonus on every state-action pairs (Xie et al., 2020; Chen et al., 2021), Algorithm 1 only takes optimism in the initial state. Instead, the constraint set contains every function that has low Bellman residue on all trajectories in the replay buffer. This 'global optimism' technique trades computational efficiency for better sample complexity, and can be found in many previous work in Markov Decision process (Zanette et al., 2020; Jin et al., 2021a). As we will see later, it is also useful in the coordinated setting.

Algorithm 1 Optimistic Nash Elimination for Markov Games (ONEMG)

Input: Function class $\mathcal{F}$
Initialize $\mathcal{B}_h = \emptyset, h \in [H]$, $\mathcal{V}^0 \leftarrow \mathcal{F}$
Set $\beta \leftarrow C\log(\mathcal{N}(\mathcal{F}, 1/K) \cdot HK/p)$ for some large constant C
for $k = 1, 2, \ldots, K$ **do**
  Compute $f^k \leftarrow \operatorname{argmax}_{f \in \mathcal{V}^{k-1}} f_1(x_1, \pi_f, \nu_f)$, let $\pi_k \leftarrow \pi_{f^k}$
  for step $h = 1, 2, \ldots, H$ **do**
    Interact with environment by playing action $a_h^k \sim \pi_h^k(x_h^k)$
    Observe reward r_h^k, opponent's action b_h^k and move to next state x_{h+1}^k
  end for
  Update $\mathcal{B}_h \leftarrow \mathcal{B}_h \cup \{(x_h^k, a_h^k, b_h^k, r_h^k, x_{h+1}^k)\}, \forall h \in [H]$
  For all $\xi, \zeta \in \mathcal{F}$ and $h \in [H]$, let

$$\mathcal{E}_{\mathcal{B}_h}(\xi_h, \zeta_{h+1}) = \sum_{\tau=1}^{k} [\xi_h(x_h^\tau, a_h^\tau, b_h^\tau) - r_h^\tau - \zeta_{h+1}(x_{h+1}^\tau, \pi_{\zeta_{h+1}}, \nu_{\zeta_{h+1}})]^2$$

  Set

$$\mathcal{V}^k \leftarrow \{f \in \mathcal{F} : \mathcal{E}_{\mathcal{B}_h}(f_h, f_{h+1}) \leq \inf_{g \in \mathcal{F}_h} \mathcal{E}_{\mathcal{B}_h}(g, f_{h+1}) + \beta, \forall h \in [H]\}$$

end for

3.3 Theoretical results

In this section we present the theoretical guarantee for Algorithm 1. The proof is deferred to Appendix B.

Theorem 3.6. *Under Assumption 3.1 and Assumption 3.2, the regret Eq* (3) *of Algorithm 1 is upper bounded by*

$$\mathrm{Reg}(K) \leq O\left(H\sqrt{K \cdot d_{\mathrm{ME}} \log(HK\zeta)}\right)$$

with probability at least $1 - p$*. Here* $d_{\mathrm{ME}} = \dim_{\mathrm{ME}}(\mathcal{F}, \sqrt{1/K})$ *and* $\zeta = \mathcal{N}(\mathcal{F}, 1/K)/p$*.*

As this theorem suggests, as long as the function class has finite Minimax Eluder dimension and covering number, P1 can achieve $\sqrt{K}$-regret against P2. Notice that when P2 always plays the best response of P1's policies, this regret guarantee indicates that algorithm eventually learns the Nash Equilibrium. In this case, the algorithm takes $\widetilde{O}(H^2 d \log(|\mathcal{F}|)/\epsilon^2)$ samples to learn a policy π that is ϵ-close to π^*.

When the opponent plays dummy actions, the Markov Game can be seen as an MDP and the value of the Markov Game is the maximum cumulative sum of rewards one can achieve in this MDP. In this case, our result reduces to $\widetilde{O}(H \cdot \sqrt{K \cdot d_{\mathrm{BE}} \log(\mathcal{N}(\mathcal{F}, 1/K))})$ where d_{BE} is the Bellman eluder dimension in Jin et al. (2021a).

In particular, when the function class is finite, the regret becomes $O(H\sqrt{K d_{\mathrm{ME}} \log(HK|\mathcal{F}|/p)})$ that depends logarithmically on the cardinality of $\mathcal{F}$. Moreover, when the reward and transition kernel have linear structures, i.e. $r_h(x, a, b) = \phi(x, a, b)^\top \theta_h$ and $\mathbb{P}_h(\cdot|x, a, b) = \phi(x, a, b)^\top \mu_h(\cdot)$ where $\phi(x, a, b) \in \mathbb{R}^d$, then $d_{\mathrm{ME}}(\mathcal{F}, \sqrt{1/K}) = \log \mathcal{N}(\mathcal{F}, 1/K) = \widetilde{O}(d)$ and the regret becomes $\widetilde{O}\big(Hd\sqrt{K}\big)$ [2]. Thus we improve Xie et al. (2020) by a $\sqrt{d}$ factor. We also provide a simpler algorithm tailored for this setting, see Appendix B.1 for details.

[2] Here we use $\widetilde{O}$ to omit the logarithm factors in K, d, H and $1/p$.

Algorithm 1 solves a non-convex optimization problem in Line 5 with highly non-convex constraint set (Line 12). In general, it is computationally inefficient. Even when reduced to linear MDP, i.e. the opponent plays dummy actions and the transition matrices and the rewards have low rank structures, it is not known if the same $\widetilde{O}(H\sqrt{d^2K})$ regret can be achieved with computationally efficient algorithms (Zanette et al., 2020). Designing computationally efficient algorithms with near-optimal sample efficiency is an interesting further direction.

4 COORDINATED SETTING

In the coordinated setting, the agent can control both P1 and P2. The goal is to find the ϵ-approximate Nash equilibrium (π, ν) in the sense that

$$V^{\pi^*_\nu,\nu}(x_1) - V^{\pi,\nu^*_\pi}(x_1) \leq \epsilon \tag{4}$$

by playing P1 and P2 against each other in the Markov game. In the following sections, we propose both model-based and model-free methods to deal with this problem.

4.1 MODEL-BASED ALGORITHM

In model-based methods, we make use of a model class $\mathcal{M}$ of candidate models to approximate the true model M^*. We are also given a test function class $\mathcal{G}$ to track model misfit. Given a model $M \in \mathcal{M}$, we use $Q^M(x, a, b)$, $r^M(x, a, b)$ and $\mathbb{P}^M(\cdot|x, a, b)$ to denote the Q-function, reward function and transition kernel of executing action a and b at state x in model M. For simplicity, we use $(r, x) \sim M$ to represent $r \sim r^M(x, a, b)$ and $x \sim \mathbb{P}^M(\cdot|x, a, b)$. We denote the NE policy in model M as π^M and ν^M. For policy π, ν and model M, we use ν^M_π to denote the best response of π in model M and π^M_ν is defined similarly.

4.1.1 ALTERNATE OPTIMISTIC MODEL ELIMINATION (AOME)

The model-based method, Alternate Optimistic Model Elimination (AOME), is presented in Algorithm 4. It maintains a constraint set $\mathcal{M}^k$ of candidate models. Throughout K iterations, AOME makes sure that the true model M^* always belongs $\mathcal{M}^k$ for all $k \in [K]$, and sequentially eliminates incorrect models from $\mathcal{M}^k$.

However, the idea of being optimistic only at initial state (Zanette et al., 2020; Jin et al., 2021a) is not directly applicable to this scenario. Since the objective to be optimized is the duality gap, the target policies being considered are thus policies pairs (π^*_ν, ν) and (π, ν^*_π). However, π^*_ν and ν^*_π are not available to the agents. Therefore, the performances of proposed policies are evaluated on out-of-distribution trajectories, meaning that optimism can not be guaranteed. This causes the distribution shift issue. In worst cases, low Bellman errors in the collected trajectories provide no information for the trajectories collected by π^*_ν, ν and π, ν^*_π.

To address this issue, the algorithm performs global planning by applying the idea of *alternate optimism* to model-based methods. In k-th episode, the algorithm first solves the optimization problem $M^k_1 = \arg\max_{M\in\mathcal{M}} Q^M(x_1, \pi^M, \nu^M)$ and let $\pi^k = \pi^{M^k_1}$. This optimization problem corresponds to finding the optimistic estimate of the value of the game. Indeed, notice that $M^* \in \mathcal{M}^k$ implies $Q^{M^k_1}(x_1, \pi^k, \nu^k) \geq Q^{M^k_1}(x_1, \pi^k, \nu^{M^k_1}_{\pi^k}) \geq V^*(x_1)$, by optimality of M^k_1. Next, it solves the second optimization problem $M^k_2 = \arg\min_{M\in\mathcal{M}} Q^M(x_1, \pi^{M^k_1}, \nu^M)$ and let $\nu^k \leftarrow \nu^{M^k_2}_{\pi^k}$. This corresponds to finding the pessimistic estimate of $V^{\pi^k,\nu^*_{\pi^k}}(x_1)$. In fact, we see that $M^* \in \mathcal{M}^k$ implies $Q^{M^k_2}(x_1, \pi^k, \nu^k) \leq V^{\pi^k,\nu^*_{\pi^k}}(x_1)$, by optimality of M^k_2. This approach appears in Wei et al. (2017) to guarantee convergence to Nash equilibrium, where it is referred to as 'Maximin-EVI'.

In Line 8, the algorithm checks if the values of model M^k_1 and M^k_2 are close to the value of the true model M^* when executing policy π^k and ν^k. If the condition holds, we have

$$V^*(x_1) - V^{\pi^k,\nu^*_{\pi^k}}(x_1) \leq Q^{M^k_1}(x_1, \pi^k, \nu^k) - Q^{M^k_2}(x_1, \pi^k, \nu^k) \leq \epsilon,$$

which means that AOME finds policy π^k that is ϵ-close the NE. One can also switch the order of alternate optimism and obtain ν^k that is ϵ-close the NE. We then terminate the process and output π^k

and ν^k. If the algorithm does not terminate in Line 8, it applies the witnessed model misfit checking method in Sun et al. (2019a). It starts by computing the empirical Bellman error of Markov games defined as follows

$$\widehat{\mathcal{L}}(M_1, M_2, M, h) := \sum_{i=1}^{n_1} \frac{1}{n_1}[Q_h^M(x_h^i, a_h^i, b_h^i) - (r_h^i + Q_h^M(x_{h+1}^i, \pi^{M_1}, \nu_{\pi^{M_1}}^{M_2})]. \tag{5}$$

As proved in Appendix D, in this case $\widehat{\mathcal{L}}(M_1^k, M_2^k, M_j^k, h^k)$ must be greater than $\epsilon/(8H)$ for some $j \in [2]$ and $h^k \in [H]$. Then the algorithm samples additional trajectories at level h^k and shrinks the constraint set by eliminating models with large empirical model misfit, which is defined as follow

$$\widehat{\mathcal{E}}(M_1^k, M_2^k, M, h^k) = \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \frac{1}{n} \mathop{\mathbb{E}}_{(r_h, x_h) \sim M}[g(x_h^i, a_h^i, b_h^i, r_h, x_h) - g(x_h^i, a_h^i, b_h^i, r_h^i, x_{h+1}^i)]. \tag{6}$$

We will show that this constraint set maintains more and more accurate models of the environment. Due to space constraints, the algorithm and theory of our model-based method are deferred to Appendix C.

4.2 Model free algorithm

In this section we propose a model-free algorithm. We consider a more general agnostic learning, where The algorithm is given a policy class $\Pi = \Pi_1 \times \cdots \Pi_H$ with $\Pi_h \subset \{\pi_h : \mathcal{S} \mapsto \Delta\}, h \in [H]$. Notice by letting Π to be the class of NE policies induced by the action-value function class $\mathcal{F}$, we recover the original setup. The goal is to find policy π and ν from Π to minimize the duality gap: $\max_{\pi' \in \Pi} V^{\pi', \nu} - \min_{\nu' \in \Pi} V^{\pi, \nu'}$, by playing only policies from Π. Since only policies from Π are considered, we overload the notation and define the optimal solution in Π as $\pi^* = \arg\max_{\pi \in \Pi} \min_{\nu' \in \Pi} V^{\pi, \nu'}$ and $\nu^* = \arg\min_{\nu \in \Pi} \max_{\pi' \in \Pi} V^{\pi', \nu}$. When Π contains all possible policies, π^* and ν^* is then the Nash equilibrium.

Similar to Section 3, we consider general function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h \subset \{f : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \mapsto [0, 1]\}$. We overload the notations in Section 3 and use ν_π^f to denote the best response of π from Π, namely $\nu_\pi^f(x) := \text{argmin}_{\nu \in \Pi} f(x, \pi, \nu)$. Similarly, we use $\pi_f(x) := \text{argmax}_{\pi \in \Pi} \text{min}_{\nu \in \Pi} f(x, \pi, \nu)$. We also use $\nu_\pi^*(x) := \text{argmin}_{\nu \in \Pi} V^{\pi, \nu}$ to denote the best response of π in the true model.

4.2.1 Alternate Optimistic Value Elimination (AOVE)

Notice that the duality gap can be decomposed as follows

$$\text{gap} = \underbrace{V^{\pi_\nu^*, \nu} - V^{\pi_{\nu^*}^*, \nu^*}}_{\text{P2}} + \underbrace{V^{\pi_{\nu^*}^*, \nu^*} - V^{\pi^*, \nu_{\pi^*}^*}}_{\text{Optimal}} + \underbrace{V^{\pi^*, \nu_{\pi^*}^*} - V^{\pi, \nu_\pi^*}}_{\text{P1}}$$

where the optimal part $V^{\pi_{\nu^*}^*, \nu^*} - V^{\pi^*, \nu_{\pi^*}^*}$ is fixed. Therefore we can consider only the P1 part here and the P2 follows by symmetry. We are interested in the following policy regret

$$\text{Reg}(K) := \sum_{k=1}^{K} (V^{\pi^*, \nu_{\pi^*}^*} - V^{\pi^k, \nu_{\pi^k}^*}).$$

Policy regret measures the extent that P1 can be exploited by policies in Π. Our algorithm Alternate Optimistic Value Elimination (AOVE), presented in Algorithm 2, works on minimizing this regret. Similar to Algorithm 1, it also uses optimism in the initial state to address exploration and exploitation trade-off and constructs a series of confidence sets of hypotheses with small Bellman errors. The differences primarily lie in a different choice of optimism and a different series of confidence sets constructed with the Bellman operator in Eq 2. Since the policy class Π is taken into consideration, the algorithm maintains a constraint set $\mathcal{B} \subset \Pi \times \mathcal{F}$ of policy-function pairs. We use the Bellman operator with regard to fixed P1's policy, in Eq (2). Thus intuitively, this constraint set maintains (π, f) pairs such that $f_h - \mathcal{T}_h^\pi f_{h+1}$ is small for all $h \in [H]$.

We apply the 'alternate optimistic' principle, presented in Line 5 and Line 6. Intuitively, the algorithm finds an optimistic planning of π and pessimistic planning of ν, together corresponding to a max-min

procedure. As such, we form an upper bound $f_1(x_1, \pi, \nu^f_\pi) - g_1(x_1, \pi^k, \nu^g_{\pi^k})$ of the duality gap. Note that it is different from the 'alternate optimistic' used in model-based settings Wei et al. (2017) in that the hypothesis chosen in Line 6 is constrained by the policy chosen in Line 5. This is because in model-based methods, 'plausible' models in the confidence set guarantee small simulation errors (Lemma C.6). However, this is not true for 'plausible' value functions in the confidence set.

The rest of the algorithm aims at minimizing this upper bound sequentially by eliminating bad hypothesis in the constraint set. The elimination process occurs in Line 13, where the algorithm uses history data to eliminate functions with large Bellman error. To estimate Bellman error, we use the similar method as in Algorithm 1, i.e. subtracting the variance of $\mathcal{T}^\pi_h f_{h+1}$ by $g = \arg\inf_{g \in \mathcal{F}_h} \mathcal{E}_{\mathcal{D}_h}(g, f_{h+1}, \pi)$. Notice that a smaller value of $f_h - \mathcal{T}^\pi_h f_{h+1}$ indicates that function approximation f is not easily exploited by P2 when P1 plays π.

As shown in Appendix D, (π^*, Q^*) is always in the constraint set, and the Bellman errors of the rest hypothesis keep shrinking. Thus in the presented algorithm of AOVE, the regret of P1 part $V^{\pi^*, \nu^*_{\pi^*}} - V^{\pi, \nu^*_\pi}$ is controlled. By symmetry, we can perform the same procedures on P2 part and obtain an upper bound on $V^{\pi^*_\nu, \nu} - V^{\pi^*_{\nu^*}, \nu^*}$. Combining these together, we show sublinear regret rate of $V^{\pi^*_{\nu^*}, \nu^*} - V^{\pi^*, \nu^*_{\pi^*}}$. By regret to PAC conversion, this means that we find the policies that is the best achievable in Π.

Algorithm 2 Alternate Optimistic Value Elimination (AOVE)

1: **Input:** Function class $\mathcal{F}$, policy class Π
2: Initialize $\mathcal{B}^0 \leftarrow \Pi \times \mathcal{F}$
3: Set $\beta \leftarrow C \log[\mathcal{N}(\mathcal{F}, 1/K) \cdot \mathcal{N}(\Pi, 1/K) \cdot HK/p]$ for some large constant C
4: **for** $k = 1, 2, \ldots, K$ **do**
5: Find $(\pi^k, f^k) \leftarrow \operatorname{argmax}_{(\pi, f) \in \mathcal{V}^{k-1}} f_1(x_1, \pi, \nu^f_\pi)$
6: Let $g \leftarrow \operatorname*{argmin}_{g:(\pi^k, g) \in \mathcal{V}^{k-1}} g_1(x_1, \pi^k, \nu^g_{\pi^k})$ and set $\nu^k \leftarrow \nu^g_{\pi^k}$
7: **for** step $h = 1, 2, \ldots, H$ **do**
8: P1 takes action $a^k_h \sim \pi^k_h(x^k_h)$, P2 takes action $b^k_h \sim \nu^k_h(x^k_h)$
9: Observe reward r^k_h and move to next state x^k_{h+1}
10: **end for**
11: Update $\mathcal{B}_h \leftarrow \mathcal{B}_h \cup \{(x^k_h, a^k_h, b^k_h, r^k_h, x^k_{h+1})\}, \forall h \in [H]$
12: For all $\xi, \zeta \in \mathcal{F}$, $\pi \in \Pi$ and $h \in [H]$, let

$$\mathcal{E}_{\mathcal{B}_h}(\xi_h, \zeta_{h+1}, \pi) = \sum_{\tau=1}^{k} [\xi_h(x^\tau_h, a^\tau_h, b^\tau_h) - r^\tau_h - \zeta_{h+1}(x^\tau_{h+1}, \pi, \nu^{\zeta_{h+1}}_\pi)]^2$$

13: Update

$$\mathcal{V}^k \leftarrow \{(\pi, f) \in \Pi \times \mathcal{F} : \mathcal{E}_{\mathcal{B}_h}(f_h, f_{h+1}, \pi) \leq \inf_{g \in \mathcal{F}_h} \mathcal{E}_{\mathcal{B}_h}(g, f_{h+1}, \pi) + \beta, \forall h \in [H]\}$$

14: **end for**

4.2.2 Theoretical results

This section presents our theoretical guarantees of Algorithm 2. First, we introduce two assumptions that take the policy class Π into consideration. Variants of these assumptions are also common in batch RL (Jin et al., 2021c; Xie et al., 2021).

Assumption 4.1 (Π-realizability). *For all $\pi \in \Pi$, $Q^{\pi, \nu^*_\pi} \in \mathcal{F}$.*

Assumption 4.2 (Π-completeness). *For all $h \in [H]$, $\pi \in \Pi$, and $f \in \mathcal{F}_{h+1}$, $\mathcal{T}^\pi_h f \in \mathcal{F}_h$ holds.*

Notice that Assumption 4.2 reduces to Q^* realizability in MDP. Both two assumptions hold for linear-parameterized Markov games studied in Xie et al. (2020). In fact, Xie et al. (2020) satisfies a stronger property: for any policy pair π, ν and any $h \in [H]$, there exists a vector $w \in \mathbb{R}^d$ such that

$$Q^{\pi,\nu}_h(x, a, b) = w^\top \phi(x, a, b), \ \forall (x, a, b) \in \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2.$$

Now we present the theory of Algorithm 2. First, we introduce the complexity measure in coordinated setting, which is a variant of Minimax Eluder dimension by replacing the Bellman operator Eq (1) with Eq (2). This Minimax Eluder dimension also allows a distribution family induced by function class $\mathcal{F}$ which is cheap to evaluate in large state space (Jin et al., 2021a).

Definition 4.3 (Minimax Eluder dimension in Coordinated setting)**.** *Let* $(I - \mathcal{T}_h^{\Pi_{h+1}})\mathcal{F} := \{f_h - \mathcal{T}_h^{\pi} f_{h+1} : f \in \mathcal{F}, \pi \in \Pi_{h+1}\}$ *be the class of residuals induced by the Bellman operator* $\mathcal{T}_h^{\pi}$*. Consider the following distribution families: 1)* $\mathcal{D}_{\Delta} = \{\mathcal{D}_{\Delta,h}\}_{h\in[H]}$ *where* $\mathcal{D}_{\Delta,h} := \{\delta(x,a,b) : x \in \mathcal{S}, a \in \mathcal{A}_1, b \in \mathcal{A}_2\}$ *is the set of Dirac measures on state-action pair* (x,a,b)*; 2)* $\mathcal{D}_{\mathcal{F}} = \{\mathcal{D}_{\mathcal{F},h}\}_{h\in[H]}$ *which is generated by executing* $(\pi_f, \nu^g_{\pi_f})$ *for* $f, g \in \mathcal{F}$*. The* ϵ*-Minimax Eluder dimension (in coordinated setting) of* $\mathcal{F}$ *is defined by* $\dim_{\mathrm{ME}'}(\mathcal{F},\epsilon) := \max_{h\in[H]} \min_{\mathcal{D}\in\{\mathcal{D}_{\Delta},\mathcal{D}_{\mathcal{F}}\}} \dim_{\mathrm{DE}}((I - \mathcal{T}_h^{\Pi_{h+1}})\mathcal{F}, \mathcal{D}, \epsilon)$.

Now we present our main theory of Algorithm 2, where we make use of a variant of Minimax Eluder dimension that depends on policy class Π. We use $\mathcal{N}(\Pi,\epsilon)$ to denote the covering number of Π under metric $d(\pi,\pi') := \max_{h\in[H]} \max_{f\in\mathcal{F}_h, x\in\mathcal{S}, b\in\mathcal{A}} |f(x,\pi,b) - f(x,\pi',b)|$.

Theorem 4.4. *Suppose Assumption 4.1 and Assumption 4.2 holds. With probability at least* $1-p$ *we have*

$$\mathrm{Reg}(K) \le O\left(H\sqrt{K d_{\mathrm{ME}'} \log[\mathcal{N}(\mathcal{F},1/K)\mathcal{N}(\Pi,1/K)HK/p]}\right)$$

where $d_{\mathrm{ME}'} = \dim_{\mathrm{ME}'}(\mathcal{F},\epsilon)$.

As shown in this theorem, the regret is sub-linear in number of episodes K and Eluder dimension $d_{\mathrm{ME}'}$ which match the regret bound in Markov decision process, e.g. in Jin et al. (2021a). Through regret-to-PAC conversion, with high probability the algorithm can find ϵ-optimal policy with $O\left(H^2 K d_{\mathrm{ME}'} \log[\mathcal{N}(\mathcal{F},1/K)\mathcal{N}(\Pi,1/K)HK/p]/\epsilon^2\right)$ sample complexity. When the reward and transition kernel have linear structures, we also recover the $\widetilde{O}(\sqrt{d^3K})$ regret of Xie et al. (2020).

Although our upper bound depends on the logarithm of covering number of Π, there are some cases when it is small. For example, when the value-function class $\mathcal{F}$ is finite and Π is the induced policy class of $\mathcal{F}$ as defined by $\Pi = \{\pi_f, \nu^g_{\pi^f} : f, g \in \mathcal{F}\}$, then $\log \mathcal{N}_{\Pi}(\epsilon) = \log|\mathcal{F}|$ and the regret is $\widetilde{O}(H\sqrt{K d_{\mathrm{ME}'} \log^2|\mathcal{F}|})$. Similarly, if the model class $\mathcal{M}$ is finite and Π is induced policy class of $\mathcal{M}$ as defined by $\Pi = \{\pi_M, \nu^{M'}_{\pi^M} : M, M' \in \mathcal{M}\}$, then $\log \mathcal{N}_{\Pi}(\epsilon) = \log|\mathcal{M}|$ and the regret is $\widetilde{O}(H\sqrt{K d_{\mathrm{ME}'} \log^2|\mathcal{M}|})$. In agnostic setting which allows Π to have arbitrary structure, the term $\log \mathcal{N}(\Pi, 1/K)$ reflects the difficulty of learning with complex candidate policy sets and we conjecture that this term is inevitable.

Similar to Algorithm 1, Algorithm 2 solves a non-convex optimization problem in Line 5-6 with highly non-convex constraint set (Line 13). This step is particularly difficult when the function class has complex structures. Algorithm 2 is in general computationally inefficient, even in cases of linear function approximations. Indeed, even when the model is known, solving for the NE is PPAD complete. Developing decentralized and provably efficient RL algorithm for multi-agent Markov game seems a challenging but interesting future direction.

5 Discussion

In this paper we study function approximations in zero-sum simultaneous move Markov games. We design sample efficient algorithms for both decoupled and coordinated learning settings and propose Minimax Eluder dimension to characterize the complexity of decision making with value function approximation. The analysis shows $\sqrt{T}$ regret bounds and half-order dependence on the Minimax Eluder dimension, matching the results in MDPs and improving the linear function approximations. Our algorithms for coordinated setting are based on the idea of the 'alternate optimism', which also shows applicability in model-based methods.

REFERENCES

Yasin Abbasi-Yadkori, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. Exploration-enhanced POLITEX. arXiv preprint arXiv:1908.10479, 2019.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. arXiv preprint arXiv:1908.00261, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, pp. 242–252, 2019.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. arXiv preprint arXiv:2002.04017, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. arXiv preprint arXiv:2006.12007, 2020.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. arXiv preprint arXiv:2102.11494, 2021.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, 1995.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, and Chris Hesse. Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680, 2019.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in neural information processing systems (NeurIPS)*, 2019.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player markov games with linear function approximation. arXiv preprint arXiv:2102.07404, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1675–1685. PMLR, 2019a.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019b.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019c.

Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. arXiv preprint arXiv:2103.10897, 2021.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Jordi Grau-Moya, Felix Leibfried, and Haitham Bou-Ammar. Balancing two-player stochastic games with soft Q-learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 268–274, 2018.

Amy Greenwald, Keith Hall, and Roberto Serrano. Correlated Q-learning. In *International Conference on Machine Learning*, volume 20, pp. 242, 2003.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems (NeurIPS)*, pp. 8571–8580, 2018.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. In *Journal of Machine Learning Research*, volume 11, pp. 1563–1600, 2010.

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. arXiv preprint arXiv:1610.09512, 2016.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1704–1713. JMLR. org, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. arXiv preprint arXiv:2102.00875, 2021a.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. arXiv preprint arXiv:2106.03352, 2021b.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning (ICML)*, 2021c.

Michail G. Lagoudakis and Ronald Parr. Value function approximation in zero-sum markov games. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 283–292. Morgan Kaufmann Publishers Inc., 2002.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Michael L. Littman and Csaba Szepesvari. A generalized reinforcement-learning model: Convergence and applications. In *13th International Conference on Machine Learning*, 1996.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. arXiv preprint arXiv:1906.10306, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, DaanWierstra, and Martin Riedmiller. Playing atari with deep reinforcement learningep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Julien Pérolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pp. 1321–1329, 2015.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning nash equilibrium for general-sum markov games from batch data. In *Artificial Intelligence and Statistics*, pp. 232–241. PMLR, 2017.

Julien Pérolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, pp. 919–928, 2018.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pp. 14410–14420, 2019.

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

Tuomas Sandholm. Super-human ai for strategic reasoning: Beating top pros in heads-up no-limit texas hold'em. In *International Joint Conference on Artificial Intelligence*, pp. 24–25, 2017.

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

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*, pp. 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, and Marc Lanctot. Mastering the game of go with deep neural networks and tree search. *Applied and Computational Harmonic Analysis*, 529(7587):484–489, 2015.

Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in Neural information Processing Systems*, pp. 3422–3435, 2018.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 881–888, 2006.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. arXiv preprint arXiv:1811.08540, 2019a.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory (COLT)*. arXiv preprint arXiv:1811:08540, 2019b.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. arXiv preprint arXiv:1909.01150, 2020a.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. arXiv preprint arXiv:2005.10804, 2020b.

Yining Wang, Ruosong Wang, Simon S. Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2021.

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

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 (COLT)*. arXiv preprint arXiv:2002.07066, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in neural information processing systems (NeurIPS)*, 2021.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. arXiv preprint arXiv:2003.00153, 2020.

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.

Yulai Zhao, Yuandong Tian, Jason D. Lee, and Simon S. Du. Provably efficient policy gradient methods for two-player zero-sum markov games. arXiv preprint arXiv:2102.08903, 2021.

A ADDITIONAL RELATED WORKS

There is a rich literature on sample complexity for single agent RL, mostly focusing on the tabular cases (Strehl et al., 2006; Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018; Russo, 2019; Zanette & Brunskill, 2019) and the linear function approximations (Yang & Wang, 2020; Wang et al., 2021; Abbasi-Yadkori et al., 2019; Jin et al., 2020; Du et al., 2019b). For neural network function approximations, Agarwal et al. (2020); Liu et al. (2019); Wang et al. (2020a); Yang et al. (2020) builds provably efficient algorithms on over-parameterized models. The issue is complicated for general function approximation, as the Q function is hard to learn due to the double-sample issue (Baird, 1995). In recent years, a line of work studies the structural properties and corresponding complexity measures that allow generalization in RL. These work can be roughly grouped with two catagories. One studies structures with low information gain. Jiang et al. (2017) proposes Bellman rank and designs a sample efficient algorithm for a large family of MDPs. Later, Sun et al. (2019b) proposes Witness rank and uses this notion to provide a provably efficient model-based RL algorithm. Recently, Du et al. (2021) proposes Bilinear family and the corresponding Bilinear rank that encapsulates a wide variety of structural properties, spanning from both model-based to model-free function approximations. The other studies structures with low Eluder dimension (Russo & Van Roy, 2013). Osband & Van Roy (2014) extends Eluder dimension to reinforcement learning and provides a unified analysis of model based methods where the regret is controlled by the Kolmogorov dimension and Eluder dimension of model function class. Wang et al. (2020b) designs a model-free algorithm and shows that regret can be controlled by the Eluder dimension of the value function class. Recently, Jin et al. (2021a) proposes Bellman Eluder dimension and builds a new framework to capture a large amount of structures. It is noted that although lots of structures, such as linear MDP (Jin et al., 2020), fit in both two viewpoints, they are not equivalent. Among the above work, Jin et al. (2021a); Zanette et al. (2020) are more related. The method of placing uncertainty quantification on only initial states, previously appeared in Zanette et al. (2020); Jin et al. (2021a), greatly simplifies the argument and improves sample complexity. The Minimax Eluder dimension is inspired by the Bellman Eluder dimension in Jin et al. (2021a).

In competitive RL, the setting with access to a sampling oracle or well explored policies is well studied (Littman & Szepesvari, 1996; Greenwald et al., 2003; Grau-Moya et al., 2018; Pérolat et al., 2018; Srinivasan et al., 2018; Sidford et al., 2020). Under these assumptions, many works consider function approximations (Lagoudakis & Parr, 2002; Pérolat et al., 2015; Fan et al., 2020; Jia et al., 2019; Zhao et al., 2021). However, without strong sampling model or a well explored policy, the issue of exploration-exploitation tradeoff must be addressed. Most of these work focus on tabular setting (Wei et al., 2017; Pérolat et al., 2017; Bai & Jin., 2020; Bai et al., 2021; 2020; Zhang et al., 2020; Zhao et al., 2021) or linear function approximation settings (Xie et al., 2020; Chen et al., 2021). Among them, (Wei et al., 2017; Xie et al., 2020) are more related. The regret-decomposition method used in the decoupled setting is inspired by Xie et al. (2020). The 'alternate optimism' used in Algorithm 4 was previously used as 'Maximin-EVI' in Wei et al. (2017), although the 'alternate optimism' used in Algorithm 2 is different. Moreover, the concurrent work of Jin et al. (2021b) studies competitive RL with general function approximation. They reach a sublinear $\widetilde{O}(\sqrt{K})$ regret for MGs with low minimax Eluder dimensions in the decoupled setting. Their results in the coordinated setting are based on different assumptions and complexity measures. Specifically, they assume finite function class and an additional class of exploiters. It is noted that by letting the policy class be the class of optimal policies induced by the value function class, we have $|\Pi| = |\mathcal{F}|$ and our algorithm achieves on-par guarantee in their setup in terms of the covering numbers of function classes. In addition, we propose a first provably sample efficient and model-based algorithm for the coordinated setting with general function approximation.

B PROOF OF THEOREM 3.6

We prove the Theorem 3.6 in the following five steps. In the first step, we show that Bellman error is small for function in the constraint set. In the second step, we show that Q^* stays in the constraint set throughout the algorithm. In the third step, we decompose the regret into a summation of Bellman error and Martingale difference. We can then bound the summation of Bellman error using Minimax Eluder dimension in step 4. Finally we combine the aforementioned steps and complete the proof of Theorem 3.6.

We define 'optimistic' value function in step h and episode k as follows

$$V_h^k(x_h^k) := f_h^k(x_h^k, \pi_h^k, \widehat{\nu}_h^k)$$

where $\widehat{\nu}_h^k$ is the best response of π_h^k in value function f^k, i.e., $\widehat{\nu}_h^k := \operatorname{argmin}_\nu f_h^k(x_h^k, \pi_h^k, \nu)$. We use $\nu^k = \{\nu_h^k\}_{h=1}^H$ to denote the policy adopted by P2 in the k-th episode. Notice that the agent obtains knowledge from ν only through its actions $b_h^k, h \in [H]$.

Step 1: Low Bellman error in the constraint set We first use a standard concentration procedure (Agarwal et al., 2012; Jin et al., 2021a) to bound the Bellman error in constraint sets.

Lemma B.1 (Concentration on Bellman error)**.** *Let $\rho > 0$ be an arbitrary fixed number. With probability at least $1-p$ for all $(k,h) \in [K] \times [H]$ we have*

$$\sum_{i=1}^{k-1} \big(f_h^k(x_h^i, a_h^i, b_h^i) - r_h^i - \mathop{\mathbb{E}}_{x \sim \mathbb{P}(\cdot|x_h^i, a_h^i, b_h^i)} f_{h+1}^k(x, \pi_{h+1}^k, \widehat{\nu}_{h+1}^k)\big)^2 \le O(\beta).$$

Proof. Consider a fixed (k,h,f) sample, let

$$\begin{aligned} U_t(h,f) := &\big(f_h(x_h^t, a_h^t, b_h^t) - r_h^t - \min_{\nu'} \max_{\pi'} f_{h+1}(x_{h+1}^t, \pi', \nu')\big)^2 \\ &- \big(\mathcal{T}_h f_{h+1}(x_h^t, a_h^t, b_h^t) - r_h^t - \min_{\nu'} \max_{\pi'} f_{h+1}(x_{h+1}^t, \pi', \nu')\big)^2 \end{aligned}$$

and $\mathscr{F}_{t,h}$ be the filtration induced by $\{x_1^i, a_1^i, b_1^i, r_1^i, \ldots, x_H^i\}_{i=1}^{t-1} \cup \{x_1^t, a_1^t, b_1^t, r_1^t, \ldots, x_h^t, a_h^t, b_h^t\}$. Recall the minimax Bellman operator

$$\mathcal{T}_h f_{h+1}(x,a,b) := r(x,a,b) + \mathop{\mathbb{E}}_{x' \sim \mathbb{P}_h(\cdot|x,a,b)} [\min_{\nu'} \max_{\pi'} f_{h+1}(x', \pi', \nu')].$$

We have

$$\begin{aligned} &\mathbb{E}[U_t(h,f)|\mathscr{F}_{t,h}] \\ =& [(f_h - \mathcal{T}_h f_{h+1})(x_h^t, a_h^t, b_h^t)] \\ &\cdot \mathbb{E}[f_h(x_h^t, a_h^t, b_h^t) + \mathcal{T}_h f_{h+1}(x_h^t, a_h^t, b_h^t) - 2r_h^t - 2\min_{\nu'} \max_{\pi'} f_{h+1}(x_{h+1}^t, \pi', \nu')|\mathscr{F}_{t,h}] \\ =& [(f_h - \mathcal{T}_h f_{h+1})(x_h^t, a_h^t, b_h^t)]^2 \end{aligned}$$

and

$$\begin{aligned} &\mathrm{Var}[U_t(h,f)|\mathscr{F}_{t,h}] \le \mathbb{E}[(U_t(h,f))^2|\mathscr{F}_{t,h}] \\ =& [(f_h - \mathcal{T}_h f_{h+1})(x_h^t, a_h^t, b_h^t)]^2 \\ &\cdot \mathbb{E}[\big(f_h(x_h^t, a_h^t, b_h^t) + \mathcal{T}_h f_{h+1}(x_h^t, a_h^t, b_h^t) - 2r_h^t - 2\min_{\nu'} \max_{\pi'} f_{h+1}(x_{h+1}^t, \pi', \nu')\big)^2|\mathscr{F}_{t,h}] \\ \le& 36[(f_h - \mathcal{T}_h f_{h+1})(x_h^t, a_h^t, b_h^t)]^2 = 36\,\mathbb{E}[U_t(h,f)|\mathscr{F}_{t,h}]. \end{aligned}$$

By Freedman's inequality, we have with probability at least $1-p$,

$$\left|\sum_{t=1}^k U_t(h,f) - \sum_{t=1}^k \mathbb{E}[U_t(h,f)|\mathscr{F}_{t,h}]\right| \le O\bigg(\sqrt{\log(1/p)\sum_{t=1}^k \mathbb{E}[U_t|\mathscr{F}_{t,h}]} + \log(1/p)\bigg).$$

Let $\mathscr{N}(\mathcal{F}, \rho)$ be a ρ-cover of $\mathcal{F}$. Taking a union bound for all $(k,h,g) \in [K] \times [H] \times \mathscr{N}(\mathcal{F}, \rho)$, then with probability at least $1-p$ for all $(k,h,g) \in [K] \times [H] \times \mathscr{N}(\mathcal{F}, \rho)$

$$\begin{aligned} &\left|\sum_{t=1}^k U_t(h,g) - \sum_{t=1}^k [(g_h - \mathcal{T}_h g_{h+1})(x_h^t, a_h^t, b_h^t)]^2\right| \\ \le& O\bigg(\sqrt{\iota \sum_{t=1}^k [(g_h - \mathcal{T}_h g_{h+1})(x_h^t, a_h^t, b_h^t)]^2} + \iota\bigg) \end{aligned} \tag{7}$$

where $\iota = \log(HK|\mathscr{N}(\mathcal{F},\rho)|/p)$. Let $g^k = \operatorname{argmin}_{g\in\mathscr{N}(\mathcal{F},\rho)} \max_{h\in[H]} \|f_h^k - g_h^k\|_\infty$. We immediately have

$$\begin{aligned}&\left|\sum_{t=1}^k U_t(h,g^k) - \sum_{t=1}^k [(g_h^k - \mathcal{T}_h g_{h+1}^k)(x_h^t, a_h^t, b_h^t)]^2\right| \\ \leq\ & O\left(\sqrt{\iota \sum_{t=1}^k [(g_h^k - \mathcal{T}_h g_{h+1}^k)(x_h^t, a_h^t, b_h^t)]^2 + \iota}\right). \end{aligned} \tag{8}$$

For all $(h,k) \in [H]\times[K]$, by the definition of $\mathcal{V}^k$ and $f^k \in \mathcal{V}^k$ we have

$$\begin{aligned}\sum_{t=1}^k U_t(h,f^k) =\ & \sum_{t=1}^k \left(f_h^k(x_h^t,a_h^t,b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} f_{h+1}(x_{h+1}^t,\pi',\nu')\right)^2 \\ & - \sum_{t=1}^k \left(\mathcal{T}_h f_h^k(x_h^t,a_h^t,b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} f_{h+1}(x_{h+1}^t,\pi',\nu')\right)^2 \\ \leq\ & \sum_{t=1}^k \left(f_h^k(x_h^t,a_h^t,b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} f_{h+1}(x_{h+1}^t,\pi',\nu')\right)^2 \\ & - \inf_{g\in\mathcal{F}} \sum_{t=1}^k \left(g_h(x_h^t,a_h^t,b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} f_{h+1}(x_{h+1}^t,\pi',\nu')\right)^2 \\ \leq\ & O(\iota + k\rho). \end{aligned} \tag{9}$$

It thus follows that,

$$\begin{aligned}\sum_{t=1}^k [(f_h^k - \mathcal{T}_h f_{h+1}^k)(x_h^t,a_h^t,b_h^t)]^2 &\leq \sum_{t=1}^k [(g_h^k - \mathcal{T}_h g_{h+1}^k)(x_h^t,a_h^t,b_h^t)]^2 + O(\iota + k\rho) \\ &\leq O(\sum_{t=1}^k U_t(h,g^k) + \iota + k\rho) \\ &\leq O(\sum_{t=1}^k U_t(h,f^k) + \iota + k\rho) \\ &\leq O(\iota + k\rho)\end{aligned}$$

where the first step is due to g^k is an ρ-approximation to f^k, the second step comes from Eq (8), the third step is due to g^k is an ρ-approximation to f^k, and the final step comes from Eq (9).

By definition of π^k and $\widehat{\nu}^k$, we have

$$\mathcal{T}_h f_{h+1}^k(x,a,b) := r(x,a,b) + \mathop{\mathbb{E}}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)}[f_{h+1}(x',\pi_{h+1}^k,\widehat{\nu}_{h+1}^k)]$$

thus we conclude the proof. □

Step 2: Q^* is always in constraint set.

Lemma B.2. *With probability at least $1-p$, we have $Q^* \in \mathcal{V}^k$ for all $k \in [K]$.*

Proof. Consider a fixed (k,h,f) sample, let

$$\begin{aligned}U_t(h,f) := &\left(f_h(x_h^t,a_h^t,b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} Q_{h+1}^*(x_{h+1}^t,\pi',\nu')\right)^2 \\ &- \left(Q_h^*(x_h^t,a_h^t,b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} Q_{h+1}^*(x_{h+1}^t,\pi',\nu')\right)^2\end{aligned}$$

and $\mathscr{F}_{t,h}$ be the filtration induced by $\{x_1^i,a_1^i,b_1^i,r_1^i,\ldots,x_H^i\}_{i=1}^{t-1} \cup \{x_1^t,a_1^t,b_1^t,r_1^t,\ldots,x_h^t,a_h^t,b_h^t\}$.

We have

$$\mathbb{E}[U_t(h,f)|\mathscr{F}_{t,h}] = [(f_h - Q_h^*)(x_h^t, a_h^t, b_h^t)]^2$$

and

$$\mathrm{Var}[U_t(h,f)|\mathscr{F}_{t,h}] \leq 36[(f_h^k - \mathcal{T}_h f_{h+1}^k)(x_h^t, a_h^t, b_h^t)]^2 = 36\,\mathbb{E}[U_t(h,f)|\mathscr{F}_{t,h}].$$

Let $\mathscr{N}(\mathcal{F},\rho)$ be a ρ-cover of $\mathcal{F}$. By Freedman's inequality, for all $(k,h,f) \in [K]\times[H]\times\mathscr{N}(\mathcal{F},\rho)$ we have with probability at least $1-p$,

$$\left|\sum_{t=1}^k U_t(h,f) - \sum_{t=1}^k [(f_h - Q_h^*)(x_h^t, a_h^t, b_h^t)]^2\right|$$
$$\leq O\left(\sqrt{\log(1/p)\sum_{t=1}^k [(f_h - Q_h^*)(x_h^t, a_h^t, b_h^t)]^2} + O(\iota)\right)$$

where $\iota = \log(HK|\mathscr{N}(\mathcal{F},\rho)|/p)$. Since $[(f_h - Q_h^*)(x_h^t, a_h^t, b_h^t)]^2 \geq 0$, this implies

$$-\sum_{t=1}^k U_t(h,g) \leq O(\iota).$$

Therefore for all $(k,h,g) \in [K]\times[H]\times\mathscr{N}(\mathcal{F},\rho)$

$$\sum_{t=1}^k \left(Q_h^*(x_h^t, a_h^t, b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} Q_{h+1}^*(x_{h+1}^t, \pi', \nu')\right)^2$$
$$\leq \sum_{t=1}^k \left(f_h(x_h^t, a_h^t, b_h^t) - r_h^t - \min_{\nu'}\max_{\pi'} Q_{h+1}^*(x_{h+1}^t, \pi', \nu')\right)^2 + O(\iota + k\rho).$$

Thus we have proven that $Q^* \in \mathcal{V}^k, \forall k \in [K]$ with probability at least $1-p$.

□

Remark B.3. *This step implies that*

$$V^* \leq V_1^k(x_1^k) \tag{10}$$

which ensures optimism from the beginning state ('global optimism').

Step 3: Regret decomposition The following result is key to our analysis. It decomposes the deviation from the value of the game into Bellman errors across time steps and a martingale difference sequence.

Lemma B.4. *Define*

$$\delta_h^k := V_h^k(x_h^k) - V_h^{\pi^k,\nu^k}(x_h^k)$$
$$\zeta_h^k := \mathbb{E}[\delta_{h+1}^k | x_h^k, a_h^k, b_h^k] - \delta_{h+1}^k$$
$$\gamma_h^k := \mathop{\mathbb{E}}_{a\sim\pi_h^k(x_h^k)}[f_h^k(x_h^k, a, b_h^k)] - f_h^k(x_h^k, a_h^k, b_h^k)$$
$$\widehat{\gamma}_h^k := \mathop{\mathbb{E}}_{a\sim\pi_h^k(x_h^k), b\sim\nu_h^k(x_h^k)}[Q_h^{\pi^k,\nu^k}(x_h^k, a, b)] - Q_h^{\pi^k,\nu^k}(x_h^k, a_h^k, b_h^k)$$

Then we have

$$\delta_h^k \leq \delta_{h+1}^k - \zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k + \epsilon_h^k \tag{11}$$

where ϵ_h^k is the bellman error defined by

$$\epsilon_h^k := f_h^k(x_h^k, a_h^k, b_h^k) - r_h^k - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^k, a_h^k, b_h^k)} f_{h+1}^k(x, \pi_{h+1}^k, \widehat{\nu}_{h+1}^k).$$

Proof. By definition of $\widehat{\nu}_h^k = \text{argmin}_\nu f_h^k(x_h^k, \pi_h^k, \nu)$, we have

$$\begin{aligned} f_h^k(x_h^k, \pi_h^k, \widehat{\nu}_h^k) &\leq f_h^k(x_h^k, \pi_h^k, b_h^k) \\ &= f_h^k(x_h^k, a_h^k, b_h^k) + \gamma_h^k. \end{aligned}$$

It thus follows that

$$\begin{aligned} \delta_h^k &\leq f_h^k(x_h^k, a_h^k, b_h^k) - Q_h^{\pi^k,\nu^k}(x_h^k, a_h^k, b_h^k) + \gamma_h^k - \widehat{\gamma}_h^k \\ &= r_h^k + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^k,a_h^k,b_h^k)} f_{h+1}^k(x, \pi_{h+1}^k, \widehat{\nu}_{h+1}^k) + \epsilon_h^k \\ &\quad - (r_h^k + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot,x_h^k,a_h^k,b_h^k)} [V_{h+1}^{\pi^k,\nu^k}(x)|x_h^k, a_h^k, b_h^k]) + \gamma_h^k - \widehat{\gamma}_h^k \\ &= f_h^k(x_{h+1}^k, \pi_{h+1}^k, \widehat{\nu}_{h+1}^k) - V_{h+1}^{\pi^k,\nu^k}(x_{h+1}^k) - \zeta_h^k + \epsilon_h^k + \gamma_h^k - \widehat{\gamma}_h^k \\ &= \delta_{h+1}^k - \zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k + \epsilon_h^k. \end{aligned}$$

Therefore we complete the proof. □

Step 4: Bounding cumulative Bellman error using Minimax Eluder dimension. Recall that Lemma B.1 gives:

$$\sum_{i=1}^{k-1} \left(f_h^k(x_h^i, a_h^i, b_h^i) - r_h^i - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^i,a_h^i,b_h^i)} f_{h+1}^k(x, \pi_{h+1}^k, \widehat{\nu}_{h+1}^k)\right)^2 \leq O(\beta).$$

Combining this and Lemma F.1 with

$$g_k = f_h^k(\cdot,\cdot,\cdot) - r_h^k - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|\cdot,\cdot,\cdot)} f_{h+1}^k(x, \pi_{h+1}^k, \widehat{\nu}_{h+1}^k)$$

and $\rho_k = \delta(x_n^k, a_h^k, b_h^k)$, we have come to the following result.

Lemma B.5. *For any $(k,h) \in [K] \times [H]$ we have*

$$\begin{aligned} &\sum_{k=1}^{K} |f_h^k(x_h^k, a_h^k, b_h^k) - r_h^k - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^k,a_h^k,b_h^k)} f_{h+1}^k(x, \pi_{h+1}^k, \widehat{\nu}_{h+1}^k)| \\ &\leq O\left(\sqrt{K \cdot \dim_{\text{ME}}(\mathcal{F}, \sqrt{1/K}) \log(HK\mathcal{N}(\mathcal{F}, 1/K)/p)}\right). \end{aligned}$$

Thus

$$\sum_{k=1}^{K} \epsilon_h^k \leq O\left(\sqrt{K \cdot \dim_{\text{ME}}(\mathcal{F}, \sqrt{1/K}) \log(HK\mathcal{N}(\mathcal{F}, 1/K)/p)}\right). \tag{12}$$

Step 5: Putting everything together By Eq (10) and Eq (11)

$$\begin{aligned} \sum_{k=1}^{K} [V_1^*(x_1) - V_1^{\pi^k,\nu^k}(x_1)] &\leq \sum_{k=1}^{K} [V_1^k(x_1) - V_1^{\pi^k,\nu^k}(x_1)] \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} (-\zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k + \epsilon_h^k) \\ &= \sum_{h=1}^{H}\sum_{k=1}^{K} (-\zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k) + \sum_{h=1}^{H}\sum_{k=1}^{K} \epsilon_h^k. \end{aligned}$$

Notice that $\sum_{h=1}^{H}\sum_{k=1}^{K}(-\zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k)$ is a martingale difference sequence and can be bounded by $O(H\sqrt{KH\log(HK/p)})$ with probability at least $1-p$.

By Eq (12) the term $\sum_{h=1}^{H}\sum_{k=1}^{K}\epsilon_h^k$ can be bounded by

$$O\left(H\sqrt{K\cdot\dim_{\mathrm{ME}}(\mathcal{F},\sqrt{1/K})\log(HK\mathcal{N}(\mathcal{F},1/K)/p)}\right)$$

with probability at least $1-p$.

It thus follows that regret can be bounded by

$$\mathrm{Reg}(K)\leq O\left(H\sqrt{K\cdot\dim_{\mathrm{ME}}(\mathcal{F},\sqrt{1/K})\log(HK\mathcal{N}(\mathcal{F},1/K)/p)}\right).$$

B.1 Linear function approximation

In the linear function approximation setting of Xie et al. (2020),

$$r_h(x,a,b)=\phi(x,a,b)^\top\theta_h,\ \ \mathbb{P}_h(\cdot|x,a,b)=\phi(x,a,b)^\top\mu_h(\cdot)$$

where $\phi(\cdot,\cdot,\cdot)\in\mathbb{R}^d$ is known feature vector and $\theta_h\in\mathbb{R}^d$ and $\mu_h(\cdot)\in\mathbb{R}^d$ are unknown with $\|\theta\|_2\leq\sqrt{d}$, $\|\nu_h(\cdot)\|_2\leq\sqrt{d}$ and $\|\phi(\cdot,\cdot)\|_2\leq 1$. In this case Algorithm 1 reduces to the following Algorithm 3. Notice that this algorithm can also be seen as a generalization of Zanette et al. (2020) to Markov games.

Algorithm 3 ONEMG for linear function approximation.

Input: Function class $\mathcal{F}$
Set $\beta\leftarrow C\sqrt{d}\log(HK/p)$ for some large constant C
for episode $k=1,2,\ldots,K$ **do**
  Receive initial state x_1^k
  Set $w_{H+1}^k=\theta_{H+1}^k=0$
  Find $\max_{w_h^k,\theta_h^k,h\in[H]}V_1^k(x_1^k)$ such that for all $h\in[H]$:

$$\begin{aligned}
&(1)\ \ w_h^k=(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau,b_h^\tau)[r_h^\tau+V_{h+1}^k(x_{h+1}^\tau)]\\
&(2)\ \ \|\theta_h^k-w_h^k\|_{\Lambda_h^k}\leq C\cdot H\beta\\
&(3)\ \ Q_h^k(\cdot,\cdot,\cdot)=(\theta_h^k)^\top\phi(\cdot,\cdot,\cdot)\\
&(4)\ \ (\pi_h^k,B_0)=\mathrm{NE}(Q_h^k)\\
&(5)\ \ V_h^k(\cdot)=\mathop{\mathbb{E}}_{a\sim\pi_h^k,b\sim B_0}[Q_h^k(\cdot,a,b)]
\end{aligned}$$

  for step $h=1,2,\ldots,H$ **do**
    P1 plays action $a_h^k\sim\pi_h^k(x_h^k)$
    P2 takes action b_h^k
    Observe reward r_h^k and move to next state x_{h+1}^k
  end for
end for

We will prove the following theoretical result of Algorithm 3. Notice the dependency on d of Algorithm 3 is d while the dependency on d of Xie et al. (2020) is $d^{3/2}$, thus this result improves theirs by $\sqrt{d}$ factor.

Theorem B.6. *Regret (Eq* (3)*) in Algorithm 3 is bounded by* $O(d\sqrt{H^2K\log(dHK/p)})$ *with probability at least* $1-p$.

B.1.1 Proof of Theorem B.6

We prove Theorem B.6 in the following five steps similar to the previous section.

Step 1: Low Bellman error in the constraint set We use some standard results from previous work. Notice that the $\mathcal{N}(\mathcal{F}, \epsilon) = O(d \log(1/\epsilon))$ and our inherent Bellman error is zero, thus these lemmata can be directly adapted into our setting.

Lemma B.7 (Concentration, adapted from Lemma 1 of Zanette et al. (2020))**.** *With probability at least $1 - p$, the following holds for all $(k, h) \in [K] \times [H]$,*

$$\left\| \sum_{\tau \in [k-1]} \phi_h^\tau [V_{h+1}^k(x_{h+1}^\tau) - \mathop{\mathbb{E}}_{x \sim \mathbb{P}(\cdot | x_h^\tau, a_h^\tau, b_h^\tau)} V_{h+1}^k(x)] \right\|_{(\Lambda_h^k)^{-1}} \leq H\beta$$

where $\beta = \Omega(\sqrt{d \log(dKH/p)})$.

Lemma B.8 (Least square error bound, adapted from Lemma 3 of Xie et al. (2020))**.** *On the event of Lemma B.7, the following holds for all $(x, a, b, h, k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{A} \times [H] \times [K]$ and any policy pair (π, ν)*

$$\left| \langle \phi(x, a, b), \theta_h^k \rangle - Q_h^{\pi,\nu}(x, a, b) - \mathop{\mathbb{E}}_{x \sim \mathbb{P}(\cdot | x, a, b)} (V_{h+1}^k - V_{h+1}^{\pi,\nu})(x) \right| \leq \beta \|\phi(x, a, b)\|_{(\Lambda_h^k)^{-1}}$$

Step 2: Q^* always stays in the constraint set

Lemma B.9 (Optimism)**.** *Let $Q_h^*(\cdot, \cdot, \cdot) = \phi(\cdot, \cdot, \cdot)^\top \theta_h^*$, then on the event of Lemma B.8 and Lemma B.7 there exists $w_h^*, h \in [H]$ such that $w_h^*, \theta_h^*, h \in [H]$ satisfies the constraints of Line 6 in Algorithm 3.*

Proof. We prove by backward induction on h. The claim trivially holds for $h = H + 1$. Now assume the inductive hypothesis holds at $h + 1$. Let

$$\begin{aligned} w_h^* &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau)[r_h^\tau + V_{h+1}^k(x_{h+1}^\tau)] \\ &= (\sum_{\tau=1}^{k-1} \phi_h^\tau (\phi_h^\tau)^\top + I)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau (r_h^\tau + \mathop{\mathbb{E}}_{x \sim \mathbb{P}(\cdot | x_h^\tau, a_h^\tau, b_h^\tau)} V_{h+1}^k(x) + \eta_h^\tau) \end{aligned}$$

where $\phi_h^\tau = \phi(x_h^\tau, a_h^\tau, b_h^\tau)$ and $\eta_h^\tau = V_{h+1}^k(x_{h+1}^\tau) - \mathbb{E}_{x \sim \mathbb{P}(\cdot | x_h^\tau, a_h^\tau, b_h^\tau)} V_{h+1}^k(x)$. Notice that inductive hypothesis implies $r_h^\tau + \mathbb{E}_{x \sim \mathbb{P}(\cdot | x_h^\tau, a_h^\tau, b_h^\tau)} V_{h+1}^k(x) = \phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top \theta_h^*$, therefore

$$\begin{aligned} w_h^* &= (\sum_{\tau=1}^{k-1} \phi_h^\tau (\phi_h^\tau)^\top + I)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \big((\phi_h^\tau)^\top \theta_h^* + \eta_h^\tau \big) \\ &= \theta_h^* - (\Lambda_h^k)^{-1} \theta_h^* + (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \eta_h^\tau. \end{aligned}$$

It thus suffices to bound the following

$$\begin{aligned} \|(\Lambda_h^k)^{-1} \theta_h^* + (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \eta_h^\tau\|_{(\Lambda_h^k)} &\leq \|\theta_h^*\|_{(\Lambda_h^k)^{-1}} + \| \sum_{\tau=1}^{k-1} \phi_h^\tau \eta_h^\tau \|_{(\Lambda_h^k)^{-1}} \\ &\leq \sqrt{d} + H\beta \\ &\leq C \cdot H\beta \end{aligned}$$

where the penultimate step comes from $\|\theta_h^*\|_2 \leq \sqrt{d}$ and Lemma B.7. □

Remark B.10. *This fact implies that*

$$V^*(x_1^k) \leq V_1^k(x_1^k). \tag{13}$$

Step 3: Recursive decomposition of regret

Lemma B.11 (Regret decomposition)**.** *Define*

$$\begin{aligned}
\delta_h^k &:= V_h^k(x_h^k) - V_h^{\pi^k,\nu^k}(x_h^k) \\
\zeta_h^k &:= \mathbb{E}[\delta_{h+1}^k | x_h^k, a_h^k, b_h^k] - \delta_{h+1}^k \\
\gamma_h^k &:= \mathop{\mathbb{E}}_{a\sim\pi_h^k(x_h^k)}[Q_h^k(x_h^k, a, b_h^k)] - Q_h^k(x_h^k, a_h^k, b_h^k) \\
\widehat{\gamma}_h^k &:= \mathop{\mathbb{E}}_{a\sim\pi_h^k(x_h^k), b\sim\nu_h^k(x_h^k)}[Q_h^{\pi^k,\nu^k}(x_h^k, a, b)] - Q_h^{\pi^k,\nu^k}(x_h^k, a_h^k, b_h^k)
\end{aligned}$$

Then we have

$$\delta_h^k \le \delta_{h+1}^k - \zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k + \epsilon_h^k \tag{14}$$

where ϵ_h^k is the width defined by

$$\epsilon_h^k := 2\beta\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}.$$

Proof. By definition of $B_0 = \text{argmin}_\nu f_h^k(x_h^k, \pi_h^k, \nu)$, we have

$$\begin{aligned}
V_h^k(x_h^k) &= Q_h^k(x_h^k, \pi_h^k, B_0) \\
&\le Q_h^k(x_h^k, \pi_h^k, b_h^k) \\
&= Q_h^k(x_h^k, a_h^k, b_h^k) + \gamma_h^k.
\end{aligned}$$

It thus follows that

$$\begin{aligned}
\delta_h^k &\le Q_h^k(x_h^k, a_h^k, b_h^k) - Q_h^{\pi^k,\nu^k}(x_h^k, a_h^k, b_h^k) + \gamma_h^k - \widehat{\gamma}_h^k \\
&\le \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^k,a_h^k,b_h^k)}[V_{h+1}^k(x)] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^k,a_h^k,b_h^k)}[V_{h+1}^{\pi^k,\nu^k}(x)] + \epsilon_h^k + \gamma_h^k - \widehat{\gamma}_h^k \\
&= V_{h+1}^k(x_{h+1}^k) - V_{h+1}^{\pi^k,\nu^k}(x_{h+1}^k) - \zeta_h^k + \epsilon_h^k + \gamma_h^k - \widehat{\gamma}_h^k \\
&= \delta_{h+1}^k - \zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k + \epsilon_h^k
\end{aligned}$$

where the second step comes from Lemma B.8 by letting $(x, a, b) = (x_h^k, a_h^k, b_h^k)$ (notice that $\langle\phi(x_h^k, a_h^k, b_h^k), \theta_h^k\rangle = Q_h^k(x_h^k, a_h^k, b_h^k)$). □

Step 4: Bounding width by Elliptical Potential Lemma This step directly makes use of the following standard result.

Lemma B.12 (Elliptical Potential Lemma, Lemma 10 of Xie et al. (2020))**.** *Suppose $\{\phi_t\}_{t\ge0}$ is a sequence in $\mathbb{R}^d$ satisfying $\|\phi_t\| \le 1, \forall t$. Let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix, and $\Lambda_t = \Lambda_0 + \sum_{i\in[t]}\phi_i\phi_i^\top$. If the smallest eigenvalue of Λ_0 is lower bounded by 1, then*

$$\log(\frac{\det\Lambda_t}{\det\Lambda_0}) \le \sum_{i\in[t]}\phi_i^\top\Lambda_{j-1}^{-1}\phi_i \le 2\log(\frac{\det\Lambda_t}{\det\Lambda_0}).$$

Step 5: Putting everything together By Eq (13) and Eq (14)

$$\begin{aligned}
\sum_{k=1}^K[V_1^*(x_1) - V_1^{\pi^k,\nu^k}(x_1)] &\le \sum_{k=1}^K[V_1^k(x_1) - V_1^{\pi^k,\nu^k}(x_1)] \\
&\le \sum_{k=1}^K\sum_{h=1}^H(-\zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k + \epsilon_h^k) \\
&= \sum_{h=1}^H\sum_{k=1}^K(-\zeta_h^k + \gamma_h^k - \widehat{\gamma}_h^k) + \sum_{h=1}^H\sum_{k=1}^K\epsilon_h^k.
\end{aligned}$$

Notice that $\sum_{h=1}^{H}\sum_{k=1}^{K}(-\zeta_h^k+\gamma_h^k-\widehat{\gamma}_h^k)$ is a martingale difference sequence and can be bounded by $H\sqrt{KH\log(HK/p)}$ with probability at least $1-p$.

By Lemma B.12 the term $\sum_{h=1}^{H}\sum_{k=1}^{K}\epsilon_h^k$ can be bounded as follows

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}\epsilon_h^k &= \sum_{h=1}^{H}\sum_{k=1}^{K}2\beta\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}\\ &\leq \sum_{h=1}^{H}\sqrt{K}\cdot\sqrt{\sum_{k=1}^{K}2\beta(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}\\ &\leq 2\beta\sum_{h=1}^{H}\sqrt{K}\cdot\sqrt{2\log(\frac{\det\Lambda_h^k}{\det\Lambda_h^0})}\\ &\leq 2\beta\sum_{h=1}^{H}\sqrt{K}\cdot\sqrt{2\log\frac{(\lambda+K\max_k\|\phi_h^k\|^2)^d}{\lambda^d}}\\ &\leq O(dH\sqrt{K\log(dHK/p)}).\end{aligned}$$

with probability at least $1-p$.

It thus follows that with probability at least $1-p$ regret can be upper bounded by

$$O(d\sqrt{H^2K\log(dHK/p)}).$$

C MODEL-BASED METHOD FOR THE COORDINATED SETTING

C.1 ALTERNATE OPTIMISTIC MODEL ELIMINATION (AOME)

Algorithm 4 Alternate Optimistic Model Elimination (AOME)

1: **Input:** Model class $\mathcal{M}$, test function class $\mathcal{G}$, precision ϵ, failure probability p
2: Initialize $\mathcal{M}^0 \leftarrow \mathcal{M}$
3: **for** $k=1,2,\ldots,K$ **do**
4: Find $M_1^k \leftarrow \text{argmax}_{M\in\mathcal{M}} Q^M(x_1,\pi^M,\nu^M)$, let $\pi^k \leftarrow \pi^{M_1^k}$
5: Find $M_2^k \leftarrow \text{argmin}_{M\in\mathcal{M}} Q^M(x_1,\pi^k,\nu_{\pi^k}^M)$, let $\nu^k \leftarrow \nu_{\pi^k}^{M_1^k}$
6: Execute π^k,ν^k and collect n_1 rewards $\{r_h^i\}_{h\in[H],i\in[n_1]}$
7: Let $\widehat{V}^k \leftarrow \frac{1}{n_1}\sum_{i=1}^{n_1}(\sum_{h=1}^{H}r_h^i)$
8: **if** $\max\{|\widehat{V}^k - Q^{M_1^k}(x_1,\pi^k,\nu^k)|, |\widehat{V}^k - Q^{M_2^k}(x_1,\pi^k,\nu^k)|\} \leq \epsilon/2$ **then**
9: Terminate and output π^k,ν^k.
10: **else**
11: Find h^k such that $\max_{i\in[2]}|\widehat{\mathcal{L}}(M_1^k,M_2^k,M_i^k,h^k)| \geq \epsilon/(4H)$ by Eq (5).
12: Collect n trajectories $\{(x_h^i,a_h^i,b_h^i,r_h^i)_{h=1}^H\}_{i=1}^n$ by executing $a_h,b_h\sim\pi_h^k,\nu_h^k,\forall h\in[H]$
13: Update constraint set, where $\widehat{\mathcal{E}}$ is given by Eq (6)

$$\mathcal{M}^k \leftarrow \{M\in\mathcal{M}^{k-1} : \widehat{\mathcal{E}}(M_1^k,M_2^k,M,h^k)\leq\phi\}$$

14: **end if**
15: **end for**

C.2 THEORY FOR ALTERNATE OPTIMISTIC MODEL ELIMINATION (AOME)

In this section we prove that Algorithm 4 terminates in finite rounds. The technique bulk is then showing that in Step 13 the constraint set can only shrink for finite times. This is dependent on the complexity of model class $\mathcal{M}$ and discriminator function class $\mathcal{G}$. We first generalize Witness rank (Sun et al., 2019b) to Markov games.

Definition C.1 (Witnessed model misfit in Markov games)**.** *For discriminator class $\mathcal{G} : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times \mathbb{R} \times \mathcal{S} \mapsto \mathbb{R}$, models $M_1, M_2, M \in \mathcal{M}$ and level $h \in [H]$, the witnessed model misfit of Markov game at level h is defined as follow*

$$\mathcal{E}(M_1, M_2, M, h) := \sup_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{(x,a,b)\sim\mathcal{P}_{M_1}^{M_2}} \mathop{\mathbb{E}}_{(r,x)\sim M} [g(x_h, a_h, b_h, r, x) - g(x_h, a_h, b_h, r_h, x_{h+1})]. \quad (15)$$

where $(x, a, b) \sim \mathcal{P}_{M_1}^{M_2}$ denotes $x_{\tau+1} \sim \mathbb{P}(\cdot|x_\tau, a_\tau, b_\tau)$, $a_\tau \sim \pi^{M_1}$ and $b_\tau \sim \nu_{\pi^{M_1}}^{M_2}$ for all $\tau \in [h]$.

Using $g(x_h, a_h, b_h, r, x) = r(x_h, a_h, b_h) + Q_M(x, \pi^{M_1}, \nu_{\pi^{M_1}}^{M_2})$, we can also give a generalization of Bellman error (Jiang et al., 2016) in Markov games by the following

$$\mathcal{L}(M_1, M_2, M, h) := \mathop{\mathbb{E}}_{(x,a,b)\sim\mathcal{P}_{M_1}^{M_2}} \left[\mathop{\mathbb{E}}_{(r,x)\sim M} [g(x_h, a_h, b_h, r, x))] - \mathop{\mathbb{E}}_{(r,x)\sim M^*} [g(x_h, a_h, b_h, r, x)] \right]. \quad (16)$$

The final components in our theory are two assumptions from Sun et al. (2019b). Define $\mathrm{rank}(A, \beta)$ to be the smallest integer k such that $A = UV^\top$ with $U, V \in \mathbb{R}^{n\times k}$ and $\|U_{i,*}\|_2 \cdot \|V_{j,*}\|_2 \le \beta$ for any pair of rows $U_{i,*}, V_{j,*}$.

Assumption C.2 (Realizability)**.** *Assume $M^* \in \mathcal{M}$.*

Assumption C.3 (Witness misfit domination)**.** *Assume $\mathcal{G}$ is finite, $\|g\|_\infty \le 2, \forall g \in \mathcal{G}$ and $\forall M_1, M_2, M \in \mathcal{M} : \mathcal{E}(M_1, M_2, M, h) \ge \mathcal{L}(M_1, M_2, M, h)$.*

Now we introduce the Witness rank for Markov games.

Definition C.4 (Witness rank for Markov games)**.** *Given a model class $\mathcal{M}$, a test function class $\mathcal{G}$ and $\kappa \in (0, 1]$. For $h \in [H]$, we define $\mathcal{N}_{\kappa,h} \subset \mathbb{R}^{|\mathcal{M}|^2\times|\mathcal{M}|}$ as the set of matrices such that any matrix $A \in \mathcal{N}_{\kappa,h}$ satisfies*

$$\kappa|\mathcal{L}(M_1, M_2, M, h)| \le A((M_1, M_2), M) \le \mathcal{E}(M_1, M_2, M, h), \ \forall M_1, M_2, M \in \mathcal{M}.$$

The Witness rank for Markov games is defined by

$$W(\kappa, \beta, \mathcal{M}, \mathcal{G}, h) := \min_{A\in\mathcal{N}_{\kappa,h}} \mathrm{rank}(A, \beta).$$

We are in the position to state the main theorem of this section.

Theorem C.5. *Under Assumption C.2 and Assumption C.3, set $\phi = \frac{\kappa\epsilon}{100H\sqrt{W_\kappa}}$ where $W_\kappa = \max_{h\in[H]} W(\kappa, \beta, \mathcal{M}, \mathcal{G}, h)$, let $T = HW_\kappa \log(\beta/2\phi)$, set $n_1 = C \cdot H^2 \log(HT/p)/\epsilon^2$ and $n = C \cdot H^2 W_\kappa |\mathcal{A}| \log(T|\mathcal{M}||\mathcal{G}|/p)/(\kappa\epsilon)^2$ for large enough constant C. Then for all $\epsilon, p, \kappa \in (0, 1]$, with probability at least $1 - p$ Algorithm 4 outputs a policy π such that $V^*(x_1) - V^{\pi,\nu_\pi^*}(x_1) \le \epsilon$ within at most $O(\frac{H^3 W_\kappa^2 |\mathcal{A}|}{\kappa^2\epsilon^2} \log(T|\mathcal{G}||\mathcal{M}|/p))$ trajectories.*

As seen in the above theorem, Algorithm 4 has sample complexity that scales quadratic with Witness rank and logarithmically with cardinality of model class. This matches the sample complexity results in the Markov decision process setting (Sun et al., 2019b). When the test function is Q-function, the Witnessed model misfit reduced to Bellman error as seen in Eq (16) and witness rank reduce to Bellman rank.

C.3 Proofs

We prove Theorem C.5 in the following steps. In the first step, we show a simulation lemma that generalize Jiang et al. (2017) to Markov games. In the second step we show concentration of empirical estimates of Bellman error, model misfit and value functions. In the third step we show that M^* always stays in the constraint set. We examine the conditions when Algorithm 4 terminates or not in the fourth step. Using these conditions, we bound the number of rounds in the fifth step. Finally we combine everything and complete the proof of Theorem C.5.

From Definition C.4 we know that there exists $A_h \in \mathcal{N}_{\kappa,h}$ such that

$$\kappa|\mathcal{L}(M_1, M_2, M, h)| \le A_h((M_1, M_2), M) \le \mathcal{E}(M_1, M_2, M, h), \ \forall M_1, M_2, M \in \mathcal{M}.$$

and we can factorize $A_h((M_1, M_2), M) = \langle \zeta_h(M_1, M_2), \chi_h(M)\rangle$ where $\|\zeta_h(M_1, M_2)\|_2, \|\chi_h(M)\|_2 \le \beta$.

Step 1: Simulation Lemma The analysis begins with a useful Lemma that decouples the value difference into a sum of Bellman errors across time steps.

Lemma C.6 (Simulation Lemma)**.** *Fix model $M_1, M_2, M \in \mathcal{M}$. Let $\pi = \pi^{M_1}, \nu = \nu^{M_2}_{\pi^{M_1}}$. Under Assumption C.3, we have*

$$Q_M(x_1, \pi, \nu) - V_1^{\pi,\nu}(x_1) = \sum_{h=1}^{H} \mathcal{L}(M_1, M_2, M, h)$$

Proof. Let $(x, a, b) \sim \mathcal{P}^{M_2}_{M_1}$ denote $x_{\tau+1} \sim \mathbb{P}(\cdot|x_\tau, a_\tau, b_\tau)$, $a_\tau \sim \pi^{M_1}$ and $b_\tau \sim \nu^{M_2}_{\pi^{M_1}}$ for all $\tau \in [h]$. Using definition of Bellman error in Eq (16), we have

$$\begin{aligned}
& Q_M(x_1, \pi, \nu) - V_1^{\pi,\nu}(x_1) \\
= & \mathop{\mathbb{E}}_{(x,a,b)\sim\mathcal{P}^{M_2}_{M_1}} \left[\mathop{\mathbb{E}}_{(r_1,x_2)\sim M} [r_1 + Q_M(x_2, \pi, \nu)] - \mathop{\mathbb{E}}_{(r_1,x_2)\sim M^*} [r_1 + Q_{M^*}(x_2, \pi, \nu)] \right] \\
= & \mathop{\mathbb{E}}_{(x,a,b)\sim\mathcal{P}^{M_2}_{M_1}} \left[\mathop{\mathbb{E}}_{(r_1,x_2)\sim M} [r_1 + Q_M(x_2, \pi, \nu)] - \mathop{\mathbb{E}}_{(r_1,x_2)\sim M^*} [r_1 + Q_M(x_2, \pi, \nu)] \right] \\
& + \mathop{\mathbb{E}}_{(x,a,b)\sim\mathcal{P}^{M_2}_{M_1}} \left[\mathop{\mathbb{E}}_{(r_1,x_2)\sim M^*} [r_1 + Q_M(x_2, \pi, \nu)] - \mathop{\mathbb{E}}_{(r_1,x_2)\sim M^*} [r_1 + Q_{M^*}(x_2, \pi, \nu)] \right] \\
= & \mathcal{L}(M_1, M_2, M, 1) + \mathop{\mathbb{E}}_{(x,a,b)\sim\mathcal{P}^{M_2}_{M_1}} [Q_M(x_2, \pi, \nu) - V_2^{\pi,\nu}(x_2)] \\
= & \cdots \\
= & \sum_{h=1}^{H} \mathcal{L}(M_1, M_2, M, h).
\end{aligned}$$

We thus complete the proof. □

Step 2: Concentrations In this part we show some standard concentration results.

Lemma C.7. *With probability at least $1 - p$, we have the following event*

1. $\left| \mathcal{E}(M_1^k, M_2^k, M, h) - \widehat{\mathcal{E}}(M_1^k, M_2^k, M, h) \right| \leq \phi$ *for all $M \in \mathcal{M}$ and $h \in [H]$,*
2. $\left| \widehat{\mathcal{L}}(M_1^k, M_2^k, M_i^k, h) - \mathcal{L}(M_1^k, M_2^k, M_i^k, h) \right| \leq \frac{\epsilon}{8H}$ *for all $h \in [H]$ and $i \in [2]$,*
3. $\left| V^{\pi^k,\nu^k} - \widehat{V}^k \right| \leq \epsilon/8$

holds for all $k \leq T$ rounds.

Proof. Fix one iteration. By Hoeffding's inequality, with probability at least $1 - p/3$,

$$\left| \widehat{\mathcal{L}}(M_1^k, M_2^k, M_i^k, h) - \mathcal{L}(M_1^k, M_2^k, M_i^k, h) \right| \leq \sqrt{\frac{\log(2H/p)}{n_1}}. \tag{17}$$

Thus (2) follows from $n_1 = C \cdot H^2 \log(HT/p)/\epsilon^2$ and union bound on all T iterations. Similarly we can verify (3). For (1), fix model $M \in \mathcal{M}$ and $g \in \mathcal{G}$, by Hoeffding's inequality, with probability at least $1 - p/3$,

$$\Big| \mathop{\mathbb{E}}_{(x,a,b)\sim\mathcal{P}^{M_2}_{M_1}} \mathop{\mathbb{E}}_{(r,x)\sim M} [g(x_h, a_h, b_h, r, x) - g(x_h, a_h, b_h, r_h, x_{h+1})] \tag{18}$$

$$- \sum_{i=1}^{n} \frac{1}{n} \mathop{\mathbb{E}}_{(r,x)\sim M} [g(x_h^i, a_h^i, b_h^i, r, x') - g(x_h^i, a_h^i, b_h^i, r_h^i, x_{h+1}^i)] \Big| \tag{19}$$

$$\leq \sqrt{\frac{\log(2/p)}{n}}. \tag{20}$$

Thus (1) follows from $n = C \cdot H^2 W_\kappa |\mathcal{A}| \log(T|\mathcal{M}||\mathcal{G}|/p)/(\kappa\epsilon)^2$ and union bound on all $k \in [T]$, $M \in \mathcal{M}$ and $g \in \mathcal{G}$. We complete the proof. □

Step 3: M^* stays in the constraint set while it shrinks In this step, we show that the constraint set always contains the true environment. The result is displayed in the following.

Lemma C.8. *Suppose for each round k, the event in Lemma C.7 holds. Then $M^* \in \mathcal{M}^k$ for all k. Furthermore, let $\widehat{\mathcal{M}}^k = \{M \in \widehat{\mathcal{M}}^{k-1} : A_{h^k}(M_1^k, M_2^k, M) \leq 2\phi\}$ with $\widehat{\mathcal{M}}^0 = \mathcal{M}$. We have $\mathcal{M}^k \subset \widehat{\mathcal{M}}^k$ for all k.*

Proof. Since $\mathcal{E}(M_1^k, M_2^k, M^*, h) = 0, \forall h \in [H]$, we have $\widehat{\mathcal{E}}(M_1^k, M_2^k, M^*, h^k) \leq \mathcal{E}(M_1^k, M_2^k, M^*, h^k) + \phi \leq \phi$, Thus $\mathcal{M}^*$ is not eliminated. We prove the second result by induction. Firstly $\widehat{\mathcal{M}}^0 = \mathcal{M}^0$. Assume $\mathcal{M}^{k-1} \subset \widehat{\mathcal{M}}^{k-1}$. For all $M \in \mathcal{M}^k$, we have $A_{h^k}(M_1^k, M_2^k, M) \leq \mathcal{E}(M_1^k, M_2^k, M, h^k) \leq \widehat{\mathcal{E}}(M_1^k, M_2^k, M, h^k) + \phi \leq 2\phi$ by Line 13 of Algorithm 4. Therefore $M \in \widehat{\mathcal{M}}^k$, we complete the proof. □

Step 4: Terminate or rank increase The following Lemma is key to the 'alternate pessimism principle'. Intuitively, it indicates that the algorithm either terminates and outputs a pair of policies with small duality gap, or detects a time step that provides useful information (large Bellman error).

Lemma C.9. *Suppose for round k, the event in Lemma C.7 holds and M^* is not eliminated. Then if the algorithm terminate, it outputs a policy pair π^k such that $V^*(x_1) - V^{\pi^k, \nu^*_{\pi^k}}(x_1) \leq \epsilon$; otherwise $A_{h^k}((M_1^k, M_2^k), M_i^k) \geq \frac{\kappa\epsilon}{8H}$ for some $i \in [2]$.*

Proof. Consider the situation that the algorithm does not terminate. Then with out loss of generality we can assume $|\widehat{V}^k - Q^{M_1^k}(x_1, \pi^k, \nu^k)| \geq \epsilon/2$. By Lemma C.7 and Lemma C.6, we have

$$\begin{aligned}\left|\sum_{h=1}^{H} \mathcal{L}(M_1^k, M_2^k, M_1^k, h)\right| &= |Q_M(x_1, \pi, \nu) - V_1^{\pi,\nu}(x_1)| \\ &\geq |\widehat{V}^k - Q^{M_1^k}(x_1, \pi^k, \nu^k)| - \left|V^{\pi^k,\nu^k} - \widehat{V}^k\right| \\ &\geq 3\epsilon/8.\end{aligned}$$

By pigeonhole principle, these exists $h \in [H]$ such that $|\mathcal{L}(M_1^k, M_2^k, M_1^k, h)| \geq \frac{3\epsilon}{8H}$. For this h we have $|\widehat{\mathcal{L}}(M_1^k, M_2^k, M_1^k, h)| \geq \frac{3\epsilon}{8H} - \frac{\epsilon}{8H} = \frac{\epsilon}{4H}$, thus the h^k in Line 11 is well defined and we have $|\widehat{\mathcal{L}}(M_1^k, M_2^k, M_1^k, h^k)| \geq \epsilon/(4H)$. Using Lemma C.7 again, we have

$$A_{h^k}((M_1^k, M_2^k), M_i^k) \geq \kappa \cdot |\mathcal{L}(M_1^k, M_2^k, M_1^k, h^k)| \geq \frac{\kappa\epsilon}{8H}.$$

If the algorithm terminates, we have

$$Q^{M_1^k}(x_1, \pi^k, \nu^k) \geq Q^{M_1^k}(x_1, \pi^k, \nu^{M_1^k}_{\pi^k}) \geq V^*(x_1)$$

and $Q^{M_2^k}(x_1, \pi^k, \nu^k) \leq V^{\pi^k, \nu^*_{\pi^k}}(x_1)$. Therefore

$$\begin{aligned}V^*(x_1) - V^{\pi^k, \nu^*_{\pi^k}}(x_1) &\leq Q^{M_1^k}(x_1, \pi^k, \nu^k) - Q^{M_2^k}(x_1, \pi^k, \nu^k) \\ &\leq |Q^{M_1^k}(x_1, \pi^k, \nu^k) - \widehat{V}^{\pi^k,\nu^k}| + |\widehat{V}^{\pi^k,\nu^k} - Q^{M_2^k}(x_1, \pi^k, \nu^k)| \\ &\leq \epsilon.\end{aligned}$$

We thus complete the proof. □

Step 5: Bounding the number of iterations. This step uses the technique of Jiang et al. (2017); Sun et al. (2019a), which uses the volume of minimum volume enclosing ellipsoid as a potential function to bound the number of iterations. The key result is presented in Lemma C.10.

Lemma C.10. *Suppose for every round k, the event in Lemma C.7 holds, then the number of iterations of Algorithm 4 is at most $HW_\kappa \log(\frac{\beta}{2\phi})/\log(5/3)$ for certain $i \in [2]$.*

Proof. If the algorithm does not terminate at round k, then by Lemma C.9, $\langle \zeta_{h^k}(M_1^k, M_2^k), \chi_{h^k}(M_i^k)\rangle = A_{h^k}((M_1^k, M_2^k), M_i^k) \geq \frac{\kappa\epsilon}{8H} = 6\sqrt{W_\kappa}\phi$. Let O_h^k denote the origin-centered minimum volume enclosing ellipsoid of $\{\chi_h(M) : M \in \widehat{\mathcal{M}}_k\}$. Let $O_h^{k-1,+}$ denote the origin-centered minimum volume enclosing ellipsoid of $\{v \in O_h^{k-1} : \langle \zeta_{h^k}(M_1^k, M_2^k), v\rangle \leq 2\phi\}$. By Lemma F.2, we have

$$\frac{\mathrm{vol}(O_{h^k}^k)}{\mathrm{vol}(O_{h^k}^{k-1})} \leq \frac{\mathrm{vol}(O_{h^k}^{k-1,+})}{\mathrm{vol}(O_{h^k}^{k-1})} \leq 3/5.$$

Define $\Phi = \sup_{M_1, M_2 \in \mathcal{M}, h\in[H]} \|\zeta_h(M_1, M_2)\|_2$ and $\Psi = \sup_{M\in\mathcal{M}, h\in[H]} \|\chi_h(M)\|_2$. Then we have $\mathrm{vol}(O_h^0) \leq V_{W_\kappa}(1) \cdot \Psi^{W_\kappa}$ where $V_{W_\kappa}(R)$ is the volume of Euclidean ball in $\mathbb{R}^{W_\kappa}$ with radius R. On the other hand, we have

$$\mathrm{vol}(O_h^k) \geq \mathrm{vol}(\{q \in \mathbb{R}^{W_\kappa} : \|q\|_2 \leq 2\phi/\Phi\}) \geq V_{W_\kappa}(1) \cdot (2\phi/\Phi)^{W_\kappa}.$$

It thus follows that the number of iterations for each h is at most $W_\kappa \log(\frac{\Phi\Psi}{2\phi})/\log(5/3)$. Using $\beta \geq \Phi\Psi$, this completes the proof. □

Step 6: Putting everything together. From Lemma C.7, with high probability the events holds. Therefore for the first T rounds, we can apply Lemma C.10 and know that the algorithm indeed terminates in at most T rounds. The number of trajectories is thus upper bounded by $(n_1 + n)T = O(\frac{H^3 W_\kappa^2 |\mathcal{A}|}{\kappa^2\epsilon^2} \log(T|\mathcal{G}||\mathcal{M}|/p))$.

D THEORY FOR ALTERNATE OPTIMISTIC VALUE ELIMINATION (AOVE)

We prove Theorem 4.4 in the following five steps. In the first step we show that the functions in the constraint set has low Bellman error. Notice that different from Lemma B.1, here we consider a different Bellman operator in Eq (2) and consider all policies in policy class Π. In the second step we show that Q^{π,ν_π^*} belongs to the constraint set for all $\pi \in \Pi$ and in particular Q^* belongs to the constraint set. In the third step we perform a regret decomposition. Notice that here we consider two Bellman errors. We then bound the summation of these Bellman errors by Minimax Eluder dimension in step 4 and complete the proof in step 5.

Step 1: Low Bellman errors in the constraint set.

Lemma D.1 (Concentration on Bellman error)**.** *Let $\rho > 0$ be an arbitrary fixed number. With probability at least $1 - p$ for all $(k, h) \in [K] \times [H]$ we have*

$$(a) \sum_{i=1}^{k-1} \Big(f_h^k(x_h^i, a_h^i, b_h^i) - r_h^i - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^i,a_h^i,b_h^i)} f_{h+1}^k(x, \pi_{h+1}^k, (\nu_{\pi^k}^{f^k})_{h+1}))\Big)^2 \leq O(\beta)$$

$$(b) \sum_{i=1}^{k-1} \Big(g_h^k(x_h^i, a_h^i, b_h^i) - r_h^i - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^i,a_h^i,b_h^i)} g_{h+1}^k(x, \pi_{h+1}^k, (\nu_{\pi^k}^{g^k})_{h+1}))\Big)^2 \leq O(\beta)$$

Proof. For simplicity we only proof (a), and (b) follows similarly.

Consider a fixed (k, h, f, π) tuple, let

$$\begin{aligned} U_t(h, f, \pi) := &\big(f_h(x_h^t, a_h^t, b_h^t) - r_h^t - f_{h+1}(x_{h+1}^t, \pi, \nu_\pi^f)\big)^2 \\ &- \big(r_h^t + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} f_{h+1}(x, \pi, \nu_\pi^f) - r_h^t - f_{h+1}(x_{h+1}^t, \pi, \nu_\pi^f)\big)^2 \end{aligned}$$

and $\mathscr{F}_{t,h}$ be the filtration induced by $\{x_1^i, a_1^i, b_1^i, r_1^i, \ldots, x_H^i\}_{i=1}^{t-1} \cup \{x_1^t, a_1^t, b_1^t, r_1^t, \ldots, x_h^t, a_h^t, b_h^t\}$.

We have

$$
\begin{aligned}
&\mathbb{E}[U_t(h,f,\pi)|\mathscr{F}_{t,h}] \\
=& [(f_h(x_h^t,a_h^t,b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|\cdot,\cdot,\cdot)} f_{h+1}(x,\pi,\nu_\pi^f))(x_h^t,a_h^t,b_h^t)] \\
&\cdot \mathbb{E}[f_h(x_h^t,a_h^t,b_h^t) + r_h^t + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} f_{h+1}(x,\pi,\nu_\pi^f) - 2r_h^t - 2f_{h+1}(x_{h+1}^t,\pi,\nu_\pi^f)|\mathscr{F}_{t,h}] \\
=& [f_h(x_h^t,a_h^t,b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} f_{h+1}(x,\pi,\nu_\pi^f)]^2
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{Var}[U_t(h,f,\pi)|\mathscr{F}_{t,h}] \le \mathbb{E}[(U_t(h,f,\pi))^2|\mathscr{F}_{t,h}] \\
=& [(f_h - r_h - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|\cdot,\cdot,\cdot)} f_{h+1}(x,\pi,\nu_\pi^f))(x_h^t,a_h^t,b_h^t)]^2 \\
&\cdot \mathbb{E}[\big(f_h(x_h^t,a_h^t,b_h^t) + r_h^t + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} f_{h+1}(x,\pi,\nu_\pi^f) - 2r_h^t - 2f_{h+1}(x_{h+1}^t,\pi,\nu_\pi^f)\big)^2|\mathscr{F}_{t,h}] \\
\le& 36[f_h(x_h^t,a_h^t,b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} f_{h+1}(x,\pi,\nu_\pi^f)]^2 = 36\,\mathbb{E}[U_t(h,f,\pi)|\mathscr{F}_{t,h}].
\end{aligned}
$$

By Freedman's inequality, we have with probability at least $1-p$,

$$
\left|\sum_{t=1}^k U_t(h,f,\pi) - \sum_{t=1}^k \mathbb{E}[U_t(h,f,\pi)|\mathscr{F}_{t,h}]\right| \le O\bigg(\sqrt{\log(1/p)\sum_{t=1}^k \mathbb{E}[U_t|\mathscr{F}_{t,h}]} + \log(1/p)\bigg).
$$

Let $\mathscr{N}_\rho \times \mathscr{Y}_\rho$ be a ρ-cover of $\mathcal{F}\times\Pi$. The covering of Π is in-terms of the distance of two policies defined as

$$
d(\pi,\pi') := \max_{f\in\mathcal{F},x\in\mathcal{S},b\in\mathcal{A}} |f(x,\pi,b) - f(x,\pi',b)|.
$$

Taking a union bound for all $(k,h,\mathfrak{f},\mathfrak{p}) \in [K]\times[H]\times\mathscr{N}_\rho\times\mathscr{Y}_\rho$, we have that with probability at least $1-p$ for all $(k,h,\mathfrak{f},\mathfrak{p}) \in [K]\times[H]\times\mathscr{N}_\rho\times\mathscr{Y}_\rho$

$$
\begin{aligned}
&\left|\sum_{t=1}^k U_t(h,\mathfrak{f},\mathfrak{p}) - \sum_{t=1}^k [\mathfrak{f}_h(x_h^t,a_h^t,b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} \mathfrak{f}_{h+1}(x,\mathfrak{p},\nu_\mathfrak{p}^\mathfrak{f})]^2\right| \\
\le& O\bigg(\sqrt{\iota\sum_{t=1}^k [\mathfrak{f}_h(x_h^t,a_h^t,b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} \mathfrak{f}_{h+1}(x,\mathfrak{p},\nu_\mathfrak{p}^\mathfrak{f})]^2} + \iota\bigg) \qquad (21)
\end{aligned}
$$

where $\iota = \log(HK|\mathscr{N}_\rho||\mathscr{Y}_\rho|/p)$. For an arbitrary $(h,k)\in[H]\times[K]$ pair, by the definition of $\mathcal{V}^k$ and $(\pi^k,f^k)\in\mathcal{V}^k$ we have

$$
\begin{aligned}
&\sum_{t=1}^k U_t(h,f^k,\pi^k) \\
&- \sum_{t=1}^k \big(r_h^t + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t,a_h^t,b_h^t)} f_{h+1}^k(x,\pi^k,\nu_{\pi^k}^{f^k}) - r_h^t - f_{h+1}^k(x_{h+1}^t,\pi^k,\nu_{\pi^k}^{f^k})\big)^2 \\
\le& \sum_{t=1}^k \big(f_h^k(x_h^t,a_h^t,b_h^t) - r_h^t - f_{h+1}^k(x_{h+1}^t,\pi^k,\nu_{\pi^k}^{f^k})\big)^2 \\
&- \inf_{g\in\mathcal{F}} \sum_{t=1}^k \big(g_h^k(x_h^t,a_h^t,b_h^t) - r_h^t - f_{h+1}^k(x_{h+1}^t,\pi^k,\nu_{\pi^k}^{f^k})\big)^2 \\
\le& O(\iota + k\rho) \qquad (22)
\end{aligned}
$$

where in the first step we use Assumption 3.2.

Define $\mathfrak{f}^k = \text{argmin}_{\mathfrak{f}\in\mathscr{N}_\rho} \max_{h\in[H]} \|f_h^k - \mathfrak{f}_h^k\|_\infty, \mathfrak{p}^k = \text{argmin}_{\mathfrak{p}\in\mathscr{Y}_\rho} d(\pi^k, \mathfrak{p})$. Since $\mathscr{N}_\rho$ is a ρ-cover of $\mathcal{F}$,

$$\left|\sum_{t=1}^k U_t(h, f^k, \pi^k) - \sum_{t=1}^k U_t(h, \mathfrak{f}^k, \mathfrak{p}^k)\right| \leq O(k\rho). \tag{23}$$

Therefore combining Eq (22) and Eq (23)

$$\sum_{t=1}^k U_t(h, \mathfrak{f}^k, \mathfrak{p}^k) \leq O(\iota + k\rho). \tag{24}$$

Recall Eq (21) indicates

$$\begin{aligned}&\left|\sum_{t=1}^k U_t(h, \mathfrak{f}^k, \mathfrak{p}^k) - \sum_{t=1}^k [\mathfrak{f}_h^k(x_h^t, a_h^t, b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t, a_h^t, b_h^t)} \mathfrak{f}_{h+1}^k(x, \mathfrak{p}^k, \nu_{\mathfrak{p}^k}^{\mathfrak{f}^k})]^2\right| \\ &\leq O\left(\sqrt{\iota \sum_{t=1}^k [\mathfrak{f}_h^k(x_h^t, a_h^t, b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t, a_h^t, b_h^t)} \mathfrak{f}_{h+1}^k(x, \mathfrak{p}^k, \nu_{\mathfrak{p}^k}^{\mathfrak{f}^k})]^2 + \iota}\right)\end{aligned} \tag{25}$$

Combining Eq (24) and Eq (25) we have

$$\sum_{t=1}^k [\mathfrak{f}_h^k(x_h^t, a_h^t, b_h^t) - r_h^t - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^t, a_h^t, b_h^t)} \mathfrak{f}_{h+1}^k(x, \mathfrak{p}^k, \nu_{\mathfrak{p}^k}^{\mathfrak{f}^k})]^2 \leq O(\iota + k\rho).$$

Since $\mathfrak{f}^k$ is an ρ-approximation to f^k and $\mathfrak{p}^k$ is an ρ-approximation to π^k, we can conclude the proof of (a) with

$$\sum_{i=1}^{k-1} \left(f_h^k(x_h^i, a_h^i, b_h^i) - r_h^i - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x_h^i, a_h^i, b_h^i)} f_{h+1}^k(x, \pi_{h+1}^k, (\nu_{\pi^k}^{f^k})_{h+1}))\right)^2 \leq O(\beta).$$

□

Step 2: $(\pi, Q^{\pi,\nu_\pi^*}), \forall\pi\in\Pi$ **is always in constraint set.**

Lemma D.2. *With probability at least* $1-p$*, we have* (π^*, Q^*) *and* $(\pi, Q^{\pi,*}), \forall\pi\in\Pi$ *all belong to* $\mathcal{V}^k$ *for all* $k\in[K]$.

Proof. Consider a fixed (k, h, f, π) tuple, let

$$\begin{aligned}U_t(h, f, \pi) := &\left(f_h(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\pi,\nu_\pi^*}(x_{h+1}^t, \pi, \nu_\pi^*)\right)^2 \\ &- \left(Q_h^{\pi,\nu_\pi^*}(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\pi,\nu_\pi^*}(x_{h+1}^t, \pi, \nu_\pi^*\right)^2\end{aligned}$$

and $\mathscr{F}_{t,h}$ be the filtration induced by $\{x_1^i, a_1^i, b_1^i, r_1^i, \ldots, x_H^i\}_{i=1}^{t-1} \cup \{x_1^t, a_1^t, b_1^t, r_1^t, \ldots, x_h^t, a_h^t, b_h^t\}$.

We have

$$\mathbb{E}[U_t(h, f, \pi)|\mathscr{F}_{t,h}] = [(f_h - Q_h^{\pi,\nu_\pi^*})(x_h^t, a_h^t, b_h^t)]^2$$

and

$$\text{Var}[U_t(h, f, \pi)|\mathscr{F}_{t,h}] \leq 36[(f_h - Q_h^{\pi,\nu_\pi^*})(x_h^t, a_h^t, b_h^t)]^2 = 36\,\mathbb{E}[U_t(h, f, \pi)|\mathscr{F}_{t,h}].$$

By Freedman's inequality, we have with probability at least $1-p$,

$$\begin{aligned}&\left|\sum_{t=1}^k U_t(h, f, \pi) - \sum_{t=1}^k [(f_h - Q_h^{\pi,\nu_\pi^*})(x_h^t, a_h^t, b_h^t)]^2\right| \\ &\leq O\left(\sqrt{\log(1/p) \sum_{t=1}^k [(f_h - Q_h^{\pi,\nu_\pi^*})(x_h^t, a_h^t, b_h^t)]^2} + \log(1/p)\right).\end{aligned}$$

Let $\mathscr{N}_\rho \times \mathscr{Y}_\rho$ be a ρ-cover of $\mathcal{F} \times \Pi$. Taking a union bound for al $(k, h, \mathfrak{f}, \mathfrak{p}) \in [K] \times [H] \times \mathscr{N}_\rho \times \mathscr{Y}_\rho$, we have that with probability at least $1 - p$ for all $(k, h, \mathfrak{f}, \mathfrak{p}) \in [K] \times [H] \times \mathscr{N}_\rho \times \mathscr{Y}_\rho$

$$-\sum_{t=1}^{k} U_t(h, \mathfrak{f}, \mathfrak{p}) \leq O(\iota)$$

where $\iota = \log(HK|\mathscr{N}_\rho|/p)$. This implies for all $(k, h, \mathfrak{f}, \mathfrak{p}) \in [K] \times [H] \times \mathscr{N}_\rho \times \mathscr{Y}_\rho$

$$\begin{aligned}
&\sum_{t=1}^{k} \left(Q_h^{\mathfrak{p},\nu_{\mathfrak{p}}^*}(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\mathfrak{p},\nu_{\mathfrak{p}}^*}(x_{h+1}^t, \mathfrak{p}, \nu_{\mathfrak{p}}^*)\right)^2 \\
\leq &\sum_{t=1}^{k} \left(\mathfrak{f}_h(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\mathfrak{p},\nu_{\mathfrak{p}}^*}(x_{h+1}^t, \mathfrak{p}, \nu_{\mathfrak{p}}^*)\right)^2 + O(\iota + k\rho).
\end{aligned}$$

Since $\mathscr{N}_\rho \times \mathscr{Y}_\rho$ be a ρ-cover of $\mathcal{F} \times \Pi$, for all $\mathfrak{p}^k \in \mathcal{Y}_\rho$

$$\begin{aligned}
&\sum_{t=1}^{k} \left(Q_h^{\mathfrak{p},\nu_{\mathfrak{p}}^*}(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\mathfrak{p},\nu_{\mathfrak{p}}^*}(x_{h+1}^t, \mathfrak{p}, \nu_{\mathfrak{p}}^*)\right)^2 \\
\leq &\inf_{g \in \mathcal{F}} \sum_{t=1}^{k} \left(g_h(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\mathfrak{p},\nu_{\mathfrak{p}}^*}(x_{h+1}^t, \mathfrak{p}, \nu_{\mathfrak{p}}^*)\right)^2 + O(\iota + k\rho).
\end{aligned}$$

Again using $\mathscr{N}_\rho \times \mathscr{Y}_\rho$ be a ρ-cover of $\mathcal{F} \times \Pi$, we have for all $\pi \in \Pi$

$$\begin{aligned}
&\sum_{t=1}^{k} \left(Q_h^{\pi,\nu_{\pi}^*}(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\pi,\nu_{\pi}^*}(x_{h+1}^t, \pi, \nu_{\pi}^*)\right)^2 \\
\leq &\inf_{g \in \mathcal{F}} \sum_{t=1}^{k} \left(g_h(x_h^t, a_h^t, b_h^t) - r_h^t - Q_{h+1}^{\pi,\nu_{\pi}^*}(x_{h+1}^t, \pi, \nu_{\pi}^*)\right)^2 + O(\iota + k\rho).
\end{aligned}$$

Therefore we have $(\pi, Q^{\pi,\nu_\pi^*}), \forall \pi \in \Pi$ all belong to $\mathcal{V}^k$ for all $k \in [K]$ with probability at least $1 - p$. □

Remark D.3. *An important implication of this fact is the following*

$$V^* - V^{\pi^k, \nu_{\pi^k}^*} \leq f_1^k(x_1^k, \pi^k, \nu_{\pi^k}^{f^k}) - g_1^k(x_1^k, \pi^k, \nu^k)$$

which gives an upper bound of the one-side duality gap.

Step 3: Bounding the regret by Bellman error and martingale.

Lemma D.4. *Define*

$$\begin{aligned}
\delta_h^k &:= f_h^k(x_h^k, \pi_h^k, (\nu_{\pi^k}^{f^k})_h) - g_h^k(x_h^k, \pi_h^k, \nu_h^k) \\
\zeta_h^k &:= \mathbb{E}[\delta_{h+1}^k | x_h^k, a_h^k, b_h^k] - \delta_{h+1}^k \\
\gamma_h^k &:= \mathop{\mathbb{E}}_{a \sim \pi_h^k, b \sim \nu_h^k} [f_h^k(x_h^k, a, b)] - f_h^k(x_h^k, a_h^k, b_h^k) \\
\widehat{\gamma}_h^k &:= \mathop{\mathbb{E}}_{a \sim \pi_h^k, b \sim \nu_h^k} [g_h^k(x_h^k, a, b)] - g_h^k(x_h^k, a_h^k, b_h^k)
\end{aligned}$$

Then we have

$$\delta_h^k \leq \delta_{h+1}^k + \zeta_h^k + \epsilon_h^k - \widehat{\epsilon}_h^k + \gamma_h^k - \widehat{\gamma}_h^k$$

where $\epsilon_h^k, \widehat{\epsilon}_h^k$ are the bellman error defined by

$$\begin{aligned}
\epsilon_h^k &:= f_h^k(x_h^k, a_h^k, b_h^k) - r_h^k - \mathop{\mathbb{E}}_{x \sim \mathbb{P}(\cdot | x_h^k, a_h^k, b_h^k)} f_h^k(x, \pi_{h+1}^k, (\nu_{\pi^k}^{f^k})_{h+1}) \\
\widehat{\epsilon}_h^k &:= g_h^k(x_h^k, a_h^k, b_h^k) - r_h^k - \mathop{\mathbb{E}}_{x \sim \mathbb{P}(\cdot | x_h^k, a_h^k, b_h^k)} g_h^k(x, \pi_{h+1}^k, \nu_{h+1}^k).
\end{aligned}$$

Proof. By definition of $(\nu^{f^k}_{\pi^k})_h = \text{argmin}_{\nu\in\Pi_h} f^k_h(x^k_h, \pi^k_h, \nu)$, we have

$$\begin{aligned} f^k_h(x^k_h, \pi^k_h, (\nu^{f^k}_{\pi^k})_h) &\leq f^k_h(x^k_h, \pi^k_h, \nu^k_h) \\ &= f^k_h(x^k_h, a^k_h, b^k_h) + \gamma^k_h \end{aligned}$$

It thus follows that

$$\begin{aligned} \delta^k_h &= f^k_h(x^k_h, \pi^k_h, (\nu^{f^k}_{\pi^k})_h) - g^k_h(x^k_h, \pi^k_h, \nu^k_h) \\ &\leq f^k_h(x^k_h, a^k_h, b^k_h) - g^k_h(x^k_h, a^k_h, b^k_h) + \gamma^k_h - \widehat{\gamma}^k_h \\ &= \big(r^k_h + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x^k_h,a^k_h,b^k_h)} f^k_h(x, \pi^k_{h+1}, (\nu^{f^k}_{\pi^k})_{h+1}) + \epsilon^k_h\big) \\ &\quad - \big(r^k_h + \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x^k_h,a^k_h,b^k_h)} g^k_h(x, \pi^k_{h+1}, \nu^k_{h+1}) + \widehat{\epsilon}^k_h\big) + \gamma^k_h - \widehat{\gamma}^k_h \\ &= \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x^k_h,a^k_h,b^k_h)} f^k_h(x, \pi^k_{h+1}, (\nu^{f^k}_{\pi^k})_{h+1}) - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot|x^k_h,a^k_h,b^k_h)} g^k_h(x, \pi^k_{h+1}, \nu^k_{h+1}) \\ &\quad + \epsilon^k_h - \widehat{\epsilon}^k_h + \gamma^k_h - \widehat{\gamma}^k_h \\ &= \delta^k_{h+1} + \zeta^k_h + \epsilon^k_h - \widehat{\epsilon}^k_h + \gamma^k_h - \widehat{\gamma}^k_h. \end{aligned}$$

Thus we complete the proof. □

Step 4: Bounding cumulative Bellman error using Minimax Eluder dimension.

Lemma D.5 (Bounding Bellman error)**.** *For any* $(k, h) \in [K] \times [H]$ *we have*

$$\begin{aligned} \sum_{t=1}^k |\epsilon^t_h| &\leq O\left(\sqrt{k \cdot \dim_{\text{ME}'}(\mathcal{F}, \sqrt{1/K}) \log[\mathcal{N}(\mathcal{F}, 1/K) \cdot \mathcal{N}(\Pi, 1/K) \cdot HK/p]}\right) \\ \sum_{t=1}^k |\widehat{\epsilon}^t_h| &\leq O\left(\sqrt{k \cdot \dim_{\text{ME}'}(\mathcal{F}, \sqrt{1/K}) \log[\mathcal{N}(\mathcal{F}, 1/K) \cdot \mathcal{N}(\Pi, 1/K) \cdot HK/p]}\right). \end{aligned}$$

Proof. Let

$$l_t(\cdot,\cdot,\cdot) = f^t_h(\cdot,\cdot,\cdot) - r_h(\cdot,\cdot,\cdot) - \mathop{\mathbb{E}}_{x\sim\mathbb{P}(\cdot,\cdot,\cdot)} f^t_{h+1}(x, \pi^t, \nu^{f^t}_{\pi^t})$$

and $\rho_t = \delta(x^t_h, a^t_h, b^t_h)$. Similarly, we can also let ρ_t be generated by (π^t, ν^t). Then Lemma D.1 indicates

$$\sum_{t=1}^k l^2_k(\rho_t) \leq O(\beta).$$

Based on this, we can apply Lemma F.1 to obtain the first inequality. The second follows similarly. □

Step 5: Putting them all together. We can upper bound the regret as follows:

$$\begin{aligned} \sum_{k=1}^K [V^* - V^{\pi^k, \nu^*_{\pi^k}}] &\leq \sum_{k=1}^K [f^k_1(x^k_1, \pi^k, \nu^{f^k}_{\pi^k}) - g^k_1(x^k_1, \pi^k, \nu^k)] \\ &\leq \sum_{k=1}^K \sum_{h=1}^H (\zeta^k_h + \epsilon^k_h - \widehat{\epsilon}^k_h + \gamma^k_h - \widehat{\gamma}^k_h) \\ &= \sum_{h=1}^H \sum_{k=1}^K (\zeta^k_h + \gamma^k_h - \widehat{\gamma}^k_h) + \sum_{h=1}^H \sum_{k=1}^K (\epsilon^k_h - \widehat{\epsilon}^k_h) \end{aligned} \tag{26}$$

where the first step comes from Lemma D.2, the second step comes from Lemma D.4 and the last step comes from Fubini's theorem. Notice that the first term in Eq (26) is a martingale difference sequence and can be bounded by $O(H\sqrt{KH\log(HK/p)})$ with probability at least $1-p$ and the second term can be bounded by Lemma D.5. We thus complete the proof.

E EXAMPLES

In this section we give several examples of Markov games that possess low Minimax Eluder dimension, including *tabular Markov Games, linear function approximation, reproducing kernel Hilbert space (RKHS) function approximation, overparameterized neural networks, and generalized linear models*. We first consider reproducing kernel Hilbert space (RKHS) function approximation, which captures many existing models.

Reproducing kernel Hilbert space (RKHS) function approximation. In this setting, the function class of interest is given by

$$\mathcal{F}_h = \{\langle \phi_h(x,a,b), \theta_h \rangle : \theta, \phi(\cdot,\cdot,\cdot) \in B_R(\mathcal{H})\}.$$

Here $\phi(\cdot,\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \times \mathcal{A} \mapsto \mathcal{H}$ is a feature map to $\mathcal{H}$, $\theta \in \mathcal{H}$, and $B_R(\mathcal{H})$ is the ball centered at zero with radius R in a Hilbert space $\mathcal{H}$. The effective dimension of a set $\mathcal{D} \subset \mathcal{H}$ is characterized by the critical information gain $\widetilde{d}(\gamma, \mathcal{D})$. Specifically, for $\gamma > 0$ and $\mathcal{D} \subset \mathcal{H}$, we define $d_n(\gamma, \mathcal{D}) = \max_{x_1,\ldots,x_n \in \mathcal{D}} \log\det(I + \frac{1}{\gamma}\sum_{i=1}^n x_i x_i^\top)$. The critical information gain (Du et al., 2021) is thus defined by $\widetilde{d}(\gamma, \mathcal{D}) = \min_{k \in \mathbb{N}: k \geq d_n(\gamma,\mathcal{D})} k$. The main result is given in the following theorem.

Theorem E.1. *In kernel function approximation, the Minimax Eluder dimensions* $\dim_{\mathrm{ME}}(\mathcal{F}, \epsilon)$ *and* $\dim_{\mathrm{ME}'}(\mathcal{F}, \epsilon)$ *are upper bounded by* $\max_{h \in [H]} \widetilde{\gamma}(\frac{\epsilon}{9R}, \Phi_h)$*, where* $\Phi_h = \{\phi_h(x,a,b) : x \in \mathcal{S}, a, b \in \mathcal{A}\}$.

Proof. We only show $\dim_{\mathrm{ME}'}(\mathcal{F}, \epsilon) \leq \max_{h \in [H]} \widetilde{\gamma}(\frac{\epsilon}{9R}, \Phi_h)$, and the other part is totally similar. Fix an $h \in [H]$. Due to completeness, for any $\pi \in \Pi$ and $f \in \mathcal{F}$ we have $(f_h - \mathcal{T}_h^\pi f_{h+1})(x,a,b) = \langle \phi_h(x,a,b), \theta_h - \mathcal{T}_h^\pi \theta_{h+1} \rangle$ where $\mathcal{T}_h^\pi \theta_{h+1}$ is the unique feature in $\mathcal{H}$ such that $\mathcal{T}_h^\pi f_{h+1} = \langle \phi_h(x,a,b), \mathcal{T}_h^\pi \theta_{h+1} \rangle$. Consider the longest sequence $\nu_1, \ldots, \nu_k \in \mathcal{D}_\Delta$ such that each one item is ϵ-independent of its precedents, distinguished by a sequence of features $\theta_h^{(1)}, \ldots, \theta_h^{(k)}$. To simplify notations, we define $\phi_i = \mathbb{E}_{\nu_i}[\phi_h(x,a,b)]$ and $\theta_i = \theta_h^{(i)} - \mathcal{T}_h^\pi \theta_{h+1}^{(i)}$ for all $i \in [k]$. Let $\Lambda_i = \sum_{\tau=1}^{i-1} \phi_\tau \phi_\tau^\top + \frac{\epsilon^2}{9R^2} I$. Then Definition 3.3 gives

$$\|\phi_i\|_{\Lambda_i^{-1}} \geq \frac{1}{\epsilon} \cdot \|\phi_i\|_{\Lambda_i^{-1}} \cdot \|\theta_i\|_{\Lambda_i} \geq \frac{1}{\epsilon} \cdot |\langle \phi_i, \theta_i \rangle| \geq 1.$$

where the first step comes from $\theta_i \in B_R(\mathcal{H})$ and $\sum_{j=1}^{i-1} \langle \phi_j, \theta_j \rangle^2 \leq \epsilon$, the second step comes from Hölder's inequality and the final step comes from $|\langle \phi_i, \theta_i \rangle| \geq \epsilon$. Thus we have

$$k \leq \sum_{i=1}^k \log(1 + \|\phi_i\|_{\Lambda_i^{-1}}) = \log\det(I + \frac{9R^2}{\epsilon^2} \sum_{i=1}^k \phi_i \phi_i^\top).$$

By the definition of critical information gain, k must be less or equal than $\widetilde{\gamma}(\frac{\epsilon}{9R}, \Phi_h)$. □

This result implies several important cases in general function approximation.

- **Linear function approximation.** In this setting $r_h(x,a,b) = \phi(x,a,b)^\top \theta_h$ and $\mathbb{P}_h(\cdot|x,a,b) = \phi(x,a,b)^\top \mu_h(\cdot)$ where $\phi(x,a,b) \in \mathbb{R}^d$. We have shown that this setting satisfies all the realizability and completeness assumptions in the previous sections. By standard Ellipsoid potential arguments, the critical information gain is upper bounded by $O(d \cdot \log(R/\epsilon))$. Therefore, this setting also has low Minimax Eluder dimensions.
- **Tabular Markov Games.** This setting can be seen as a special case of linear function approximation where the feature vectors are chosen to be one-hot representations of state-actions pairs. In this case, $d = |\mathcal{S}| \cdot |\mathcal{A}|^2$ and thus Minimax Eluder dimensions is upper bounded by $\widetilde{O}(|\mathcal{S}| \cdot |\mathcal{A}|^2)$.
- **Overparameterized neural networks.** In overparameterized neural network function approximations, the functions are given by $h(\phi(x,a,b))$ where h is the Neural Tangent Random Feature space defined in Cao & Gu (2019). We rewrite the information gain $d_n(\gamma, \mathcal{D})$ as $\max_{x_1,\ldots,x_n \in \mathcal{D}} \log\det(I + \frac{1}{\gamma} K_n)$ where $(K_n)_{i,j} = K(x_i, x_j)$ and K is the Neural Tangent Kernel. It has been shown (Jacot et al., 2018; Du et al., 2019c;a; Allen-Zhu

et al., 2019) that with proper initialization and learning rates, the function approximations approximately lie in the Reproducing kernel Hilbert space given by the neural tangent feature as above. Therefore the Minimax Eluder dimensions are bounded by $\widetilde{\gamma}(\frac{\epsilon}{9R}, \Phi_h) + \lambda$, where $\widetilde{\gamma}$ is the critical information gain defined by the $d_n(\gamma, \mathcal{D})$ rewritten above and λ is a term that vanishes as width tends to infinity.

Generalized linear models. Finally, we consider generalized linear model as an extension of the linear function approximations.

Theorem E.2. *Suppose* $(f_h - \mathcal{T}_h^\pi f_{h+1})(x, a, b) = g(\phi(x, a, b)^\top \theta_h)$ *for any* $\pi \in \Pi$ *and* $f \in \mathcal{F}$*, where* g *is a differentiable and strictly increasing function and* $\phi, \theta_h \in \mathbb{R}^d$*. Assume* $g' \in (c_1, c_2)$ *and* $\|\phi\|_2, \|\theta_h\|_2 \leq R$ *for* $c_1, c_2, R > 0$*. Then* $\dim_{\mathrm{ME}'}(\mathcal{F}, \epsilon) \leq O(d(c_2/c_1)^2 \log(\frac{c_2 R}{c_1 \epsilon}))$.

Proof. The proof mirrors those in Theorem E.1. For the longest sequence $\nu_1, \ldots, \nu_k \in \mathcal{D}_\Delta$ such that each one item is ϵ-independent of its precedents, distinguished by a sequence of features $\theta_h^{(1)}, \ldots, \theta_h^{(k)}$, we define $\phi_i = \mathbb{E}_{\nu_i}[\phi_h(x, a, b)]$ and $\theta_i = \theta_h^{(i)} - \mathcal{T}_h^\pi \theta_{h+1}^{(i)}$ for all $i \in [k]$. Let $\Lambda_i = \sum_{\tau=1}^{i-1} \phi_\tau \phi_\tau^\top + \frac{\epsilon^2}{9R^2 c_1^2} I$. Then Definition 3.3 gives

$$\|\phi_i\|_{\Lambda_i^{-1}} \geq \frac{c_1}{\epsilon} \cdot \|\phi_i\|_{\Lambda_i^{-1}} \cdot \|\theta_i\|_{\Lambda_i} \geq \frac{c_1}{\epsilon} \cdot |\langle \phi_i, \theta_i \rangle| \geq \frac{c_1}{c_2}.$$

This means $\det(\Lambda_k) \geq (1 + \frac{c_1^2}{c_2^2})^{k-1}$. But by Elliptic potential lemma, $\det(\Lambda_k) \leq (\frac{R^2 k}{d} + \frac{\epsilon^2}{9R^2 c_1^2})^d$. Combining these two inequalities gives $k \leq O(d(c_2/c_1)^2 \log(\frac{c_2 R}{c_1 \epsilon}))$. □

F TECHNICAL CLAIMS

This section lists the facts we use from the previous work.

Lemma F.1 (Cumulative error control, Lemma 26 of Jin et al. (2021a))**.** *Given a function class* $\mathcal{G}$ *defined on* $\mathcal{X}$ *with* $|g(x)| \leq C$ *for all* $(g, x) \in \mathcal{G} \times \mathcal{X}$ *and a family of probability measures* $\mathcal{D}$ *over* $\mathcal{X}$*. Suppose sequence* $\{g_k\}_{k=1}^K \subset \mathcal{G}$ *and* $\{\rho_k\}_{k=1}^H \subset \mathcal{D}$ *such that for all* $k \in [K]$*,* $\sum_{t=1}^{k-1} (\mathbb{E}_{\rho_t}[g_k])^2 \leq \beta$*. Then for all* $k \in [K]$ *and* $\omega > 0$*,*

$$\sum_{t=1}^{k} |\mathop{\mathbb{E}}_{\rho_t}[g_t]| \leq O\left(\sqrt{\dim_{\mathrm{DE}}(\mathcal{G}, \mathcal{D}, \epsilon)\beta k} + \min\{k, \dim_{\mathrm{DE}}(\mathcal{G}, \mathcal{D}, \epsilon)\}C + k\omega\right).$$

Lemma F.2 (Lemma 11 of Jiang et al. (2017))**.** *Consider a closed and bounded set* $V \subset \mathbb{R}^d$ *and a vector* $p \in \mathbb{R}^d$*. Let* B *be any origin-centered enclosing ellipsoid of* V*. Suppose there exists* $v \in V$ *such that* $p^\top v \geq \kappa$ *and define* B_+ *as the minimum volume enclosing ellipsoid of* $\{v \in B : |p^\top v| \leq \frac{\kappa}{3\sqrt{d}}\}$*. With vol*$(\cdot)$ *denoting the (Lebesgue) volume, we have:*

$$\frac{vol(B_+)}{vol(B)} \leq \frac{3}{5}.$$