Provably Efficient Fictitious Play Policy Optimization for Zero-Sum Markov Games with Structured Transitions

Shuang Qiu [1] **Xiaohan Wei** [2] **Jieping Ye** [1] **Zhaoran Wang** [3] **Zhuoran Yang** [4]

Abstract

While single-agent policy optimization in a fixed environment has attracted a lot of research attention recently in the reinforcement learning community, much less is known theoretically when there are multiple agents playing in a potentially competitive environment. We take steps forward by proposing and analyzing new fictitious play policy optimization algorithms for two-player zero-sum Markov games with structured but unknown transitions. We consider two classes of transition structures: factored independent transition and single-controller transition. For both scenarios, we prove tight $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bounds after T steps in a two-agent competitive game scenario. The regret of each player is measured against a potentially adversarial opponent who can choose a single best policy in hindsight after observing the full policy sequence. Our algorithms feature a combination of Upper Confidence Bound (UCB)-type optimism and fictitious play under the scope of simultaneous policy optimization in a non-stationary environment. When both players adopt the proposed algorithms, their overall optimality gap is $\widetilde{\mathcal{O}}(\sqrt{T})$.

1. Introduction

Widely applied in multi-agent reinforcement learning (Sutton & Barto, 2018; Bu et al., 2008), Policy Optimization (PO) has achieved tremendous empirical success (Foerster et al., 2016; Leibo et al., 2017; Silver et al., 2016; 2017; Berner et al., 2019; Vinyals et al., 2019), due to its high efficiency and easiness to combine with different optimization techniques. Despite these empirical successes, theoretical understanding of multi-agent policy optimization, especially the zero-sum Markov game (Littman, 1994) via policy optimization, lags rather behind. Most recent works studying zero-sum Markov games (e.g. Xie et al. (2020); Bai & Jin (2020)) focus on value-based methods achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regrets and they assume there is a central controller available solving for coarse correlated equilibrium or Nash equilibrium at each step, which introduces extra computational cost. On the other hand, although there has been great progress on understanding single-agent PO algorithms (Sutton et al., 2000; Kakade, 2002; Schulman et al., 2015; Papini et al., 2018; Cai et al., 2019; Bhandari & Russo, 2019; Liu et al., 2019), directly extending single-agent PO to multi-agent setting encounters a main challenge of non-stationary environments caused by agents changing their own policies simultaneously (Bu et al., 2008; Zhang et al., 2019a). In this paper, we aim to answer the following open question:

Can policy optimization probably solve zero-sum Markov games to achieve $\mathcal{O}(\sqrt{T})$ regrets?

As an initial attempt to tackle the problem, in this work, we focus on two *non-trivial* classes of zero-sum Markov games with structured transitions: *factored independent transition* and *single-controller transition*. For the game with the factored independent transition, the transition model is factored into two independent parts, and each player makes transition following their own transition model. The single-controller zero-sum game assumes that the transition model is entirely controlled by the actions of Player 1. In both settings, the rewards received are decided jointly by the actions of both players. These two problems capture the non-stationarity of the multi-agent reinforcement learning in the following aspects: (1) the rewards depend on both players' potentially adversarial actions and policies in both settings; (2) Player 2 in the single-controller setting faces non-stationary states determined by Player 1's policies. In addition to the non-stationarity, practically, the true transition model of the environment could be unknown to players and only bandit feedback is accessible to players. Thus, the non-stationarity, as well as the unknown transition model and the full reward function, poses great challenges to the design and theoretical analysis of the multi-agent PO algorithms.

In this paper, we propose two novel optimistic Fictitious

[1] University of Michigan [2] Facebook, Inc. [3] Northwestern University [4] Princeton University. Correspondence to: Shuang Qiu <qiush@umich.edu>, Xiaohan Wei <ubimeteor@fb.com>, Jieping Ye <jpye@umich.edu>, Zhaoran Wang <zhaoranwang@gmail.com>, Zhuoran Yang <zy6@princeton.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Play (FP) policy optimization algorithms for the games with factored independent transition and single-controller zero-sum games respectively. Our algorithms are motivated by the close connection between multi-agent PO and FP framework. Specifically, FP (Robinson, 1951) is a classical model for solving games based on simultaneous policy updates, which includes two major steps: inferring the opponent (including learning the opponent's policy), and taking the best response policy against the estimated policy of the opponent. As an extension of FP to Markov games, our proposed PO algorithms possess two phases of learning, namely policy evaluation and policy improvement. The policy evaluation phase involves exchanging the policies of the previous step, which is motivated by the step of inferring the opponent in FP. By making use of the policies from the previous step, the algorithms further compute the value function and the Q-function with the estimated reward function and transition model. By the principle of "optimism in the face of uncertainty" (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012), their estimation incorporates bonus terms to handle the non-stationarity of the environment as well as the uncertainty arising from only observing finite historical data. Furthermore, the policy improvement phase corresponds to taking the (regularized) best response policy via a mirror descent/ascent step (where the regularization comes from KL divergence), which can be viewed as soft-greedy step based on the historical information about the opponent and the environment. This step resembles the smoothed FP (Fudenberg & Levine, 1995; Perolat et al., 2018; Zhang et al., 2019a) for normal form games (or matrix games). During this phase, both players in the factored independent transition setting and Player 2 in the single-controller setting demand to estimate the opponent's state reaching probability to handle the non-stationarity.

For each player, we measure the performance of its algorithm by the regret of the learned policy sequence comparing against the best policy in hindsight after T steps. In the two settings, our proposed algorithms can achieve an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for both players, matching the regret of value-based algorithms. Furthermore, with both players running the proposed PO algorithms, they have $\widetilde{\mathcal{O}}\sqrt{T}$ optimality gap. To the best of our knowledge, this seems the first provably sample-efficient fictitious play policy optimization algorithm for zero-sum Markov games (with structured transitions). Our work also partially solves one open question in Bai & Jin (2020) that how to solve a zero-sum Markov game of multiple steps ($H \geq 2$) with an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret via mirror descent-type (policy optimization) algorithms.

Related Work. There have been a large number of classical works studying the games with the independent transition model, e.g., Altman et al. (2005; 2008); Flesch et al. (2008); Singh & Hemachandra (2014). In addition, the single-controller games are also broadly investigated in many existing works, .e.g, Parthasarathy & Raghavan (1981); Filar & Raghavan (1984); Rosenberg et al. (2004); Guan et al. (2016). Most of the aforementioned works do not focus on the non-asymptotic regret analysis. Guan et al. (2016) studies the regret of the single-controller zero-sum game but with an assumption that the transition model is known to players. In contrast, our work provide a regret analysis for both transition models under a more realistic setting that the transition model is unknown. The games with the two structured transition models is closely associated with the applications in communications. The game with the factored independent transition (Altman et al., 2005) finds applications in wireless communications. An application example of the single-controller game is the attack-defense modeling in communications (Eldosouky et al., 2016).

Recently, there are many works focusing on the non-asymptotic analysis of Markov games (Heinrich & Silver, 2016; Guan et al., 2016; Wei et al., 2017; Perolat et al., 2018; Zhang et al., 2019b; Xie et al., 2020; Bai & Jin, 2020). Some of them aim to propose sample-efficient algorithms with theoretical regret guarantees for zero-sum games. Wei et al. (2017) proposes an algorithm extending single-agent UCRL2 algorithm (Jaksch et al., 2010), which requires solving an constrained optimization problem each round. Zhang et al. (2019b) also studies PO algorithms but does not provide regret analysis, which also assume extra linear quadratic structure and known transition model. In addition, recent works on Markov games (Xie et al., 2020; Bai & Jin, 2020; Liu et al., 2020; Bai et al., 2020) propose value-based algorithms under the assumption that there exists a central controller that specifies the policies of agents by finding the coarse correlated equilibrium or Nash equilibrium for a set of matrix games in each episode. Bai & Jin (2020) also makes an attempt to investigate PO algorithms in zero-sum games. However, this work shows restrictive results where each player only play one step in each episode. A concurrent work (Tian et al., 2020) studies zero-sum games under a different online agnostic setting with PO methods and achieves an $\widetilde{\mathcal{O}}(T^{3/4})$ regret. Comparing with aforementioned works, motivated by classical fictitious play works (Robinson, 1951; Fudenberg & Levine, 1995; Heinrich et al., 2015; Perolat et al., 2020), recently, we focus on the setting where there is no central controller which determines the policies of the two players and we propose an fictitious play policy optimization algorithm where each player updates its own policy based solely on the historical information at hand. Moreover, our result matches the $\mathcal{O}(\sqrt{T})$ regret upper bounds in Xie et al. (2020); Bai & Jin (2020) that are obtained by value-based methods .

Furthermore, we note that the game for each individual player can be viewed as a special case of MDP with adversarial rewards and bandit feedback due to the adversarial actions of opponents. For such a class of MDP models in general, Jin & Luo (2019) proposes an algorithm based on mirror descent involving occupancy measures and attains an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. However, each update step of the algorithm requires solving another optimization problem which is more computationally demanding than our PO method. Besides, it is also unclear whether the algorithm in Jin & Luo (2019) can be extended to zero-sum games. Moreover, for the same MDP model, Efroni et al. (2020) proposes a optimistic policy optimization algorithm that achieves an $\widetilde{\mathcal{O}}(T^{2/3})$ regret. Thus, directly applying this result would yield an $\widetilde{\mathcal{O}}(T^{2/3})$ regret. In fact, regarding the problem as an MDP with adversarial rewards neglects the fact that such "adversarial reward functions" are determined by the actions and policies of the opponent. Thus, since each player knows the past actions taken and policies executed by the opponent under the FP framework, both players can construct accurate estimators of the reward functions after a sufficiently large number of episodes. As we will show in Sections 3 and 4, the proposed PO methods explicitly utilizes the information of the opponent in the policy evaluation step, which is critical for the method to obtain an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

2. Background and Preliminaries

In this section, we formally introduce notations and setups. Then, we describe the two transition structures in detail .

2.1. Notations and Setups

We define a tabular episodic two-player zero-sum Markov game (MG) by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \mathcal{P}, r)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ and $\mathcal{B}$ are finite action spaces of Player 1 and Player 2 respectively, H is the length of each episode, $\mathcal{P}_h(s' \,|\, s, a)$ denotes the transition probability at the h-th step to the state s' in the $(h+1)$-th step when Player 1 takes action $a \in \mathcal{A}$ in an episode, $r_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \mapsto [0,1]$ denotes the reward function at the h-step, with the value normalized in the range $[0,1]$. In this paper, we let $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^H$ be the *true* transition model, which is *unknown* to both players. Throughout this paper, we let $\langle\cdot,\cdot\rangle_{\mathcal{S}}$, $\langle\cdot,\cdot\rangle_{\mathcal{A}}$, and $\langle\cdot,\cdot\rangle_{\mathcal{B}}$ denote the inner product over $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{B}$ respectively.

The policy of Player 1 is a collection of probability distributions $\mu = \{\mu_h\}_{h=1}^H$ where $\mu_h(a|s) \in \Delta_{\mathcal{A}}$ with $\Delta_{\mathcal{A}}$ denoting a probability simplex defined on space $\mathcal{A}$. Analogously, we have the policy of Player 2 as a collection of probability distributions $\nu = \{\nu_h\}_{h=1}^H$, where $\nu_h(b|s) \in \Delta_{\mathcal{B}}$ with $\Delta_{\mathcal{B}}$ denoting the probability simplex on space $\mathcal{B}$. We denote $\mu^k = \{\mu_h^k\}_{h=1}^H$ and $\nu^k = \{\nu_h^k\}_{h=1}^H$ as the policies at episode k for Players 1 and 2.

Fictitious Play. At the beginning of the k-th episode, each player observes the opponent's policy during the $(k-1)$-th episode. For simplicity of theoretical analysis, we assume there exists an oracle to exchange players' past policies. Then, they take regularized best response policies via a mirror descent/ascent step for the current episode and make simultaneous moves. By the end of the k-th episodes, each player observes only the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$ and the bandit feedback along the trajectory. The bandit setting is more challenging than the full-information setting, where only the reward values $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$ on the trajectory are observed rather than the exact value function $r_h(s,a,b)$ for all $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$. Moreover, the rewards $r_h^k(\cdot,\cdot,\cdot) \in [0,1]$ is time-varying with its expectation $r_h = \mathbb{E}[r_h^k]$ which can be adversarially affected by the opponent's action or policy, indicating the non-stationarity of the environment.

Value Function. We define the value function $V_h^{\mu,\nu} : \mathcal{S} \mapsto \mathbb{R}$ under any policies $\mu = \{\mu_h\}_{h=1}^H$, $\nu = \{\nu_h\}_{h=1}^H$ and the transition model $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^H$ by $V_h^{\mu,\nu}(s) := \mathbb{E}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}, b_{h'}) \,|\, s_h = s]$, where the expectation is taken over the random state-action pairs $\{(s_{h'}, a_{h'}, b_{h'})\}_{h'=h}^H$. The corresponding action-value function (Q-function) $Q_h^{\mu,\nu} : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \mapsto \mathbb{R}$ is then defined as $Q_h^{\mu,\nu}(s,a,b) := \mathbb{E}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}, b_{h'}) \,|\, s_h = s, a_h = a, b_h = b]$. Therefore, according to the above definitions, we have the following Bellman equation

$$V_h^{\mu,\nu}(s) = [\mu_h(\cdot|s)]^\top Q_h^{\mu,\nu}(s,\cdot,\cdot)\nu_h(\cdot|s), \tag{1}$$

$$Q_h^{\mu,\nu}(s,a,b) = r_h(s,a,b) + \big\langle \mathcal{P}_h(\cdot|s,a,b), V_{h+1}^{\mu,\nu}(\cdot)\big\rangle_{\mathcal{S}}, \tag{2}$$

where $\mu_h(\cdot|s)$ and $\nu_h(\cdot|s)$ are column vectors over the space $\mathcal{A}$ and the space $\mathcal{B}$ respectively, $V_{h+1}^{\mu,\nu}(\cdot)$ is a column vector over the space $\mathcal{S}$, and $Q_h^{\mu,\nu}(s,\cdot,\cdot)$ is a matrix over the space $\mathcal{A}\times\mathcal{B}$. The above Bellman equation holds for all $h \in [H]$ with setting $V_{H+1}^{\mu,\nu}(s) = 0, \forall s \in \mathcal{S}$. Hereafter, to simplify the notation, we let $\mathcal{P}V(s,a,b) := \langle \mathcal{P}(\cdot|s,a,b), V(\cdot)\rangle_{\mathcal{S}}$ for any value function V and transition $\mathcal{P}$.

Nash Equilibrium. We define the Nash equilibrium (NE) as $(\mu^\dagger, \nu^\dagger)$ as a solution to $\max_\mu \min_\nu V_1^{\mu,\nu}(s_1)$. Then, we further have the following relation

$$V_1^{\mu^\dagger,\nu^\dagger}(s_1) = \max_\mu \min_\nu V_1^{\mu,\nu}(s_1) = \min_\nu \max_\mu V_1^{\mu,\nu}(s_1).$$

Thus, we say a policy pair $(\widetilde{\mu}, \widetilde{\nu})$ is an ε-approximate Nash equilibrium if it satisfies

$$\max_\mu V_1^{\mu,\widetilde{\nu}}(s_1) - \min_\nu V_1^{\widetilde{\mu},\nu}(s_1) \le \varepsilon,$$

where the weak duality $\max_\mu V_1^{\mu,\widetilde{\nu}}(s_1) \ge V_1^{\mu^\dagger,\nu^\dagger}(s_1) \ge \min_\nu V_1^{\widetilde{\mu},\nu}(s_1)$ always holds.

Regret and Optimality Gap. The goal for Player 1 is to learn a sequence of policies, $\{\mu^k\}_{k>0}$, to have a small regret as possible in K episodes, which is defined as

$$\mathrm{Regret}_1(K) := \sum_{k=1}^K \Big[V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1)\Big], \tag{3}$$

and $\{\nu^k\}_{k=1}^K$ is any possible and potentially adversarial policy sequence of Player 2. The policy μ^* is *the best policies in hindsight*, which is defined as $\mu^* := \mathrm{argmax}_\mu \sum_{k=1}^K V_1^{\mu,\nu^k}(s_1)$ for any specific $\{\nu^k\}_{k=1}^K$. Similarly, Player 2 aims to learn a sequence of policies, $\{\nu^k\}_{k>0}$, to have a small regret defined as

$$\mathrm{Regret}_2(K) := \sum_{k=1}^K \Big[V_1^{\mu^k,\nu^k}(s_1) - V_1^{\mu^k,\nu^*}(s_1)\Big]. \quad (4)$$

where $\{\mu^k\}_{k=1}^K$ is any possible policy sequence of Player 1. The policies ν^* is also *the best policies in hindsight* which is defined as $\nu^* := \mathrm{argmin}_\nu \sum_{k=1}^K V_1^{\mu^k,\nu}(s_1)$ for any specific $\{\mu^k\}_{k=1}^K$. Note that μ^* and ν^* depend on opponents' policy sequence and is non-deterministic, and we drop such a dependency in the notation for simplicity. We further define the *optimality gap* $\mathrm{Gap}(K)$ as follows

$$\mathrm{Gap}(K) := \mathrm{Regret}_1(K) + \mathrm{Regret}_2(K). \quad (5)$$

Our definition of optimality gap is consistent with a certain form of the regret to measure the learning performance of zero-sum games defined in Bai & Jin (2020, Definition 8). Specifically, when the two players executes their algorithms to have small regrets, i.e., $\mathrm{Regret}_1(K)$ and $\mathrm{Regret}_2(K)$ are small, then their optimality gap $\mathrm{Gap}(K)$ is small as well.

On the other hand, letting the uniform mixture policies $\widehat{\mu} \sim \mathrm{Unif}(\mu^1, \ldots, \mu^K)$ and $\widehat{\nu} \sim \mathrm{Unif}(\nu^1, \ldots, \nu^K)$ be random policies sampled uniformly from the learned policies, then $(\widehat{\mu}, \widehat{\nu})$ can be viewed as an ε-approximate NE if $\mathrm{Regret}(K)/K \le \varepsilon$. This build a connection between the approximate NE and the optimality gap.

2.2. Structured Transition Models

Factored Independent Transition. Consider a two-player MG where the state space are factored as $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ such that a state can be represented as $s = (s^1, s^2)$ with $s^1 \in \mathcal{S}_1$ and $s^2 \in \mathcal{S}_2$. Under this setting, the transition model is factored into two independent components, i.e.,

$$\mathcal{P}_h(s' \mid s, a, b) = \mathcal{P}_h^1(s^{1\prime} \mid s^1, a)\mathcal{P}_h^2(s^{2\prime} \mid s^2, b), \quad (6)$$

where we also have $s' = (s^{1\prime}, s^{2\prime})$, and $\mathcal{P}_h(s^{1\prime} \mid s^1, a)$ is the transition model for Player 1 and $\mathcal{P}_h(s^{2\prime} \mid s^2, b)$ for Player 2. Additionally, we consider the case where the policy of Player 1 only depends on its own state s^1 such that we have $\mu(a|s) = \mu(a|s^1)$ and meanwhile Player 2 similarly has the policy of the form $\nu(b|s) = \nu(b|s^2)$. Though the transitions, policies, and state spaces of two players are independent of each other, the reward function still depends on both players' actions and states, i.e., $r_h(s, a, b) = r_h(s^1, s^2, a, b)$.

Single-Controller Transition. In this setting, we take steps forward by not assuming the relatively independent structures of the policies and state spaces for two players. For the single-controller game, we consider that the transition model is controlled by the action of one player, e.g., Player 1 in this paper, which is thus characterized by

$$\mathcal{P}_h(s' \mid s, a, b) = \mathcal{P}_h(s' \mid s, a). \quad (7)$$

In addition, the policies remain to be $\mu(a|s)$ and $\nu(b|s)$ that depend on the state s jointly decided by both players, and reward $r_h(s, a, b)$ is determined by both players as well.

Remark 2.1 (Misspecification)**.** *The above assumptions are made only for ease of theoretical analysis. When the above transition models may not be ideally satisfied, one can potentially consider scenarios that the transitions satisfy, for example,* $\|\mathcal{P}_h(\cdot \mid s, a, b) - \mathcal{P}_h(s^{1\prime} \mid s^1, a)\mathcal{P}_h(s^{2\prime} \mid s^2, b)\|_\infty \le \varrho$ *or* $\|\mathcal{P}_h(\cdot \mid s, a, b) - \mathcal{P}_h(\cdot \mid s, a)\|_\infty \le \varrho$, $\forall (s, a, b)$, *with a misspecification error* ϱ. *One can still follow the techniques in this paper to analyze such misspecified scenarios, and obtain regrets with an extra bias term* ϱT, *as the misspecification error* ϱ *will be accumulated across* T *episodes. When* ϱ *is small, it implies that the MG has approximately factored independent transition or single-controller transition structures, and then the extra bias term* ϱT *should be small.*

3. MG with Factored Independent Transition

In this section, we propose and analyze optimistic policy optimization algorithms for both players under the setting of the factored independent transition.

Algorithm for Player 1. The algorithm for Player 1 is illustrated in Algorithm 1. Assume that the game starts from a fixed state $s_1 = (s_1^1, s_1^2)$ each round. We also assume that the true transition model $\mathcal{P}$ is not known to Player 1, and Player 1 can only access the bandit feedback of the rewards along this trajectory instead of the full information. Thus, Player 1 needs to empirically estimate the reward function and the transition model for all (s, a, b, s') and $h \in [H]$ via

$$\begin{aligned}
\widehat{r}_h^k(s, a, b) &= \frac{\sum_{\tau=1}^k \mathbf{1}_{\{(s,a,b)=(s_h^\tau, a_h^\tau, b_h^\tau)\}} r_h^k(s, a, b)}{\max\{N_h^k(s, a, b), 1\}}, \\
\widehat{\mathcal{P}}_h^{1,k}(s^{1\prime}|s^1, a) &= \frac{\sum_{\tau=1}^k \mathbf{1}_{\{(s^1,a,s^{1\prime})=(s_h^{1,\tau}, a_h^\tau, s_{h+1}^{1,\tau})\}}}{\max\{N_h^k(s^1, a), 1\}}, \quad (8) \\
\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime}|s^2, b) &= \frac{\sum_{\tau=1}^k \mathbf{1}_{\{(s^2,b,s^{2\prime})=(s_h^{2,\tau}, b_h^\tau, s_{h+1}^{2,\tau})\}}}{\max\{N_h^k(s^2, b), 1\}},
\end{aligned}$$

where we denote $\mathbf{1}_{\{\cdot\}}$ as an indicator function, and $N_h^k(s, a, b)$ counts the empirical number of observation for a certain tuple (s, a, b) at step h until k-th iteration as well as $N_h^k(s^1, a)$ for (s^1, a) and $N_h^k(s^2, b)$ for (s^2, b). Then, we have the estimation of the overall transition as $\widehat{\mathcal{P}}_h^k(s'|s, a, b) = \widehat{\mathcal{P}}_h^{1,k}(s^{1\prime}|s^1, a)\widehat{\mathcal{P}}_h^{2,k}(s^{2\prime}|s^2, b)$. For simplicity of presentation, in this section, we let $s = (s^1, s^2)$ and we use s^1, s^2 separately when necessary.

Based on the estimation of the transition model and reward function, we further estimate the Q-function and value-function as shown in Line 7 and 8 in Algorithm 1. In terms of the principle of *"optimism in the face of uncertainty"*, bonus terms are introduced to construct a UCB update for Q-function as shown in Line 7 of Algorithm 1. We set the bonus term as

$$\beta_h^k(s,a,b) = \beta_h^{r,k}(s,a,b) + \beta_h^{\mathcal{P},k}(s,a,b), \qquad (9)$$

where we define $\beta_h^{r,k}(s,a,b) := \sqrt{\frac{4\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}$ as well as $\beta_h^{\mathcal{P},k}(s,a,b) := \sqrt{\frac{2H^2|\mathcal{S}_1|\log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s^1,a),1\}}} + \sqrt{\frac{2H^2|\mathcal{S}_2|\log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^2,b),1\}}}$ with $\delta \in (0,1)$. Here, we decompose $\beta_h^k(s,a,b)$ into two terms where $\beta_h^{r,k}(s,a,b)$ is the bonus term for the reward and $\beta_h^{\mathcal{P},k}(s,a)$ for the transition estimation. As shown in Lemmas B.3 and B.4 of the supplementary material, the bonus terms $\beta_h^{r,k}(s,a,b)$ and $\beta_h^{\mathcal{P},k}(s,a,b)$ are obtained by using Hoeffding's inequality. Note that the two terms in the definition of $\beta_h^{\mathcal{P},k}$ stem from the uncertainties of estimating both transitions $\mathcal{P}_h^1(s^{1\prime} \mid s^1, a)$ and $\mathcal{P}_h^2(s^{2\prime} \mid s^2, b)$.

Next, we introduce the notion of the state reaching probability $q^{\nu^k,\mathcal{P}^2}(s^2)$ for any state $s^2 \in \mathcal{S}_2$ under the policy ν^k and the true transition $\mathcal{P}^2$, which is defined as

$$q_h^{\nu^k,\mathcal{P}^2}(s^2) := \Pr(s_h^2 = s^2 \mid \nu^k, \mathcal{P}^2, s_1^2), \forall h \in [H].$$

To handle non-stationarity of the opponent, as in Line 10, Player 1 needs to estimate the state reaching probability of Player 2 by the empirical reaching probability under the empirical transition model $\widehat{\mathcal{P}}^{2,k}$ for Player 2, i.e.,

$$d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2) = \Pr(s_h^2 = s^2 \mid \nu^k, \widehat{\mathcal{P}}^{2,k}, s_1^2), \forall h \in [H].$$

The empirical reaching probability can be simply computed dynamically from $h = 1$ to H by $d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2) = \sum_{s^{2\prime}\in\mathcal{S}_2}\sum_{a'\in\mathcal{A}} d_{h-1}^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^{2\prime})\nu_{h-1}^k(b'|s^{2\prime})\widehat{\mathcal{P}}_{h-1}^{2,k}(s^2|s^{2\prime},b')$. Based on the estimated state reaching probability, the policy improvement step is associated with solving the following optimization problem (denoting by D_{KL} the KL divergence)

$$\max_{\mu}\sum_{h=1}^{H}[\overline{G}_h^{k-1}(\mu_h) - \eta^{-1}D_{\mathrm{KL}}(\mu_h(\cdot|s^1), \mu_h^k(\cdot|s^1))], \quad (10)$$

where we define the linear function as $\overline{G}_h^{k-1}(\mu_h) := \langle \mu_h(\cdot|s^1) - \mu_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s,\cdot)d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\rangle_{\mathcal{A}}$ with $F_h^{1,k}(s,a) = \langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot|s^2)\rangle_{\mathcal{B}}$. One can see that (10) is a mirror ascent step and admits a closed-form solution as $\mu_h^k(a|s^1) = (Y_h^{k-1})^{-1}\mu_h^{k-1}(a \mid s^1) \cdot \exp\{\eta\sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s,a)d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)\rangle_{\mathcal{B}}\}$, where Y_h^{k-1} is a probability normalization term.

Algorithm 1 Optimistic Policy Optimization for Player 1

```
1: Initialize: For all $h \in [H]$, $(s^1, s^2, a, b) \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{A} \times \mathcal{B}$: $\mu_h^0(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$, $\widehat{\mathcal{P}}_h^{1,0}(\cdot|s^1,a) = \mathbf{1}/|\mathcal{S}_1|$, $\widehat{\mathcal{P}}_h^{2,0}(\cdot|s^2,b) = \mathbf{1}/|\mathcal{S}_2|$, $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^0(\cdot,\cdot,\cdot) = \mathbf{0}$.
2: for episode $k = 1, \ldots, K$ do
3:     Observe Player 2's policy $\{\nu_h^{k-1}\}_{h=1}^H$.
4:     Start from state $s_1 = (s_1^1, s_1^2)$, set $\overline{V}_{H+1}^{k-1}(\cdot) = \mathbf{0}$.
5:     for step $h = H, H-1, \ldots, 1$ do
6:         Estimate the transition and reward function by $\widehat{\mathcal{P}}_h^{k-1}(\cdot|\cdot,\cdot)$ and $\widehat{r}_h^{k-1}(\cdot,\cdot,\cdot)$ as (11).
7:         Update Q-function $\forall(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$:
           $\overline{Q}_h^{k-1}(s,a,b) = \min\{(\widehat{r}_h^{k-1} + \widehat{\mathcal{P}}_h^{k-1}\overline{V}_{h+1}^{k-1} + \beta_h^{k-1})(s,a,b), H-h+1\}^+$.
8:         Update value-function $\forall s \in \mathcal{S}$:
           $\overline{V}_h^{k-1}(s) = [\mu_h^{k-1}(\cdot|s)]^\top \overline{Q}_h^{k-1}(s,\cdot,\cdot)\nu_h^{k-1}(\cdot|s)$.
9:     end for
10:    Compute the empirical state reaching probability $d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s^2)$ of Player 2 under $\nu^k, \widehat{\mathcal{P}}^{2,k}$, $\forall h \in [H]$.
11:    Update policy $\mu_h^k(a|s^1)$ by solving (10), $\forall(s^1,a,h)$.
12:    Take actions following $a_h^k \sim \mu_h^k(\cdot|s_h^{1,k})$, $\forall h \in [H]$.
13:    Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
14: end for
```

Algorithm for Player 2. For the setting of MG with factored independent transition, the algorithm for Player 2 is trying to minimize the expected cumulative reward w.r.t. $r_h(\cdot,\cdot,\cdot)$. In another word, Player 2 is maximizing the expected cumulative reward w.r.t. $-r_h(\cdot,\cdot,\cdot)$. From this perspective, one can view the algorithm for Player 2 as a *'symmetric'* version of Algorithm 1. Due to the limit of space here, we present Algorithm 4 in Section A of the supplementary material. Specifically, in this algorithm, Player 2 also estimates the transition model and the reward function the same as (11). Since Player 2 is minimizing the expected cumulative reward, the bonus terms as (9) are subtracted in the Q-function estimation step by the UCB optimism principle. The algorithm further estimates the state reaching probability of Player 1, $q_h^{\mu^k,\mathcal{P}^1}(s^1)$, by the empirical one $d_h^{\mu^k,\widehat{\mathcal{P}}^{1,k}}(s^1)$, which can be dynamically computed. For the policy improvement step, Algorithm 1 performs a mirror descent step based on the empirical reaching probability. Please see more details in Section A.

3.1. Theoretical Results

Theorem 3.1. *By setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, with probability at least $1-4\delta$, Algorithm 1 ensures the sublinear regret bound for Player 1*[1] *i.e.,* $\mathrm{Regret}_1(K) \le \widetilde{\mathcal{O}}\big(C\sqrt{T}\big)$,

[1] Hereafter, we use $\widetilde{\mathcal{O}}$ to hide the logarithmic factors on $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{B}|, H, K$, and $1/\delta$.

where $T = HK$ is the number of steps, and the constant $C = \sqrt{(|\mathcal{S}_1|^2|\mathcal{A}| + |\mathcal{S}_2|^2|\mathcal{B}|)H^3} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H}$.

Theorem 3.1 shows that Player 1 can obtain an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret by Algorithm 1, when the opponent, Player 2, performs potentially adversarial policies.

Theorem 3.2. *By setting $\gamma = \sqrt{\log|\mathcal{B}|/(KH^2)}$, with probability at least $1-4\delta$, Algorithm 4 ensures the sublinear regret bound for Player 2, i.e., $\mathrm{Regret}_2(K) \le \widetilde{\mathcal{O}}(C\sqrt{T})$, where $T = HK$ is the number of steps, and the constant $C = \sqrt{(|\mathcal{S}_1|^2|\mathcal{A}| + |\mathcal{S}_2|^2|\mathcal{B}|)H^3} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H}$.*

Theorem 3.2 shows that $\mathrm{Regret}_2(K)$ admits the same $\widetilde{\mathcal{O}}(\sqrt{T})$ regret as Theorem 3.1 given any arbitrary and adversarial policies of the opponent Player 1, due to the symmetric nature of the two algorithms.

From the perspective of each individual player, the game can be viewed as a special case of an MDP with adversarial bandit feedback due to the potentially adversarial actions or policies of the opponent. For MDPs with adversarial bandit feedback, Jin & Luo (2019) attains an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret via an occupancy measure based method, which requires solving a constrained optimization problem in each update step that is more computationally demanding than PO. Efroni et al. (2020) proposes a PO method for the same MDP model, achieving an $\widetilde{\mathcal{O}}(T^{2/3})$ regret. Thus, directly applying this result would yield an $\widetilde{\mathcal{O}}(T^{2/3})$ regret. However, for the problem of zero-sum games, regarding the problem faced by one player as an MDP with adversarial rewards neglects the fact that such "adversarial reward functions" are determined by the actions and policies of the opponent. Thus, under the FP framework, by utilizing the past actions and policies of the opponent, Algorithm 1 and 4 obtain an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

In particular, if Player 1 runs Algorithm 1 and Player 2 runs Algorithm 4 *simultaneously*, then we have the following corollary of Theorems 3.1 and 3.2.

Corollary 3.3. *By setting η and γ as in Theorem 3.1 and Theorem 3.2, letting $T = HK$, with probability at least $1-8\delta$, Algorithm 1 and Algorithm 4 ensures the following optimality gap $\mathrm{Gap}(K) \le \widetilde{\mathcal{O}}(\sqrt{T})$.*

4. MG with Single-Controller Transition

In this section, we propose and analyze optimistic policy optimization algorithms for the single-controller game.

Algorithm for Player 1. The algorithm for Player 1 is illustrated in Algorithm 2. Since transition model is unknown and only bandit feedback of the rewards is available, Player 1 needs to empirically estimate the reward function and the transition model for all (s, a, b, s') and $h \in [H]$ via

$$\begin{aligned} \widehat{r}_h^k(s,a,b) &= \frac{\sum_{\tau=1}^k \mathbf{1}_{\{(s,a,b)=(s_h^\tau,a_h^\tau,b_h^\tau)\}} r_h^k(s,a,b)}{\max\{N_h^k(s,a,b),1\}}, \\ \widehat{\mathcal{P}}_h^k(s'|s,a) &= \frac{\sum_{\tau=1}^k \mathbf{1}_{\{(s,a,s')=(s_h^\tau,a_h^\tau,s_{h+1}^\tau)\}}}{\max\{N_h^k(s,a),1\}}. \end{aligned} \tag{11}$$

Based on the estimations, Algorithm (2) further estimates the Q-function and value-function for policy evaluation. In terms of the optimism principle, bonus terms are added to construct a UCB update for Q-function as shown in Line 7 of Algorithm 2. The bonus terms are computed as

$$\beta_h^k(s,a,b) = \beta_h^{r,k}(s,a,b) + \beta_h^{\mathcal{P},k}(s,a), \tag{12}$$

where the two bonus terms above are expressed as $\beta_h^{r,k}(s,a,b) := \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}$ and $\beta_h^{\mathcal{P},k}(s,a) := \sqrt{\frac{2H^2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}}$ for $\delta \in (0,1)$. Here we also decompose $\beta_h^k(s,a,b)$ into two terms with $\beta_h^{r,k}(s,a,b)$ denoting the bonus term for the reward and $\beta_h^{\mathcal{P},k}(s,a)$ for the transition estimation. Note that the transition bonus are only associated with (s,a) due to the single-controller structure. The bonus terms are derived in Lemmas C.4 and C.5 of the supplementary material.

Different from Algorithm 1, in this algorithm for Player 1, there is no need to estimate the state reaching probability of the opponent as the transition only depends on Player 1. The policy improvement step is then associated with solving the following optimization problem

$$\max_{\mu} \sum_{h=1}^H [\overline{L}_h^{k-1}(\mu_h) - \eta^{-1} D_{\mathrm{KL}}(\mu_h(\cdot|s), \mu_h^{k-1}(\cdot|s))], \tag{13}$$

where we define the function $\overline{L}_h^{k-1}(\mu_h) := [\mu_h(\cdot|s) - \mu_h^{k-1}(\cdot|s)]^\top \overline{Q}_h^{k-1}(s,\cdot,\cdot)\nu_h^{k-1}(\cdot|s)$. This is a mirror ascent step and admits the closed-form solution as $\mu_h^k(a|s) = (Z_h^{k-1})^{-1}\mu_h^{k-1}(a\,|\,s)\exp\{\eta\langle\overline{Q}_h^{k-1}(s,a,\cdot), \nu_h^k(\cdot\,|\,s)\rangle_{\mathcal{B}}\}$, where Z_h^{k-1} is a probability normalization term.

Algorithm for Player 2. The algorithm for Player 2 is illustrated in Algorithm 3. Player 2 also estimates the transition model and the reward function the same as (11). However, due to the *asymmetric* nature of the single-controller transition model, Player 2 has a different way to learning the policy. The main differences to Algorithm 2 are summarized in the following three aspects: First, according to our theoretical analysis shown in Lemma C.2, no transition model estimation are involved. Instead, only a reward function estimation is considered in Line 7 of Algorithm 3. Second, in the policy improvement step, Player 2 needs to approximate the state reaching probability $q_h^{\mu^k,\mathcal{P}}(s) := \Pr(s_h = s\,|\,\mu^k, \mathcal{P}, s_1)$ under μ^k and

Algorithm 2 Optimistic Policy Optimization for Player 1

```
1:  Initialize: $\mu_h^0(\cdot|s) = \mathbf{1}/|\mathcal{A}|$ for all $s \in \mathcal{S}$ and $h \in [H]$.
    $\widehat{\mathcal{P}}_h^0(\cdot|s,a) = \mathbf{1}/|\mathcal{S}|$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$.
    $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^0(\cdot,\cdot,\cdot) = \mathbf{0}$ for all $h \in [H]$.
2:  for episode $k = 1, \dots, K$ do
3:      Observe Player 2's policy $\{\nu_h^{k-1}\}_{h=1}^H$.
4:      Start from $s_1^k = s_1$, and set $\overline{V}_{H+1}^{k-1}(\cdot) = \mathbf{0}$.
5:      for step $h = H, H-1, \dots, 1$ do
6:          Estimate the transition and reward function by $\widehat{\mathcal{P}}_h^{k-1}(\cdot|\cdot,\cdot)$ and $\widehat{r}_h^{k-1}(\cdot,\cdot,\cdot)$ as (11).
7:          Update Q-function $\forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$:
            $$\overline{Q}_h^{k-1}(s,a,b) = \min\{\widehat{r}_h^{k-1}(s,a,b) + \widehat{\mathcal{P}}_h^{k-1}\overline{V}_{h+1}^{k-1}(s,a) + \beta_h^{k-1}(s,a,b), H-h+1\}^+$$
8:          Update value-function $\forall s \in \mathcal{S}$:
            $$\overline{V}_h^{k-1}(s) = \left[\mu_h^{k-1}(\cdot|s)\right]^\top \overline{Q}_h^{k-1}(s,\cdot,\cdot)\nu_h^{k-1}(\cdot|s).$$
9:      end for
10:     Update policy $\mu_h^k(a|s)$ by solving (13), $\forall (s,a,h)$.
11:     Take actions following $a_h^k \sim \mu_h^k(\cdot|s_h^k)$, $\forall h \in [H]$.
12:     Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
13: end for
```

true transition $\mathcal{P}$ by the empirical reaching probability $d_h^k(s) = \Pr(s_h = s \,|\, \mu^k, \widehat{\mathcal{P}}^k, s_1)$ with the empirical transition model $\widehat{\mathcal{P}}^k$, which can also be computed dynamically from $h = 1$ to H. Third, we subtract a reward bonus term $\beta_h^{r,k-1}$ in Line 7 instead of adding the bonus. Similar to our discussion in Section 3, it is still a UCB step if viewing Player 2 is maximizing the cumulative reward w.r.t. $-r_h(\cdot,\cdot,\cdot)$.

Particularly, the policy improvement step of Algorithm 3 is associated with solving the following minimization problem

$$\min_\nu \sum_{h=1}^H \{\underline{L}_h^{k-1}(\nu_h) + \gamma^{-1} D_{\mathrm{KL}}\big(\nu_h(\cdot|s), \nu_h^{k-1}(\cdot|s)\big)\}, \quad (14)$$

where we define $\underline{L}_h^{k-1}(\nu_h) := d_h^{k-1}(s)[\mu_h^{k-1}(\cdot|s)]^\top \cdot \widetilde{r}_h^{k-1}(s,\cdot,\cdot)[\nu_h(\cdot|s) - \nu_h^{k-1}(\cdot|s)]$. This is a mirror descent step with the closed-form solution $\nu_h^k(a|s) = (\widetilde{Z}_h^{k-1})^{-1} \cdot \nu_h^k(b\,|\,s)\exp\{-\gamma d_h^{k-1}(s)\langle \widetilde{r}_h^{k-1}(s,\cdot,b), \mu_h^{k-1}(\cdot\,|\,s)\rangle_{\mathcal{A}}\}$, with the denominator $\widetilde{Z}_h^{k-1}$ being a normalization term.

4.1. Theoretical Results

Theorem 4.1. *By setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, with probability at least $1 - 3\delta$, Algorithm 2 ensures the following regret bound for Player 1 $\mathrm{Regret}_1(K) \le \widetilde{\mathcal{O}}\big(C\sqrt{T}\big)$, where $T = HK$ is the total number of steps, and the constant*

$$C = \sqrt{|\mathcal{S}|^2|\mathcal{A}|H^3} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H}.$$

Algorithm 3 Optimistic Policy Optimization for Player 2

```
1:  Initialize: $\nu_h^0(\cdot|s) = \mathbf{1}/|\mathcal{B}|$ for all $s \in \mathcal{S}$ and $h \in [H]$.
    $\widehat{\mathcal{P}}_h^0(\cdot|s,a) = \mathbf{1}/|\mathcal{S}|$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$.
    $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^{r,0}(\cdot,\cdot,\cdot) = \mathbf{0}$ for all $h \in [H]$.
2:  for episode $k = 1, \dots, K$ do
3:      Observe Player 1's policy $\{\mu_h^{k-1}\}_{h=1}^H$.
4:      Start from the initial state $s_1^k = s_1$.
5:      for step $h = 1, 2, \dots, H$ do
6:          Estimate the transition and reward function by $\widehat{\mathcal{P}}_h^{k-1}$ and $\widehat{r}_h^{k-1}$ as (11).
7:          Update $\widetilde{r}_h^{k-1}$, $\forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$:
            $$\widetilde{r}_h^{k-1}(s,a,b) = \max\left\{\widehat{r}_h^{k-1}(s,a,b) - \beta_h^{r,k-1}(s,a,b), 0\right\}.$$
8:          Estimate the reaching probability by $d_h^{k-1}(s)$ with $\mu_h^{k-1}$ and $\widehat{\mathcal{P}}_h^{k-1}$, $\forall s \in \mathcal{S}$.
9:      end for
10:     Update policy $\nu_h^k(b|s)$ by solving (14), $\forall (s,b,h)$.
11:     Take actions following $b_h^k \sim \nu_h^k(\cdot|s_h^k)$, $\forall h \in [H]$.
12:     Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
13: end for
```

Theorem 4.1 shows that $\mathrm{Regret}_1(K)$ is in the level of $\widetilde{\mathcal{O}}(\sqrt{T})$, for arbitrary policies of Player 2. Similar to the discussion after Theorem 3.2, from the perspective of Player 1, the game can also be viewed as a special case of an MDP with adversarial bandit feedback. Under the FP framework, by utilizing the past policies of Player 2, Algorithm 2 can achieve an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, comparing to $\widetilde{\mathcal{O}}(T^{2/3})$ regret by the PO method (Efroni et al., 2020) and $\widetilde{\mathcal{O}}(T^{1/2})$ regret by a computationally demanding non-PO method (Jin & Luo, 2019) for MDP with adversarial rewards.

Theorem 4.2. *By setting $\gamma = \sqrt{|\mathcal{S}|\log|\mathcal{B}|/K}$, with probability at least $1 - 2\delta$, Algorithm 3 ensures the sublinear regret bound for Player 2, i.e., $\mathrm{Regret}_2(K) \le \widetilde{\mathcal{O}}(C\sqrt{T})$, where $T = HK$ is the total number of steps, and the constant $C = \sqrt{|\mathcal{S}|^2|\mathcal{A}|H^3} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H}$.*

Interestingly, Theorem 4.2 also shows that $\mathrm{Regret}_2(K)$ has the same bound (including the constant factor C) as $\mathrm{Regret}_1(K)$ given any opponent's policy, though the transition model bonus is not involved in Algorithm 3 and the learning process for two players are essentially different. In fact, although the bonus term for estimating the transition is not involved in this algorithm, approximating the state reaching probability of Player 1 implicitly reflects the gap between the empirical transition $\widehat{\mathcal{P}}^k$ and the true transition $\mathcal{P}$, which can explain the same upper bound in Theorems 4.1 and 4.2.

Moreover, if Player 1 runs Algorithm 1 and Player 2 runs Algorithm 4 *simultaneously*, we have the following corollary of the above two theorems.

Corollary 4.3. *By setting η and γ as in Theorem 4.1 and Theorem 4.2, letting $T = HK$, with probability at least $1 - 5\delta$, Algorithm 2 and Algorithm 3 ensures the optimality gap* $\mathrm{Gap}(K) \le \widetilde{\mathcal{O}}(\sqrt{T})$.

5. Theoretical Analysis

5.1. Proofs of Theorems 3.1 and 3.2

Proof. To bound $\mathrm{Regret}_1(K)$, we need to analyze the value function difference for the instantaneous regret at the k-th episode, i.e., $V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1)$. By Lemma B.1, we decompose the difference between $V_1^{\mu^*,\nu^k}(s_1)$ and $V_1^{\mu^k,\nu^k}(s_1)$ into four terms

$$
\begin{aligned}
V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1) \le{}& \underbrace{\overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1)}_{\mathrm{Err}_k(\mathrm{I.1})} \\
&+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\{[\mu_h^*(\cdot|s_h)]^\top \bar{\iota}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,|\, s_1\}}_{\mathrm{Err}_k(\mathrm{I.2})} \\
&+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}^1}\{\langle \mu_h^*(\cdot|s_h^1) - \mu_h^k(\cdot|s_h^1), M_h^k(s_h^1,\cdot)\rangle_{\mathcal{A}} \,|\, s_1\}}_{\mathrm{Err}_k(\mathrm{I.3})} \\
&+ \underbrace{2H \sum_{h=1}^{H} \sum_{s_h^2 \in \mathcal{S}_2} |q_h^{\nu^k,\mathcal{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)|,}_{\mathrm{Err}_k(\mathrm{I.4})}
\end{aligned}
$$

where $M_h^k(s_h^1,\cdot) := \sum_{s_h^2\in\mathcal{S}_2} F_h^k(s_h^1, s_h^2,\cdot) d_h^{\nu^k,\widehat{\mathcal{P}}^{2,k}}(s_h^2)$. Here we define the model prediction error of Q-function as $\bar{\iota}_h^k(s,a,b) = r_h(s,a,b) + \mathcal{P}_h\overline{V}_{h+1}^k(s,a,b) - \overline{Q}_h^k(s,a,b)$. Let s_h^1, s_h^2, a_h, b_h be random variables for state and actions.

Specifically, $\mathrm{Err}_k(\mathrm{I.1})$ is the difference between the estimated value function and the true value function, $\mathrm{Err}_k(\mathrm{I.2})$ is associated with the model prediction error $\bar{\iota}_h^k(s,a,b)$ for Q-function, $\mathrm{Err}_k(\mathrm{I.3})$ is the error from the policy mirror ascent step, and $\mathrm{Err}_k(\mathrm{I.4})$ is the error for reaching probability estimation. As shown in Lemmas B.6, B.5, B.2, B.8, we have that $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.1}) \le \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}_1|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}_2|^2|\mathcal{B}|H^4K} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H^2K})$, then the second error term is bounded as $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.2}) \le 0$, the third error term is bounded as $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.3}) \le \mathcal{O}(\sqrt{H^4K\log|\mathcal{A}|})$, and the last error term is bounded as $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.4}) \le \widetilde{\mathcal{O}}(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K})$. The above inequalities hold with probability at least $1 - 4\delta$ by union bound. As shown above, by the UCB optimism, the second error term is always bounded by 0, which shows the significance of the principle of "optimism in the face of uncertainty. Therefore, letting $T = HK$, by the relation that $\mathrm{Regret}_1(K) = \sum_{k=1}^K [V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1)] \le \sum_{k=1}^K [\mathrm{Err}_k(\mathrm{I.1}) + \mathrm{Err}_k(\mathrm{I.2}) + \mathrm{Err}_k(\mathrm{I.3}) + \mathrm{Err}_k(\mathrm{I.4})]$, we can obtain the result in Theorem 3.1.

Due to the symmetric nature of Algorithm 1 and Algorithm 4 as we discussed in Section 3, the proof for Theorem 3.2 exactly follows the proof of Theorem 3.2. This completes the proof. □

5.2. Proofs of Theorems 4.1 and 4.2

Proof. We first show the proof of Theorem 4.1. By lemma C.1, we have that the difference between value functions $V_1^{\mu^*,\nu^k}(s_1)$ and $V_1^{\mu^k,\nu^k}(s_1)$ is bounded by three terms

$$
\begin{aligned}
V_1^{\mu^*,\nu^k}(s_1) - V_1^{\mu^k,\nu^k}(s_1) \le{}& \underbrace{\overline{V}_1^k(s_1) - V_1^{\mu^k,\nu^k}(s_1)}_{\mathrm{Err}_k(\mathrm{II.1})} \\
&+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P}}[\langle \mu_h^*(\cdot|s_h) - \mu_h^k(\cdot|s_h), U_h^k(s_h,\cdot)\rangle_{\mathcal{A}} \,|\, s_1]}_{\mathrm{Err}_k(\mathrm{II.2})} \\
&+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\mathcal{P},\nu^k}\big[\bar{\varsigma}_h^k(s_h,a_h,b_h) \,\big|\, s_1\big],}_{\mathrm{Err}_k(\mathrm{II.3})}
\end{aligned}
$$

where s_h, a_h, b_h are random variables for state and actions, $U_h^k(s,a) := \langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot\,|\,s)\rangle_{\mathcal{B}}$, and we define the model prediction error of Q-function as $\bar{\varsigma}_h^k(s,a,b) = r_h(s,a,b) + \mathcal{P}_h\overline{V}_{h+1}^k(s,a) - \overline{Q}_h^k(s,a,b)$.

Particularly, $\mathrm{Err}_k(\mathrm{II.1})$ is the difference between the estimated $\overline{V}_1^k(s_1)$ and the true value function $V_1^{\mu^k,\nu^k}(s_1)$, $\mathrm{Err}_k(\mathrm{II.2})$ characterizes the error from the policy mirror ascent step, and $\mathrm{Err}_k(\mathrm{II.3})$ is associated with the model prediction error $\bar{\varsigma}_h^k(s,a,b)$ for Q-function. As shown in Lemma C.7, $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{II.1}) \le \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K})$ with probability at least $1-\delta$. In addition, Lemma C.3 shows the cumulative error for the mirror ascent step is $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{II.2}) \le \mathcal{O}(\sqrt{H^4K\log|\mathcal{A}|})$ with setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$. Furthermore, by the UCB optimism, we have $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{II.3}) \le 0$ with probability at least $1 - 2\delta$ as shown in Lemma C.6, which shows the significance of the principle of "optimism in the face of uncertainty. Therefore, letting $T = HK$, further by the relation that $\mathrm{Regret}_1(K) \le \sum_{k=1}^K [\mathrm{Err}_k(\mathrm{II.1}) + \mathrm{Err}_k(\mathrm{II.2}) + \mathrm{Err}_k(\mathrm{II.3})]$, we can obtain the result in Theorem 4.1 with probability at least $1 - 3\delta$ by the union bound.

Next, we show the proof of Theorem 4.2. By Lemma C.2, we can decompose the difference between $V_1^{\mu^k,\nu^k}(s_1)$ and

$V_1^{\mu^k,\nu^*}(s_1)$ into three terms

$$
\begin{aligned}
V_1^{\mu^k,\nu^k}(s_1) - V_1^{\mu^k,\nu^*}(s_1) \leq{}& \underbrace{\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} |q_h^{\mu^k,\mathcal{P}}(s) - d_h^k(s)|}_{\mathrm{Err}_k(\text{III.1})} \\
&+ \underbrace{\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^k(s) \big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}}}_{\mathrm{Err}_k(\text{III.2})} \\
&+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\mu^k,\mathcal{P},\nu^k}[\beta_h^{r,k}(s_h,a_h,b_h)\,|\,s_1]}_{\mathrm{Err}_k(\text{III.3})},
\end{aligned}
$$

with $W_h^k(s,b) = \langle \widetilde{r}_h^k(s,\cdot,b), \mu_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}}$. Due to the single-controller structure, distinct from the decomposition for Theorem 4.1, here we have that $\mathrm{Err}_k(\text{III.1})$ is the difference between the true state reaching probability and the empirical one, $\mathrm{Err}_k(\text{III.2})$ is the error from the policy mirror descent step, and $\mathrm{Err}_k(\text{III.3})$ is the expectation of reward bonus term. Technically, in the proof of this lemma, we can show $V_1^{\mu^k,\nu^k}(s_1) - V_1^{\mu^k,\nu^*}(s_1) = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} q_h^{\mu^k,\mathcal{P}}(s)[\mu_h^k(\cdot|s)]^\top r_h(s,\cdot,\cdot)(\nu_h^k - \nu_h^*)(\cdot|s)$, where the value function difference is only related to the reward function $r_h(s,\cdot,\cdot)$ instead of Q-function, which is the reason that only the reward bonus and reward-based mirror descent step appear in Algorithm 3.

As shown in Lemmas C.10, C.8, and C.11, we can obtain upper bounds that $\sum_{k=1}^{K}\mathrm{Err}_k(\text{III.1}) \leq \widetilde{\mathcal{O}}(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K})$, $\sum_{k=1}^{K}\mathrm{Err}_k(\text{III.2}) \leq \mathcal{O}(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|})$, and also $\sum_{k=1}^{K}\mathrm{Err}_k(\text{III.3}) \leq \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K})$ by taking summation of the three error terms from $k=1$ to K. The above inequalities hold with probability at least $1-2\delta$ by union bound. Therefore, letting $T = HK$, further by the relation that $\mathrm{Regret}_1(K) \leq \sum_{k=1}^{K}[\mathrm{Err}_k(\text{II.1}) + \mathrm{Err}_k(\text{II.2}) + \mathrm{Err}_k(\text{II.3})]$, we can obtain the result in Theorem 4.2. This completes the proof. □

6. Conclusion and Discussion

In this paper, we propose and analyze new fictitious play policy optimization algorithms for two-player zero-sum Markov games with structured but unknown transitions. We consider two classes of transition structures: factored independent transition and single-controller transition. For both scenarios, we prove $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bounds for each player after T steps in a two-agent competitive game scenario. When both players adopt the proposed algorithms, their overall optimality gap is $\widetilde{\mathcal{O}}(\sqrt{T})$.

Our proposed algorithms and the associated analysis can be potentially extended to different game settings, e.g., the extensions to the multi-player or general-sum game with the factored independent transition, and the extensions from the two-player single controller game to the multi-player game with a single controller. We leave them as our future work.

Acknowledgements

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

Altman, E., Avrachenkov, K., Marquez, R., and Miller, G. Zero-sum constrained stochastic games with independent state processes. *Mathematical Methods of Operations Research*, 62(3):375–386, 2005.

Altman, E., Avrachenkov, K., Bonneau, N., Debbah, M., El-Azouzi, R., and Menasche, D. S. Constrained cost-coupled stochastic games with independent state processes. *Operations Research Letters*, 36(2):160–164, 2008.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

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

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

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Bu, L., Babu, R., De Schutter, B., et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pp. 12203–12213, 2019.

Efroni, Y., Shani, L., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

Eldosouky, A., Saad, W., and Niyato, D. Single controller stochastic games for optimized moving target defense. In *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6. IEEE, 2016.

Filar, J. A. and Raghavan, T. A matrix game solution of the single-controller stochastic game. *Mathematics of Operations Research*, 9(3):356–362, 1984.

Flesch, J., Schoenmakers, G., and Vrieze, K. Stochastic games on a product state space. *Mathematics of Operations Research*, 33(2):403–420, 2008.

Foerster, J., Assael, I. A., De Freitas, N., and Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, pp. 2137–2145, 2016.

Fudenberg, D. and Levine, D. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 1995.

Guan, P., Raginsky, M., Willett, R., and Zois, D.-S. Regret minimization algorithms for single-controller zero-sum stochastic games. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7075–7080. IEEE, 2016.

Heinrich, J. and Silver, D. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

Heinrich, J., Lanctot, M., and Silver, D. Fictitious self-play in extensive-form games. In *International Conference on Machine Learning*, pp. 805–813, 2015.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Jin, T. and Luo, H. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Kakade, S. M. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

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

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

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

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

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1806.05618*, 2018.

Parthasarathy, T. and Raghavan, T. An orderfield property for stochastic games when one player controls transition probabilities. *Journal of Optimization Theory and Applications*, 33(3):375–392, 1981.

Perolat, J., Piot, B., and Pietquin, O. Actor-critic fictitious play in simultaneous move multistage games. 2018.

Perolat, J., Munos, R., Lespiau, J.-B., Omidshafiei, S., Rowland, M., Ortega, P., Burch, N., Anthony, T., Balduzzi, D., De Vylder, B., et al. From poincar\'e recurrence to convergence in imperfect information games: Finding equilibrium via regularization. *arXiv preprint arXiv:2002.08456*, 2020.

Robinson, J. An iterative method of solving a game. *Annals of mathematics*, pp. 296–301, 1951.

Rosenberg, D., Solan, E., and Vieille, N. Stochastic games with a single controller and incomplete information. *SIAM journal on control and optimization*, 43(1): 86–110, 2004.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Singh, V. V. and Hemachandra, N. A characterization of stationary nash equilibria of constrained stochastic games with independent state processes. *Operations Research Letters*, 42(1):48–52, 2014.

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

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Tian, Y., Wang, Y., Yu, T., and Sra, S. Provably efficient online agnostic learning in markov games. *arXiv preprint arXiv:2010.15020*, 2020.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 1, 2008.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pp. 4987–4997, 2017.

Wei, X., Yu, H., and Neely, M. J. Online primal-dual mirror descent under stochastic constraints. *arXiv preprint arXiv:1908.00305*, 2019.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019a.

Zhang, K., Yang, Z., and Basar, T. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pp. 11598–11610, 2019b.