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. Our algorithms are based on new techniques of alternate optimism.

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 give an affirmative answer to this question. 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 arbitrary 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 design a new 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. Our algorithms use the new idea of alternate optimism to address difficulties from the subtle game-theoretic considerations.

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; Perolat 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 in the optimistic planning procedure. Since 'global optimism' only obtains optimism along the trajectories of behaviour policies, this causes the distribution shift from the target policies (i.e. NE). As a result, using CCE to plan will cause the regret to diverge. To deal with this problem, we propose a new method called 'alternate optimism'. We prove two regret decomposition lemmata, laying a theoretical footing for its strong applicability in MGs.

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 (e.g. Perolat et al. (2015b)), 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

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

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 the value function of NE is learned. 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 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 $f \in \mathcal{F}_{h+1}$, $\mathcal{T}_h f \in \mathcal{F}_h$.*

These assumptions are common in value function approximation literature (e.g. Wang et al. (2020b)). 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 $f \in \mathcal{F}_h$, 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 low complexity measure. We introduce Minimax Eluder dimension to capture the structural complexity of a function class in zero-sum games. To define it, we use the notion of 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.2).

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)

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

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

12: Set

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

13: **end for**

3.3 Theoretical results

In this section we present the theoretical guarantee for Algorithm 1. Due to space constraint, we refer the readers to Appendix B.1 for the formal statement and detailed discussions.

Theorem 3.6 (Informal statement of Theorem B.1)**.** *Under Assumption 3.1 and Assumption 3.2, the regret Eq* (3) *of Algorithm 1 is upper bounded by* $\widetilde{O}\left(H\sqrt{K\dim_{\mathrm{ME}}(\mathcal{F}, \sqrt{1/K})\log[\mathcal{N}(\mathcal{F}, 1/K)/p]}\right)$ *with probability at least* $1 - p$.

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

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

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_π^M to denote the best response of π in model M and π_ν^M is defined similarly.

4.1.1 ALTERNATE OPTIMISTIC MODEL ELIMINATION (AOME)

The model-based method, Alternate Optimistic Model Elimination (AOME), is presented in Algorithm 2. 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. This causes the distribution shift between the trajectories of target policies and those collected by the behavior policies. 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 via *alternate optimism*. In k-th episode, the algorithm first solves the optimization problem $M_1^k = \arg\max_{M\in\mathcal{M}} Q^M(x_1, \pi^M, \nu^M)$ and let $\pi^k = \pi^{M_1^k}$. This optimization problem corresponds to finding the optimistic estimate of the value of the game. Indeed, notice that $M^* \in \mathcal{M}^k$ implies $Q^{M_1^k}(x_1, \pi^k, \nu^k) \geq Q^{M_1^k}(x_1, \pi^k, \nu_{\pi^k}^{M_1^k}) \geq V^*(x_1)$, by optimality of M_1^k. Next, it solves the second optimization problem $M_2^k = \arg\min_{M\in\mathcal{M}} Q^M(x_1, \pi^{M_1^k}, \nu^M)$ and let $\nu^k \leftarrow \nu_{\pi^k}^{M_2^k}$. 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 $Q^{M_2^k}(x_1, \pi^k, \nu^k) \leq V^{\pi^k, \nu_{\pi^k}^*}(x_1)$, by optimality of M_2^k.

In Line 8, the algorithm checks if the values of model M_1^k and M_2^k are close to the value of the true model M^* when executing policy π^k and ν^k. If the condition holds, we have

$$V^*(x_1) - V^{\pi^k, \nu_{\pi^k}^*}(x_1) \leq Q^{M_1^k}(x_1, \pi^k, \nu^k) - Q^{M_2^k}(x_1, \pi^k, \nu^k) \leq \epsilon,$$

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

$$\widehat{\mathcal{L}}(M_1, M_2, M, h) := \sum_{i=1}^{n_1} \frac{1}{n_1} [Q_h^M(x_h^i, a_h^i, b_h^i) - (r_h^i + Q_h^M(x_{h+1}^i, \pi^{M_1}, \nu_{\pi^{M_1}}^{M_2})]. \tag{4}$$

As proved in Appendix C, 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

$$\widehat{\mathcal{E}}(M_1^k, M_2^k, M, h^k) = \sup_{g\in\mathcal{G}} \sum_{i=1}^{n} \frac{1}{n} \mathop{\mathbb{E}}_{(r_h, x_h)\sim M} [g(x_h^i, a_h^i, b_h^i, r_h, x_h) - g(x_h^i, a_h^i, b_h^i, r_h^i, x_{h+1}^i)]. \tag{5}$$

We will show that this constraint set maintains more and more accurate models of the environment.

4.1.2 THEORETICAL RESULTS

In this section we prove that Algorithm 2 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.

Algorithm 2 Alternate Optimistic Model Elimination (AOME)

Input: Model class $\mathcal{M}$, test function class $\mathcal{G}$, precision ϵ, failure probability p
Initialize $\mathcal{M}^0 \leftarrow \mathcal{M}$
for $k = 1, 2, \ldots, K$ **do**
  Find $M_1^k \leftarrow \text{argmax}_{M \in \mathcal{M}} Q^M(x_1, \pi^M, \nu^M)$, let $\pi^k \leftarrow \pi^{M_1^k}$
  Find $M_2^k \leftarrow \text{argmin}_{M \in \mathcal{M}} Q^M(x_1, \pi^k, \nu^M_{\pi^k})$, let $\nu^k \leftarrow \nu^{M_1^k}_{\pi^k}$
  Execute π^k, ν^k and collect n_1 rewards $\{r_h^i\}_{h \in [H], i \in [n_1]}$
  Let $\widehat{V}^k \leftarrow \frac{1}{n_1} \sum_{i=1}^{n_1} (\sum_{h=1}^{H} r_h^i)$
  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**
    Terminate and output π^k, ν^k.
  else
    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 (4).
    Collect n trajectories $\{(x_h^i, a_h^i, b_h^i, r_h^i)_{h=1}^H\}_{i=1}^n$ by executing $a_h, b_h \sim \pi_h^k, \nu_h^k, \forall h \in [H]$
    Update constraint set, where $\widehat{\mathcal{E}}$ is given by Eq (5)

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

  end if
end for

Definition 4.1 (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*

$$\mathcal{E}(M_1, M_2, M, h) := \sup_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{(x,a,b) \sim \mathcal{P}_{M_1}^{M_2}} \mathop{\mathbb{E}}_{(r,x) \sim M} [g(x_h, a_h, b_h, r, x) - g(x_h, a_h, b_h, r_h, x_{h+1})]. \tag{6}$$

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 $b_\tau \sim \nu^{M_2}_{\pi^{M_1}}$ for all $\tau \in [h]$.

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

$$\mathcal{L}(M_1, M_2, M, h) := \mathop{\mathbb{E}}_{(x,a,b) \sim \mathcal{P}_{M_1}^{M_2}} \left[\mathop{\mathbb{E}}_{(r,x) \sim M} [g(x_h, a_h, b_h, r, x))] - \mathop{\mathbb{E}}_{(r,x) \sim M^*} [g(x_h, a_h, b_h, r, x)] \right]. \tag{7}$$

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

Assumption 4.2 (Realizability)**.** *Assume $M^* \in \mathcal{M}$.*

Assumption 4.3 (Witness misfit domination)**.** *Assume $\mathcal{G}$ is finite, $\|g\|_\infty \leq 2, \forall g \in \mathcal{G}$ and $\forall M_1, M_2, M \in \mathcal{M} : \mathcal{E}(M_1, M_2, M, h) \geq \mathcal{L}(M_1, M_2, M, h)$.*

Now we introduce the Witness rank for Markov games.

Definition 4.4 (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*

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

The Witness rank for Markov games is defined by

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

We are in the position to state the main theorem of this section. Due to space constraint, we defer the formal statements and discussions to Appendix C.1.1.

Theorem 4.5 (Informal statement of Theorem C.1)**.** *Under Assumption 4.2 and Assumption 4.3, with probability at least $1-p$ Algorithm 2 outputs a ϵ-optimal policy π within at most $\widetilde{O}(\frac{H^3 W_\kappa^2 |\mathcal{A}|}{\kappa^2 \epsilon^2} \log(T|\mathcal{G}||\mathcal{M}|/p))$ trajectories, where W_κ is the witness rank.*

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 approximation with 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 3, works on minimizing this regret. 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 also recall the Bellman operator with regard to fixed P1's policy from Eq (2):

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

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 $f_1(x_1,\pi,\nu_\pi^f) - g_1(x_1,\pi^k,\nu_{\pi^k}^g)$ of the duality gap.

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 $\mathcal{T}_h^\pi f_{h+1}$ 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 $f_h - \mathcal{T}_h^\pi f_{h+1}$ indicates that function approximation f is not easily exploited by P2 when P1 plays π.

As shown in Appendix C, (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 3 Alternate Optimistic Value Elimination (AOVE)

Input: Function class $\mathcal{F}$, policy class Π
Initialize $\mathcal{B}^0 \leftarrow \Pi \times \mathcal{F}$
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
for $k = 1, 2, \ldots, K$ **do**
  Find $(\pi^k, f^k) \leftarrow \operatorname{argmax}_{(\pi,f)\in\mathcal{V}^{k-1}} f_1(x_1, \pi, \nu^f_\pi)$
  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 $\nu^k \leftarrow \nu^g_{\pi^k}$
  for step $h = 1, 2, \ldots, H$ **do**
    P1 takes action $a^k_h \sim \pi^k_h(x^k_h)$, P2 takes action $b^k_h \sim \nu^k_h(x^k_h)$
    Observe reward r^k_h and move to next state x^k_{h+1}
  end for
  Update $\mathcal{B}_h \leftarrow \mathcal{B}_h \cup \{(x^k_h, a^k_h, b^k_h, r^k_h, x^k_{h+1})\}, \forall h \in [H]$
  For all $\xi, \zeta \in \mathcal{F}$, $\pi \in \Pi$ and $h \in [H]$, let

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

  Update

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

end for

4.2.2 Theoretical results

This section presents our theoretical guarantees of Algorithm 3. First, we introduce two assumptions that take the policy class Π into consideration.

Assumption 4.6 (Π-realizability)**.** *For all $\pi \in \Pi$, $Q^{\pi,\nu^*_\pi} \in \mathcal{F}$.*

Assumption 4.7 (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 4.7 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 state the regret guarantee of Algorithm 3. Due to space constraints, the formal statement and detailed discussions are deferred to Appendix C.2.1.

Theorem 4.8 (Informal statement of Theorem C.8)**.** *Suppose Assumption 4.6 and Assumption 4.7 holds. With probability at least $1 - p$, the regret of Algorithm 3 is upper bounded by $\widetilde{O}\left(H\sqrt{K d_{\mathrm{ME}'} \log[\mathcal{N}(\mathcal{F}, 1/K)\mathcal{N}(\Pi, 1/K)/p]}\right)$ where $d_{\mathrm{ME}'}$ is a variant of Minimax Eluder dimension defined in Definition C.7.*

5 Discussion

In this paper we give a comprehensive discussion of general 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 novel 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, 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. 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)*. `https://arxiv.org/pdf/1810.02054`, 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. *Journal of Machine Learning Research*, 11(Apr):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.

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. *arXiv preprint arXiv:1406.1853*, 2014.

Julien Perolat, 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, 2015a.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum markov games. *International Conference on Machine Learning*, 37:1321–1329, 2015b.

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 Perolat, 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.

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, 2021.