Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence

Dongsheng Ding [*1] **Chen-Yu Wei** [*1] **Kaiqing Zhang** [*2] **Mihailo R. Jovanović** [1]

Abstract

We examine global non-asymptotic convergence properties of policy gradient methods for multi-agent reinforcement learning (RL) problems in Markov potential games (MPGs). To learn a Nash equilibrium of an MPG in which the size of state space and/or the number of players can be very large, we propose new independent policy gradient algorithms that are run by all players in tandem. When there is no uncertainty in the gradient evaluation, we show that our algorithm finds an ϵ-Nash equilibrium with $O(1/\epsilon^2)$ iteration complexity which does not explicitly depend on the state space size. When the exact gradient is not available, we establish $O(1/\epsilon^5)$ sample complexity bound in a potentially infinitely large state space for a sample-based algorithm that utilizes function approximation. Moreover, we identify a class of independent policy gradient algorithms that enjoy convergence for both zero-sum Markov games and Markov cooperative games with the players that are oblivious to the types of games being played. Finally, we provide computational experiments to corroborate the merits and the effectiveness of our theoretical developments.

1. Introduction

Multi-agent reinforcement learning (RL) studies how multiple players learn to maximize their long-term returns in a setup where players' actions influence the environment and other agents' returns (Busoniu et al., 2008; Zhang et al., 2021a). Recently, multi-agent RL has achieved significant success in various multi-agent learning scenarios, e.g., competitive game-playing (Silver et al., 2016; 2018; Vinyals et al., 2019), autonomous robotics (Shalev-Shwartz et al., 2016; Levine et al., 2016), and economic policy-making (Zheng et al., 2020; Trott et al., 2021). In the framework of stochastic games (Shapley, 1953; Fink, 1964), most results are established for *fully-competitive* (i.e., two-player zero-sum) games; e.g., see Daskalakis et al. (2020); Wei et al. (2021b); Cen et al. (2021). However, to achieve social welfare for AI (Dafoe et al., 2020; 2021; Stastny et al., 2021), it is imperative to establish theoretical guarantees for multi-agent RL in Markov games with cooperation.

Policy gradient methods (Williams, 1992; Sutton et al., 2000) have received significant attention for both single-agent (Bhandari & Russo, 2019; Agarwal et al., 2021) and multi-agent RL problems (Zhang et al., 2019; Daskalakis et al., 2020; Wei et al., 2021b). Independent policy gradient (Zhang et al., 2021a; Ozdaglar et al., 2021) is probably the most practical protocol in multi-agent RL, where each player behaves myopically by only observing her own rewards and actions (as well as the system states), while individually optimizing its own policy. More importantly, independent learning dynamics do not scale exponentially with the number of players in the game. Recently, Daskalakis et al. (2020); Leonardos et al. (2022); Zhang et al. (2021b) have in fact shown that multi-agent RL players could perform policy gradient updates independently, while enjoying global non-asymptotic convergence. However, these results are only focused on the basic tabular setting in which the value functions are represented by tables; they do not carry over to large-scale multi-agent RL problems in which the state space size is potentially infinite and the number of players is large. This motivates the following question:

Can we design independent policy gradient methods for large-scale Markov games, with non-asymptotic global convergence guarantees?

In this paper, we provide the first affirmative answer to this question for a class of mixed cooperative/competitive Markov games: Markov potential games (MPGs) (Macua et al., 2018; Leonardos et al., 2022; Zhang et al., 2021b). In particular, we make the following contributions:

- We propose an independent policy gradient algorithm

[*]Alphabetical order [1]University of Southern California [2]Massachusetts Institute of Technology. Correspondence to: Dongsheng Ding <dongshed@usc.edu>, Chen-Yu Wei <chenyu.wei@usc.edu>, Kaiqing Zhang <kaiqing@mit.edu>, Mihailo R. Jovanović <mihailo@usc.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

– Algorithm 1 – for learning an ϵ-Nash equilibrium of MPGs with $O(1/\epsilon^2)$ iteration complexity. In contrast to the existing results (Leonardos et al., 2022; Zhang et al., 2021b), such iteration complexity does not explicitly depend on the *state space size*.

- We consider a linear function approximation setting and design an independent sample-based policy gradient algorithm – Algorithm 2 – that learns an ϵ-Nash equilibrium with $O(1/\epsilon^5)$ sample complexity. This appears to be the first result for learning MPGs with function approximation.

- We establish the convergence of an independent optimistic policy gradient algorithm – Algorithm 3 that has been proved to converge in learning zero-sum Markov games (Wei et al., 2021b) – for learning a subclass of MPGs: Markov cooperative games. We show that the same type of optimistic policy learning algorithm provides an ϵ-Nash equilibrium in both zero-sum Markov games and Markov cooperative games while the players are oblivious to the types of games being played. To the best of our knowledge, this appears to be the first *game-agnostic* convergence result in Markov games.

We next discuss some related work.

Markov potential games (MPGs). In stochastic optimal control, the MPG model dates back to Dechert & O'Donnell (2006); González-Sánchez & Hernández-Lerma (2013). More recent studies include Zazo et al. (2016); Mazalov et al. (2017); Macua et al. (2018); Mguni et al. (2018) and all of these studies focus on systems with known dynamics. MPGs have also attracted attention in multi-agent RL. In the infinite-horizon setting, Leonardos et al. (2022); Zhang et al. (2021b) extended the policy gradient method (Agarwal et al., 2021; Kakade, 2001) for multiple players and established the iteration/sample complexity that scales with the size of state space; Fox et al. (2022) generalized the natural policy gradient method (Kakade, 2001; Agarwal et al., 2021) and established the global asymptotic convergence. In the finite-horizon setting, Song et al. (2022) built on the single-agent Nash-VI (Liu et al., 2021) to propose a sample efficient turn-based algorithm and Mao et al. (2022) studied the policy gradient method. Earlier, Wang & Sandholm (2002); Lowe et al. (2017) studied Markov cooperative games and Kleinberg et al. (2009); Palaiopanos et al. (2017); Cohen et al. (2017a) studied one-state MPGs; both of these are special cases of MPGs. We note that the term: Markov potential game has also been used to refer to state-based potential MDPs (Marden, 2012; Mguni et al., 2021), which are different from the MPGs that we study; see counterexamples in Leonardos et al. (2022).

Policy gradient methods for Markov games. Despite recent advances on the theory of policy gradient (Bhandari & Russo, 2019; Agarwal et al., 2021), the theory of policy gradient methods for multi-agent RL is relatively less studied. In the basic two-player zero-sum Markov games, Zhang et al. (2019); Bu et al. (2019); Daskalakis et al. (2020); Zhao et al. (2021) established global convergence guarantees for policy gradient methods for learning an (approximate) Nash equilibrium. More recently, Cen et al. (2021); Wei et al. (2021b) examined variants of policy gradient methods and provided last-iterate convergence guarantees. However, it is much harder for the policy gradient methods to work in general Markov games (Mazumdar et al., 2020; Hambly et al., 2021). The effectiveness of (natural) policy gradient methods for tabular MPGs was demonstrated in Leonardos et al. (2022); Zhang et al. (2021b); Fox et al. (2022); Zhang et al. (2022). Moreover, Xie & Zhong (2020); Wang et al. (2021a); Yu et al. (2021); Peng et al. (2021) reported impressive empirical performance of multi-agent policy gradient methods with function approximation in cooperative Markov games, but the theoretical foundation has not been provided.

Independent learning recently received attention in multi-agent RL (Daskalakis et al., 2020; Zhang et al., 2021a; Ozdaglar et al., 2021; Sayin et al., 2021; Jin et al., 2021a; Song et al., 2022; Kao et al., 2022), because it only requires local information for learning and naturally yields algorithms that scale to a large number of players. The algorithms in Leonardos et al. (2022); Zhang et al. (2021b); Fox et al. (2022); Zhang et al. (2022) can also be generally categorized as independent learning algorithms for MPGs.

Game-agnostic convergence. Being game-agnostic is a desirable property for independent learning in which players are oblivious to the types of games being played. In particular, classical fictitious-play warrants average-iterate convergence for several games (Robinson, 1951; Monderer & Shapley, 1996; Hofbauer & Sandholm, 2002). Although online learning algorithms, e.g., the one based on multiplicative weight updates (MWU) (Cesa-Bianchi & Lugosi, 2006), offer average-iterate convergence in zero-sum matrix games, they often do not provide last-iterate convergence guarantees (Bailey & Piliouras, 2018), which motivates recent studies (Daskalakis & Panageas, 2018; Mokhtari et al., 2020; Wei et al., 2020). Interestingly, while MWU converges in last-iterate for potential games (Palaiopanos et al., 2017; Cohen et al., 2017a), this is not the case for zero-sum matrix games (Cheung & Piliouras, 2020). Recently, Leonardos et al. (2021); Leonardos & Piliouras (2022) established last-iterate convergence of Q-learning dynamics for both zero-sum and potential/cooperative matrix games. However, it is open question whether an algorithm can have last-iterate convergence for both zero-sum and potential/cooperative Makov games.

2. Preliminaries

In this section, we introduce Markov potential games (MPGs), define the Nash equilibrium, and describe the problem setting.

We consider an N-player, infinite-horizon, discounted Markov potential game (Macua et al., 2018; Leonardos et al., 2022; Zhang et al., 2021b),

$$\text{MPG}\,(\,\mathcal{S},\,\{\mathcal{A}_i\}_{i=1}^N,\,\mathbb{P},\,\{r_i\}_{i=1}^N,\,\gamma,\,\rho\,) \tag{1}$$

where $\mathcal{S}$ is the state space, $\mathcal{A}_i$ is the action space for the ith player, with the joint action space of $N \geq 2$ players denoted as $\mathcal{A} := \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$, $\mathbb{P}$ is the transition probability measure specified by a distribution $\mathbb{P}(\,\cdot \,|\, s, a)$ over $\mathcal{S}$ if N players jointly take an action a from $\mathcal{A}$ in state s, r_i: $\mathcal{S} \times \mathcal{A} \to [0,1]$ is the ith player immediate reward function, $\gamma \in [0,1)$ is the discount factor, and ρ is the initial state distribution over $\mathcal{S}$. We assume that all action spaces are finite with the same size $A = A_i = |\mathcal{A}_i|$ for all $i = 1, \ldots, N$. It is straightforward to apply our analysis to the general case in which players' finite action spaces have different sizes.

For the ith player, $\Delta(\mathcal{A}_i)$ represents the probability simplex over the action set $\mathcal{A}_i$. A stochastic policy for player i is given by π_i: $\mathcal{S} \to \Delta(\mathcal{A}_i)$ that specifies the action distribution $\pi_i(\cdot \,|\, s) \in \Delta(\mathcal{A}_i)$ for each state $s \in \mathcal{S}$. The set of stochastic policies for player i is denoted by $\Pi_i := (\Delta(\mathcal{A}_i))^{\mathcal{S}}$, the joint probability simplex is given by $\Delta(\mathcal{A}) := \Delta(\mathcal{A}_1) \times \ldots \times \Delta(\mathcal{A}_N)$, and the joint policy space is $\Pi := (\Delta(\mathcal{A}))^{\mathcal{S}}$. A Markov product policy $\pi := \{\pi_i\}_{i=1}^N \in \Pi$ for N players consists of the policy $\pi_i \in \Pi$ for all players $i = 1, \ldots, N$. We use the shorthand $\pi_{-i} = \{\pi_k\}_{k=1,\,k\neq i}^N$ to represent the policy of all but the ith player. We denote by $V_i^\pi : \mathcal{S} \to \mathbb{R}$ the ith player value function under the joint policy π, starting from an initial state $s^{(0)} = s$:

$$V_i^\pi(s) \;:=\; \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \,\middle|\, s^{(0)} = s \right]$$

where the expectation $\mathbb{E}^\pi$ is over $a^{(t)} \sim \pi(\cdot \,|\, s^{(t)})$ and $s^{(t+1)} \sim \mathbb{P}(\cdot \,|\, s^{(t)}, a^{(t)})$. Finally, $V_i^\pi(\mu)$ denotes the expected value function of $V_i^\pi(s)$ over a state distribution μ, $V_i^\pi(\mu) := \mathbb{E}_{s\sim\mu}[V_i^\pi(s)]$.

In a MPG, at any state $s \in \mathcal{S}$, there exists a global function – the potential function $\Phi^\pi(s)$: $\Pi \times \mathcal{S} \to \mathbb{R}$ – that captures the incentive of all players to vary their policies at state s,

$$V_i^{\pi_i,\,\pi_{-i}}(s) \;-\; V_i^{\pi_i',\,\pi_{-i}}(s) \;=\; \Phi^{\pi_i,\,\pi_{-i}}(s) \;-\; \Phi^{\pi_i',\,\pi_{-i}}(s)$$

for any policies $\pi_i, \pi_i' \in \Pi_i$ and $\pi_{-i} \in \Pi_{-i}$. Let $\Phi^\pi(\mu) := \mathbb{E}_{s\sim\mu}[\Phi^\pi(s)]$ be the expected potential function over a state distribution μ. Thus, $V_i^{\pi_i,\,\pi_{-i}}(\mu) - V_i^{\pi_i',\,\pi_{-i}}(\mu) = \Phi^{\pi_i,\,\pi_{-i}}(\mu) - \Phi^{\pi_i',\,\pi_{-i}}(\mu)$. There always exists a constant $C_\Phi > 0$ such that $|\Phi^\pi(\mu) - \Phi^{\pi'}(\mu)| \leq C_\Phi$ for any π, π', μ; see a trivial upper bound in Lemma 18 in Appendix E. An important subclass of Markov potential games is given by Markov cooperative games (MCG) in which all players share the same reward function $r = r_i$ for all $i = 1, \ldots, N$.

We also denote by $Q_i^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the action-value function under policy π, starting from an initial state-action pair $(s^{(0)}, a^{(0)}) = (s, a)$:

$$Q_i^\pi(s,a) := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \,\middle|\, s^{(0)} = s, a^{(0)} = a \right].$$

The value function can be equivalently expressed as $V_i^\pi(s) = \sum_{a'\,\in\,\mathcal{A}} \pi(a' \,|\, s) Q_i^\pi(s, a')$. For each player i, by averaging out π_{-i}, we can define the averaged action-value function $\bar{Q}_i^{\pi_i,\,\pi_{-i}}$: $\mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$,

$$\bar{Q}_i^{\pi_i,\,\pi_{-i}}(s, a_i) := \sum_{a_{-i}\,\in\,\mathcal{A}_{-i}} \pi_{-i}(a_{-i} \,|\, s)\, Q_i^{\pi_i,\,\pi_{-i}}(s, a_i, a_{-i})$$

where $\mathcal{A}_{-i}$ is the set of actions of all but the ith player. We use the shorthand $\bar{Q}_i^\pi$ for $\bar{Q}_i^{\pi_i,\,\pi_{-i}}$ when π_i and π_{-i} are from the same joint policy π. It is straightforward to see that V_i^π, Q_i^π, and $\bar{Q}_i^\pi$ are bounded between 0 and $1/(1-\gamma)$.

We recall the notion of (Markov perfect stationary) Nash equilibrium (Fink, 1964). A joint policy $\pi^\star$ is called a Nash equilibrium if for each player $i = 1, \ldots, N$,

$$V_i^{\pi_i^\star,\,\pi_{-i}^\star}(s) \;\geq\; V_i^{\pi_i,\,\pi_{-i}^\star}(s), \quad \text{for all } \pi_i \in \Pi_i,\; s \in \mathcal{S},$$

and called an ϵ-Nash equilibrium if for $i = 1, \ldots, N$,

$$V_i^{\pi_i^\star,\,\pi_{-i}^\star}(s) \;\geq\; V_i^{\pi_i,\,\pi_{-i}^\star}(s) - \epsilon, \quad \text{for all } \pi_i \in \Pi_i,\; s \in \mathcal{S}.$$

Nash equilibria for MPGs with finite states and actions always exist (Fink, 1964). When the state space is infinite, we assume the existence of a Nash equilibrium; see Takahashi (1962); Maitra & Parthasarathy (1970; 1971); Altman et al. (1997) for cases with countable or compact state spaces.

Given policy π and initial state $s^{(0)}$, we define the discounted state visitation distribution,

$$d_{s^{(0)}}^\pi(s) \;=\; (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \,\text{Pr}^\pi(\,s^{(t)} = s \,|\, s^{(0)}\,).$$

For a state distribution μ, define $d_\mu^\pi(s) = \mathbb{E}_{s^{(0)}\sim\mu}[\,d_{s^{(0)}}^\pi(s)\,]$. By definition, $d_\mu^\pi(s) \geq (1-\gamma)\mu(s)$ for any μ and s.

It is useful to introduce a variant of the performance difference lemma (Agarwal et al., 2021) for multiple players; for other versions, see Zhang et al. (2019); Daskalakis et al. (2020); Zhang et al. (2021b); Leonardos et al. (2022).

Lemma 1 (Performance difference)**.** *For the ith player, if we fix the policy π_{-i} and any state distribution μ, then for*

any two policies $\widehat{\pi}_i$ and $\bar{\pi}_i$,

$$V_i^{\widehat{\pi}_i,\pi_{-i}}(\mu) - V_i^{\bar{\pi}_i,\pi_{-i}}(\mu) = \frac{1}{1-\gamma}\sum_{s,a_i} d_\mu^{\widehat{\pi}_i,\pi_{-i}}(s)\cdot(\widehat{\pi}_i-\bar{\pi}_i)(a_i\,|\,s)\bar{Q}_i^{\bar{\pi}_i,\pi_{-i}}(s,a_i)$$

where $\bar{Q}_i^{\bar{\pi}_i,\pi_{-i}}(s,a_i)=\sum_{a_{-i}}\pi_{-i}(a_{-i}|s)Q_i^{\bar{\pi}_i,\pi_{-i}}(s,a_i,a_{-i})$.

It is common to use the distribution mismatch coefficient to measure the exploration difficulty in policy optimization (Agarwal et al., 2021). We next define a distribution mismatch coefficient for MPGs (Leonardos et al., 2022) in Definition 1, and its minimax variant in Definition 2.

Definition 1 (Distribution mismatch coefficient)**.** *For any distribution $\mu\in\Delta(\mathcal{S})$ and policy $\pi\in\Pi$, the distribution mismatch coefficient κ_μ is the maximum distribution mismatch of π relative to μ, $\kappa_\mu := \sup_{\pi\in\Pi}\left\|d_\mu^\pi/\mu\right\|_\infty$, where the division d_μ^π/μ is evaluated in a componentwise manner.*

Definition 2 (Minimax distribution mismatch coefficient)**.** *For any distribution $\mu\in\Delta(\mathcal{S})$, the minimax distribution mismatch coefficient $\widetilde{\kappa}_\mu$ is the minimax value of the distribution mismatch of π relative to ν, $\widetilde{\kappa}_\mu := \inf_{\nu\in\Delta(\mathcal{S})}\sup_{\pi\in\Pi}\left\|d_\mu^\pi/\nu\right\|_\infty$, where the division d_μ^π/ν is evaluated in a componentwise manner.*

Other notation. We denote by $\|\cdot\|$ the ℓ_2-norm of a vector or the spectral norm of a matrix. The inner product of a function $f\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ with $p\in\Delta(\mathcal{A})$ at fixed $s\in\mathcal{S}$ is given by $\langle f(s,\cdot),p(\cdot)\rangle_{\mathcal{A}} := \sum_{a\in\mathcal{A}} f(s,a)p(a)$. The ℓ_2-norm projection operator onto a convex set Ω is defined as $\mathcal{P}_\Omega(x) := \operatorname{argmin}_{x'\in\Omega}\|x'-x\|$. For functions f and g, we write $f(n)=O(g(n))$ if there exists $N<\infty$ and $C<\infty$ such that $f(n)\le Cg(n)$ for $n\ge N$, and write $f(n)=\widetilde{O}(g(n))$ if $\log g(n)$ appears in $O(\cdot)$. We use "$\lesssim$" and "$\gtrsim$" to denote "$\le$" and "$\ge$" up to a constant.

3. Independent Learning Setting

We examine an independent learning setting (Zhang et al., 2021a; Daskalakis et al., 2020; Ozdaglar et al., 2021) for Markov potential games in which all players repeatedly execute their own policy and update rules individually. At each time t, all players propose their own polices $\pi_i^{(t)}\colon \mathcal{S}\to\Delta(\mathcal{A}_i)$ with the player index $i=1,\dots,N$, while a game oracle can either evaluate each player's policy or generate a set of sample trajectories for each player. In repeating such protocol for T times, each player behaves myopically in optimizing its own policy, but does not observe actions or policies from any other player.

To evaluate the learning performance, we introduce a notion of regret,

$$\text{Nash-Regret}(T) := \frac{1}{T}\sum_{t=1}^{T}\max_i\left(\max_{\pi_i'} V_i^{\pi_i',\pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho)\right)$$

which averages the worst player's local gaps in T iterations: $\max_{\pi_i'} V_i^{\pi_i',\pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho)$ for $t=1,\dots,T$, where $\max_{\pi_i'} V_i^{\pi_i',\pi_{-i}^{(t)}}(\rho)$ is the ith player best response given $\pi_{-i}^{(t)}$. In Nash-Regret(T), we compare the learnt joint policy $\pi^{(t)}$ with the best policy that the ith player can take by fixing $\pi_{-i}^{(t)}$. We notice that Nash-Regret is closely related to the notion of dynamic regret (Zinkevich, 2003) in which the regret comparator changes over time. This is a suitable notion because the environment is non-stationary from the perspective of an independent learner (Matignon et al., 2012; Zhang et al., 2021a).

To obtain an ϵ-Nash equilibrium $\pi^{(t^\star)}$ with a tolerance $\epsilon>0$, our goal is to show the following average performance,

$$\text{Nash-Regret}(T) = \epsilon.$$

The existence of such $t^\star$ is straightforward,

$$t^\star := \operatorname*{argmin}_{1\le t\le T}\max_i\left(\max_{\pi_i'} V_i^{\pi_i',\pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho)\right).$$

Since each summand above is non-negative, $V_i^{\pi^{(t^\star)}}(\rho)\ge V_i^{\pi_i',\pi_{-i}^{(t^\star)}}(\rho)-\epsilon$ for any π_i' and $i=1,\dots,N$, which implies that $\pi^{(t^\star)}$ is an ϵ-Nash equilibrium.

For an independent learning setting without uncertainty in gradient evaluation, we introduce a policy gradient method for Markov potential/cooperative games in Section 4. In Section 5, we utilize a sample-based approach with function approximation to address the scenario in which true gradient is not available and, in Section 6, we provide the game-agnostic convergence analysis.

4. Independent Policy Gradient Methods

In this section, we assume that we have access to exact gradient and examine a gradient-based method for learning a Nash equilibrium in Markov potential/cooperative games.

4.1. Policy gradient for Markov potential games

A natural independent learning scheme for MPGs is to let every player independently perform policy gradient ascent (Leonardos et al., 2022; Zhang et al., 2021b). In this approach, the ith player updates its policy according the gradient of the value function with respect to the policy parameters,

$$\pi_i^{(t+1)}(\cdot\,|\,s) \leftarrow \mathcal{P}_{\Delta(\mathcal{A}_i)}\left(\pi_i^{(t)}(\cdot\,|\,s) + \eta\frac{\partial V_i^\pi(\rho)}{\partial\pi_i(a_i\,|\,s)}\bigg|_{\pi=\pi^{(t)}}\right)$$

$$\frac{\partial V_i^\pi(\rho)}{\partial\pi_i(a_i\,|\,s)} = \frac{1}{1-\gamma}d_\rho^\pi(s)\bar{Q}_i^\pi(s,a_i) \tag{3}$$

where the calculation for the gradient in (3) can be found in Agarwal et al. (2021); Leonardos et al. (2022); Zhang et al.

Algorithm 1 Independent policy gradient ascent

1: **Parameters:** $\eta > 0$.
Initialization: Let $\pi_i^{(1)}(a_i \,|\, s) = 1/A$ for $s \in \mathcal{S}$, $a_i \in \mathcal{A}_i$ and $i = 1, \ldots, N$.
2: **for** step $t = 1, \ldots, T$ **do**
3: **for** player $i = 1, \ldots, N$ (in parallel) **do**
4: Define player i's policy on $s \in \mathcal{S}$,

$$\pi_i^{(t+1)}(\cdot \,|\, s) \;:=\; \underset{\pi_i(\cdot \,|\, s) \,\in\, \Delta(\mathcal{A}_i)}{\operatorname{argmax}} \left\{ \big\langle \pi_i(\cdot \,|\, s),\, \bar{Q}_i^{(t)}(s, \cdot) \big\rangle_{\mathcal{A}_i} \;-\; \frac{1}{2\eta} \big\| \pi_i(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \big\|^2 \right\}. \tag{2}$$

where $\bar{Q}_i^{(t)}(s,\, a_i)$ is a shorthand for $\bar{Q}_i^{\pi_i^{(t)},\, \pi_{-i}^{(t)}}(s, a_i)$ (defined in Section 2).
5: **end for**
6: **end for**

(2021b).

Update rule (3) may suffer from a slow learning rate for some states. Since the gradient with respect to $\pi_i(a_i \,|\, s)$ scales with $d_\rho^\pi(s)$ – which may be small if the current policy π has small visitation frequency to s – the corresponding states may experience slow learning progress. To address this issue, we propose the following update rule (equivalent to (2) in Algorithm 1):

$$\pi_i^{(t+1)}(\cdot \,|\, s) \;\leftarrow\; \mathcal{P}_{\Delta(\mathcal{A}_i)} \Big(\pi_i^{(t)}(\cdot \,|\, s) \;+\; \eta \bar{Q}_i^{\pi^{(t)}}(s, \cdot) \Big) \tag{4}$$

which essentially removes the $d_\rho^\pi(s)/(1-\gamma)$ factor in standard policy gradient (3) and alleviates the slow-learning issue. Interestingly, update rule (4) for the single-player MDP has also been studied in Xiao (2022), concurrently. However, since the optimal value is not unique, the analysis of Xiao (2022) does not apply to our multi-player case for which many Nash policies exist and the set that contains them is non-convex (Leonardos et al., 2022). We also note that regularized variants of (4) for the single-player MDP appeared in Lan (2022); Zhan et al. (2021).

Furthermore, in contrast to (3), our update rule (4) is invariant to the initial state distribution ρ. This allows us to establish performance guarantees simultaneously for all ρ in a similar way as typically done for natural policy gradient (NPG) and other policy mirror descent algorithms for single-player MDPs (Agarwal et al., 2021; Lan, 2022; Zhan et al., 2021).

Theorem 1 establishes performance guarantees for Algorithm 1; see Appendix B.1 for proof.

Theorem 1 (Nash-Regret bound for Markov potential games)**.** *For MPG (1) with an initial state distribution ρ, if all players independently perform the policy update in Algorithm 1 then, for two different choices of stepsize η, we have*

$$\text{Nash-Regret}(T) \;\lesssim\; \mathcal{R}(\eta)$$

$$\mathcal{R}(\eta) = \begin{cases} \dfrac{\sqrt{\widetilde{\kappa}_\rho\, AN}\,(C_\Phi)^{\frac{1}{4}}}{(1-\gamma)^{\frac{9}{4}}\, T^{\frac{1}{4}}}, & \eta = \dfrac{(1-\gamma)^{\frac{5}{2}}\sqrt{C_\Phi}}{NA\sqrt{T}} \\[2ex] \dfrac{\min(\kappa_\rho, S)^2 \sqrt{ANC_\Phi}}{(1-\gamma)^3\sqrt{T}}, & \eta = \dfrac{(1-\gamma)^4}{8\min(\kappa_\rho, S)^3 NA}. \end{cases}$$

Depending on the stepsize η, Theorem 1 provides two rates for the average Nash regret: $T^{-1/4}$ and $T^{-1/2}$. The technicalities behind these choices will be explained later and, to obtain an ϵ-Nash equilibrium, our two bounds suggest respective iteration complexities,

$$\frac{\widetilde{\kappa}_\rho^2\, A^2 N^2\, C_\Phi}{(1-\gamma)^9\, \epsilon^4} \quad \text{and} \quad \frac{\min(\kappa_\rho, S)^4\, AN\, C_\Phi}{(1-\gamma)^6\, \epsilon^2}.$$

Compared with the iteration complexity guarantees in Leonardos et al. (2022); Zhang et al. (2021b), our bounds in Theorem 1 improve the dependence on the distribution mismatch coefficient κ_ρ and the state space size S. Since our minimax distribution mismatch coefficient $\widetilde{\kappa}_\rho$ satisfies

$$\widetilde{\kappa}_\rho \;\le\; \min(\kappa_\rho, S) \;\le\; \kappa_\rho$$

our $\widetilde{\kappa}_\rho$-dependence or $\min(\kappa_\rho, S)$-dependence are less restrictive than the explicit S-dependence in Leonardos et al. (2022); Zhang et al. (2021b). Importantly, this permits our bounds to work for systems with large number of states, and makes Algorithm 1 suitable for sample-based scenario with function approximation (see Section 5). With polynomial dependence on the number of players N instead of exponential, Algorithm 1 overcomes the *curse of multiagents* (Jin et al., 2021a; Song et al., 2022). In terms of problem parameters (γ, A, N, C_Φ), our iteration complexity either improves or becomes slightly worse.

Remark 1 (Infinite state space)**.** *When the state space is infinite, explicit S-dependence disappears in our iteration complexities. Implicit S-dependence only exists in the distribution mismatch coefficient κ_ρ or $\widetilde{\kappa}_\rho$. However, it is easy to bound κ_ρ by devising an initial state distribution without*

introducing constraints on the MDP dynamics. For instance, in MPGs with agent-independent transitions (in which every state is a potential game and transitions do not depend on actions (Leonardos et al., 2022)), if we select ρ to be the stationary state distribution d_ρ^π then $\kappa_\rho = 1$ regardless of the state-space size S.

Remark 2 (Our key techniques)**.** *A key step of the analysis is to quantify the policy improvement regarding the potential function Φ in each iteration. Similar to the standard descent lemma in optimization (Arora, 2008), applying the projected policy gradient algorithm to a smooth Φ yields the following ascent property (cf. Eq. (9) in Leonardos et al. (2022) and Lemmas 11 and 12 in Zhang et al. (2021b)),*

$$\Phi^{\pi^{(t+1)}}(\mu) - \Phi^{\pi^{(t)}}(\mu) \gtrsim \frac{1}{\beta} \sum_{i=1}^{N} \sum_{s} \left\| \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \right\|^2$$

where $\beta > 0$ is related to the smoothness constant (or the second-order derivative) of the potential function. However, since the search direction in our policy update is not the standard search direction utilized in policy gradient, this ascent analysis does not apply to our algorithm.

To obtain such improvement bound, it is crucial to analyze the joint policy improvement*. Let us consider two players i and j: player i changes its policy from π_i to π_i' to maximize its own reward based on the* current *policy profile (π_i, π_j) and player j changes its policy from π_j to π_j' in its own interest. What is the overall progress after they independently change their policies from (π_i, π_j) to (π_i', π_j')? One method of capturing the joint policy improvement exploits the smoothness of the potential function, which is useful in the standard policy gradient ascent method (Leonardos et al., 2022; Zhang et al., 2021b). In our analysis, we connect the joint policy improvement with the individual policy improvement via the performance difference lemma. In particular, as shown in Lemma 3, Lemma 2 and Lemma 21 provide an effective means for analyzing the joint policy improvement. The proposed approach could be of independent interests for analyzing other Markov games.*

In Lemma 3, we obtain two different joint policy improvement bounds by dealing with the cross terms in two different ways (see the proofs for details). Hence, we establish two different Nash-Regret bounds in Theorem 1: one has better dependence on T while the other has better dependence on κ_ρ. Even though, it is an open issue how to achieve the best of the two, we next show that this is indeed possible for a special case: Markov cooperative games.

4.2. Faster rates for Markov cooperative games

When all players use the same reward function, i.e., $r = r_i$ for all $i = 1, \ldots, N$, MPG (1) reduces to a Markov cooperative game. In this case, $V_i^\pi = V^\pi$ and $Q_i^\pi = Q^\pi$ for all $i = 1, \ldots, N$ and Algorithm 1 works immediately. Thus, we continue to use Nash-Regret$_i(T)$ that is defined through $V_i^{\pi_i', \pi_{-i}^{(t)}} = V^{\pi_i', \pi_{-i}^{(t)}}$ and $V_i^{\pi^{(t)}} = V^{\pi^{(t)}}$.

Theorem 2 provides a Nash-Regret bound for Markov cooperative games; see Appendix B.2 for proof.

Theorem 2 (Nash-Regret bound for Markov cooperative games)**.** *For MPG (1) with identical rewards and an initial state distribution ρ, if all players independently perform the policy update in Algorithm 1 with stepsize $\eta = (1-\gamma)/(2NA)$ then,*

$$\text{Nash-Regret}(T) \lesssim \frac{\sqrt{\widetilde{\kappa}_\rho \, AN}}{(1-\gamma)^2 \sqrt{T}}.$$

For Markov cooperative games, Theorem 2 achieves the best of the two bounds in Theorem 1 and an ϵ-Nash equilibrium is achieved with the following iteration complexity,

$$\frac{\widetilde{\kappa}_\rho \, AN}{(1-\gamma)^4 \, \epsilon^2}.$$

This iteration complexity improves the ones provided in Leonardos et al. (2022); Zhang et al. (2021b) in several aspects. In particular, we have introduced the minimax distribution mismatch coefficient $\widetilde{\kappa}_\rho$, which is upper bounded by κ_ρ. When we take this upper bound, our bound improves the κ_ρ-dependence in Leonardos et al. (2022); Zhang et al. (2021b) from κ_ρ^2 to κ_ρ. We note that if we view the Markov cooperative game as an MPG, then the value function V^π serves as a potential function Φ which is bounded between 0 and $1/(1-\gamma)$. Thus, our $(1-\gamma)$-dependence matches the one in Zhang et al. (2021b) and improves the one in Leonardos et al. (2022) by $(1-\gamma)^2$.

5. Independent Policy Gradient with Function Approximation

We next remove the exact gradient requirement and apply Algorithm 1 to the linear function approximation setting. In what follows, we assume that the averaged action value function is linear in a given feature map.

Assumption 1 (Linear averaged Q)**.** *In MPG (1), for each player i, there is a feature map $\phi_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}^d$, such that for any $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ and any policy $\pi \in \Pi$,*

$$\bar{Q}_i^\pi(s, a_i) = \langle \phi_i(s, a_i), w_i^\pi \rangle, \text{ for some } w_i^\pi \in \mathbb{R}^d.$$

Moreover, $\|\phi_i\| \le 1$ for all s, a_i, and $\|w_i^\pi\| \le W$ for all π.

Without loss of generality, we can assume $W \le \sqrt{d}/(1-\gamma)$; see Lemma 8 in Wei et al. (2021a). Assumption 1 is a multi-agent generalization of the standard linear Q assumption (Abbasi-Yadkori et al., 2019) for single-player MDPs. It is more general than the multi-agent linear MDP assumption (Xie et al., 2020; Dubey & Pentland, 2021) in which

both transition and reward functions are linear in given feature maps. A special case of Assumption 1 is the tabular case in which the sizes of state/action spaces are finite, and where we can select ϕ_i to be an indicator function. Since the feature map ϕ_i is locally-defined coordination between players is avoided (Zhao et al., 2021).

Remark 3 (Function approximation). *Since RL with function approximation is statistically hard in general, e.g., see Weisz et al. (2021); Wang et al. (2021b) for hardness results, assuming regularity of underlying MDPs is necessary for the application of function approximation to multi-agent RL in which either the value function (Xie et al., 2020; Dubey & Pentland, 2021; Jin et al., 2021b; Huang et al., 2022) or the policy (Zhao et al., 2021) is approximated. Because of restrictive function approximation power, the main challenge is the entanglement of policy improvement (or optimization) and policy evaluation (or approximation) errors. In Theorem 3 and Theorem 4, we show that optimization and approximation errors are decoupled under Assumption 1 so that we can control them, separately. Our analysis can be generalized to some neural networks, e.g., overparametrized neural networks (Liu et al., 2019), a rich function class that allows splitting optimization and approximation errors, which we leave for future work.*

We formally present our algorithm in Algorithm 2 (see it in Appendix A). At each step t, there are two phases. In Phase 1, the players begin with the initial state $\bar{s}^{(0)} \sim \rho$ and simultaneously execute their current policies $\{\pi_i^{(t)}\}_{i=1}^N$ to interact with the environment for K rounds. In each round k, we terminate the interaction at step $H = \max_i(h_i + h_i')$, where h_i and h_i' are sampled from a geometric distribution $\text{GEOMETRIC}(1-\gamma)$, independently; the state at h_i naturally follows $\bar{s}^{(h_i)} \sim d_\rho^{\pi^{(t)}}$. By collecting rewards from step h_i to $h_i + h_i' - 1$, as shown in (6), we can justify $\mathbb{E}[R_i^{(k)}] = \bar{Q}_i^{(t)}(\bar{s}^{(h_i)}, \bar{a}_i^{(h_i)})$ where $\bar{Q}_i^{(t)}(\cdot,\cdot) := \bar{Q}_i^{\pi^{(t)}}(\cdot,\cdot)$ and $\bar{a}_i^{(h_i)} \sim \pi_i^{(t)}(\cdot \mid \bar{s}^{(h_i)})$, in Appendix C.1. In the end of round k, we collect a sample tuple: $(s_i^{(k)}, a_i^{(k)}, R_i^{(k)})$ in (6) for each player i.

After each player collects K samples, in Phase 2, they use these samples to estimate $\bar{Q}_i^{(t)}(\cdot,\cdot)$, which is required for policy updates. By Assumption 1,

$$\bar{Q}_i^{(t)}(s,a_i) \;=\; \langle \phi_i(s,a_i), w_i^{(t)} \rangle,\; \forall (s,a_i) \in \mathcal{S} \times \mathcal{A}_i$$

where $w_i^{(t)}$ represents $w_i^{\pi^{(t)}}$. Our goal is to obtain a solution $\widehat{w}_i^{(t)} \approx w_i^{(t)}$ using samples, and estimate $\bar{Q}_i^{(t)}(s,a_i)$ via

$$\widehat{Q}_i^{(t)}(s,a_i) \;:=\; \langle \phi_i(s,a_i), \widehat{w}_i^{(t)} \rangle,\; \forall (s,a_i) \in \mathcal{S} \times \mathcal{A}_i. \quad (5)$$

To obtain $\widehat{w}_i^{(t)}$, the standard approach is to solve linear regression (7) since $\mathbb{E}[R_i^{(k)}] = \bar{Q}_i^{(t)}(s_i^{(k)}, a_i^{(k)}) = \langle \phi_i(s_i^{(k)}, a_i^{(k)}), w_i^{(t)} \rangle$. We measure the estimation quality of $\widehat{w}_i^{(t)}$ via the expected regression loss,

$$L_i^{(t)}(w_i) = \mathbb{E}_{(s,a_i)\sim\nu_i^{(t)}} \left[\left(\bar{Q}_i^{(t)}(s,a_i) - \langle \phi_i(s,a_i), w_i \rangle \right)^2 \right]$$

where $\nu_i^{(t)}(s,a_i) := d_\rho^{(t)}(s) \circ \pi_i^{(t)}(a_i \mid s)$ and $L_i^{(t)}(w_i^{(t)}) = 0$ by Assumption 1. We make the following assumption for the expected regression loss of $\widehat{w}_i^{(t)}$.

Assumption 2 (Bounded statistical error). *Fix a state distribution ρ. For any sequence of iterates $\widehat{w}_i^{(1)}, \ldots, \widehat{w}_i^{(T)}$ for $i = 1, \ldots, N$ that are generated by Algorithm 2, there exists an $\epsilon_{\text{stat}} < \infty$ such that*

$$\mathbb{E}\big[\, L_i^{(t)}(\, \widehat{w}_i^{(t)}\,)\,\big] \;\le\; \epsilon_{\text{stat}}$$

for all i and t, where the expectation is on randomness in generating $\widehat{w}_i^{(t)}$.

The bound for ϵ_{stat} can be established using standard linear regression analysis (Audibert & Catoni, 2009) and it is given by $\epsilon_{\text{stat}} = O\big(\frac{dW^2}{K(1-\gamma)^2}\big)$. This bound can be achieved by applying the stochastic projected gradient descent method (Hsu et al., 2012; Cohen et al., 2017b) to the regression problem.

After obtaining $\widehat{Q}_i^{(t)}(\cdot,\cdot)$, we update the polices in (8) which is different from the update in Algorithm 1 in two aspects: (i) the gradient direction $\widehat{Q}_i^{(t)}(\cdot,\cdot)$ is the estimated version of $\bar{Q}_i^{(t)}(\cdot,\cdot)$; and (ii) the Euclidean projection set becomes $\Delta_\xi(\mathcal{A}_i) := \{\,(1-\xi)\,\pi_i(\cdot \mid s) + \xi\,\text{Unif}_{\mathcal{A}_i}, \forall \pi_i(\cdot \mid s)\,\}$ that introduces ξ-greedy policies for exploration (Leonardos et al., 2022; Zhang et al., 2021b), where $\xi \in (0,1)$.

Theorem 3 establishes performance guarantees for Algorithm 2; see Appendix C.2 for proof.

Theorem 3 (Nash-Regret bound for Markov potential games with function approximation). *Let Assumption 1 hold for MPG (1) with an initial state distribution ρ. If all players independently run Algorithm 2 with $\xi = \min\Big(\Big(\frac{\kappa_\rho^2 NA\epsilon_{\text{stat}}}{(1-\gamma)^2 W^2}\Big)^{\frac{1}{3}}, \frac{1}{2}\Big)$ and Assumption 2 holds, then*

$$\mathbb{E}\,[\,\text{Nash-Regret}(T)\,] \;\lesssim\; \mathcal{R}(\eta) \;+\; \left(\frac{\kappa_\rho^2\, WAN\epsilon_{\text{stat}}}{(1-\gamma)^5}\right)^{\frac{1}{3}}$$

$$\mathcal{R}(\eta) \;=\; \begin{cases} \dfrac{\sqrt{\kappa_\rho\, WN}\,(AC_\Phi)^{\frac{1}{4}}}{(1-\gamma)^{\frac{7}{4}}\, T^{\frac{1}{4}}}, & \eta \;=\; \dfrac{(1-\gamma)^{\frac{3}{2}}\sqrt{C_\Phi}}{WN\sqrt{AT}} \\ \dfrac{\kappa_\rho^2\,\sqrt{AN\,C_\Phi}}{(1-\gamma)^3\,\sqrt{T}}, & \eta \;=\; \dfrac{(1-\gamma)^4}{16\,\kappa_\rho^3\, NA}. \end{cases}$$

Theorem 3 shows the additive effect of the function approximation error ϵ_{stat} on the Nash regret of Algorithm 2. When

$\epsilon_{\text{stat}} = 0$, Theorem 3 matches the rates in Theorem 1 in the exact gradient case. As in Algorithm 1, even though update rule (8) iterates over all $s \in \mathcal{S}$, we do not need to assume a finite state space $\mathcal{S}$. In fact, (8) only "defines" a function $\pi_i^{(t)}(\cdot \mid s)$ instead of "calculating" it. This is commonly used in policy optimization with function approximation, e.g., Cai et al. (2020); Luo et al. (2021). To execute this algorithm, $\pi_i^{(t)}(\cdot \mid s)$ only needs to be evaluated if necessary, e.g., when the state s is visited in Phase 1 of Algorithm 2.

When we apply stochastic projected gradient updates to (7), Algorithm 2 becomes a sample-based algorithm and existing stochastic projected gradient results directly apply. Depending on the stepsize choice, an ϵ-Nash equilibrium is achieved with sample complexities (see Corollary 1 in Appendix C.4),

$$TK \;=\; O\left(\frac{1}{\epsilon^7}\right) \;\text{ and }\; O\left(\frac{1}{\epsilon^5}\right), \;\text{ respectively.}$$

Compared with the sample complexity guarantees for the tabular MPG case (Leonardos et al., 2022; Zhang et al., 2021b), our sample complexity guarantees hold for MPGs with potentially infinitely large state spaces. When we specialize Assumption 1 to the tabular case, our second sample complexity improves the sample complexity in Leonardos et al. (2022); Zhang et al. (2021b) from $O(1/\epsilon^6)$ to $O(1/\epsilon^5)$.

As before, we get improved performance guarantees when we apply Algorithm 2 to Markov cooperative games.

Theorem 4 (Nash-Regret bound for Markov cooperative games with function approximation)**.** *Let Assumption 1 hold for MPG (1) with identical rewards and an initial state distribution $\rho > 0$. If all players independently perform the policy update in Algorithm 2 with stepsize $\eta = (1-\gamma)/(2NA)$ and exploration rate $\xi = \min\left(\left(\frac{\kappa_\rho^2 NA\epsilon_{\text{stat}}}{(1-\gamma)^2 W^2}\right)^{\frac{1}{3}}, \frac{1}{2}\right)$, with Assumption 2,*

$$\mathbb{E}\left[\,\text{Nash-Regret}(T)\,\right] \;\lesssim\; \mathcal{R}(\eta) \;+\; \left(\frac{\kappa_\rho^2 WAN\epsilon_{\text{stat}}}{(1-\gamma)^5}\right)^{\frac{1}{3}}$$

where $\mathcal{R}(\eta) = \frac{\sqrt{\kappa_\rho AN}}{(1-\gamma)^2\sqrt{T}}$.

We prove Theorem 4 in Appendix C.3 and show sample complexity $TK = O(1/\epsilon^5)$ in Corollary 2 of Appendix C.4.

6. Game-Agnostic Convergence

In Section 4 and Section 5, we have shown that our independent policy gradient method converges (in best-iterate sense) to a Nash equilibrium of MPGs. For the same algorithm in two-player case, however, (Bailey & Piliouras, 2019) showed that players' policies can diverge for zero-sum matrix games (a single-state case of zero-sum Markov games). A natural question arises:

Does there exist a simple gradient-based algorithm that provably converges to a Nash equilibrium in both potential/cooperative and zero-sum games?

Unfortunately, classical MWU and optimistic MWU updates do not converge to a Nash equilibrium in zero-sum and coordination games simultaneously (Cheung & Piliouras, 2020). Recently, this question was partially answered by Leonardos et al. (2021); Leonardos & Piliouras (2022) in which the authors established last-iterate convergence of Q-learning dynamics to a quantal response equilibrium for both zero-sum and potential/cooperative matrix games. In this work, we provide an affirmative answer to this question for general Markov games that cover matrix games. Specifically, we next show that optimistic gradient descent/ascent with a smoothed critic (see Algorithm 3 in Appendix A) – an algorithm that converges to a Nash equilibrium in two-player zero-sum Markov games (Wei et al., 2021b) – also converges to a Nash equilibrium in Markov cooperative games.

We now setup notation for tabular two-player Markov cooperative games with $N = 2$, $r = r_1 = r_2$, $A = |\mathcal{A}_1| = |\mathcal{A}_2|$, and $S = |\mathcal{S}|$. For convenience, we use $x_s \in \mathbb{R}^A$ and $y_s \in \mathbb{R}^A$ to denote policies $\pi_1(\cdot \mid s)$ and $\pi_2(\cdot \mid s)$ taken at state $s \in \mathcal{S}$, and $Q_s^\pi \in \mathbb{R}^{A\times A}$ to denote $Q^\pi(s, a_1, a_2)$ with $a_1 \in \mathcal{A}_1$ and $a_2 \in \mathcal{A}_2$. We describe our policy update (9) in Algorithm 3: the next iterate $(x_s^{(t+1)}, y_s^{(t+1)})$ is obtained from two steps of policy gradient ascent with an intermediate iterate $(\bar{x}_s^{(t+1)}, \bar{y}_s^{(t+1)})$. Motivated by Wei et al. (2021b), we introduce a critic $\mathcal{Q}_s^{(t)}$ to learn the value function at each state s using the learning rate $\alpha^{(t)}$. When the critic is ideal, i.e., $\mathcal{Q}_s^{(t)} = Q_s^{(t)}$, where $Q_s^{(t)}$ is a matrix form of $Q^{(t)}(s, a_1, a_2)$ for $a_1 \in \mathcal{A}_1$ and $a_2 \in \mathcal{A}_2$, we can view Algorithm 3 as a two-player case of Algorithm 1.

In Theorem 5, we establish asymptotic last-iterate convergence of Algorithm 3 in Markov cooperative games; see Appendix D.1 for proof.

Theorem 5 (Last-iterate convergence for two-player Markov cooperative games)**.** *For MPG (1) with two players and identical rewards, if both players run Algorithm 3 with $0 < \eta < (1-\gamma)/(32\sqrt{A})$ and a non-increasing $\{\alpha^{(t)}\}_{t=1}^\infty$ that satisfies $0 < \alpha^{(t)} < 1/6$ and $\sum_{t=t'}^\infty \alpha^{(t)} = \infty$ for any $t' \geq 0$, then the policy pair $(x^{(t)}, y^{(t)})$ converges to a Nash equilibrium when $t \to \infty$.*

Last-iterate convergence in Theorem 5 is measured by the local gaps $\max_{x'}(V^{x', y^{(t)}}(\rho) - V^{x^{(t)}, y^{(t)}}(\rho))$ and $\max_{y'}(V^{x^{(t)}, y'}(\rho) - V^{x^{(t)}, y^{(t)}}(\rho))$, i.e., a policy pair $(x^{(t)}, y^{(t)})$ constitutes an approximate Nash policy for large t. The condition on algorithm parameters η and $\alpha^{(t)}$ in Theorem 5 is mild in sense that it is straightforward to take a pair of such parameters that ensures last-iterate convergence in zero-sum Markov games (Wei et al., 2021b). Hence, Al-

gorithm 3 enjoys last-iterate convergence in both two-player Markov cooperative and zero-sum competitive games. Compared with the result (Fox et al., 2022), our proof of Theorem 5 utilizes gap convergence instead of point-wise policy convergence that is restricted to isolated fixed points of the algorithm dynamics. Moreover, our algorithm works for both cooperative and competitive Markov games.

In the following Theorem 6, we further strengthen our result of Theorem 5 and show the sublinear Nash-Regret bounds for Algorithm 3 in both two-player Markov cooperative and zero-sum competitive games; see Appendix D.2 for proof.

Theorem 6 (Nash-Regret bound for two-player Markov cooperative/competitive games)**.** (i) *For MPG (1) with two players and identical rewards ($r_1 = r_2 = r$), if both players independently run Algorithm 3 with $\alpha^{(t)} = \frac{1}{6\sqrt[3]{t}}$ and $\eta = \frac{(1-\gamma)^2}{32\sqrt{SA}}$, then*

$$\frac{1}{T}\sum_{t=1}^{T}\max_{x',y'}\left(V^{x',y^{(t)}}(\rho)+V^{x^{(t)},y'}(\rho)-2V^{x^{(t)},y^{(t)}}(\rho)\right) \lesssim \frac{(S^3A)^{\frac{1}{4}}}{(1-\gamma)^{\frac{7}{2}}T^{\frac{1}{6}}}$$

(ii) *For a two-player zero-sum Markov game ($r_1 = -r_2 = r$), if both players independently run Algorithm 3 with the same choice of $\alpha^{(t)}$ and η, then*

$$\frac{1}{T}\sum_{t=1}^{T}\max_{x',y'}\left(V^{x',y^{(t)}}(\rho)-V^{x^{(t)},y'}(\rho)\right) \lesssim \frac{(S^3A)^{\frac{1}{2}}}{(1-\gamma)^{\frac{15}{4}}T^{\frac{1}{6}}}.$$

For two-player Markov cooperative/competitive games, Theorem 6 establishes the same rate $T^{-1/6}$ for the average Nash regret and the average duality gap, respectively. Alternatively, independent players in Algorithm 3 can find an ϵ-Nash equilibrium after $O(1/\epsilon^6)$ iterations, no matter which types of games are being played. To the best of our knowledge, Theorem 6 appears to be the first game-agnostic convergence for Markov cooperative/competitive games with finite-time performance guarantees. We leave the extension to more general Markov games for future work.

7. Experimental Results

To demonstrate the merits and the effectiveness of our approach, we examine an MDP in which every state defines a congestion game. This example is borrowed from Bistritz & Bambos (2020) and it includes MPG as a special case.

Figure 1 shows that our independent policy gradient with a large stepsize (green curve) quickly converges to a Nash equilibrium. We note that stepsize $\eta \geq 0.001$ does not provide convergence of the projected stochastic gradient ascent (Leonardos et al., 2022). In contrast, our approach allows large stepsizes for a broad range of initial distributions; see Appendix G for additional details.

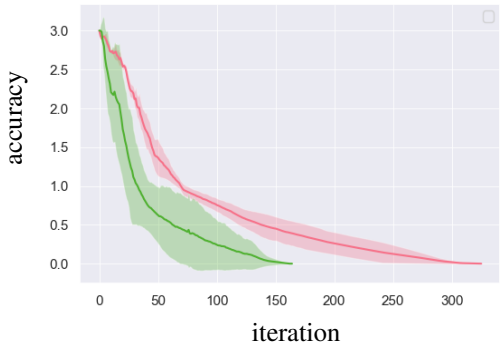

Figure 1. Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.002$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ (Leonardos et al., 2022). The accuracy measures the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N}\sum_{i=1}^{N}\|\pi_i^{(t)} - \pi_i^{\text{Nash}}\|_1$. Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval.

8. Concluding Remarks

We have proposed new independent policy gradient algorithms for learning a Nash equilibrium of Markov potential games when the size of state space and/or the number of players are large. In the exact gradient case, we show that our algorithm finds an ϵ-Nash equilibrium with $O(1/\epsilon^2)$ iteration complexity. Such iteration complexity does not explicitly depend on the state space size. In the sample-based case, our algorithm works in the function approximation setting, and we prove $O(1/\epsilon^5)$ sample complexity in a potentially infinitely large state space. This appears to be the first result for learning MPGs with function approximation. Moreover, we identify a class of independent policy gradient algorithms that enjoys last-iterate convergence and sublinear Nash regret for both zero-sum Markov games and Markov cooperative games (a special case of MPGs). This finding sheds light on an open question in the literature on the existence of such an algorithm.

Future directions include extending techniques that offer faster rates for the single-agent policy gradient methods (Lan, 2022; Zhan et al., 2021; Xiao, 2022) to independent multi-agent learning and applying independent policy gradient for other large-scale Markov games.

Acknowledgements

The work of D. Ding and M. R. Jovanović is supported in part by the National Science Foundation under awards ECCS-1708906 and 1809833. The work of C.-Y. Wei is supported by NSF Award IIS-1943607. The work of K. Zhang is supported in part by the Simons-Berkeley Research Fellowship. Part of this work was done while K. Zhang was visiting Simons Institute for the Theory of Computing.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702, 2019.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Altman, E., Hordijk, A., and Spieksma, F. Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards. *Mathematics of Operations Research*, 22(3):588–618, 1997.

Arora, S. Lecture 6 in toward theoretical understanding of deep learning, October 2008. https://www.cs.princeton.edu/courses/archive/fall18/cos597G/lecnotes/lecture6.pdf.

Audibert, J.-Y. and Catoni, O. Risk bounds for linear regression, 2009. http://imagine.enpc.fr/~audibert/Mes%20articles/Chevaleret09.pdf.

Bailey, J. and Piliouras, G. Fast and furious learning in zero-sum games: Vanishing regret with non-vanishing step sizes. *Advances in Neural Information Processing Systems*, 32, 2019.

Bailey, J. P. and Piliouras, G. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 321–338, 2018.

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

Bistritz, I. and Bambos, N. Cooperative multi-player bandit optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Bu, J., Ratliff, L. J., and Mesbahi, M. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

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

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Cen, S., Wei, Y., and Chi, Y. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Cheung, Y. K. and Piliouras, G. Chaos, extremism and optimism: Volume analysis of learning in games. *Advances in Neural Information Processing Systems*, 33:9039–9049, 2020.

Cohen, J., Héliou, A., and Mertikopoulos, P. Learning with bandit feedback in potential games. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6372–6381, 2017a.

Cohen, K., Nedić, A., and Srikant, R. On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. *IEEE Transactions on Automatic Control*, 62(11):5974–5981, 2017b.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., and Graepel, T. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., and Graepel, T. Cooperative AI: machines must learn to find common ground, 2021.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. *arXiv preprint arXiv:1807.03907*, 2018.

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Dechert, W. D. and O'Donnell, S. The stochastic lake game: A numerical solution. *Journal of Economic Dynamics and Control*, 30(9-10):1569–1587, 2006.

Dubey, A. and Pentland, A. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021.

Fink, A. M. Equilibrium in a stochastic n-person game. *Journal of science of the hiroshima university, series ai (mathematics)*, 28(1):89–93, 1964.

Fox, R., Mcaleer, S. M., Overman, W., and Panageas, I. Independent natural policy gradient always converges in Markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4414–4425, 2022.

González-Sánchez, D. and Hernández-Lerma, O. *Discrete–time stochastic control and dynamic potential games: the Euler–Equation approach*. Springer Science & Business Media, 2013.

Hambly, B. M., Xu, R., and Yang, H. Policy gradient methods find the Nash equilibrium in N-player general-sum linear-quadratic games. *arXiv preprint arXiv:2107.13090*, 2021.

Hofbauer, J. and Sandholm, W. H. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6): 2265–2294, 2002.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1, 2012.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. Towards general function approximation in zero-sum Markov games. In *International Conference on Learning Representations*, 2022.

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning – A simple, efficient, decentralized algorithm for multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021a.

Jin, C., Liu, Q., and Yu, T. The power of exploiter: Provable multi-agent RL in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021b.

Kakade, S. M. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

Kao, H., Wei, C.-Y., and Subramanian, V. Decentralized cooperative reinforcement learning with hierarchical information structure. In *International Conference on Algorithmic Learning Theory*, pp. 573–605, 2022.

Kleinberg, R., Piliouras, G., and Tardos, É. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 533–542, 2009.

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 2022.

Leonardos, S. and Piliouras, G. Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory. *Artificial Intelligence*, 304:103653, 2022.

Leonardos, S., Piliouras, G., and Spendlove, K. Exploration-exploitation in multi-agent competition: Convergence with bounded rationality. *Advances in Neural Information Processing Systems*, 34, 2021.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. Global convergence of multi-agent policy gradient in Markov potential games. In *International Conference on Learning Representations*, 2022.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010, 2021.

Lowe, R., WU, Y., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30:6379–6390, 2017.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34, 2021.

Macua, S. V., Zazo, J., and Zazo, S. Learning parametric closed-loop policies for Markov potential games. In *International Conference on Learning Representations*, 2018.

Maitra, A. and Parthasarathy, T. On stochastic games. *Journal of Optimization Theory and Applications*, 5(4):289–300, 1970.

Maitra, A. and Parthasarathy, T. On stochastic games, II. *Journal of Optimization Theory and Applications*, 8(2): 154–160, 1971.

Mao, W., Basar, T., Yang, L. F., and Zhang, K. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2022.

Marden, J. R. State based potential games. *Automatica*, 48 (12):3075–3088, 2012.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Mazalov, V. V., Rettieva, A. N., and Avrachenkov, K. E. Linear-quadratic discrete-time dynamic potential games. *Automation and Remote Control*, 78(8):1537–1544, 2017.

Mazumdar, E., Ratliff, L. J., Jordan, M. I., and Sastry, S. S. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. In *AAMAS*, 2020.

Mguni, D., Jennings, J., and de Cote, E. M. Decentralised learning in systems with many, many strategic agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Mguni, D. H., Wu, Y., Du, Y., Yang, Y., Wang, Z., Li, M., Wen, Y., Jennings, J., and Wang, J. Learning in nonzero-sum stochastic games with potentials. In *International Conference on Machine Learning*, pp. 7688–7699, 2021.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

Monderer, D. and Shapley, L. Fictitious play property for games with identical interests. *Journal of Economic Theory*, 68:258–265, 1996.

Ozdaglar, A., Sayin, M. O., and Zhang, K. Independent learning in stochastic games. *arXiv preprint arXiv:2111.11743*, 2021.

Palaiopanos, G., Panageas, I., and Piliouras, G. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. *Advances in Neural Information Processing Systems*, 30, 2017.

Peng, B., Rashid, T., Schroeder de Witt, C., Kamienny, P.-A., Torr, P., Böhmer, W., and Whiteson, S. FACMAC: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems*, 34: 12208–12221, 2021.

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

Sayin, M. O., Zhang, K., Leslie, D. S., Basar, T., and Ozdaglar, A. E. Decentralized Q-learning in zero-sum Markov games. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

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

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

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):484–489, 2016.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Song, Z., Mei, S., and Bai, Y. When can we learn general-sum Markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022.

Stastny, J., Riché, M., Lyzhov, A., Treutlein, J., Dafoe, A., and Clifton, J. Normative disagreement as a challenge for cooperative AI. *arXiv preprint arXiv:2111.13872*, 2021.

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.

Takahashi, M. Stochastic games with infinitely many strategies. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, 26(2):123–134, 1962.

Trott, A., Srinivasa, S., van der Wal, D., Haneuse, S., and Zheng, S. Building a foundation for data-driven, interpretable, and robust policy design using the AI economist. *arXiv preprint arXiv:2108.02904*, 2021.

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.

Wang, X. and Sandholm, T. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. *Advances in neural information processing systems*, 15: 1603–1610, 2002.

Wang, Y., Han, B., Wang, T., Dong, H., and Zhang, C. DOP: Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations*, 2021a.

Wang, Y., Wang, R., and Kakade, S. An exponential lower bound for linearly realizable MDP with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34, 2021b.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations*, 2020.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021a.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on learning theory*, 2021b.

Weisz, G., Amortila, P., and Szepesvári, C. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264, 2021.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Xiao, L. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

Xie, D. and Zhong, X. Semicentralized deep deterministic policy gradient in cooperative StarCraft games. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

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

Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of PPO in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955*, 2021.

Zazo, S., Macua, S. V., Sánchez-Fernández, M., and Zazo, J. Dynamic potential games with constraints: Fundamentals and applications in communications. *IEEE Transactions on Signal Processing*, 64(14):3806–3821, 2016.

Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *arXiv preprint arXiv:2105.11066*, 2021.

Zhang, K., Yang, Z., and Basar, T. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32:11602–11614, 2019.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021a.

Zhang, R., Ren, Z., and Li, N. Gradient play in multi-agent Markov stochastic games: Stationary points and convergence. *arXiv preprint arXiv:2106.00198*, 2021b.

Zhang, R., Mei, J., Dai, B., Schuurmans, D., and Li, N. On the effect of log-barrier regularization in decentralized softmax gradient play in multiagent systems. *arXiv preprint arXiv:2202.00872*, 2022.

Zhao, Y., Tian, Y., Lee, J. D., and Du, S. S. Provably efficient policy gradient methods for two-player zero-sum Markov games. *arXiv preprint arXiv:2102.08903*, 2021.

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

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning*, pp. 928–936, 2003.

Supplementary Materials for "Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence"

A. Algorithms in Section 5 and Section 6

Algorithm 2 Independent policy gradient with linear function approximation

Parameters: K, W, and $\eta > 0$.
Initialization: Let $\pi_i^{(1)}(a_i \,|\, s) = 1/A$ for $s \in \mathcal{S}$, $a_i \in \mathcal{A}_i$ and $i = 1, \ldots, N$.
for step $t = 1, \ldots, T$ **do**
  `// Phase 1 (data collection)`
  for round $k = 1, \ldots, K$ **do**
    For each $i \in [N]$, sample $h_i \sim \textsc{Geometric}(1-\gamma)$ and $h_i' \sim \textsc{Geometric}(1-\gamma)$.
    Draw an initial state $\bar{s}^{(0)} \sim \rho$.
    Continuing from $\bar{s}^{(0)}$, let all players interact with each other using $\{\pi_i^{(t)}\}_{i=1}^N$ for $H = \max_i(h_i + h_i')$ steps, which generates a state-joint-action-reward trajectory $\bar{s}^{(0)}$, $\bar{a}^{(0)}, \bar{r}^{(0)}$, $\bar{s}^{(1)}$, $\bar{a}^{(1)}$, $\bar{r}^{(1)}$, $\ldots, \bar{s}^{(H)}$, $\bar{a}^{(H)}$, $\bar{r}^{(H)}$.
    Define for every player $i \in [N]$:

$$s_i^{(k)} \;=\; \bar{s}^{(h_i)}, \qquad a_i^{(k)} \;=\; \bar{a}_i^{(h_i)}, \qquad R_i^{(k)} \;=\; \sum_{h=h_i}^{h_i+h_i'-1} \bar{r}_i^{(h)}. \tag{6}$$

  end for
  `// Phase 2 (policy update)`
  for player $i = 1, \ldots, N$ (in parallel) **do**
    Compute $\widehat{w}_i^{(t)}$ as

$$\widehat{w}_i^{(t)} \;\approx\; \operatorname*{argmin}_{\|w_i\| \le W} \sum_{k=1}^{K} \Big(R_i^{(k)} \;-\; \big\langle \phi_i(s_i^{(k)}, a_i^{(k)}), w_i \big\rangle \Big)^2. \tag{7}$$

    Define $\widehat{Q}_i^{(t)}(s, \cdot) := \big\langle \phi_i(s, \cdot), \widehat{w}_i^{(t)} \big\rangle$ and player i's policy for $s \in \mathcal{S}$,

$$\pi_i^{(t+1)}(\cdot \,|\, s) \;=\; \operatorname*{argmax}_{\pi_i(\cdot \,|\, s) \in \Delta_\xi(\mathcal{A}_i)} \left\{ \big\langle \pi_i(\cdot \,|\, s),\, \widehat{Q}_i^{(t)}(s, \cdot) \big\rangle_{\mathcal{A}_i} \;-\; \frac{1}{2\eta} \big\| \pi_i(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \big\|^2 \right\}. \tag{8}$$

  end for
end for

Algorithm 3 Independent optimistic policy gradient ascent

1: **Parameters**: $0 < \eta \leq \frac{1-\gamma}{32\sqrt{A}}$ and a non-increasing sequence $\{\alpha^{(t)}\}_{t=1}^{\infty}$ that satisfies

$$0 \;<\; \alpha^{(t)} \;\leq\; \frac{1}{6} \quad \text{for all } t \qquad\qquad \text{and} \qquad\qquad \sum_{t=t'}^{\infty} \alpha^{(t)} \;=\; \infty \quad \text{for any } t'.$$

2: **Initialization**: Let $x_s^{(1)} = \bar{x}_s^{(1)} = y_s^{(1)} = \bar{y}_s^{(1)} = 1/A$ and $\mathcal{V}_s^{(0)} = 0$ for all $s \in \mathcal{S}$.

3: **for** step $t = 1, 2, \ldots$ **do**

4: Define $\mathcal{Q}_s^{(t)} \in \mathbb{R}^{A \times A}$ for all $s \in \mathcal{S}$,

$$\mathcal{Q}_s^{(t)}(a_1, a_2) \;=\; r(s, a_1, a_2) \;+\; \gamma\, \mathbb{E}_{s' \sim \mathbb{P}(\cdot \,|\, s, a_1, a_2)} \left[\mathcal{V}_{s'}^{(t-1)} \right].$$

5: Define two players' policies for $s \in \mathcal{S}$,

$$\begin{aligned}
\bar{x}_s^{(t+1)} &= \operatorname*{argmax}_{x_s \,\in\, \Delta(\mathcal{A}_1)} \left\{ x_s^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \;-\; \frac{1}{2\eta} \big\| x_s - \bar{x}_s^{(t)} \big\|^2 \right\} \\
x_s^{(t+1)} &= \operatorname*{argmax}_{x_s \,\in\, \Delta(\mathcal{A}_1)} \left\{ x_s^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \;-\; \frac{1}{2\eta} \big\| x_s - \bar{x}_s^{(t+1)} \big\|^2 \right\} \\
\bar{y}_s^{(t+1)} &= \operatorname*{argmax}_{y_s \,\in\, \Delta(\mathcal{A}_2)} \left\{ (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s \;-\; \frac{1}{2\eta} \big\| y_s - \bar{y}_s^{(t)} \big\|^2 \right\} \\
y_s^{(t+1)} &= \operatorname*{argmax}_{y_s \,\in\, \Delta(\mathcal{A}_2)} \left\{ (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s \;-\; \frac{1}{2\eta} \big\| y_s - \bar{y}_s^{(t+1)} \big\|^2 \right\}
\end{aligned} \tag{9}$$

$$\mathcal{V}_s^{(t)} \;=\; (1 - \alpha^{(t)}) \mathcal{V}_s^{(t-1)} \;+\; \alpha^{(t)} (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}.$$

6: **end for**

B. Proofs for Section 4

In this section, we provide proofs of Theorem 1 and Theorem 2 in Appendix B.1 and Appendix B.2, respectively.

B.1. Proof of Theorem 1

We first seek to decompose the difference of a potential function $\Phi^{\pi}(\mu)$ at two different policies for any state distribution μ.

Let $\Psi^{\pi}:\Pi\to\mathbb{R}$ be any multivariate function mapping a policy $\pi\in\Pi$ to a real number. In Lemma 2, we show that the difference $\Psi^{\pi'}-\Psi^{\pi}$ at any two policies π,π' equals to a sum of several partial differences. For $i,j\in\{1,\dots,N\}$ with $i<j$, we denote by "$i\sim j$" the set of indices $\{k\,|\,i<k<j\}$, "$<i$" the set of indices $\{k\,|\,k=1,\dots,i-1\}$, and "$>j$" the set of indices $\{k\,|\,k=j+1,\dots,N\}$. We use the shorthand $\pi_I:=\{\pi_k\}_{k\in I}$ to represent the joint policy for all players $k\in I$. For example, when $I=i\sim j$, $\pi_I=\{\pi_k\}_{k=i+1}^{j-1}$ is a joint policy for players from $i+1$ to $j-1$; $\pi_{<i,\,i\sim j}$, $\pi_{<i}$, and $\pi_{>j}$ can be introduced similarly.

Lemma 2 (Multivariate function difference)**.** *For any function $\Psi^{\pi}:\Pi\to\mathbb{R}$, and any two policies $\pi,\pi'\in\Pi$,*

$$\begin{aligned}\Psi^{\pi'}-\Psi^{\pi} \;=\; &\sum_{i=1}^{N}\Big(\Psi^{\pi'_i,\,\pi_{-i}}-\Psi^{\pi}\Big)\\ &+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Big(\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi'_i,\,\pi'_j}-\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi_i,\,\pi'_j}\\ &\qquad\qquad-\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi'_i,\,\pi_j}+\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi_i,\,\pi_j}\Big).\end{aligned}\tag{10}$$

Proof of Lemma 2. We prove (10) by induction on the number of players N. In the basic step: $N=2$, the right-hand side of (10) becomes

$$\Big(\Psi^{\pi'_1,\,\pi_2}-\Psi^{\pi_1,\,\pi_2}\Big)+\Big(\Psi^{\pi_1,\,\pi'_2}-\Psi^{\pi_1,\,\pi_2}\Big)+\Big(\Psi^{\pi'_1,\,\pi'_2}-\Psi^{\pi_1,\,\pi'_2}-\Psi^{\pi'_1,\,\pi_2}+\Psi^{\pi_1,\,\pi_2}\Big)$$

which equals to the left-hand side: $\Psi^{\pi'_1,\pi'_2}-\Psi^{\pi_1,\pi_2}$.

Assume the equality (10) holds for N players. We next consider the induction step for $N+1$ players . By subtracting and adding $\Psi^{\pi_{\le N},\,\pi'_{N+1}}$,

$$\Psi^{\pi'}-\Psi^{\pi}=\underbrace{\Big(\Psi^{\pi'_{\le N},\,\pi'_{N+1}}-\Psi^{\pi_{\le N},\,\pi'_{N+1}}\Big)}_{\textbf{Diff}_{\le N}}+\underbrace{\Big(\Psi^{\pi_{\le N},\,\pi'_{N+1}}-\Psi^{\pi_{\le N},\,\pi_{N+1}}\Big)}_{\textbf{Diff}_{N+1}}.\tag{11}$$

In (11), we use the shorthand $\pi'_{\le N}$ and $\pi_{\le N}$ for $\{\pi'_k\}_{k=1}^N$ and $\{\pi_k\}_{k=1}^N$, respectively. We note that $\textbf{Diff}_{\le N}$ or $\textbf{Diff}_{N+1}$ can be viewed as a function for N players if we fix the $(N+1)$th policy. By the induction assumption, for the first term $\textbf{Diff}_{\le N}$,

$$\begin{aligned}\textbf{Diff}_{\le N} \;=\; &\sum_{i=1}^{N}\Big(\Psi^{\pi'_i,\,\pi_{<i,i\sim N+1},\,\pi'_{N+1}}-\Psi^{\pi_{\le N},\,\pi'_{N+1}}\Big)\\ &+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Big(\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi'_i,\,\pi'_j,\,\pi'_{N+1}}-\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi_i,\,\pi'_j,\,\pi'_{N+1}}\\ &\qquad\qquad-\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi'_i,\,\pi_j,\,\pi'_{N+1}}+\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi_i,\,\pi_j,\,\pi'_{N+1}}\Big)\\ =\; &\sum_{i=1}^{N}\Big(\Psi^{\pi'_i,\,\pi_{<i,i\sim N+1},\,\pi_{N+1}}-\Psi^{\pi_{\le N},\,\pi_{N+1}}\Big)\\ &+\sum_{i=1}^{N}\Big(\Psi^{\pi'_i,\,\pi_{<i,i\sim N+1},\,\pi'_{N+1}}-\Psi^{\pi_{\le N},\,\pi'_{N+1}}-\Psi^{\pi'_i,\,\pi_{<i,i\sim N+1},\,\pi_{N+1}}+\Psi^{\pi_{\le N},\,\pi_{N+1}}\Big)\\ &+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Big(\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi'_i,\,\pi'_j,\,\pi'_{N+1}}-\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi_i,\,\pi'_j,\,\pi'_{N+1}}\\ &\qquad\qquad-\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi'_i,\,\pi_j,\,\pi'_{N+1}}+\Psi^{\pi_{<i,i\sim j},\,\pi'_{>j},\,\pi_i,\,\pi_j,\,\pi'_{N+1}}\Big)\end{aligned}$$

where we use $\pi'_{>j}$ to represent $\{\pi'_k\}_{k=j+1}^N$.

Adding $\mathbf{Diff}_{N+1}$ to the last equivalent expression of $\mathbf{Diff}_{\le N}$ above yields

$$\begin{aligned}
\mathbf{Diff}_{\le N} + \mathbf{Diff}_{N+1} &= \sum_{i=1}^{N+1}\Big(\Psi^{\pi'_i,\,\pi_{-i}} - \Psi^{\pi}\Big) \\
&\quad + \sum_{i=1}^{N}\sum_{j=N+1}^{N+1}\Big(\Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi'_j} - \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi'_j} \\
&\qquad\qquad - \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi_j} + \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi_j}\Big) \\
&\quad + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\Big(\Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi'_j,\,\pi'_{N+1}} - \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi'_j,\,\pi'_{N+1}} \\
&\qquad\qquad - \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi_j,\,\pi'_{N+1}} + \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi_j,\,\pi'_{N+1}}\Big) \\
&= \sum_{i=1}^{N+1}\Big(\Psi^{\pi'_i,\,\pi_{-i}} - \Psi^{\pi}\Big) \\
&\quad + \sum_{i=1}^{N+1}\sum_{j=i+1}^{N+1}\Big(\Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi'_j} - \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi'_j} \\
&\qquad\qquad - \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi_j} + \Psi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi_j}\Big).
\end{aligned}$$

where the first equality has a slight abuse of the notation: $\pi'_{>j}$ represents $\{\pi'_k\}_{k=j+1}^{N+1}$ in the first double sum and $\pi'_{>j}$ represents $\{\pi'_k\}_{k=j+1}^{N}$ in the second double sum. Therefore, (10) holds for $N+1$ players. The proof is completed by induction. □

We apply Lemma 2 to the potential function $\Phi^{\pi}(\mu)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 1, where μ is an initial state distribution. We use the shorthand $\Phi^{(t)}(\mu)$ for $\Phi^{\pi^{(t)}}(\mu)$, the value of potential function at policy $\pi^{(t)}$.

Lemma 3 (Policy improvement: Markov potential games)**.** *For MPG (1) with any state distribution μ, the potential function $\Phi^{\pi}(\mu)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 1 satisfies*

$$\begin{aligned}
\text{(i)}\ \ \Phi^{(t+1)}(\mu) - \Phi^{(t)}(\mu) &\ge \frac{1}{2\eta(1-\gamma)}\sum_{i=1}^{N}\sum_{s} d_{\mu}^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s)\right\|^2 - \frac{4\eta^2A^2N^2}{(1-\gamma)^5} \\
\text{(ii)}\ \ \Phi^{(t+1)}(\mu) - \Phi^{(t)}(\mu) &\ge \frac{1}{2\eta(1-\gamma)}\sum_{i=1}^{N}\sum_{s} d_{\mu}^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)\left(1 - \frac{4\eta\kappa_{\mu}^3AN}{(1-\gamma)^4}\right)\left\|\pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s)\right\|^2
\end{aligned}$$

where η is the stepsize, N is the number of players, A is the size of one player's action space, and κ_{μ} is the distribution mismatch coefficient relative to μ (see κ_{μ} in Definition 1).

Proof of Lemma 3. We let $\pi' = \pi^{(t+1)}$ and $\pi = \pi^{(t)}$ for brevity. By Lemma 2 with $\Psi^{\pi} = \Phi^{\pi}(\mu)$, it is equivalent to analyze

$$\Phi^{(t+1)}(\mu) - \Phi^{(t)}(\mu) = \mathbf{Diff}_{\alpha} + \mathbf{Diff}_{\beta} \tag{12}$$

where

$$\mathbf{Diff}_{\alpha} = \sum_{i=1}^{N}\Big(\Phi^{\pi'_i,\,\pi_{-i}}(\mu) - \Phi^{\pi}(\mu)\Big)$$

$$\begin{aligned}
\mathbf{Diff}_{\beta} &= \sum_{i=1}^{N}\sum_{j=i+1}^{N}\Big(\Phi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi'_j}(\mu) - \Phi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi'_j}(\mu) \\
&\qquad - \Phi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi'_i,\,\pi_j}(\mu) + \Phi^{\pi_{<i},\,i\sim j,\,\pi'_{>j},\,\pi_i,\,\pi_j}(\mu)\Big).
\end{aligned}$$

Bounding Diff$_\alpha$. By the property of the potential function $\Phi^\pi(\mu)$,

$$\begin{aligned}\Phi^{\pi_i', \pi_{-i}}(\mu) - \Phi^{\pi}(\mu) &= V_i^{\pi_i', \pi_{-i}}(\mu) - V_i^{\pi}(\mu) \\ &= \frac{1}{1-\gamma} \sum_{s,\, a_i} d_\mu^{\pi_i', \pi_{-i}}(s) \left(\pi_i'(a_i \,|\, s) - \pi_i(a_i \,|\, s)\right) \bar{Q}_i^{\pi_i, \pi_{-i}}(s, a_i)\end{aligned} \tag{13}$$

where the second equality is due to Lemma 1 using $\widehat{\pi}_i = \pi_i'$ and $\bar{\pi}_i = \pi_i$. The optimality of $\pi_i' = \pi_i^{(t+1)}$ in line 4 of Algorithm 1 leads to

$$\left\langle \pi_i'(\cdot \,|\, s), \bar{Q}_i^{\pi_i, \pi_{-i}}(s, \cdot) \right\rangle_{\mathcal{A}_i} - \frac{1}{2\eta} \left\| \pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s) \right\|^2 \geq \left\langle \pi_i(\cdot \,|\, s), \bar{Q}_i^{\pi_i, \pi_{-i}}(s, \cdot) \right\rangle_{\mathcal{A}_i}. \tag{14}$$

Combining (13) and (14), we get

$$\Phi^{\pi_i', \pi_{-i}}(\mu) - \Phi^{\pi}(\mu) \geq \frac{1}{2\eta(1-\gamma)} \sum_s d_\mu^{\pi_i', \pi_{-i}}(s) \left\| \pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s) \right\|^2.$$

Therefore,

$$\textbf{Diff}_\alpha \geq \frac{1}{2\eta(1-\gamma)} \sum_{i=1}^N \sum_s d_\mu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2. \tag{15}$$

Bounding Diff$_\beta$. For simplicity, we denote $\widetilde{\pi}_{-ij}$ as the joint policy of players $N \backslash \{i, j\}$ where players $< i$ and $i \sim j$ use π and players $> j$ use π'. For each summand in **Diff**$_\beta$,

$$\begin{aligned}
&\Phi^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(\mu) - \Phi^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j'}(\mu) - \Phi^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j}(\mu) + \Phi^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(\mu) \\
\overset{(a)}{=}\ & V_i^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(\mu) - V_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j'}(\mu) - V_i^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j}(\mu) + V_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(\mu) \\
\overset{(b)}{=}\ & \frac{1}{1-\gamma} \sum_{s,\, a_i} d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(s) \left(\pi_i'(a_i \,|\, s) - \pi_i(a_i \,|\, s)\right) \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j'}(s, a_i) \\
& - \frac{1}{1-\gamma} \sum_{s,\, a_i} d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j}(s) \left(\pi_i'(a_i \,|\, s) - \pi_i(a_i \,|\, s)\right) \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(s, a_i) \\
=\ & \frac{1}{1-\gamma} \sum_{s,\, a_i} d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(s) \left(\pi_i'(a_i \,|\, s) - \pi_i(a_i \,|\, s)\right) \left(\bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j'}(s, a_i) - \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(s, a_i)\right) \\
& + \frac{1}{1-\gamma} \sum_{s,\, a_i} \left(d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(s) - d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j}(s)\right) \left(\pi_i'(a_i \,|\, s) - \pi_i(a_i \,|\, s)\right) \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(s, a_i) \\
\geq\ & - \frac{1}{1-\gamma} \sum_s d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(s) \left\| \pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s) \right\|_1 \left\| \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j'}(s, \cdot) - \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(s, \cdot) \right\|_\infty \\
& - \frac{1}{1-\gamma} \sum_s \left| d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(s) - d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j}(s) \right| \left\| \pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s) \right\|_1 \left\| \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(s, \cdot) \right\|_\infty \\
\overset{(c)}{\geq}\ & - \frac{1}{(1-\gamma)^3} \left(\max_s \left\| \pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s) \right\|_1\right) \left(\max_s \left\| \pi_j'(\cdot \,|\, s) - \pi_j(\cdot \,|\, s) \right\|_1\right) \\
& - \frac{1}{(1-\gamma)^2} \left(\max_s \left\| \pi_j'(\cdot \,|\, s) - \pi_j(\cdot \,|\, s) \right\|_1\right) \left(\max_s \left\| \pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s) \right\|_1\right) \\
\overset{(d)}{\geq}\ & - \frac{8\eta^2 A^2}{(1-\gamma)^5}
\end{aligned}$$

where (a) is due to the property of the potential function, (b) is due to Lemma 1; for (c), we use Lemma 4, Lemma 20, and the fact that $\sum_s d_\mu^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(s) = 1$ and $\left\| \bar{Q}_i^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(s, \cdot) \right\|_\infty \leq \frac{1}{1-\gamma}$; The last inequality (d) follows a direct result from

the optimality of $\pi_i' = \pi_i^{(t+1)}$ given by (14) and $\|\cdot\| \le \sqrt{A}\|\cdot\|_\infty$ and $\|\cdot\|_1 \le \sqrt{A}\|\cdot\|$:

$$\begin{aligned}
\|\pi_i'(\cdot\,|\,s) - \pi_i(\cdot\,|\,s)\|^2 \;&\le\; 2\eta\big\langle \pi_i^{(t+1)}(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s), \bar{Q}_i^{\pi_i,\pi_{-i}}(s,\cdot)\big\rangle_{\mathcal{A}_i} \\
&\le\; 2\eta \left\|\pi_j^{(t+1)}(\cdot\,|\,s) - \pi_j^{(t)}(\cdot\,|\,s)\right\| \left\|\bar{Q}_i^{\pi_i,\pi_{-i}}(s,\cdot)\right\| \\
&\implies \left\|\pi_j^{(t+1)}(\cdot\,|\,s) - \pi_j^{(t)}(\cdot\,|\,s)\right\| \;\le\; 2\eta \left\|\bar{Q}_i^{\pi_i,\pi_{-i}}(s,\cdot)\right\| \;\le\; \frac{2\eta\sqrt{A}}{1-\gamma} \\
&\implies \left\|\pi_j^{(t+1)}(\cdot\,|\,s) - \pi_j^{(t)}(\cdot\,|\,s)\right\|_1 \;\le\; \frac{2\eta A}{1-\gamma}.
\end{aligned}$$

Therefore,

$$\textbf{Diff}_\beta \;\ge\; -\frac{N(N-1)}{2} \times \frac{8\eta^2 A^2}{(1-\gamma)^5} \;\ge\; -\frac{4\eta^2 A^2 N^2}{(1-\gamma)^5}. \tag{16}$$

We now complete the proof of (i) by combining (12), (15), and (16).

Alternatively, by Lemma 21, we can bound each summand of $\textbf{Diff}_\beta$ by

$$\begin{aligned}
&\Phi^{\widetilde{\pi}_{-ij},\,\pi_i',\,\pi_j'}(\mu) \;-\; \Phi^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j'}(\mu) \;-\; \Phi^{\widetilde{\pi}_{-ij},\,\pi_i',\,\pi_j}(\mu) \;+\; \Phi^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j}(\mu) \\
&=\; V_i^{\widetilde{\pi}_{-ij},\,\pi_i',\,\pi_j'}(\mu) \;-\; V_i^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j'}(\mu) \;-\; V_i^{\widetilde{\pi}_{-ij},\,\pi_i',\,\pi_j}(\mu) \;+\; V_i^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j}(\mu) \\
&\ge\; -\frac{2\kappa_\mu^2 A}{(1-\gamma)^4} \sum_s d_\mu^{\widetilde{\pi}_{-ij},\pi_i,\pi_j}(s) \left(\|\pi_i(\cdot\,|\,s) - \pi_i'(\cdot\,|\,s)\|^2 + \|\pi_j(\cdot\,|\,s) - \pi_j'(\cdot\,|\,s)\|^2\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\textbf{Diff}_\beta \;&\ge\; -\frac{2\kappa_\mu^2 A}{(1-\gamma)^4} \sum_{i=1}^N \sum_{j=i+1}^N \sum_s d_\mu^{\widetilde{\pi}_{-ij},\pi_i,\pi_j}(s) \left(\|\pi_i(\cdot\,|\,s) - \pi_i'(\cdot\,|\,s)\|^2 + \|\pi_j(\cdot\,|\,s) - \pi_j'(\cdot\,|\,s)\|^2\right) \\
&\ge\; -\frac{2\kappa_\mu^3 NA}{(1-\gamma)^5} \sum_{i=1}^N \sum_s d_\mu^{\pi_i^{(t+1)},\,\pi_{-i}^{(t)}}(s) \left\|\pi_i^{(t)}(\cdot\,|\,s) - \pi_i^{(t+1)}(\cdot\,|\,s)\right\|^2. \qquad \left(\text{since } \tfrac{d_\mu^\pi(s)}{d_\mu^{\pi'}(s)} \le \tfrac{\kappa_\mu}{1-\gamma} \text{ for any } \pi, \pi', s\right)
\end{aligned}$$

Combining the inequality above with (12) and (15) finishes the proof of (ii). □

Lemma 4. *Suppose $i < j$ for $i, j = 1, \ldots, N$. Let $\widetilde{\pi}_{-ij}$ be the policy for all players but i, j and π_i be the policy for player i. For any two policies for player j: π_j and π_j', we have*

$$\max_s \left\|\bar{Q}_i^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j'}(s,\cdot) - \bar{Q}_i^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j}(s,\cdot)\right\|_\infty \;\le\; \frac{1}{(1-\gamma)^2} \max_s \left\|\pi_j'(\cdot\,|\,s) - \pi_j(\cdot\,|\,s)\right\|_1.$$

Proof of Lemma 4. We note that $\bar{Q}_i^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j'}(s,\cdot)$ and $\bar{Q}_i^{\widetilde{\pi}_{-ij},\,\pi_i,\,\pi_j}(s,\cdot)$ are averaged action value functions for player i using policy π_i, but they have different underlying averaged MDPs because of different policies executed by player j. Hence, we can directly apply Lemma 19. Specifically, let (r, p) be the averaged reward and transition functions for player i induced by $(\widetilde{\pi}_{-ij}, \pi_j)$, and $(\widetilde{r}, \widetilde{p})$ be those induced by $(\widetilde{\pi}_{-ij}, \pi_j')$. Then,

$$\begin{aligned}
&|r(s,a_i) - \widetilde{r}(s,a_i)| \\
&= \left|\sum_{a_j,a_{-ij}} r(s,a_i,a_j,a_{-ij})\pi_j(a_j\,|\,s)\widetilde{\pi}_{-ij}(a_{-ij}\,|\,s) - \sum_{a_j,a_{-ij}} r(s,a_i,a_j,a_{-ij})\pi_j'(a_j\,|\,s)\widetilde{\pi}_{-ij}(a_{-ij}\,|\,s)\right| \\
&\le \left\|\pi_j(\cdot\,|\,s) - \pi_j'(\cdot\,|\,s)\right\|_1
\end{aligned}$$

and

$$\begin{aligned}
&\|p(\cdot|s,a_i)-\widetilde{p}(\cdot|s,a_i)\|_1\\
=\;&\sum_{s'}\left|\sum_{a_j,\,a_{-ij}} p(s'\,|\,s,a_i,a_j,a_{-ij})\left(\pi_j(a_j\,|\,s)-\pi_j'(a_j\,|\,s)\right)\widetilde{\pi}_{-ij}(a_{-ij}\,|\,s)\right|\\
\le\;&\sum_{s'}\sum_{a_j,\,a_{-ij}} p(s'\,|\,s,a_i,a_j,a_{-ij})\widetilde{\pi}_{-ij}(a_{-ij}\,|\,s)\left|\pi_j(a_j\,|\,s)-\pi_j'(a_j\,|\,s)\right|\\
\le\;&\left\|\pi_j(\cdot|s)-\pi_j'(\cdot|s)\right\|_1.
\end{aligned}$$

Application of two inequalities above to Lemma 19 competes the proof. □

Proof of Theorem 1. By the optimality of $\pi_i^{(t+1)}$ in line 4 of Algorithm 1,

$$\left\langle \pi_i'(\cdot\,|\,s)-\pi_i^{(t+1)}(\cdot\,|\,s),\,\eta\bar{Q}_i^{(t)}(s,\cdot)-\pi_i^{(t+1)}(\cdot\,|\,s)+\pi_i^{(t)}(\cdot\,|\,s)\right\rangle_{\mathcal{A}_i}\;\le\;0,\quad\text{for any }\pi_i'\in\Pi_i.$$

Hence, if $\eta\le\frac{1-\gamma}{\sqrt{A}}$, then for any $\pi_i'\in\Pi_i$,

$$\begin{aligned}
&\left\langle \pi_i'(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s),\,\bar{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i}\\
=\;&\left\langle \pi_i'(\cdot\,|\,s)-\pi_i^{(t+1)}(\cdot\,|\,s),\,\bar{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i}+\left\langle \pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s),\,\bar{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i}\\
\le\;&\frac{1}{\eta}\left\langle \pi_i'(\cdot\,|\,s)-\pi_i^{(t+1)}(\cdot\,|\,s),\,\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\rangle_{\mathcal{A}_i}+\left\langle \pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s),\,\bar{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i}\\
\overset{(a)}{\le}\;&\frac{2}{\eta}\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|+\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|\left\|\bar{Q}_i^{(t)}(s,\cdot)\right\|\\
\overset{(b)}{\le}\;&\frac{3}{\eta}\left\|\pi_i^{(t+1)}(\cdot|s)-\pi_i^{(t)}(\cdot|s)\right\|.
\end{aligned}$$

where in (a) we apply the Cauchy-Schwarz inequality and that $\|p-p'\|\le\|p-p'\|_1\le 2$ for any two distributions p and p'; (b) is because of $\|\bar{Q}_i^{(t)}(s,\cdot)\|\le\frac{\sqrt{A}}{1-\gamma}$ and $\eta\le\frac{1-\gamma}{\sqrt{A}}$. Therefore, for any initial distribution ρ,

$$\begin{aligned}
&\sum_{t=1}^{T}\max_i\left(\max_{\pi_i'}V_i^{\pi_i',\,\pi_{-i}^{(t)}}(\rho)-V_i^{\pi^{(t)}}(\rho)\right)\\
\overset{(a)}{=}\;&\frac{1}{1-\gamma}\sum_{t=1}^{T}\max_{\pi_i'}\sum_{s,\,a_i}d_\rho^{\pi_i',\pi_{-i}^{(t)}}(s)\left(\pi_i'(a_i\,|\,s)-\pi_i^{(t)}(a_i\,|\,s)\right)\bar{Q}_i^{(t)}(s,a_i)\\
\overset{(b)}{\le}\;&\frac{3}{\eta(1-\gamma)}\sum_{t=1}^{T}\sum_s d_\rho^{\pi_i',\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|\\
\overset{(c)}{\lesssim}\;&\frac{\sqrt{\sup_{\pi\in\Pi}\|d_\rho^\pi/\nu\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}}\sum_{t=1}^{T}\sum_s\sqrt{d_\rho^{\pi_i',\pi_{-i}^{(t)}}(s)\times d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)}\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|\\
\overset{(d)}{\le}\;&\frac{\sqrt{\sup_{\pi\in\Pi}\|d_\rho^\pi/\nu\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}}\sqrt{\sum_{t=1}^{T}\sum_s d_\rho^{\pi_i',\pi_{-i}^{(t)}}(s)}\times\sqrt{\sum_{t=1}^{T}\sum_s d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|^2}\\
\overset{(e)}{\le}\;&\frac{\sqrt{\sup_{\pi\in\Pi}\|d_\rho^\pi/\nu\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}}\sqrt{\sum_{t=1}^{T}\sum_s d_\rho^{\pi_i',\pi_{-i}^{(t)}}(s)}\times\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_s d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|^2}
\end{aligned}\tag{17}$$

where (a) is due to Lemma 1 and we slightly abuse the notation i to represent argmax_i, in (b) we slightly abuse the notation π_i' to represent $\mathrm{argmax}_{\pi_i'}$, in (c) we choose an arbitrary $\nu\in\Delta(\mathcal{S})$ and use the following inequality:

$$\frac{d_\rho^{\pi_i',\,\pi_{-i}^{(t)}}(s)}{d_\nu^{\pi_i^{(t+1)},\,\pi_{-i}^{(t)}}(s)}\;\le\;\frac{d_\rho^{\pi_i',\,\pi_{-i}^{(t)}}(s)}{(1-\gamma)\nu(s)}\;\le\;\frac{\sup_\pi\|d_\rho^{\pi\,\in\,\Pi}/\nu\|_\infty}{1-\gamma}.$$

We apply the Cauchy–Schwarz inequality in (d), and finally we replace i (argmax_i in (a)) in the last square root term in (e) by the sum over all players.

If we proceed (17) with $\nu = \operatorname{argmin}_{\nu \in \Delta(\mathcal{S})} \max_{\pi \in \Pi} \|d_\rho^\pi / \nu\|_\infty$, then,

$$\begin{aligned}
&\sum_{t=1}^{T} \max_i \left(\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho) \right) \\
&\overset{(a)}{\leq} \frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{T} \times \sqrt{2\eta(1-\gamma)\left(\Phi^{(T+1)}(\nu) - \Phi^{(1)}(\nu)\right) + \frac{4\eta^3 A^2 N^2}{(1-\gamma)^4} T} \\
&\overset{(b)}{\lesssim} \sqrt{\frac{\widetilde{\kappa}_\rho T C_\Phi}{\eta(1-\gamma)^2}} + \sqrt{\frac{\widetilde{\kappa}_\rho \eta T^2 A^2 N^2}{(1-\gamma)^7}}
\end{aligned}$$

where in (a) we apply the first bound (i) in Lemma 3 (with $\mu = \nu$) and use Definition 2: $\widetilde{\kappa}_\rho = \min_{\nu \in \Delta(\mathcal{S})} \max_{\pi \in \Pi} \|d_\rho^\pi / \nu\|_\infty$, and in (b) we use $|\Phi^\pi(\nu) - \Phi^{\pi'}(\nu)| \leq C_\Phi$ for any π, π', and further simplify the bound in (b). We complete the proof for the first bound by taking stepsize $\eta = \frac{(1-\gamma)^{2.5}\sqrt{C_\Phi}}{NA\sqrt{T}}$ (by the upper bound of C_Φ given in Lemma 18, the condition $\eta \leq \frac{1-\gamma}{\sqrt{A}}$ is satisfied).

If we proceed (17) with the second bound (ii) in Lemma 3 with the choice of $\eta \leq \frac{(1-\gamma)^4}{8\kappa_\nu^3 NA}$, then,

$$\begin{aligned}
&\sum_{t=1}^{T} \max_i \left(\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho) \right) \\
&\leq \frac{\sqrt{\sup_{\pi \in \Pi} \|d_\rho^\pi / \nu\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{T} \times \sqrt{4\eta(1-\gamma)\left(\Phi^{(T+1)}(\nu) - \Phi^{(1)}(\nu)\right)} \\
&\lesssim \sqrt{\frac{\sup_{\pi \in \Pi} \|d_\rho^\pi / \nu\|_\infty T C_\Phi}{\eta(1-\gamma)^2}}.
\end{aligned}$$

We next discuss two special choices of ν for proving our bound. First, if $\nu = \rho$, then $\eta \leq \frac{(1-\gamma)^4}{8\kappa_\rho^3 NA}$. By letting $\eta = \frac{(1-\gamma)^4}{8\kappa_\rho^3 NA}$, the last square root term can be bounded by $O\left(\sqrt{\frac{\kappa_\rho^4 NATC_\Phi}{(1-\gamma)^6}}\right)$. Second, if $\nu = \frac{1}{S}\mathbf{1}$, the uniform distribution over $\mathcal{S}$, then $\kappa_\nu \leq \frac{1}{S}$, which allows a valid choice $\eta = \frac{(1-\gamma)^4}{8S^3 NA} \leq \frac{(1-\gamma)^4}{8\kappa_\nu^3 NA}$. Hence, we can bound the last square root term by $O\left(\sqrt{\frac{S^4 NATC_\Phi}{(1-\gamma)^6}}\right)$. Since ν is arbitrary, combining these two special choices completes the proof. □

B.2. Proof of Theorem 2

We first establish policy improvement regarding the Q-function at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 1.

Lemma 5 (Policy improvement: Markov cooperative games)**.** *For MPG (1) with identical rewards and an initial state distribution $\rho > 0$, if all players independently perform the policy update in Algorithm 1 with stepsize $\eta \leq \frac{1-\gamma}{2N}$, then for any t and any s,*

$$\mathbb{E}_{a \sim \pi^{(t+1)}(\cdot \mid s)}\left[Q^{(t)}(s,a)\right] - \mathbb{E}_{a \sim \pi^{(t)}(\cdot \mid s)}\left[Q^{(t)}(s,a)\right] \geq \frac{1}{4\eta} \sum_{i=1}^{N} \left\| \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \right\|^2$$

where η is the stepsize and N is the number of players.

Proof of Lemma 5. Fixing the time t and the state s, we apply Lemma 2 to

$$\Psi^\pi = \mathbb{E}_{a \sim \pi(\cdot \mid s)}\left[Q^{(t)}(s,a)\right]$$

where $Q^{(t)} := Q^{\pi^{(t)}}$ (recall that π is a joint policy of all players). By Lemma 2, for any two policies π' and π,

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi'(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] - \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
=\;& \sum_{i=1}^{N}\left(\mathbb{E}_{a_i\sim\pi_i'(\cdot\,|\,s),\,a_{-i}\sim\pi_{-i}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] - \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right]\right) \\
&+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Bigg(\mathbb{E}_{a_i\sim\pi_i'(\cdot\,|\,s),\,a_j\sim\pi_j'(\cdot\,|\,s),\,a_{-ij}\sim\widetilde{\pi}_{-ij}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
&\qquad -\mathbb{E}_{a_i\sim\pi_i(\cdot\,|\,s),\,a_j\sim\pi_j'(\cdot\,|\,s),\,a_{-ij}\sim\widetilde{\pi}_{-ij}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
&\qquad -\mathbb{E}_{a_i\sim\pi_i'(\cdot\,|\,s),\,a_j\sim\pi_j(\cdot\,|\,s),\,a_{-ij}\sim\widetilde{\pi}_{-ij}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
&\qquad +\mathbb{E}_{a_i\sim\pi_i(\cdot\,|\,s),\,a_j\sim\pi_j(\cdot\,|\,s),\,a_{-ij}\sim\widetilde{\pi}_{-ij}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right]\Bigg)
\end{aligned} \tag{18}$$

where $\widetilde{\pi}_{-ij}$ is a joint policy of players $N\backslash\{i,j\}$ in which players $< i$ and $i\sim j$ use π, and players $> j$ use π'. Particularly, we choose $\pi' = \pi^{(t+1)}$ and $\pi = \pi^{(t)}$. Thus, we can reduce (18) into

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi'(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] - \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
=\;& \sum_{i=1}^{N}\sum_{a_i}\left(\pi_i'(a_i\,|\,s)-\pi_i(a_i\,|\,s)\right)\bar{Q}_i^{(t)}(s,a_i) \\
&+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{a_i,\,a_j}\left(\pi_i'(a_i\,|\,s)-\pi_i(a_i\,|\,s)\right)\left(\pi_j'(a_j\,|\,s)-\pi_j(a_j\,|\,s)\right)\mathbb{E}_{a_{-ij}\sim\widetilde{\pi}_{-ij}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
\overset{(a)}{\geq}\;& \sum_{i=1}^{N}\frac{1}{2\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 - \frac{1}{1-\gamma}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{a_i,\,a_j}\left|\pi_i'(a_i\,|\,s)-\pi_i(a_i\,|\,s)\right|\left|\pi_j'(a_j\,|\,s)-\pi_j(a_j\,|\,s)\right| \\
\overset{(b)}{\geq}\;& \sum_{i=1}^{N}\frac{1}{2\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 - \frac{A}{2(1-\gamma)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2+\left\|\pi_j'(\cdot\,|\,s)-\pi_j(\cdot\,|\,s)\right\|^2\right) \\
=\;& \sum_{i=1}^{N}\frac{1}{2\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 - \frac{(N-1)A}{2(1-\gamma)}\sum_{i=1}^{N}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 \\
\overset{(c)}{\geq}\;& \sum_{i=1}^{N}\frac{1}{4\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2
\end{aligned}$$

where (a) is due to the optimality condition (14) and $Q^{(t)}(s,a)\leq\frac{1}{1-\gamma}$, (b) is due to $\langle x,y\rangle\leq\frac{\|x\|^2+\|y\|^2}{2}$, and (c) follows the choice of $\eta\leq\frac{1-\gamma}{2NA}$. □

Proof of Theorem 2. By Lemma 1 and Lemma 5, we have for any $\nu\in\Delta(\mathcal{S})$,

$$\begin{aligned}
V^{(t+1)}(\nu) - V^{(t)}(\nu) &= \frac{1}{1-\gamma}\sum_{s,\,a} d_\nu^{\pi^{(t+1)}}(s)\left(\pi^{(t+1)}(a\,|\,s)-\pi^{(t)}(a\,|\,s)\right)Q^{(t)}(s,a) \\
&\geq \frac{1}{4\eta(1-\gamma)}\sum_{i=1}^{N}\sum_{s} d_\nu^{\pi^{(t+1)}}(s)\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|^2.
\end{aligned} \tag{19}$$

By the same argument as the proof of Theorem 1,

$$
\begin{aligned}
&\sum_{t=1}^{T} \max_i \left(\max_{\pi_i'} V^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V^{\pi^{(t)}}(\rho) \right) \\
\overset{(a)}{\leq} \quad & \frac{3}{\eta(1-\gamma)} \sum_{t=1}^{T} \sum_{s} d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\| \\
\overset{(b)}{\lesssim} \quad & \frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_{s} \sqrt{d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \times d_\nu^{\pi^{(t+1)}}(s)} \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\| \\
\leq \quad & \frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_{s} d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T} \sum_{s} d_\nu^{\pi^{(t+1)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2} \\
\overset{(c)}{\leq} \quad & \frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_{s} d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{s} d_\nu^{\pi^{(t+1)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2} \\
\overset{(d)}{\leq} \quad & \frac{\sqrt{\widetilde{\kappa}_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{T} \times \sqrt{4\eta(1-\gamma)\left(V^{(T+1)}(\nu) - V^{(1)}(\nu)\right)}.
\end{aligned}
$$

where in (a) we slightly abuse the notation i to represent argmax_i as in (17), in (b) we take $\nu = \mathrm{argmin}_{\nu \in \Delta(\mathcal{S})} \max_{\pi \in \Pi} \|d_\rho^\pi / \nu\|_\infty$ and use the definition of $\widetilde{\kappa}_\rho$ from Definition 2, and we replace i (argmax_i in (a)) in the last square root term in (c) by the sum over all players, and we apply (19) in (d).

Finally, we complete the proof by taking stepsize $\eta = \frac{1-\gamma}{2NA}$ and using $V^{(T+1)}(\nu) - V^{(1)}(\nu) \leq \frac{1}{1-\gamma}$. □

C. Proofs for Section 5

In this section, we provide proofs of Theorem 3 and Theorem 4 in Appendix C.2 and Appendix C.3, respectively.

C.1. Unbiased estimate

We consider the kth sampling in the data collection phase of Algorithm 2. By the sampling model in lines 6-8 of Algorithm 2, it is straightforward to see that $\bar{s}^{(h_i)} \sim d_\rho^{\pi^{(t)}}$ for player i. Then, we take $\bar{a}_i^{(h_i)} \sim \pi_i^{(k)}(\cdot \,|\, s^{(h_i)})$ at step h_i for player i. Each player i begins with such $(\bar{s}^{(h_i)}, \bar{a}_i^{(h_i)})$ while all players execute the policy $\{\pi_i^{(t)}\}_{i=1}^N$ with the termination probability $1-\gamma$.

Once terminated, we add all rewards collected in $R_i^{(k)}$. We next show that $\mathbb{E}[\, R_i^{(k)}\,] = \bar{Q}_i^{\pi}(\bar{s}^{(h_i)}, \bar{a}_i^{(h_i)})$,

$$
\begin{aligned}
\mathbb{E}[\, R_i^{(k)}\,] &= \mathbb{E}\left[\sum_{h=h_i}^{h_i+h_i'-1} \bar{r}_i^{(h)} \,\middle|\, \bar{s}^{(h_i)}, \bar{a}_i^{(h_i)}, \bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)}), h_i' \sim \textsc{Geometric}(1-\gamma)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{h=0}^{h_i'-1} \bar{r}_i^{(h+h_i)} \,\middle|\, \bar{s}^{(h_i)}, \bar{a}_i^{(h_i)}, \bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)}), h_i' \sim \textsc{Geometric}(1-\gamma)\right] \\
&= \mathbb{E}\left[\sum_{h=0}^{\infty} \mathbf{1}_{\{0\le h\le h_i'-1\}} \bar{r}_i^{(h+h_i)} \,\middle|\, \bar{s}^{(h_i)}, \bar{a}_i^{(h_i)}, \bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)}), h_i' \sim \textsc{Geometric}(1-\gamma)\right] \\
&\overset{(b)}{=} \sum_{h=0}^{\infty} \mathbb{E}\left[\mathbb{E}_{h_i'}\left[\mathbf{1}_{\{0\le h\le h_i'-1\}}\right] \bar{r}_i^{(h+h_i)} \,\middle|\, \bar{s}^{(h_i)}, \bar{a}_i^{(h_i)}, \bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)})\right] \\
&\overset{(c)}{=} \sum_{h=0}^{\infty} \mathbb{E}\left[\gamma^h \bar{r}_i^{(h+h_i)} \,\middle|\, \bar{s}^{(h_i)}, \bar{a}_i^{(h_i)}, \bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)})\right] \\
&= \mathbb{E}_{\bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)})} \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h \bar{r}_i^{(h+h_i)} \,\middle|\, \bar{s}^{(h_i)}, \bar{a}_i^{(h_i)}, \bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)})\right] \\
&= \mathbb{E}_{\bar{a}_{-i}^{(h_i)} \sim \pi_{-i}^{(t)}(\cdot \,|\, \bar{s}^{(h_i)})}\left[Q_i^{(t)}(\bar{s}^{(h_i)}, \bar{a}_i^{(h_i)}, \bar{a}_{-i}^{(h_i)})\right] \\
&= \bar{Q}_i^{(t)}(\bar{s}^{(h_i)}, \bar{a}_i^{(h_i)})
\end{aligned}
$$

where in (a) we change the range of index h while using the same initial state and action, (b) is due to the tower property, (c) follows that $\mathbb{E}_{h_i'}\left[\mathbf{1}_{\{0\le h\le h_i'-1\}}\right] = 1-(1-(1-p)^h) = \gamma^h$, where $p = 1-\gamma$, and we also apply the monotone convergence and dominated convergence theorems for swapping the sum and the expectation.

C.2. Proof of Theorem 3

We apply Lemma 2 to the potential function $\Phi^{\pi}(\rho)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 2, where ρ is the initial state distribution. We use the shorthand $\Phi^{(t)}(\rho)$ for $\Phi^{\pi^{(t)}}(\rho)$, the value of potential function at policy $\pi^{(t)}$. The proof extends Lemma 3 by accounting for the statistical error in Assumption 2.

Lemma 6 (Policy improvement: Markov potential games)**.** *Let Assumption 1 hold. In Algorithm 2, the potential function $\Phi^{\pi}(\rho)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ satisfies*

$$
\begin{aligned}
\text{(i)}\ \ \Phi^{(t+1)}(\rho) - \Phi^{(t)}(\rho) &\ge \frac{1}{4\eta(1-\gamma)} \sum_{i=1}^{N} \sum_{s} d_{\rho}^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2 - \frac{4\eta^2 A W^2 N^2}{(1-\gamma)^3} \\
&\quad - \frac{\eta \kappa_{\rho} A}{(1-\gamma)^2 \xi} \sum_{i=1}^{N} L_i^{(t)}(\widehat{w}_i^{(t)}) \\
\text{(ii)}\ \ \Phi^{(t+1)}(\rho) - \Phi^{(t)}(\rho) &\ge \frac{1}{4\eta(1-\gamma)} \sum_{i=1}^{N} \sum_{s} d_{\rho}^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left(1 - \frac{4\eta \kappa_{\rho}^3 N A}{(1-\gamma)^4}\right) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2 \\
&\quad - \frac{\eta \kappa_{\rho} A}{(1-\gamma)^2 \xi} \sum_{i=1}^{N} L_i^{(t)}(\widehat{w}_i^{(t)})
\end{aligned}
$$

where η is the stepsize, N is the number of players, A is the size of one player's action space, W is the 2-norm bound of $\widehat{w}_i^{(t)}$, and κ_{ρ} is the distribution mismatch coefficient relative to ρ (see κ_{ρ} in Definition 1).

Proof of Lemma 6. We let $\pi' = \pi^{(t+1)}$ and $\pi = \pi^{(t)}$ for brevity. We first express $\Phi^{(t+1)}(\rho) - \Phi^{(t)}(\rho) = \textbf{Diff}_{\alpha} + \textbf{Diff}_{\beta}$, where $\textbf{Diff}_{\alpha}$ and $\textbf{Diff}_{\beta}$ are given as those in (12).

Bounding Diff$_\alpha$. By the property of the potential function $\Phi^\pi(\rho)$ and Lemma 1,

$$\begin{aligned}\Phi^{\pi_i', \pi_{-i}}(\rho) - \Phi^{\pi}(\rho) &= V_i^{\pi_i', \pi_{-i}}(\rho) - V_i^{\pi}(\rho) \\ &= \frac{1}{1-\gamma}\sum_{s, a_i} d_\rho^{\pi_i', \pi_{-i}}(s)\left(\pi_i'(a_i \,|\, s) - \pi_i(a_i \,|\, s)\right) \bar{Q}_i^{\pi_i, \pi_{-i}}(s, a_i).\end{aligned}$$

The optimality of $\pi_i' = \pi_i^{(t+1)}$ in line 14 of Algorithm 2 leads to

$$\left\langle \pi_i'(\cdot \,|\, s), \widehat{Q}_i^{(t)}(s, \cdot)\right\rangle_{\mathcal{A}_i} - \frac{1}{2\eta}\left\|\pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s)\right\|^2 \geq \left\langle \pi_i(\cdot \,|\, s), \widehat{Q}_i^{(t)}(s, \cdot)\right\rangle_{\mathcal{A}_i}. \tag{20}$$

Hence,

$$\begin{aligned}\Phi^{\pi_i', \pi_{-i}}(\rho) - \Phi^{\pi}(\rho) \geq\ & \frac{1}{2\eta(1-\gamma)}\sum_s d_\rho^{\pi_i', \pi_{-i}}(s)\left\|\pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s)\right\|^2 \\ & + \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_i', \pi_{-i}}(s)\left\langle \pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s), \bar{Q}_i^{\pi_i, \pi_{-i}}(s, \cdot) - \widehat{Q}_i^{(t)}(s, \cdot)\right\rangle_{\mathcal{A}_i}.\end{aligned}$$

Therefore,

$$\begin{aligned}\textbf{Diff}_\alpha \geq\ & \frac{1}{2\eta(1-\gamma)}\sum_{i=1}^N\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s)\right\|^2 \\ & + \frac{1}{1-\gamma}\sum_{i=1}^N\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left\langle \left(\pi_i^{(t+1)} - \pi_i^{(t)}\right)(\cdot \,|\, s), \bar{Q}_i^{(t)}(s, \cdot) - \widehat{Q}_i^{(t)}(s, \cdot)\right\rangle_{\mathcal{A}_i}.\end{aligned}$$

However,

$$\begin{aligned}&\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left\langle \left(\pi_i^{(t+1)} - \pi_i^{(t)}\right)(\cdot \,|\, s), \bar{Q}_i^{(t)}(s, \cdot) - \widehat{Q}_i^{(t)}(s, \cdot)\right\rangle_{\mathcal{A}_i} \\ \overset{(a)}{\geq}\ & -\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left(\frac{1}{2\eta'}\left\|\left(\pi_i^{(t+1)} - \pi_i^{(t)}\right)(\cdot \,|\, s)\right\|^2 + \frac{\eta'}{2}\left\|\bar{Q}_i^{(t)}(s, \cdot) - \widehat{Q}_i^{(t)}(s, \cdot)\right\|^2\right) \\ \overset{(b)}{=}\ & -\frac{1}{4\eta}\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left\|\left(\pi_i^{(t+1)} - \pi_i^{(t)}\right)(\cdot \,|\, s)\right\|^2 - \eta\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left\|\bar{Q}_i^{(t)}(s, \cdot) - \widehat{Q}_i^{(t)}(s, \cdot)\right\|^2\end{aligned}$$

where (a) follows the inequality $\langle x, y\rangle \leq \frac{\|x\|^2}{2\eta'} + \frac{\eta'\|y\|^2}{2}$ for $\eta' > 0$, and we choose $\eta' = 2\eta$ in (b).

Therefore,

$$\begin{aligned}\textbf{Diff}_\alpha \geq\ & \frac{1}{4\eta(1-\gamma)}\sum_{i=1}^N\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s)\right\|^2 \\ & - \frac{\eta}{1-\gamma}\sum_{i=1}^N\sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)\left\|\bar{Q}_i^{(t)}(s, \cdot) - \widehat{Q}_i^{(t)}(s, \cdot)\right\|^2.\end{aligned} \tag{21}$$

Bounding Diff$_\beta$. For simplicity, we denote $\widetilde{\pi}_{-ij}$ as the joint policy of players $N\backslash\{i, j\}$ where players $< i$ and $i \sim j$ use π and players $> j$ use π'. As done in the proof of Lemma 3, we can bound each summand in Diff$_\beta$ except for the last step from (c) to (d),

$$\begin{aligned}&\Phi^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j'}(\rho) - \Phi^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j'}(\rho) - \Phi^{\widetilde{\pi}_{-ij}, \pi_i', \pi_j}(\rho) + \Phi^{\widetilde{\pi}_{-ij}, \pi_i, \pi_j}(\rho) \\ \overset{(c)}{\geq}\ & -\frac{1}{(1-\gamma)^3}\left(\max_s\left\|\pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s)\right\|_1\right)\left(\max_s\left\|\pi_j'(\cdot \,|\, s) - \pi_j(\cdot \,|\, s)\right\|_1\right) \\ & -\frac{1}{(1-\gamma)^2}\left(\max_s\left\|\pi_j'(\cdot \,|\, s) - \pi_j(\cdot \,|\, s)\right\|_1\right)\left(\max_s\left\|\pi_i'(\cdot \,|\, s) - \pi_i(\cdot \,|\, s)\right\|_1\right) \\ \overset{(d)}{\geq}\ & -\frac{8\eta^2 AW^2}{(1-\gamma)^3}\end{aligned}$$

where (d) follows a direct result from the optimality of $\pi_j^{(t+1)}$ given by (20),

$$\left\|\pi_j^{(t+1)}(\cdot\,|\,s) - \pi_j^{(t)}(\cdot\,|\,s)\right\| \;\le\; 2\eta \left\|\widehat{Q}_i^{(t)}(s,\cdot)\right\| \;\le\; 2\eta\, W$$

and that $\|\cdot\|_1 \le \sqrt{A}\|\cdot\|$. Therefore,

$$\textbf{Diff}_\beta \;\ge\; -\frac{4\eta^2 A W^2 N^2}{(1-\gamma)^3}. \tag{22}$$

We now complete the proof of (i) by combining (21) and (22) and we also employ that

$$\begin{aligned}
&\sum_s d_\rho^{\pi_i^{(t+1)},\,\pi_{-i}^{(t)}}(s)\left\|\bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot)\right\|^2 \\
\overset{(a)}{\le}\;& \frac{\kappa_\rho}{1-\gamma}\sum_s d_\rho^{\pi_i^{(t)},\,\pi_{-i}^{(t)}}(s)\left\|\bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot)\right\|^2 \\
\overset{(b)}{\le}\;& \frac{\kappa_\rho A}{(1-\gamma)\xi} L_i^{(t)}(\widehat{w}_i^{(t)})
\end{aligned}$$

where (a) follows the definition of κ_ρ and (b) is the definition of $L_i^{(t)}(\widehat{w}_i^{(t)})$:

$$L_i^{(t)}(\widehat{w}_i^{(t)}) \;:=\; \mathbb{E}_{s\sim d_\rho^{(t)},\,a_i\sim\pi_i^{(t)}(\cdot|s)}\left[\left(\bar{Q}_i^{(t)}(s,a_i) - \widehat{Q}_i^{(t)}(s,a_i)\right)^2\right] \;\ge\; \frac{\xi}{A}\mathbb{E}_{s\sim d_\rho^{(t)}}\sum_{a_i}\left(\bar{Q}_i^{(t)}(s,a_i) - \widehat{Q}_i^{(t)}(s,a_i)\right)^2.$$

Alternatively, as done in Lemma 3, we can apply Lemma 21 to each summand of $\textbf{Diff}_\beta$ and show that

$$\textbf{Diff}_\beta \;\ge\; -\frac{2\kappa_\rho^3 N A}{(1-\gamma)^5}\sum_{i=1}^N\sum_s d_\rho^{\pi_i^{(t+1)},\,\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t)}(\cdot\,|\,s) - \pi_i^{(t+1)}(\cdot\,|\,s)\right\|^2.$$

Combining the inequality above with (21) finishes the proof of (ii). □

Proof of Theorem 3. By the optimality of $\pi_i^{(t+1)}$ in line 14 of Algorithm 2,

$$\left\langle (1-\xi)\pi_i'(\cdot\,|\,s) + \frac{\xi}{A}\mathbf{1} - \pi_i^{(t+1)}(\cdot\,|\,s),\, \eta\widehat{Q}_i^{(t)}(s,\cdot) - \pi_i^{(t+1)}(\cdot\,|\,s) + \pi_i^{(t)}(\cdot\,|\,s)\right\rangle_{\mathcal{A}_i} \;\le\; 0, \quad \text{for any } \pi_i' \in \Pi_i.$$

which leads to

$$\begin{aligned}
&\left\langle \pi_i'(\cdot\,|\,s) - \pi_i^{(t+1)}(\cdot\,|\,s),\, \eta\widehat{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i} \\
&\le \left\langle \pi_i'(\cdot\,|\,s) - \pi_i^{(t+1)}(\cdot\,|\,s),\, \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s)\right\rangle_{\mathcal{A}_i} \qquad (23)\\
&\quad + \frac{\xi}{1-\xi}\left\langle \pi^{(t+1)}(\cdot|s) - \frac{1}{A}\mathbf{1}, \eta\widehat{Q}_i^{(t)}(s,\cdot) - \pi_i^{(t+1)}(\cdot\,|\,s) + \pi_i^{(t)}(\cdot\,|\,s)\right\rangle \\
&\lesssim \left\|\pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s)\right\| + \eta\xi W \qquad (24)
\end{aligned}$$

where the last inequality is because of $\|\widehat{Q}_i^{(t)}(s,\cdot)\| \le W$ and $\xi \le \frac{1}{2}$. Hence, if $\eta \le \frac{1}{W}$, then for any $\pi_i' \in \Pi_i$,

$$\begin{aligned}
&\left\langle \pi_i'(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s),\, \bar{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i} \\
=\;& \left\langle \pi_i'(\cdot\,|\,s) - \pi_i^{(t+1)}(\cdot\,|\,s),\, \widehat{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i} + \left\langle \pi_i^{(t+1)}(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s),\, \widehat{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i} \\
&+ \left\langle \pi_i'(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s),\, \bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i} \\
\overset{(a)}{\lesssim}\;& \frac{1}{\eta}\left\|\pi_i^{(t+1)}(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s)\right\| + \xi W + \left\|\pi_i^{(t+1)}(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s)\right\|\left\|\widehat{Q}_i^{(t)}(s,\cdot)\right\| \\
&+ \left\langle \pi_i'(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s),\, \bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i} \\
\overset{(b)}{\lesssim}\;& \frac{1}{\eta}\left\|\pi_i^{(t+1)}(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s)\right\| + \xi W \\
&+ \left\langle \pi_i'(\cdot\,|\,s) - \pi_i^{(t)}(\cdot\,|\,s),\, \bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot)\right\rangle_{\mathcal{A}_i}.
\end{aligned}$$

where we apply (24) and the Cauchy-Schwarz inequality in (a), and (b) is because $\|\widehat{Q}_i^{(t)}(s,\cdot)\| \le W$ and $\eta \le \frac{1}{W}$. As done in the proof of Theorem 1, the different steps begin from (b) in (17),

$$
\begin{aligned}
&\sum_{t=1}^{T} \max_i \left(\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho) \right) \\
&\overset{(b)}{\lesssim} \frac{1}{\eta(1-\gamma)} \sum_{t=1}^{T} \sum_s d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\| + \frac{\xi TW}{1-\gamma} \\
&\quad + \frac{1}{1-\gamma} \sum_{t=1}^{T} \sum_s d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s) \big\langle \pi_i'(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s),\, \bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot) \big\rangle_{\mathcal{A}_i} \\
&\overset{(c)}{\lesssim} \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_s \sqrt{d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \times d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s)} \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\| + \frac{\xi TW}{1-\gamma} \\
&\quad + \frac{\kappa_\rho}{1-\gamma} \left| \sum_{t=1}^{T} \sum_s d_\rho^{\pi^{(t)}}(s) \big\langle \pi_i'(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s),\, \bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot) \big\rangle_{\mathcal{A}_i} \right| \\
&\overset{(d)}{\le} \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T} \sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2} \\
&\quad + \frac{\xi TW}{1-\gamma} + \frac{\kappa_\rho}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{AL_i^{(t)}(\widehat{w}_i^{(t)})}{\xi}} \\
&\overset{(e)}{\le} \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_s d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2} \\
&\quad + \frac{\xi TW}{1-\gamma} + \frac{\kappa_\rho}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{AL_i^{(t)}(\widehat{w}_i^{(t)})}{\xi}}
\end{aligned}
\tag{25}
$$

where we slightly abuse the notation π_i' in (b) to represent $\operatorname{argmax}_{\pi_i'}$ and i represents argmax_i as in (17), (c) is due to the definition of the distribution mismatch coefficient (see it in Definition 1):

$$
\frac{d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s)}{d_\rho^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \le \frac{d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s)}{(1-\gamma)\rho(s)} \le \frac{\kappa_\rho}{1-\gamma},
$$

(d) follows the Cauchy–Schwarz inequality, the inequality $\sqrt{\sum_i x_i} \le \sum_i \sqrt{x_i}$ for any $x_i \ge 0$, the Jensen's inequality, and the definition of $L_i^{(t)}(\widehat{w}_i^{(t)})$,

$$
\begin{aligned}
&\left| \sum_s d_\rho^{\pi^{(t)}}(s) \big\langle \pi_i'(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s),\, \bar{Q}_i^{(t)}(s,\cdot) - \widehat{Q}_i^{(t)}(s,\cdot) \big\rangle_{\mathcal{A}_i} \right| \\
&\lesssim \sqrt{\sum_s d_\rho^{\pi^{(t)}}(s)} \sqrt{\sum_s d_\rho^{\pi^{(t)}}(s) \sum_{a_i} \left(\bar{Q}_i^{(t)}(s,a_i) - \widehat{Q}_i^{(t)}(s,a_i) \right)^2} \\
&\le \sqrt{\frac{AL_i^{(t)}(\widehat{w}_i^{(t)})}{\xi}}
\end{aligned}
$$

where $L_i^{(t)}(\widehat{w}_i^{(t)}) := \mathbb{E}_{s \sim d_\rho^{(t)}, a \sim \pi_i^{(t)}(\cdot|s)} \big[\big(\bar{Q}_i^{(t)}(s,a_i) - \widehat{Q}_i^{(t)}(s,a_i) \big)^2 \big]$, and $\widehat{Q}_i^{(t)}(s,a_i) = \langle \phi_i(s,a_i), \widehat{w}_i^{(t)} \rangle$, and we replace i (argmax_i in (b)) in the square root term in (e) by the sum over all players.

If we proceed (25) with the first bound (i) in Lemma 6, then,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\max_{i}\left(\max_{\pi_i'}V_i^{\pi_i',\pi_{-i}^{(t)}}(\rho)-V_i^{\pi^{(t)}}(\rho)\right)\right]\\
&\overset{(a)}{\lesssim}\ \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}}\sqrt{T}\sqrt{\eta(1-\gamma)(\Phi^{(N+1)}-\Phi^{(1)})+\frac{\eta^3AW^2N^2}{(1-\gamma)^2}T+\frac{\eta^2\kappa_\rho A}{(1-\gamma)\xi}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[L_i^{(t)}(\widehat{w}_i^{(t)})\right]}\\
&\quad+\frac{\xi TW}{1-\gamma}+\frac{\kappa_\rho}{1-\gamma}\sum_{t=1}^{T}\sqrt{\frac{A\mathbb{E}\left[L_i^{(t)}(\widehat{w}_i^{(t)})\right]}{\xi}}\\
&\overset{(b)}{\lesssim}\ \sqrt{\frac{\kappa_\rho TC_\Phi}{\eta(1-\gamma)^2}}+T\sqrt{\frac{\eta\kappa_\rho AW^2N^2}{(1-\gamma)^5}}+\frac{\kappa_\rho T}{(1-\gamma)^2}\sqrt{\frac{AN\epsilon_{\text{stat}}}{\xi}}+\frac{\xi TW}{1-\gamma}
\end{aligned}$$

where we apply the first bound (i) in Lemma 6 and the telescoping sum for (a), and we use the boundedness of the potential function: $|\Phi^{\pi}-\Phi^{\pi'}|\leq C_\Phi$ for any π and π', and further simplify the bound in (f) by Assumption Assumption 2. We complete the proof of (i) by taking stepsize $\eta=\frac{(1-\gamma)^{3/2}\sqrt{C_\Phi}}{WN\sqrt{AT}}$ and exploration rate $\xi\leq\left(\frac{\kappa_\rho^2NA\epsilon_{\text{stat}}}{(1-\gamma)^2W^2}\right)^{\frac{1}{3}}$.

If we proceed (25) with the first bound (ii) in Lemma 6 with the choice of $\eta\leq\frac{(1-\gamma)^4}{16\kappa_\rho^3NA}$, then,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\max_{i}\left(\max_{\pi_i'}V_i^{\pi_i',\pi_{-i}^{(t)}}(\rho)-V_i^{\pi^{(t)}}(\rho)\right)\right]\\
&\overset{(a)}{\lesssim}\ \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}}\sqrt{T}\sqrt{\eta(1-\gamma)(\Phi^{(N+1)}-\Phi^{(1)})+\frac{\eta^2\kappa_\rho A}{(1-\gamma)\xi}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[L_i^{(t)}(\widehat{w}_i^{(t)})\right]}\\
&\quad+\frac{\xi TW}{1-\gamma}+\frac{\kappa_\rho}{1-\gamma}\sum_{t=1}^{T}\sqrt{\frac{A\mathbb{E}\left[L_i^{(t)}(\widehat{w}_i^{(t)})\right]}{\xi}}\\
&\overset{(b)}{\lesssim}\ \sqrt{\frac{\kappa_\rho TC_\Phi}{\eta(1-\gamma)^2}}+\frac{\kappa_\rho T}{(1-\gamma)^2}\sqrt{\frac{AN\epsilon_{\text{stat}}}{\xi}}+\frac{\xi TW}{1-\gamma}
\end{aligned}$$

which completes the proof if we choose $\eta=\frac{(1-\gamma)^4}{16\kappa_\rho^3NA}$ and exploration rate $\xi\leq\left(\frac{\kappa_\rho^2NA\epsilon_{\text{stat}}}{(1-\gamma)^2W^2}\right)^{\frac{1}{3}}$. □

C.3. Proof of Theorem 4

We first establish policy improvement regarding the Q-function at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 2.

Lemma 7 (Policy improvement: Markov cooperative games)**.** *For MPG (1) with identical rewards and an initial state distribution $\rho>0$, if all players independently perform the policy update in Algorithm 2 with stepsize $\eta\leq\frac{1-\gamma}{2N}$, then for any t and any s,*

$$\begin{aligned}
\mathbb{E}_{a\sim\pi^{(t+1)}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right]-\mathbb{E}_{a\sim\pi^{(t)}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right]\ \geq\ &\frac{1}{8\eta}\sum_{i=1}^{N}\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|^2\\
&-\eta\sum_{i=1}^{N}\left\|Q_i^{(t)}(s,\cdot)-\widehat{Q}_i^{(t)}(s,\cdot)\right\|^2
\end{aligned}$$

where η is the stepsize and N is the number of players,

Proof of Lemma 7. As done in the proof of Lemma 5, we let $\Psi^{\pi}:=\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right]$ and (18) holds, where

$Q^{(t)} := Q^{\pi^{(t)}}$. By taking $\pi' = \pi^{(t+1)}$ and $\pi = \pi^{(t)}$ for (18),

$$
\begin{aligned}
&\mathbb{E}_{a\sim\pi'(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \;-\; \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
=\;& \sum_{i=1}^{N}\sum_{a_i}\left(\pi_i'(a_i\,|\,s)-\pi_i(a_i\,|\,s)\right)\widehat{Q}_i^{(t)}(s,a_i) \;+\; \sum_{i=1}^{N}\sum_{a_i}\left(\pi_i'(a_i\,|\,s)-\pi_i(a_i\,|\,s)\right)\left(\bar{Q}_i^{(t)}(s,a_i)-\widehat{Q}_i^{(t)}(s,a_i)\right) \\
&+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{a_i,\,a_j}\left(\pi_i'(a_i\,|\,s)-\pi_i(a_i\,|\,s)\right)\left(\pi_j'(a_j\,|\,s)-\pi_j(a_j\,|\,s)\right)\mathbb{E}_{a_{-ij}\sim\widetilde{\pi}_{-ij}(\cdot\,|\,s)}\left[Q^{(t)}(s,a)\right] \\
\overset{(a)}{\geq}\;& \sum_{i=1}^{N}\frac{1}{2\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 \;-\; \sum_{i=1}^{N}\left(\frac{1}{2\eta'}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2+\frac{\eta'}{2}\left\|\bar{Q}_i^{(t)}(s,\cdot)-\widehat{Q}_i^{(t)}(s,\cdot)\right\|^2\right) \\
&-\frac{1}{1-\gamma}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{a_i,\,a_j}\left|\pi_i'(a_i\,|\,s)-\pi_i(a_i\,|\,s)\right|\left|\pi_j'(a_j\,|\,s)-\pi_j(a_j\,|\,s)\right| \\
\overset{(b)}{\geq}\;& \sum_{i=1}^{N}\frac{1}{4\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 \;-\; \sum_{i=1}^{N}\eta\left\|\bar{Q}_i^{(t)}(s,\cdot)-\widehat{Q}_i^{(t)}(s,\cdot)\right\|^2 \\
&-\frac{1}{2(1-\gamma)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2+\left\|\pi_j'(\cdot\,|\,s)-\pi_j(\cdot\,|\,s)\right\|^2\right) \\
=\;& \sum_{i=1}^{N}\frac{1}{4\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 \;-\; \sum_{i=1}^{N}\eta\left\|\bar{Q}_i^{(t)}(s,\cdot)-\widehat{Q}_i^{(t)}(s,\cdot)\right\|^2 \;-\; \frac{N-1}{2(1-\gamma)}\sum_{i=1}^{N}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 \\
\overset{(c)}{\geq}\;& \sum_{i=1}^{N}\frac{1}{8\eta}\left\|\pi_i'(\cdot\,|\,s)-\pi_i(\cdot\,|\,s)\right\|^2 \;-\; \sum_{i=1}^{N}\eta\left\|\bar{Q}_i^{(t)}(s,\cdot)-\widehat{Q}_i^{(t)}(s,\cdot)\right\|^2
\end{aligned}
$$

where (a) is due to the optimality condition (20), the inequality $\langle x,y\rangle \leq \frac{\|x\|^2}{2\eta'}+\frac{\eta'\|y\|^2}{2}$ for $\eta'>0$, and $Q^{(t)}(s,a)\leq\frac{1}{1-\gamma}$, (b) is due to $\langle x,y\rangle\leq\frac{\|x\|^2+\|y\|^2}{2}$ and $\eta'=2\eta$, and (c) follows the choice of $\eta\leq\frac{1-\gamma}{4N}$. □

Proof of Theorem 4. By Lemma 1 and Lemma 7,

$$
\begin{aligned}
V^{(t+1)}(\rho)\;-\;V^{(t)}(\rho) \;&=\; \frac{1}{1-\gamma}\sum_{s,\,a}d_\rho^{\pi^{(t+1)}}(s)\left(\pi^{(t+1)}(a\,|\,s)-\pi^{(t)}(a\,|\,s)\right)Q^{(t)}(s,a) \\
&\geq\; \frac{1}{8\eta(1-\gamma)}\sum_{i=1}^{N}\sum_{s}d_\rho^{\pi^{(t+1)}}(s)\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|^2 \\
&\quad-\frac{\eta}{1-\gamma}\sum_{i=1}^{N}\sum_{s}d_\rho^{\pi^{(t+1)}}(s)\left\|\bar{Q}_i^{(t)}(s,\cdot)-\widehat{Q}_i^{(t)}(s,\cdot)\right\|^2 \\
&\geq\; \frac{1}{8\eta(1-\gamma)}\sum_{i=1}^{N}\sum_{s}d_\rho^{\pi^{(t+1)}}(s)\left\|\pi_i^{(t+1)}(\cdot\,|\,s)-\pi_i^{(t)}(\cdot\,|\,s)\right\|^2 \;-\; \frac{\eta\kappa_\rho A}{\xi(1-\gamma)^2}\sum_{i=1}^{N}L_i^{(t)}(\widehat{w}_i^{(t)}).
\end{aligned}
$$

where the last inequality is due to that

$$
\begin{aligned}
&\sum_{s}d_\rho^{\pi^{(t+1)}}(s)\left\|\bar{Q}_i^{\pi^{(t)}}(s,\cdot)-\widehat{Q}_i^{\pi^{(t)}}(s,\cdot)\right\|^2 \\
\overset{(a)}{\leq}\;& \frac{\kappa_\rho}{1-\gamma}\sum_{s}d_\rho^{\pi^{(t)}}(s)\left\|\bar{Q}_i^{\pi^{(t)}}(s,\cdot)-\widehat{Q}_i^{\pi^{(t)}}(s,\cdot)\right\|^2 \\
=\;& \frac{\kappa_\rho A}{(1-\gamma)\xi}\sum_{s}d_\rho^{(t)}(s)\sum_{a_i}\frac{\xi}{A}\left(\bar{Q}_i^{(t)}(s,a_i)-\langle\phi_i(s,a_i),\widehat{w}_i^{(t)}\rangle\right)^2 \\
\overset{(b)}{\leq}\;& \frac{\kappa_\rho A}{(1-\gamma)\xi}\,\mathbb{E}_{s\sim d_\rho^{(t)},\,a_i\sim\pi_i^{(t)}(\cdot\,|\,s)}\left[\left(\bar{Q}_i^{(t)}(s,a_i)-\langle\phi_i(s,a_i),\widehat{w}_i^{(t)}\rangle\right)^2\right] \\
=\;& \frac{\kappa_\rho A}{(1-\gamma)\xi}L_i^{(t)}(\widehat{w}_i^{(t)})
\end{aligned}
$$

where (a) follows the definition of κ_ρ and (b) is the definition of $L_i^{(t)}(\widehat{w}_i^{(t)})$.

By the same argument as the proof of Theorem 3,

$$\begin{aligned}
&\sum_{t=1}^{T} \max_i \left(\max_{\pi_i'} V^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V^{\pi^{(t)}}(\rho) \right) \\
&\lesssim \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_s d_\rho^{\pi_i', \pi_{-i}^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_s d_\rho^{\pi^{(t+1)}}(s) \left\| \pi_i^{(t+1)}(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\|^2} \\
&\quad + \frac{\xi T W}{1-\gamma} + \frac{\kappa_\rho}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{A L_i^{(t)}(\widehat{w}_i^{(t)})}{\xi}}.
\end{aligned}$$

By taking expectation and the Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \max_i \left(\max_{\pi_i'} V^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V^{\pi^{(t)}}(\rho) \right) \right] \\
&\lesssim \frac{\sqrt{\kappa_\rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{T} \sqrt{8\eta(1-\gamma)(V^{(N+1)} - V^{(1)}) + \frac{8\eta^2 \kappa_\rho A}{(1-\gamma)\xi} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}\left[L_i^{(t)}(\widehat{w}_i^{(t)}) \right]} \\
&\quad + \frac{\xi T W}{1-\gamma} + \frac{\kappa_\rho}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{A \mathbb{E}\left[L_i^{(t)}(\widehat{w}_i^{(t)}) \right]}{\xi}} \\
&\lesssim \sqrt{\frac{8\kappa_\rho T}{\eta(1-\gamma)^3}} + \kappa_\rho T \sqrt{\frac{8AN}{(1-\gamma)^4 \xi} \epsilon_{\text{stat}}} + \frac{\xi T W}{1-\gamma}.
\end{aligned}$$

We complete the proof by taking stepsize $\eta = \frac{1-\gamma}{2NA}$, exploration rate $\xi \le \left(\frac{\kappa_\rho^2 NA \epsilon_{\text{stat}}}{(1-\gamma)^2 W^2} \right)^{\frac{1}{3}}$, and using $V^{(N+1)} - V^{(1)} \le \frac{1}{1-\gamma}$. □

C.4. Sample complexity

We present our sample complexity guarantees for Algorithm 2 in which the regression problem (7) in each iteration is approximately solved by the stochastic projected gradient descent (38). We measure the sample complexity by the total number of trajectory samples TK, where T is the number of iterations and K is the batch size of trajectories.

Corollary 1 (Sample complexity for Markov potential games)**.** *Assume the setting in Theorem 3 except for Assumption 2. Suppose we compute $\widehat{w}_i^{(t)} := \frac{1}{K} \sum_{k=1}^{K} \beta_k^{(K)} w_i^{(k)}$ via a stochastic projected gradient descent (38) with stepsize $\lambda^{(k)} = \frac{2}{2+k}$ and $\beta_k^{(K)} = \frac{1/\lambda^{(k)}}{\sum_{r=1}^{K} 1/\lambda^{(r)}}$. Then, if we choose stepsize $\eta = \frac{(1-\gamma)^{3/2}\sqrt{C_\Phi}}{WN\sqrt{AT}}$ and exploration rate $\xi = \min\left(\left(\frac{\kappa_\rho^2 AN d}{(1-\gamma)^4 K} \right)^{\frac{1}{3}}, \frac{1}{2} \right)$, then,*

$$\mathbb{E}\left[\text{Nash-Regret}(T) \right] \lesssim \frac{\sqrt{\kappa_\rho W N} (AC_\Phi)^{\frac{1}{4}}}{(1-\gamma)^{\frac{7}{2}} T^{\frac{1}{4}}} + \frac{W(\kappa_\rho^2 AN d)^{\frac{1}{3}}}{(1-\gamma)^{\frac{7}{3}} K^{\frac{1}{3}}}.$$

Furthermore, if we choose stepsize $\eta = \frac{(1-\gamma)^4}{16\kappa_\rho^3 NA}$ and exploration rate $\xi = \min\left(\left(\frac{\kappa_\rho^2 AN d}{(1-\gamma)^4 K} \right)^{\frac{1}{3}}, \frac{1}{2} \right)$, then,

$$\mathbb{E}\left[\text{Nash-Regret}(T) \right] \lesssim \frac{\kappa_\rho^2 \sqrt{ANC_\Phi}}{(1-\gamma)^3 \sqrt{T}} + \frac{W(\kappa_\rho^2 AN d)^{\frac{1}{3}}}{(1-\gamma)^{\frac{7}{3}} K^{\frac{1}{3}}}.$$

Moreover, their sample complexity guarantees are $TK = O(\frac{1}{\epsilon^7})$ or $TK = O(\frac{1}{\epsilon^5})$, respectively, for obtaining an ϵ-Nash equilibrium.

Proof of Corollary 1. By the unbiased estimate in Appendix C.1, the stochastic gradient $\widehat{\nabla}_i^{(t)}$ in (38) is also unbiased. We note the variance of the stochastic gradient is bounded by $\frac{1}{(1-\gamma)^2}$. By Lemma 23, if we choose $\lambda^{(k)} = \frac{2}{2+k}$ and $\beta_k^{(K)} = \frac{1/\lambda^{(k)}}{\sum_{r=1}^K 1/\lambda^{(r)}}$, then

$$\mathbb{E}\left[L_i^{(t)}(\widehat{w}_i^{(t)})\right] - L_i^{(t)}(w_i^{(t)}) \leq \frac{dW^2}{(1-\gamma)^2 K}$$

where $L_i^{(t)}(w_i^{(t)}) = 0$. by Assumption 1. Therefore, substitution of $\epsilon_{\text{stat}} \leq \frac{dW^2}{(1-\gamma)^2 K}$ into Theorem 3 yields desired results.

Finally, we let the upper bound on Nash-Regret(T) be $\epsilon > 0$ and calculate the sample complexity $TK = O(\frac{1}{\epsilon^7})$ or $TK = O(\frac{1}{\epsilon^5})$, respectively. □

Corollary 2 (Sample complexity for Markov cooperative games)**.** *Assume the setting in Theorem 4 except for Assumption Assumption 2. Suppose we compute $\widehat{w}_i^{(t)} := \frac{1}{K}\sum_{k=1}^K \beta_k^{(K)} w_i^{(k)}$ via a stochastic projected gradient descent (38) with stepsize $\lambda^{(k)} = \frac{2}{2+k}$ and $\beta_k^{(K)} = \frac{1/\lambda^{(k)}}{\sum_{r=1}^K 1/\lambda^{(r)}}$. Then, if we choose stepsize $\eta = \frac{1-\gamma}{WNA\sqrt{T}}$ and exploration rate $\xi = \min\left(\left(\frac{\kappa_\rho^2 AN}{(1-\gamma)^4 K}\right)^{\frac{1}{3}}, \frac{1}{2}\right)$, then,*

$$\mathbb{E}\left[\,\text{Nash-Regret}(T)\,\right] \lesssim \frac{\sqrt{\kappa_\rho\, AN}}{(1-\gamma)^2\sqrt{T}} + \frac{W(\,\kappa_\rho^2\, ANd\,)^{\frac{1}{3}}}{(1-\gamma)^{\frac{7}{3}} K^{\frac{1}{3}}}.$$

Moreover, the sample complexity guarantee is $TK = O(\frac{1}{\epsilon^5})$ for obtaining an ϵ-Nash equilibrium.

Proof of Corollary 2. The proof follows the proof steps of Corollary 1 above. □

D. Proofs for Section 6

In this section, we prove Theorem 5 and Theorem 6 in Appendix D.1 and Appendix D.2, respectively.

D.1. Proof of Theorem 5

It is convenient to introduce an auxiliary sequence $\{\alpha^{(t,\tau)}\}_{\tau=0}^{\infty}$ associated with the learning rate $\{\alpha^{(t)}\}_{t=1}^{\infty}$,

$$\alpha^{(t,\tau)} := \begin{cases} \prod_{j=1}^{t}(1-\alpha^{(j)}), & \text{for } \tau = 0 \\ \alpha^{(\tau)} \prod_{j=\tau+1}^{t}(1-\alpha^{(j)}), & \text{for } 1 \leq \tau \leq t \\ 0, & \text{for } \tau > t. \end{cases} \tag{26}$$

It is straightforward to verify that $\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} = 1$ for $t \geq 1$.

Lemma 8. *In Algorithm 3, $\mathcal{V}_s^{(t)} = \sum_{\tau=1}^{t} \alpha^{(t,\tau)} (x_s^{(\tau)})^\top \mathcal{Q}_s^{(\tau)} y_s^{(\tau)}$ for all s, t.*

Proof of Lemma 8. We prove it by induction. When $t = 0$ and $t = 1$, it holds trivially by noting that $\mathcal{V}_s^{(0)} = 0$ and $\alpha^{(1,1)} = \alpha^{(1)}$. Assume that it holds for $0, 1, \ldots, t-1$. By the update rule for $\mathcal{V}_s^{(t)}$ in Algorithm 3,

$$\begin{aligned} \mathcal{V}_s^{(t)} &= (1-\alpha^{(t)})\mathcal{V}_s^{(t-1)} + \alpha^{(t)}(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \\ &\overset{(a)}{=} (1-\alpha^{(t)}) \sum_{\tau=1}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(\tau)})^\top \mathcal{Q}_s^{(\tau)} y_s^{(\tau)} + \alpha^{(t,t)} (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \\ &\overset{(b)}{=} \sum_{\tau=1}^{t} \alpha^{(t,\tau)} (x_s^{(\tau)})^\top \mathcal{Q}_s^{(\tau)} y_s^{(\tau)} \end{aligned}$$

where (a) follows the induction hypothesis and (b) is due to the definition of $\alpha^{(t,\tau)}$. □

Lemma 9. *In Algorithm 3, for every state s and time $t \geq 1$,*

$$(x_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t+1)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \;\geq\; \frac{15}{16\eta}\left\|z_s^{(t+1)} - \bar{z}_s^{(t+1)}\right\|^2 + \frac{7}{16\eta}\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 - \frac{9}{16\eta}\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2$$

where $z_s^{(t)} = (x_s^{(t)}, y_s^{(t)})$ and $\bar{z}_s^{(t)} = (\bar{x}_s^{(t)}, \bar{y}_s^{(t)})$.

Proof of Lemma 9. We decompose the difference into three terms:

$$\begin{aligned}
&(x_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t+1)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \\
&= \underbrace{(x_s^{(t+1)} - x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}}_{\textbf{Diff}_x} + \underbrace{(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} (y_s^{(t+1)} - y_s^{(t)})}_{\textbf{Diff}_y} + \underbrace{(x_s^{(t+1)} - x_s^{(t)})^\top \mathcal{Q}_s^{(t)} (y_s^{(t+1)} - y_s^{(t)})}_{\textbf{Diff}_{xy}}.
\end{aligned}$$

We next deal with $\textbf{Diff}_x$, $\textbf{Diff}_y$, and $\textbf{Diff}_{xy}$, separately.

Bounding $\textbf{Diff}_x$. The optimality of $x_s^{(t+1)}$ implies that for any $x_s' \in \Delta(\mathcal{A}_1)$,

$$(x_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} - \frac{1}{2\eta}\left\|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\right\|^2 \;\geq\; (x_s')^\top \mathcal{Q}_s^{(t)} y_s^{(t)} - \frac{1}{2\eta}\left\|x_s' - \bar{x}_s^{(t+1)}\right\|^2 + \frac{1}{2\eta}\left\|x_s' - x_s^{(t+1)}\right\|^2$$

which implies that, by taking $x_s' = \bar{x}_s^{(t+1)}$,

$$(x_s^{(t+1)} - \bar{x}_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \;\geq\; \frac{1}{\eta}\left\|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\right\|^2. \tag{27}$$

The optimality of $\bar{x}_s^{(t+1)}$ implies that for any $x_s' \in \Delta(\mathcal{A}_1)$,

$$(\bar{x}_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} - \frac{1}{2\eta}\left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\|^2 \;\geq\; (x_s')^\top \mathcal{Q}_s^{(t)} y_s^{(t)} - \frac{1}{2\eta}\left\|x_s' - \bar{x}_s^{(t)}\right\|^2 + \frac{1}{2\eta}\left\|x_s' - \bar{x}_s^{(t+1)}\right\|^2$$

which implies that, by taking $x_s' = x_s^{(t)}$,

$$(\bar{x}_s^{(t+1)} - x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \;\geq\; \frac{1}{2\eta}\left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\|^2 - \frac{1}{2\eta}\left\|x_s^{(t)} - \bar{x}_s^{(t)}\right\|^2. \tag{28}$$

Combining the two inequalities above yields

$$\textbf{Diff}_x \;=\; (x_s^{(t+1)} - x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \;\geq\; \frac{1}{\eta}\left\|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\right\|^2 + \frac{1}{2\eta}\left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\|^2 - \frac{1}{2\eta}\left\|x_s^{(t)} - \bar{x}_s^{(t)}\right\|^2. \tag{29}$$

Bounding $\textbf{Diff}_y$. Similarly,

$$\textbf{Diff}_y \;=\; (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} (y_s^{(t+1)} - y_s^{(t)}) \;\geq\; \frac{1}{\eta}\left\|y_s^{(t+1)} - \bar{y}_s^{(t+1)}\right\|^2 + \frac{1}{2\eta}\left\|\bar{y}_s^{(t+1)} - \bar{y}_s^{(t)}\right\|^2 - \frac{1}{2\eta}\left\|y_s^{(t)} - \bar{y}_s^{(t)}\right\|^2. \tag{30}$$

Bounding $\textbf{Diff}_{xy}$. By the AM-GM and Cauchy-Schwarz inequalities,

$$\begin{aligned}
\textbf{Diff}_{xy} \;&\geq\; -\frac{\sqrt{A}}{2(1-\gamma)}\left\|x_s^{(t+1)} - x_s^{(t)}\right\|^2 - \frac{\sqrt{A}}{2(1-\gamma)}\left\|y_s^{(t+1)} - y_s^{(t)}\right\|^2 \\
&\overset{(a)}{\geq}\; -\frac{3\sqrt{A}}{2(1-\gamma)}\Bigg(\left\|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\right\|^2 + \left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\|^2 + \left\|\bar{x}_s^{(t)} - x_s^{(t)}\right\|^2 \\
&\qquad\qquad + \left\|y_s^{(t+1)} - \bar{y}_s^{(t+1)}\right\|^2 + \left\|\bar{y}_s^{(t+1)} - \bar{y}_s^{(t)}\right\|^2 + \left\|\bar{y}_s^{(t)} - y_s^{(t)}\right\|^2\Bigg) \\
&\overset{(b)}{\geq}\; -\frac{1}{16\eta}\Bigg(\left\|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\right\|^2 + \left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\|^2 + \left\|\bar{x}_s^{(t)} - x_s^{(t)}\right\|^2 \\
&\qquad\qquad + \left\|y_s^{(t+1)} - \bar{y}_s^{(t+1)}\right\|^2 + \left\|\bar{y}_s^{(t+1)} - \bar{y}_s^{(t)}\right\|^2 + \left\|\bar{y}_s^{(t)} - y_s^{(t)}\right\|^2\Bigg)
\end{aligned}$$

where (a) follows $\|x+y+z\|^2 \leq 3\|x\|^2 + 3\|y\|^2 + 3\|z\|^2$ and (b) is by $\eta \leq \frac{1-\gamma}{32\sqrt{A}}$.

Finally, we complete the proof by summing up the bounds above for $\textbf{Diff}_x$, $\textbf{Diff}_y$, and $\textbf{Diff}_{xy}$. □

Lemma 10. *In Algorithm 3, for all t and s, the following two inequalities hold:*

(i) $\mathcal{V}_s^{(t)} \geq \mathcal{V}_s^{(t-1)}$;

(ii) $(x_s^{(t+1)})^\top \mathcal{Q}_s^{(t+1)} y_s^{(t+1)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \geq \frac{15}{16\eta} \left\| z_s^{(t+1)} - \bar{z}_s^{(t+1)} \right\|^2 + \frac{7}{16\eta} \left\| \bar{z}_s^{(t+1)} - \bar{z}_s^{(t)} \right\|^2 - \frac{9}{16\eta} \left\| z_s^{(t)} - \bar{z}_s^{(t)} \right\|^2$.

Proof of Lemma 10. We first note that (ii) is a consequence of Lemma 9 and (i),

$$
\begin{aligned}
&(x_s^{(t+1)})^\top \mathcal{Q}_s^{(t+1)} y_s^{(t+1)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \\
&= (x_s^{(t+1)})^\top \mathcal{Q}_s^{(t+1)} y_s^{(t+1)} - (x_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t+1)} + (x_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t+1)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} \\
&\overset{(a)}{\geq} \min_{s'} \gamma \left(\mathcal{V}_{s'}^{(t)} - \mathcal{V}_{s'}^{(t-1)} \right) + \frac{15}{16\eta} \left\| z_s^{(t+1)} - \bar{z}_s^{(t+1)} \right\|^2 + \frac{7}{16\eta} \left\| \bar{z}_s^{(t+1)} - \bar{z}_s^{(t)} \right\|^2 - \frac{9}{16\eta} \left\| z_s^{(t)} - \bar{z}_s^{(t)} \right\|^2 \\
&\overset{(b)}{\geq} \frac{15}{16\eta} \left\| z_s^{(t+1)} - \bar{z}_s^{(t+1)} \right\|^2 + \frac{7}{16\eta} \left\| \bar{z}_s^{(t+1)} - \bar{z}_s^{(t)} \right\|^2 - \frac{9}{16\eta} \left\| z_s^{(t)} - \bar{z}_s^{(t)} \right\|^2
\end{aligned}
$$

where (a) is due to Lemma 9, and the update of $\mathcal{Q}_s^{(t)}$ in Algorithm 3,

$$
\mathcal{Q}_s^{(t+1)}(a_1, a_2) - \mathcal{Q}_s^{(t)}(a_1, a_2) = \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s, a_1, a_2)} \left[\mathcal{V}_{s'}^{(t)} - \mathcal{V}_{s'}^{(t-1)} \right]
$$

and (b) follows (i).

Therefore, it suffices to prove (i). We prove it by induction. Define $\zeta_s^{(t)} := \left\| z_s^{(t)} - \bar{z}_s^{(t)} \right\|^2$ and $\lambda_s^{(t)} := \left\| \bar{z}_s^{(t+1)} - \bar{z}_s^{(t)} \right\|^2$. For notational simplicity, define $\mathcal{Q}_s^{(0)} = \mathbf{0}_{A \times A}$, $z_s^{(0)} = \bar{z}_s^{(0)} = \frac{1}{A}\mathbf{1} = z_s^{(1)} = \bar{z}_s^{(1)}$. Thus, (ii) holds for $t = 0$ and (i) holds for $t = 1$. We note that for $t \geq 2$,

$$
\begin{aligned}
&\mathcal{V}_s^{(t)} - \mathcal{V}_s^{(t-1)} \\
&\overset{(a)}{=} \alpha^{(t)} \left(x_s^{(t)} \mathcal{Q}_s^{(t)} y_s^{(t)} - \mathcal{V}_s^{(t-1)} \right) \\
&\overset{(b)}{=} \alpha^{(t)} \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left(x_s^{(t)} \mathcal{Q}_s^{(t)} y_s^{(t)} - x_s^{(\tau)} \mathcal{Q}_s^{(\tau)} y_s^{(\tau)} \right) \right) \\
&= \alpha^{(t)} \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \sum_{i=\tau}^{t-1} \left(x_s^{(i+1)} \mathcal{Q}_s^{(i+1)} y_s^{(i+1)} - x_s^{(i)} \mathcal{Q}_s^{(i)} y_s^{(i)} \right) \right) \\
&= \alpha^{(t)} \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \sum_{i=\tau}^{t-1} \left(x_s^{(i+1)} \mathcal{Q}_s^{(i+1)} y_s^{(i+1)} - x_s^{(i)} \mathcal{Q}_s^{(i)} y_s^{(i)} - \frac{15}{16\eta} \zeta_s^{(i+1)} - \frac{7}{16\eta} \lambda_s^{(i)} + \frac{9}{16\eta} \zeta_s^{(i)} \right) \right) \\
&\quad + \alpha^{(t)} \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \sum_{i=\tau}^{t-1} \left(\frac{15}{16\eta} \zeta_s^{(i+1)} + \frac{7}{16\eta} \lambda_s^{(i)} - \frac{9}{16\eta} \zeta_s^{(i)} \right) \right) \\
&\overset{(c)}{\geq} \alpha^{(t)} \sum_{i=0}^{t-1} \left(\sum_{\tau=0}^{i} \alpha^{(t-1,\tau)} \right) \left(\frac{15}{16\eta} \zeta_s^{(i+1)} - \frac{9}{16\eta} \zeta_s^{(i)} \right) \\
&= \alpha^{(t)} \sum_{i=1}^{t} \zeta_s^{(i)} \left(\frac{15}{16\eta} \sum_{\tau=0}^{i-1} \alpha^{(t-1,\tau)} - \frac{9}{16\eta} \sum_{\tau=0}^{i} \alpha^{(t-1,\tau)} \right) - \alpha^{(t)} \left(\sum_{\tau=0}^{0} \alpha^{(t-1,\tau)} \right) \frac{9\eta}{16} \zeta_s^{(0)} \\
&\overset{(d)}{=} \alpha^{(t)} \sum_{i=2}^{t} \zeta_s^{(i)} \left(\frac{15}{16\eta} \sum_{\tau=0}^{i-1} \alpha^{(t-1,\tau)} - \frac{9}{16\eta} \sum_{\tau=0}^{i} \alpha^{(t-1,\tau)} \right) \\
&\overset{(e)}{\geq} 0
\end{aligned}
$$

where (a) follows the update of $\mathcal{V}_s^{(t)}$ in Algorithm 3, we apply Lemma 8 and $\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} = 1$ in (b), (c) follows the induction hypothesis (ii), (d) is due to that $\zeta_s^{(0)} = \zeta_s^{(1)} = 0$, and we apply Lemma 14 for (e). □

Lemma 11. *For every $s \in \mathcal{S}$, the following quantities in Algorithm 3 all converge to some fixed values when $t \to \infty$:*

(i) $\mathcal{V}_s^{(t)}$;

(ii) $\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2 + \left\|\bar{z}_s^{(t)} - \bar{z}_s^{(t-1)}\right\|^2$ *(converges to zero);*

(iii) $(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}$.

Proof. **Establishing (i)**. By (i) in Lemma 10, $\{\mathcal{V}_s^{(t)}\}_{t=0}^{\infty}$ is a bounded increasing sequence. By the monotone convergence theorem, it is convergent. Therefore, (i) holds.

Establishing (ii). By summing up the inequality (ii) in Lemma 10 over t and using the fact that $z_s^{(1)} = \bar{z}_s^{(1)}$,

$$\sum_{\tau=1}^{t}\left(\frac{6}{16\eta}\left\|z_s^{(\tau+1)} - \bar{z}_s^{(\tau+1)}\right\|^2 + \frac{7}{16\eta}\left\|\bar{z}_s^{(\tau+1)} - \bar{z}_s^{(\tau)}\right\|^2\right) \le (x_s^{(t+1)})^\top \mathcal{Q}_s^{(t+1)} y_s^{(t+1)} - (x_s^{(1)})^\top \mathcal{Q}_s^{(1)} y_s^{(1)} \le \frac{1}{1-\gamma}$$

which implies that $\frac{6}{16\eta}\left\|z_s^{(\tau+1)} - \bar{z}_s^{(\tau+1)}\right\|^2 + \frac{7}{16\eta}\left\|\bar{z}_s^{(\tau+1)} - \bar{z}_s^{(\tau)}\right\|^2$ must converge to zero when $\tau \to \infty$, which further implies (ii).

Establishing (iii). By (ii) in Lemma 10,

$$\begin{aligned}
&(x_s^{(t+1)})^\top \mathcal{Q}_s^{(t+1)} y_s^{(t+1)} - \frac{15}{16\eta}\left\|z_s^{(t+1)} - \bar{z}_s^{(t+1)}\right\|^2 \\
&\ge \left((x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} - \frac{15}{16\eta}\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2\right) + \frac{7}{16\eta}\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 + \frac{6}{16\eta}\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2
\end{aligned}$$

Therefore,

$$(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} - \frac{15}{16\eta}\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2$$

converges to a fixed value (increasing and upper bounded). In (ii), we have shown that $\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2$ converges to zero. Therefore, $(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}$ must also converge. Therefore, (iii) holds. □

Lemma 12. *In Algorithm 3, for every $s \in \mathcal{S}$, $\lim_{t\to\infty} V_s^{x^{(t)},y^{(t)}}$ exists, and*

$$\lim_{t\to\infty} \mathcal{V}_s^{(t)} = \lim_{t\to\infty} V_s^{x^{(t)},y^{(t)}}.$$

Proof of Lemma 12. By Lemma 11, $\mathcal{V}_s^{(t)}$ and $(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}$ both are convergent. Let $\mathcal{V}_s^\star := \lim_{t\to\infty} \mathcal{V}_s^{(t)}$ and $\sigma_s^\star := \lim_{t\to\infty} (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}$. We next show $\mathcal{V}_s^\star = \sigma_s^\star$ by contradiction. Assume that there exists $\epsilon > 0$ such that $|\mathcal{V}_s^\star - \sigma_s^\star| = \epsilon$. Since $(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}$ converges to $\sigma_s^\star$, there exists some $t_0 > 0$ such that for all $t \ge t_0$,

$$\left|(x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} - \sigma_s^\star\right| \le \frac{\epsilon}{3}. \tag{31}$$

By our choice of $\alpha^{(t)}$, $\sum_{t=t'}^{\infty} \alpha^{(t)} = \infty$ for any t'. Thus, there exists $t_1 > 0$ such that for all $t \ge t_1$ and all $\tau \le t_0$,

$$\alpha^{(t,\tau)} \le \prod_{i=\tau+1}^{t}(1-\alpha^{(i)}) \overset{(a)}{\le} \exp\left(-\sum_{i=t_0+1}^{t}\alpha^{(i)}\right) \le \frac{\epsilon(1-\gamma)}{3t_0} \tag{32}$$

where $\log(1-x) \le -x$ for $x \in (0,1)$ is used in (a). By the update of $\mathcal{V}_s^{(t)}$ in Algorithm 3, for all $t \ge \max(t_0, t_1)$,

$$\begin{aligned}
\left|\mathcal{V}_s^{(t)} - \sigma_s^\star\right| &= \left|\sum_{\tau=0}^{t} \alpha^{(t,\tau)} \left((x_s^{(\tau)})^\top \mathcal{Q}_s^{(\tau)} y_s^{(\tau)} - \sigma_s^\star\right)\right| \\
&\overset{(a)}{\le} \left|\sum_{\tau=0}^{t_0-1} \alpha^{(t,\tau)} \left((x_s^{(\tau)})^\top \mathcal{Q}_s^{(\tau)} y_s^{(\tau)} - \sigma_s^\star\right)\right| + \left|\sum_{\tau=t_0}^{t} \alpha^{(t,\tau)} \left((x_s^{(\tau)})^\top \mathcal{Q}_s^{(\tau)} y_s^{(\tau)} - \sigma_s^\star\right)\right| \\
&\overset{(b)}{\le} \left(\sum_{\tau=0}^{t_0-1} \alpha^{(t,\tau)}\right) \times \frac{1}{1-\gamma} + \left(1 - \sum_{\tau=1}^{t_0-1} \alpha^{(t,\tau)}\right) \times \frac{\epsilon}{3} \\
&\le t_0 \max_{\tau \le t_0} \alpha^{(t,\tau)} \times \frac{1}{1-\gamma} + \frac{\epsilon}{3} \\
&\overset{(c)}{\le} \frac{2\epsilon}{3}
\end{aligned}$$

where we apply the triangle inequality for (a), (b) is due to (31) and $\sum_{\tau=1}^{t} \alpha^{(t,\tau)} = 1$, and (c) follows (32). Since $|\mathcal{V}_s^\star - \sigma_s^\star| = \epsilon$, it is impossible that $\mathcal{V}_s^{(t)}$ converges to $\mathcal{V}_s^\star$, and it must be that $\mathcal{V}_s^\star = \sigma_s^\star$. Therefore, $\mathcal{V}_s^{(t)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}$ converges to zero as $t \to \infty$.

Equivalently, $\mathcal{V}_s^{(t)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}$ can be expressed as

$$\left(\mathcal{V}_s^{(t)} - \mathcal{V}_s^{(t-1)}\right) + \mathcal{V}_s^{(t-1)} - \sum_{a_1,a_2} x_s^{(t)}(a_1) y_s^{(t)}(a_2) \left(r(s,a_1,a_2) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s,a_1,a_2)} \left[\mathcal{V}_{s'}^{(t-1)}\right]\right).$$

By letting $t \to 0$, since $\mathcal{V}_s^{(t)} - \mathcal{V}_s^{(t-1)} \to 0$, thus,

$$\mathcal{V}_s^{(t-1)} - \sum_{a_1,a_2} x_s^{(t)}(a_1) y_s^{(t)}(a_2) \left(r(s,a_1,a_2) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s,a_1,a_2)} \left[\mathcal{V}_{s'}^{(t-1)}\right]\right)$$

also converges to zero. Hence, $\mathcal{V}_s^{(t)}$ converges to the unique fixed point of the Bellman equation. By the uniqueness, $\mathcal{V}_s^{(t-1)} - V_s^{x^{(t)},y^{(t)}}$ converges to zero. Therefore, $\lim_{t\to\infty} V_s^{x^{(t)},y^{(t)}} = \lim_{t\to\infty} \mathcal{V}_s^{(t-1)} = \mathcal{V}_s^\star$. □

Lemma 13. *In Algorithm 3, for every s,*

$$\lim_{t\to\infty} \max_{x'} (x'_s - x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} = 0.$$

Proof of Lemma 13. By the optimality of $x_s^{(t+1)}$,

$$\left\langle x'_s - x_s^{(t+1)},\, \eta \mathcal{Q}_s^{(t)} y_s^{(t)} - x_s^{(t+1)} + \bar{x}_s^{(t+1)} \right\rangle \le 0, \quad \text{for any } x'_s.$$

Rearranging the inequality yields, for any x'_s,

$$\begin{aligned}
\langle x'_s - x_s^{(t+1)}, \mathcal{Q}_s^{(t)} y_s^{(t)} \rangle &\le \frac{1}{\eta}\left(\left\langle x'_s - x_s^{(t+1)},\, x_s^{(t+1)} - x_s^{(t)} \right\rangle + \left\langle x_s^{(t+1)} - x_s^{(t)},\, \eta \mathcal{Q}_s^{(t)} y_s^{(t)} - x_s^{(t+1)} + \bar{x}_s^{(t+1)} \right\rangle\right) \\
&\lesssim \frac{1}{\eta}\left\|x_s^{(t+1)} - x_s^{(t)}\right\| \\
&\le \frac{1}{\eta}\left(\left\|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\right\| + \left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\| + \left\|\bar{x}_s^{(t)} - x_s^{(t)}\right\|\right)
\end{aligned}$$

By (ii) of Lemma 11, the right-hand side above converges to zero, which completes the proof. □

Lemma 14. *Let $\{\alpha^{(t)}\}_{t=1}^\infty$ be a non-increasing sequence that satisfies $0 < \alpha^{(t)} \le \frac{1}{6}$ for all t. Then for any $t \ge i \ge 2$,*

$$\sum_{\tau=0}^{i} \alpha^{(t,\tau)} \le \frac{5}{3} \sum_{\tau=0}^{i-1} \alpha^{(t,\tau)}.$$

Proof of Lemma 14. Equivalently, we prove

$$\alpha^{(t,i)} \;\le\; \frac{2}{3}\sum_{\tau=0}^{i-1}\alpha^{(t,\tau)}.$$

If suffices to show that $\alpha^{(t,i)} \le \frac{2}{3}\alpha^{(t,i-1)} + \frac{2}{3}\alpha^{(t,i-2)}$. We have the following two cases.

Case 1: $i > 2$. By the definition of $\alpha^{(t,\tau)}$ and the monotonicity of $0 < \alpha^{(t)} \le \frac{1}{6}$,

$$\frac{\alpha^{(t,i)}}{\alpha^{(t,i-1)}} \;=\; \frac{\alpha^{(i)}\prod_{j=i+1}^{t}(1-\alpha^{(j)})}{\alpha^{(i-1)}\prod_{j=i}^{t}(1-\alpha^{(j)})} \;=\; \frac{\alpha^{(i)}}{\alpha^{(i-1)}(1-\alpha^{(i)})} \;\le\; \frac{1}{1-\alpha^{(i)}} \;\le\; \frac{1}{1-\frac{1}{6}} \;=\; \frac{6}{5}$$
$$\frac{\alpha^{(t,i)}}{\alpha^{(t,i-2)}} \;=\; \frac{\alpha^{(i)}}{\alpha^{(i-2)}(1-\alpha^{(i)})(1-\alpha^{(i-1)})} \;\le\; \frac{36}{25}$$

Therefore,

$$\frac{2}{3}\alpha^{(t,i-1)} + \frac{2}{3}\alpha^{(t,i-2)} \;\ge\; \frac{2}{3}\left(\frac{5}{6}+\frac{25}{36}\right)\alpha^{(t,i)} \;\ge\; \alpha^{(t,i)}.$$

Case 2: $i = 2$. By the definition of $\alpha^{(t,\tau)}$ and the monotonicity of $0 < \alpha^{(t)} \le \frac{1}{6}$,

$$\frac{\alpha^{(t,2)}}{\alpha^{(t,0)}} \;=\; \frac{\alpha^{(2)}\prod_{j=3}^{t}(1-\alpha^{(j)})}{\prod_{j=1}^{t}(1-\alpha^{(j)})} \;=\; \frac{\alpha^{(2)}}{(1-\alpha^{(1)})(1-\alpha^{(2)})} \;\le\; \frac{\frac{1}{6}}{\frac{5}{6}\times\frac{5}{6}} \;=\; \frac{6}{25}$$

Therefore,

$$\frac{2}{3}\alpha^{(t,1)} + \frac{2}{3}\alpha^{(t,0)} \;\ge\; \frac{2}{3}\times\frac{25}{6}\alpha^{(t,2)} \;\ge\; \alpha^{(t,2)}.$$

□

Proof of Theorem 5.

$$\begin{aligned}
&\max_{x'}\left(V^{x',y^{(t)}}(\rho) - V^{x^{(t)},y^{(t)}}(\rho)\right)\\
&= \max_{x'}\frac{1}{1-\gamma}\sum_{s} d_\rho^{x',y^{(t)}}(s)\left(x'_s - x_s^{(t)}\right)^\top Q_s^{x^{(t)},y^{(t)}} y_s^{(t)}\\
&\le \underbrace{\max_{x'}\frac{1}{1-\gamma}\sum_{s} d_\rho^{x',y^{(t)}}(s)\left(x'_s - x_s^{(t)}\right)^\top \mathcal{Q}_s^{(t)} y_s^{(t)}}_{\textbf{Diff}_P} + \underbrace{\max_{x'}\frac{1}{1-\gamma}\sum_{s} d_\rho^{x',y^{(t)}}(s)\left(x'_s - x_s^{(t)}\right)^\top \left(Q_s^{x^{(t)},y^{(t)}} - \mathcal{Q}_s^{(t)}\right) y_s^{(t)}}_{\textbf{Diff}_Q}.
\end{aligned}$$

By Lemma 13, $\textbf{Diff}_P \to 0$ when $t \to \infty$. For $\textbf{Diff}_Q$, we notice that

$$\left|Q_s^{x^{(t)},y^{(t)}} - \mathcal{Q}_s^{(t)}\right| \;\le\; \gamma\max_{s'}\left|V_{s'}^{x^{(t)},y^{(t)}} - \mathcal{V}_{s'}^{(t)}\right|$$

which converges to zero by Lemma 12. Therefore, $\textbf{Diff}_Q \to 0$ when $t \to \infty$. Therefore, $(x^{(t)}, y^{(t)})$ converges to a Nash equilibrium when $t \to \infty$. □

D.2. Proof of Theorem 6

We first introduce a corollary of Lemma 10.

Corollary 3. *In Algorithm 3, for every state s, and any $T > 0$,*

$$\sum_{t=1}^{T}\left(\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 + \left\|\bar{z}_s^{(t+1)} - z_s^{(t)}\right\|^2\right) \;\le\; \frac{8\eta}{1-\gamma}$$

where $z_s^{(t)} = (x_s^{(t)}, y_s^{(t)})$ and $\bar{z}_s^{(t)} = (\bar{x}_s^{(t)}, \bar{y}_s^{(t)})$.

Proof of Corollary 3. By (ii) of Lemma 10,

$$\begin{aligned}
&(x_s^{(t+1)})^\top \mathcal{Q}_s^{(t+1)} y_s^{(t+1)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} + \frac{15}{16\eta}\left(\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2 - \left\|z_s^{(t+1)} - \bar{z}_s^{(t+1)}\right\|^2\right) \\
\geq\ & \frac{7}{16\eta}\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 + \frac{6}{16\eta}\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2 \\
\geq\ & \frac{6}{16\eta}\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 + \frac{6}{16\eta}\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2.
\end{aligned}$$

Thus, by the inequality $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$,

$$\begin{aligned}
\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 + \left\|\bar{z}_s^{(t+1)} - z_s^{(t)}\right\|^2 &\leq 3\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 + 2\left\|\bar{z}_s^{(t)} - z_s^{(t)}\right\|^2 \\
&\leq 3\left\|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\right\|^2 + 3\left\|\bar{z}_s^{(t)} - z_s^{(t)}\right\|^2 \\
&\leq 8\eta\left((x_s^{(t+1)})^\top \mathcal{Q}_s^{(t+1)} y_s^{(t+1)} - (x_s^{(t)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)}\right) \\
&\quad + \frac{15}{2}\left(\left\|z_s^{(t)} - \bar{z}_s^{(t)}\right\|^2 - \left\|z_s^{(t+1)} - \bar{z}_s^{(t+1)}\right\|^2\right)
\end{aligned}$$

which yields our desired result if we sum it over t, use $(x_s^{(T+1)})^\top \mathcal{Q}_s^{(T+1)} y_s^{(T+1)} \leq \frac{1}{1-\gamma}$ and $z_s^{(1)} = \bar{z}_s^{(1)}$, and ignore a negative term. □

Lemma 15. *In Algorithm 3, the gap between the critic $\mathcal{Q}_s^{(t)}$ and the true $Q_s^{(t)}$ satisfies*

$$\sum_{t=1}^{T} \max_s \left\|\mathcal{Q}_s^{(t)} - Q_s^{(t)}\right\|_\infty^2 \lesssim \frac{A}{(\alpha^{(T)})^2(1-\gamma)^6} \sum_{t=1}^{T} \max_s \left(\left\|x_s^{(t)} - x_s^{(t-1)}\right\|^2 + \left\|y_s^{(t)} - y_s^{(t-1)}\right\|^2\right).$$

Proof of Lemma 15. For notational simplicity, define $\mathcal{Q}_s^{(0)} = Q_s^{(0)} = \mathbf{0}_{A\times A}$.

$$
\begin{aligned}
&\max_s \left\|Q_s^{(t)} - \mathcal{Q}_s^{(t)}\right\|_\infty^2 \\
:= \quad & \max_{s,a_1,a_2} \left|Q_s^{(t)}(a_1,a_2) - \mathcal{Q}_s^{(t)}(a_1,a_2)\right|^2 \\
\leq \quad & \max_{s,a_1,a_2} \Bigg| r(s,a_1,a_2) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot\,|\,s,a_1,a_2)}\left[(x_{s'})^\top Q_{s'}^{(t)} y_{s'}\right] \\
& \qquad - \sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left(r(s,a_1,a_2) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot\,|\,s,a_1,a_2)}\left[(x_{s'}^{(\tau)})^\top \mathcal{Q}_{s'}^{(\tau)} y_{s'}^{(\tau)}\right]\right) \Bigg|^2 \\
\leq \quad & \gamma^2 \max_{s'} \left| \sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left((x_{s'})^\top Q_{s'}^{(t)} y_{s'} - (x_{s'}^{(\tau)})^\top \mathcal{Q}_{s'}^{(\tau)} y_{s'}^{(\tau)}\right)\right|^2 \\
\overset{(a)}{\leq} \quad & \frac{6\gamma^2}{1-\gamma} \max_s \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)})^\top \left(Q_s^{(t)} - Q_s^{(\tau)}\right) y_s^{(t)} \right)^2 \\
& + \frac{2\gamma^2}{1+\gamma} \max_s \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)})^\top \left(Q_s^{(\tau)} - \mathcal{Q}_s^{(\tau)}\right) y_s^{(t)} \right)^2 \\
& + \frac{6\gamma^2}{1-\gamma} \max_s \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)})^\top \mathcal{Q}_s^{(\tau)} (y_s^{(t)} - y_s^{(\tau)}) \right)^2 \\
& + \frac{6\gamma^2}{1-\gamma} \max_s \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)} - x_s^{(\tau)})^\top \mathcal{Q}_s^{(\tau)} y_s^{(\tau)} \right)^2 \\
\overset{(b)}{\leq} \quad & \frac{2\gamma^2}{1+\gamma} \max_s \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \right) \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left((x_s^{(t)})^\top \left(\mathcal{Q}_s^{(\tau)} - Q_s^{(\tau)}\right) y_s^{(t)}\right)^2 \right) \\
& + c' \max_s \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left(\|x_s^{(t)} - x_s^{(\tau)}\|_1 + \|y_s^{(t)} - y_s^{(\tau)}\|_1 \right) \right)^2 \\
\overset{(c)}{\leq} \quad & \frac{2\gamma^2}{1+\gamma} \max_s \left[\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left\|\mathcal{Q}_s^{(\tau)} - Q_s^{(\tau)}\right\|_\infty^2 \right] + c' \max_s \left(\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \sum_{h=\tau+1}^{t} \text{diff}_s^{(h)} \right)^2 \\
\leq \quad & \gamma \max_s \left[\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left\|\mathcal{Q}_s^{(\tau)} - Q_s^{(\tau)}\right\|_\infty^2 \right] + c' \max_s \left(\sum_{h=1}^{t} \sum_{\tau=0}^{h-1} \alpha^{(t-1,\tau)} \text{diff}_s^{(h)} \right)^2 \\
\overset{(d)}{\leq} \quad & \gamma \max_s \left[\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left\|\mathcal{Q}_s^{(\tau)} - Q_s^{(\tau)}\right\|_\infty^2 \right] + c' \max_s \left(\sum_{h=1}^{t} \delta^{(t-1,h-1)} \text{diff}_s^{(h)} \right)^2 \\
\overset{(e)}{\leq} \quad & \gamma \max_s \left[\sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} \left\|\mathcal{Q}_s^{(\tau)} - Q_s^{(\tau)}\right\|_\infty^2 \right] + c' \max_s \left(\sum_{h=1}^{t} (1-\alpha^{(t)})^{t-h} \text{diff}_s^{(h)} \right)^2
\end{aligned}
$$

where in (a) we apply $(x+y+z+w)^2 \leq \frac{6x^2}{1-\gamma} + \frac{2y^2}{1+\gamma} + \frac{6z^2}{1-\gamma} + \frac{6w^2}{1-\gamma}$ from the Cauchy-Schwarz inequality, in (b) we use Lemma 17 and obtain $c' = O\left(\frac{1}{(1-\gamma)^5}\right)$, in (c) we introduce notation,

$$
\text{diff}_s^{(h)} \;:=\; \left\|x_s^{(h)} - x_s^{(h-1)}\right\|_1 + \left\|y_s^{(h)} - y_s^{(h-1)}\right\|_1,
$$

and in (d) we introduce notation,

$$
\delta^{(t,\tau)} \;=\; \prod_{i=\tau+1}^{t} (1-\alpha^{(i)}).
$$

and apply Lemma 35 of Wei et al. (2021b), (e) is due to that $\{\alpha^{(t)}\}_{t=0}^{\infty}$ is a non-increasing sequence.

Application of Lemma 33 of Wei et al. (2021b) to the recursion relation above yields

$$\begin{aligned}\max_s \left\| \mathcal{Q}_s^{(t)} - Q_s^{(t)} \right\|_\infty^2 &\le c' \sum_{\tau=1}^{t} \beta^{(t,\tau)} \max_s \left(\sum_{q=1}^{\tau} (1-\alpha^{(\tau)})^{\tau-q} \mathrm{diff}_s^{(q)} \right)^2 \\ &\le c' \sum_{\tau=1}^{t} \beta^{(t,\tau)} \left(\sum_{q=1}^{\tau} (1-\alpha^{(\tau)})^{\tau-q} \mathrm{diff}^{(q)} \right)^2 \end{aligned} \tag{33}$$

where $\beta^{(t,\tau)} := \alpha^{(\tau)} \prod_{i=\tau}^{t-1} (1-\alpha^{(i)} + \alpha^{(i)}\gamma)$ for $1 \le \tau < t$ and $\beta^{(t,t)} := 1$, and $\mathrm{diff}^{(t)} := \max_s \mathrm{diff}_s^{(t)}$.

The right-hand side of (33) can be further upper bounded by

$$\begin{aligned}& c' \sum_{\tau=1}^{t} \beta^{(t,\tau)} \left(\sum_{q=1}^{\tau} (1-\alpha^{(\tau)})^{\tau-q} \right) \left(\sum_{q=1}^{\tau} (1-\alpha^{(\tau)})^{\tau-q} (\mathrm{diff}^{(q)})^2 \right) \\ &\overset{(a)}{\le} c' \sum_{\tau=1}^{t} \frac{\beta^{(t,\tau)}}{\alpha^{(\tau)}} \sum_{q=1}^{\tau} (1-\alpha^{(\tau)})^{\tau-q} (\mathrm{diff}^{(q)})^2 \\ &= c' \sum_{q=1}^{t} \sum_{\tau=q}^{t} \frac{\beta^{(t,\tau)}}{\alpha^{(\tau)}} (1-\alpha^{(\tau)})^{\tau-q} (\mathrm{diff}^{(q)})^2 \\ &= c' \sum_{q=1}^{t} \left[\sum_{\tau=q}^{t-1} (1-\alpha^{(t)}+\alpha^{(t)}\gamma)^{t-\tau} (1-\alpha^{(t)})^{\tau-q} (\mathrm{diff}^{(q)})^2 + \frac{(1-\alpha^{(t)})^{t-q}}{\alpha^{(t)}} (\mathrm{diff}^{(q)})^2 \right] \\ &= c' \sum_{q=1}^{t} \left[(1-\alpha^{(t)}+\alpha^{(t)}\gamma)^{t-q} \sum_{\tau=q}^{t-1} \left(\frac{1-\alpha^{(t)}}{1-\alpha^{(t)}+\alpha^{(t)}\gamma} \right)^{\tau-q} + \frac{(1-\alpha^{(t)})^{t-q}}{\alpha^{(t)}} \right] (\mathrm{diff}^{(q)})^2 \\ &= c' \sum_{q=1}^{t} \left[(1-\alpha^{(t)}+\alpha^{(t)}\gamma)^{t-q} \frac{1 - \left(\frac{1-\alpha^{(t)}}{1-\alpha^{(t)}+\alpha^{(t)}\gamma} \right)^{t-q}}{\frac{\alpha^{(t)}\gamma}{1-\alpha^{(t)}+\alpha^{(t)}\gamma}} + \frac{(1-\alpha^{(t)})^{t-q}}{\alpha^{(t)}} \right] (\mathrm{diff}^{(q)})^2 \\ &\le \frac{2c'}{\alpha^{(t)}\gamma} \sum_{q=1}^{t} (1-\alpha^{(t)}+\alpha^{(t)}\gamma)^{t-q} (\mathrm{diff}^{(q)})^2, \end{aligned}$$

where (a) is due to that $\sum_{q=1}^{\tau} (1-\alpha^{(\tau)})^{\tau-q} \le \frac{1}{\alpha^{(\tau)}}$.

Substitution of the upper bound above into (33) yields,

$$\begin{aligned}\sum_{t=1}^{T} \max_s \left\| \mathcal{Q}_s^{(t)} - Q_s^{(t)} \right\|_\infty^2 &\lesssim \sum_{t=1}^{T} \frac{c'}{\alpha^{(t)}} \sum_{q=1}^{t} (1-\alpha^{(t)}+\alpha^{(t)}\gamma)^{t-q} (\mathrm{diff}^{(q)})^2 \\ &\overset{(a)}{\le} \sum_{q=1}^{T} \sum_{t=q}^{T} \frac{c'}{\alpha^{(T)}} (1-\alpha^{(T)}+\alpha^{(T)}\gamma)^{t-q} (\mathrm{diff}^{(q)})^2 \\ &\overset{(b)}{\le} \frac{c'}{(\alpha^{(T)})^2(1-\gamma)} \sum_{q=1}^{T} (\mathrm{diff}^{(q)})^2 \end{aligned}$$

where (a) is due to that $\alpha^{(t)}$ is non-increasing, and (b) is due to that $\sum_{t=q}^{T} (1-\alpha^{(T)}+\alpha^{(T)}\gamma)^{t-q} \le \frac{1}{\alpha^{(T)}(1-\gamma)}$.

Finally, using the definition of c' and applying $\|\cdot\|_1 \le \sqrt{A}\|\cdot\|$ to $\mathrm{diff}^{(q)}$ lead to the desired result. □

Proof of Theorem 6. The proof consists of two parts: Markov cooperative games and Markov competitive games, separately.

Markov cooperative games. Fix s, the optimality of $\bar{x}_s^{(t+1)}$ in Algorithm 3 yields

$$\left\langle \eta \mathcal{Q}_s^{(t)} y_s^{(t)} - (\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}), x_s' - \bar{x}_s^{(t+1)} \right\rangle \quad \le \quad 0, \quad \text{for any } x_s' \in \Delta(\mathcal{A}_1). \tag{34}$$

Thus, for any $x'_s \in \Delta(\mathcal{A}_1)$,

$$\begin{aligned}
&(x'_s - x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \\
=\;& (x'_s - \bar{x}_s^{(t+1)})^\top \mathcal{Q}_s^{(t)} y_s^{(t)} + (x'_s - \bar{x}_s^{(t+1)})^\top (Q_s^{(t)} - \mathcal{Q}_s^{(t)}) y_s^{(t)} + (\bar{x}_s^{(t+1)} - x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \\
\overset{(a)}{\lesssim}\;& \frac{1}{\eta}(x'_s - \bar{x}_s^{(t+1)})^\top (\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}) + \left\|Q_s^{(t)} - \mathcal{Q}_s^{(t)}\right\| + \frac{\sqrt{A}}{1-\gamma}\left\|\bar{x}_s^{(t+1)} - x_s^{(t)}\right\| \\
\overset{(b)}{\lesssim}\;& \frac{1}{\eta}\left(\left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\| + \left\|\bar{x}_s^{(t+1)} - x_s^{(t)}\right\|\right) + \left\|Q_s^{(t)} - \mathcal{Q}_s^{(t)}\right\|,
\end{aligned}$$

where we use (34) and $\left\|Q_s^{(t)}\right\| \leq \frac{\sqrt{A}}{1-\gamma}$ in (a), and (b) is due to the Cauchy-Schwarz inequality and the choice of $\eta \leq \frac{1-\gamma}{32\sqrt{A}}$. Hence,

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\max_{x'} V^{x', y^{(t)}}(\rho) - V^{x^{(t)}, y^{(t)}}(\rho)\right) \\
=\;& \frac{1}{1-\gamma}\sum_{t=1}^{T}\max_{x'}\sum_s d_\rho^{x', y^{(t)}}(s)\left(x'_s - x_s^{(t)}\right)^\top Q_s^{(t)} y_s^{(t)} \\
\lesssim\;& \frac{1}{\eta(1-\gamma)}\sum_{t=1}^{T}\sum_s d_\rho^{x', y^{(t)}}(s)\left(\left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\| + \left\|\bar{x}_s^{(t+1)} - x_s^{(t)}\right\|\right) + \frac{1}{(1-\gamma)}\sum_{t=1}^{T}\sum_s d_\rho^{x', y^{(t)}}(s)\left\|Q_s^{(t)} - \mathcal{Q}_s^{(t)}\right\| \\
\overset{(a)}{\lesssim}\;& \frac{1}{\eta(1-\gamma)}\sqrt{\sum_{t=1}^{T}\sum_s d_\rho^{x', y^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T}\sum_s d_\rho^{x', y^{(t)}}(s)\left(\left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\|^2 + \left\|\bar{x}_s^{(t+1)} - x_s^{(t)}\right\|^2\right)} \\
&+ \frac{1}{(1-\gamma)}\sum_{t=1}^{T}\sum_s d_\rho^{x', y^{(t)}}(s)\left\|Q_s^{(t)} - \mathcal{Q}_s^{(t)}\right\| \\
\overset{(b)}{\leq}\;& \frac{\sqrt{T}}{\eta(1-\gamma)} \times \sqrt{\sum_{t=1}^{T}\sum_s\left(\left\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\right\|^2 + \left\|\bar{x}_s^{(t+1)} - x_s^{(t)}\right\|^2\right)} + \frac{1}{(1-\gamma)}\sum_{t=1}^{T}\sum_s d_\rho^{x', y^{(t)}}(s)\left\|Q_s^{(t)} - \mathcal{Q}_s^{(t)}\right\| \\
\overset{(c)}{\lesssim}\;& \frac{\sqrt{T}}{\eta(1-\gamma)}\sqrt{\frac{\eta S}{1-\gamma}} + \frac{1}{1-\gamma}\sqrt{\sum_{t=1}^{T}\sum_s d_\rho^{x', y^{(t)}}(s)}\sqrt{\sum_{t=1}^{T}\sum_s\left\|Q_s^{(t)} - \mathcal{Q}_s^{(t)}\right\|^2} \\
\overset{(d)}{\lesssim}\;& \sqrt{\frac{ST}{\eta(1-\gamma)^3}} + \frac{1}{1-\gamma}\sqrt{T}\sqrt{\frac{SA}{(\alpha^{(T)})^2(1-\gamma)^6}\sum_{t=1}^{T}\max_s\left(\left\|x_s^{(t)} - x_s^{(t-1)}\right\|^2 + \left\|y_s^{(t)} - y_s^{(t-1)}\right\|^2\right)} \\
\overset{(e)}{\lesssim}\;& \sqrt{\frac{ST}{\eta(1-\gamma)^3}} + \frac{1}{1-\gamma}\sqrt{T}\sqrt{\frac{\eta S^2 A}{\left(\alpha^{(T)}\right)^2(1-\gamma)^7}}
\end{aligned}$$

where we apply the Cauchy-Schwarz inequality for (a), (b) follows the state distribution $d_\rho^{x', y^{(t)}}(s)$, (c) is due to Corollary 3, (d) is because of Lemma 15, and (e) again is due to Corollary 3. By taking $\eta = \frac{(1-\gamma)^2}{32\sqrt{SA}}$ and $\alpha^{(t)} = \frac{1}{6\sqrt[3]{t}}$, the last upper bound above is of order,

$$O\left(\frac{(S^3A)^{\frac{1}{4}}\sqrt{T}}{(1-\gamma)^{\frac{7}{2}}\alpha^{(T)}}\right) = O\left(\frac{(S^3A)^{\frac{1}{4}}T^{\frac{5}{6}}}{(1-\gamma)^{\frac{7}{2}}}\right).$$

Markov competitive game. We start from an intermediate step in the proof of Theorem 1 of (Wei et al., 2021b). Specifically,

they have shown that if both players use Algorithm 3 in a two-player zero-sum Markov game, then,

$$\sum_{t=1}^{T}\left(\max_{x',y'} V^{x',y^{(t)}}(\rho) - V^{x^{(t)},y'}(\rho)\right) \;=\; O\left(\frac{S\sqrt{C_\alpha C_\beta T}}{\eta(1-\gamma)}\right)$$

where $C_\alpha := 1 + \sum_{t=1}^{T}\alpha^{(t)}$ and C_β is an upper bound for $\sum_{t=\tau}^{T}\beta^{(t,\tau)}$ with $\beta^{(t,\tau)} := \alpha^{(\tau)}\prod_{i=\tau}^{t-1}(1-\alpha^{(i)}+\alpha^{(i)}\gamma)$ if $\tau < t$ and $\beta^{(t,t)} := 1$. We next calculate the upper bounds for C_α and C_β.

Bounding C_α. Recall that $\alpha^{(t)} = \frac{1}{6}t^{-\frac{1}{3}}$. By the definition of C_α,

$$C_\alpha \;=\; 1 + \frac{1}{6}\sum_{t=1}^{T} t^{-\frac{1}{3}} \;=\; O\left(T^{\frac{2}{3}}\right).$$

Bounding C_β. Using $\alpha^{(t)} = \frac{1}{6}t^{-\frac{1}{3}}$, for any $\tau \geq 1$, we have

$$\begin{aligned}
\sum_{t=\tau}^{T}\beta^{(t,\tau)} \;&\leq\; 1 + \sum_{t=\tau+1}^{T}\alpha^{(\tau)}\prod_{i=\tau}^{t-1}(1-\alpha^{(i)}+\alpha^{(i)}\gamma) \\
&=\; 1 + \frac{1}{6}\sum_{t=\tau+1}^{T}\tau^{-\frac{1}{3}}\left(1-\frac{1}{6}t^{-\frac{1}{3}}(1-\gamma)\right)^{t-\tau} \\
&\leq\; 1 + \frac{1}{6}\sum_{t=\tau+1}^{t_0}\tau^{-\frac{1}{3}} + \frac{1}{6}\sum_{t=t_0+1}^{T}\tau^{-\frac{1}{3}}\left(1-\frac{1}{6}t^{-\frac{1}{3}}(1-\gamma)\right)^{t-\tau} \qquad \text{(for some } t_0 \text{ defined below)} \\
&\leq\; 1 + \frac{1}{6}\tau^{-\frac{1}{3}}(t_0-\tau) + \frac{1}{6}\tau^{-\frac{1}{3}}\sum_{t=t_0+1}^{T}\exp\left(-\frac{1}{6}t^{-\frac{1}{3}}(1-\gamma)(t-\tau)\right). \qquad (35)
\end{aligned}$$

Define $t_0 := \tau + H(\tau+c)^{\frac{1}{3}}\ln(\tau+c) + c$, where $H := \frac{48}{1-\gamma}$ and $c := 2\left(\frac{2H}{1-\frac{1}{3}}\ln\frac{H}{1-\frac{1}{3}}\right)^{\frac{1}{1-\frac{1}{3}}}$ (if $t_0 > T$, we simply ignore the second term in (35)). By Lemma 16 with $q = \frac{1}{3}$, for all $t \geq t_0$,

$$t - \tau \;\geq\; \frac{H}{2}\left(\frac{t}{2}\right)^{\frac{1}{3}}\ln\left(\frac{t}{2}\right).$$

Hence, we can continue to bound the right-hand side of (35) by

$$\begin{aligned}
&O\left(H\left(\frac{\tau+c}{\tau}\right)^{\frac{1}{3}}\ln(\tau+c) + \frac{c}{\tau^{\frac{1}{3}}}\right) + \frac{1}{6}\tau^{-\frac{1}{3}}\sum_{t=t_0+1}^{T}\exp\left(-\frac{1}{12}t^{-\frac{1}{3}}(1-\gamma)\frac{H}{2}\left(\frac{t}{2}\right)^{\frac{1}{3}}\ln\left(\frac{t}{2}\right)\right) \\
&\leq\; \widetilde{O}\big(H(1+c)^{\frac{1}{3}} + c\big) + \frac{1}{6}\tau^{-\frac{1}{3}}\sum_{t=t_0+1}^{T}\frac{2}{t} \\
&=\; \widetilde{O}\left(\frac{1}{(1-\gamma)^{\frac{3}{2}}}\right)
\end{aligned}$$

which proves that $C_\beta = \widetilde{O}\left(\frac{1}{(1-\gamma)^{\frac{3}{2}}}\right)$.

Therefore,

$$\sum_{t=1}^{T}\left(\max_{x',y'} V^{x',y^{(t)}}(\rho) - V^{x^{(t)},y'}(\rho)\right) \;=\; O\left(\frac{S\sqrt{C_\alpha C_\beta T}}{\eta(1-\gamma)}\right) \;=\; \widetilde{O}\left(\frac{ST^{\frac{5}{6}}}{\eta(1-\gamma)^{\frac{7}{4}}}\right)$$

which completes the proof by taking $\eta = \frac{(1-\gamma)^2}{32\sqrt{SA}}$. □

Lemma 16. *Fix $\tau \in \mathbb{N}$, $0 < q < 1$, $H \geq 1$. Let*

$$t_0 \;:=\; \tau \;+\; H(\tau + c)^q \ln(\tau + c) \;+\; c$$

where $c := 2\left(\frac{2H}{1-q} \ln \frac{H}{1-q}\right)^{\frac{1}{1-q}}$. Then for all $t \geq t_0$,

$$t \;-\; H\left(\frac{t}{2}\right)^q \ln\left(\frac{t}{2}\right) \;\geq\; \tau.$$

Proof of Lemma 16. We first show that for all $t \geq c$,

- $Ht^q \ln t \leq t$;
- $t - H\left(\frac{t}{2}\right)^q \ln\left(\frac{t}{2}\right)$ is non-decreasing.

To show the two items above, we apply Lemma A.1 of (Shalev-Shwartz & Ben-David, 2014) which states that $x \geq 2a\ln(a) \Rightarrow x \geq a\ln(x)$ for any $a > 0$. By the definition of c, for all $t \geq c$, $t^{1-q} \geq \left(\frac{t}{2}\right)^{1-q} \geq \frac{2H}{1-q} \ln \frac{H}{1-q}$ and thus

$$t^{1-q} \;\geq\; \frac{H}{1-q} \ln(t^{1-q}) \;=\; H \ln t$$

which proves the first item, and that

$$\left(\frac{t}{2}\right)^{1-q} \;\geq\; \frac{H}{1-q} \ln\left(\frac{t}{2}\right)^{1-q} \;=\; H \ln \frac{t}{2}$$

which gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(t - \frac{H}{2}\left(\frac{t}{2}\right)^q \ln\left(\frac{t}{2}\right)\right) \;&=\; 1 \;-\; \frac{H}{2} \cdot \frac{q}{2}\left(\frac{t}{2}\right)^{q-1} \ln\left(\frac{t}{2}\right) \;-\; \frac{H}{2} \cdot \left(\frac{t}{2}\right)^q \frac{1}{t} \\ &\geq\; 1 \;-\; \frac{1}{2} \cdot \frac{q}{2} \;-\; \frac{1}{2} \;\geq\; 0\end{aligned}$$

which proves the second item.

By the first item and the definition of t_0, $t_0 \leq \tau + (\tau + c) + c = 2\tau + 2c$. Then by the second item, for all $t \geq t_0$ we have

$$t \;-\; \frac{H}{2}\left(\frac{t}{2}\right)^q \ln\left(\frac{t}{2}\right) \;\geq\; t_0 \;-\; \frac{H}{2}\left(\frac{t_0}{2}\right)^q \ln\left(\frac{t_0}{2}\right) \;\geq\; t_0 \;-\; \frac{H}{2}(\tau + c)^q \ln(\tau + c) \;\geq\; \tau$$

which completes the proof. □

Lemma 17. *For any two policies (x', y') and (x, y),*

$$\max_s \left\| Q_s^{x',y'} - Q_s^{x,y} \right\|_{\max} \;\leq\; \frac{\gamma}{(1-\gamma)^2} \max_{s'} \left(\|x'_{s'} - x_{s'}\|_1 + \|y'_{s'} - y_{s'}\|_1 \right).$$

Proof of Lemma 17. By the definition,

$$\begin{aligned}&\left\| Q_s^{x',y'} - Q_s^{x,y} \right\|_{\max} \\ \overset{(a)}{:=}\;& \max_{a_1,a_2} \left| Q_s^{x',y'}(a_1,a_2) - Q_s^{x,y}(a_1,a_2) \right| \\ \overset{(b)}{\leq}\;& \gamma \sum_{s'} \mathbb{P}(s' \mid s, \bar{a}_1, \bar{a}_2) \left| (x'_{s'})^\top Q_{s'}^{x',y'} y'_{s'} - (x_{s'})^\top Q_{s'}^{x,y} y_{s'} \right| \\ \leq\;& \gamma \max_{s'} \underbrace{\left| (x_{s'})^\top Q_{s'}^{x',y'} y_{s'} - (x_{s'})^\top Q_{s'}^{x,y} y_{s'} \right|}_{\textbf{Qiff}}\end{aligned} \tag{36}$$

where $\bar{a}_1$ and $\bar{a}_2$ achieve the maximum in (a), and (b) is due to the Bellman equation,

$$\begin{aligned} Q_s^{x,y}(a_1,a_2) &= r(s,a_1,a_2) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot\,|\,s,a_1,a_2)}\left[V_{s'}^{x,y}\right] \\ &= r(s,a_1,a_2) + \gamma \sum_{s'} \mathbb{P}(s'\,|\,s,a_1,a_2) \sum_{a_1',a_2'} x_{s'}(a_1') y_{s'}(a_2') Q_{s'}^{x,y}(a_1',a_2'). \end{aligned}$$

Fix s', we next subtract and add $(x_{s'})^\top Q_{s'}^{x',y'} y_{s'}$ in Qiff and apply $|a+b| \le |a| + |b|$ to reach,

$$\begin{aligned} \textbf{Qiff} &\le \left|(x'_{s'})^\top Q_{s'}^{x',y'} y'_{s'} - (x_{s'})^\top Q_{s'}^{x',y'} y_{s'}\right| + \left|(x_{s'})^\top Q_{s'}^{x',y'} y_{s'} - (x_{s'})^\top Q_{s'}^{x,y} y_{s'}\right| \\ &\le \frac{1}{1-\gamma}\left|\sum_{a_1',a_2'} \left(x'_{s'}(a_1) y'_{s'}(a_2') - x_{s'}(a_1') y_{s'}(a_2')\right) Q_{s'}^{x',y'}(a_1',a_2')\right| \\ &\quad + \left|(x_{s'})^\top \left(Q_{s'}^{x',y'} - Q_{s'}^{x,y}\right) y_{s'}\right| \\ &\le \frac{1}{1-\gamma} \sum_{a_1',a_2'} \left|x'_{s'}(a_1') y'_{s'}(a_2') - x_{s'}(a_1') y_{s'}(a_2')\right| + \left|(x_{s'})^\top \left(Q_{s'}^{x',y'} - Q_{s'}^{x,y}\right) y_{s'}\right|. \end{aligned}$$

We also notice that

$$\begin{aligned} \left\|x'_{s'} \circ y'_{s'} - x_{s'} \circ y_{s'}\right\|_1 &:= \sum_{a_1',a_2'} \left|x'_{s'}(a_1') y'_{s'}(a_2') - x_{s'}(a_1') y_{s'}(a_2')\right| \\ &\le \sum_{a_1',a_2'} \left|x'_{s'}(a_1') y'_{s'}(a_2') - x_{s'}(a_1') y'_{s'}(a_2')\right| \\ &\quad + \sum_{a_1',a_2'} \left|x_{s'}(a_1') y'_{s'}(a_2') - x_{s'}(a_1') y_{s'}(a_2')\right| \\ &\le \sum_{a_1'} \left|x'_{s'}(a_1') - x_{s'}(a_1')\right| + \sum_{a_2'} \left|y'_{s'}(a_2') - y_{s'}(a_2')\right| \\ &= \left\|x'_{s'} - x_{s'}\right\|_1 + \left\|y'_{s'} - y_{s'}\right\|_1 \end{aligned}$$

and

$$\left|(x_{s'})^\top \left(Q_{s'}^{x',y'} - Q_{s'}^{x,y}\right) y_{s'}\right| \le \max_{a_1,a_2} \left|Q_{s'}^{x',y'}(a_1,a_2) - Q_{s'}^{x,y}(a_1,a_2)\right| := \left\|Q_{s'}^{x',y'} - Q_{s'}^{x,y}\right\|_{\max}.$$

By substituting the upper bound on **Qiff** above into (36),

$$\begin{aligned} &\left\|Q_s^{x',y'} - Q_s^{x,y}\right\|_{\max} \\ &\le \gamma \max_{s'} \frac{1}{1-\gamma} \left(\left\|x'_{s'} - x_{s'}\right\|_1 + \left\|y'_{s'} - y_{s'}\right\|_1\right) + \gamma \max_{s'} \left\|Q_{s'}^{x',y'} - Q_{s'}^{x,y}\right\|_{\max}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\max_s \left\|Q_s^{x',y'} - Q_s^{x,y}\right\|_{\max} \\ &\le \gamma \max_{s'} \frac{1}{1-\gamma} \left(\left\|x'_{s'} - x_{s'}\right\|_1 + \left\|y'_{s'} - y_{s'}\right\|_1\right) + \gamma \max_{s'} \left\|Q_{s'}^{x',y'} - Q_{s'}^{x,y}\right\|_{\max}. \end{aligned}$$

which yields the desired result. □

E. Auxiliary Lemmas

In this section, we provide some auxiliary lemmas that are helpful in our analysis.

E.1. Auxiliary lemmas for potential functions

Lemma 18. *For any N-player Markov potential game with instantaneous reward bounded in $[0,1]$, it holds that*

$$\left|\Phi^{\pi}(\mu) - \Phi^{\pi'}(\mu)\right| \le \frac{N}{1-\gamma},$$

for any $\pi, \pi' \in \Pi$ and $\mu \in \Delta(\mathcal{S})$.

Proof of Lemma 18. By the potential property,

$$\begin{aligned}\Phi^{\pi}(\mu) - \Phi^{\pi'}(\mu) &= (\Phi^{\pi} - \Phi^{\pi_1', \pi_{-1}}) + (\Phi^{\pi_1', \pi_{-1}} - \Phi^{\pi'_{\{1,2\}}, \pi_{-\{1,2\}}}) + \ldots + (\Phi^{\pi_N, \pi'_{-N}} - \Phi^{\pi'}) \\ &= (V_1^{\pi} - V_1^{\pi_1', \pi_{-1}}) + (V_2^{\pi_1', \pi_{-1}} - V_2^{\pi'_{\{1,2\}}, \pi_{-\{1,2\}}}) + \ldots + (V_N^{\pi_N, \pi'_{-N}} - V_N^{\pi'}) \\ &\leq \frac{N}{1-\gamma}\end{aligned}$$

where the last inequality is due to $V_i^{\pi} - V_i^{\pi'} \leq \frac{1}{1-\gamma}$ for any π and π'. By symmetry, $\Phi^{\pi'}(\mu) - \Phi^{\pi}(\mu) \leq \frac{N}{1-\gamma}$. □

E.2. Auxiliary lemmas for single-player MDPs

We provide some auxiliary lemmas in the context of single-player MDPs.

Lemma 19 (Action value function difference)**.** *Suppose that two MDPs have the same state/action spaces, but different reward and transition functions: (r, p) and $(\widetilde{r}, \widetilde{p})$. Then, for a given policy π, two action value functions associated with two MDPs satisfy*

$$\max_{s,\,a} |Q^{\pi}(s,a) - \widetilde{Q}^{\pi}(s,a)| \;\leq\; \frac{1}{1-\gamma} \max_{s,\,a} |r(s,a) - \widetilde{r}(s,a)| \;+\; \frac{\gamma}{(1-\gamma)^2} \max_{s,\,a} \|p(\cdot \,|\, s,a) - \widetilde{p}(\cdot \,|\, s,a)\|_1 \,.$$

Proof of Lemma 19. By the Bellman equations,

$$\begin{aligned}Q^{\pi}(s,a) &= r(s,a) + \gamma \sum_{s',\,a'} p(s' \,|\, s,a)\, \pi(a' \,|\, s')\, Q^{\pi}(s',a') \\ \widetilde{Q}^{\pi}(s,a) &= \widetilde{r}(s,a) + \gamma \sum_{s',a'} \widetilde{p}(s' \,|\, s,a)\, \pi(a' \,|\, s')\, \widetilde{Q}^{\pi}(s',a').\end{aligned}$$

Subtracting equalities above on both sides yields

$$\begin{aligned}&|Q^{\pi}(s,a) - \widetilde{Q}^{\pi}(s,a)| \\ \leq\;& |r(s,a) - \widetilde{r}(s,a)| + \gamma \left| \sum_{s',\,a'} (p(s' \,|\, s,a) - \widetilde{p}(s' \,|\, s,a))\, \pi(a' \,|\, s') Q^{\pi}(s',a') \right| \\ &+ \gamma \left| \sum_{s',\,a'} \widetilde{p}(s' \,|\, s,a) \pi(a' \,|\, s') \left(Q^{\pi}(s',a') - \widetilde{Q}^{\pi}(s',a') \right) \right| \\ \leq\;& |r(s,a) - \widetilde{r}(s,a)| + \frac{\gamma}{1-\gamma} \|p(\cdot \,|\, s,a) - \widetilde{p}(\cdot \,|\, s,a)\|_1 + \gamma \max_{s',\,a'} \left| Q^{\pi}(s',a') - \widetilde{Q}^{\pi}(s',a') \right|.\end{aligned}$$

Taking the maximum over (s,a) leads to

$$\begin{aligned}&\max_{s,\,a} |Q^{\pi}(s,a) - \widetilde{Q}^{\pi}(s,a)| \\ \leq\;& \max_{s,\,a} |r(s,a) - \widetilde{r}(s,a)| + \frac{\gamma}{1-\gamma} \max_{s,\,a} \|p(\cdot \,|\, s,a) - \widetilde{p}(\cdot \,|\, s,a)\|_1 + \gamma \max_{s,\,a} |Q^{\pi}(s,a) - \widetilde{Q}^{\pi}(s,a)|\end{aligned}$$

which leads to the desired inequality after rearrangement. □

Lemma 20 (Visitation measure difference)**.** *Let π and π' be two policies for a MDP, and μ be an initial state distribution. Then,*

$$\sum_{s} \left| d_{\mu}^{\pi}(s) - d_{\mu}^{\pi'}(s) \right| \;\leq\; \max_{s} \|\pi(\cdot \,|\, s) - \pi'(\cdot \,|\, s)\|_1 \,.$$

Proof of Lemma 20. By the definition, for a fixed state $s^{\sharp}$,

$$d_{\mu}^{\pi}(s^{\sharp}) \;=\; (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}_{\{s_t = s^{\sharp}\}} \;\middle|\; s_0 \sim \mu,\; \pi \right].$$

By taking reward function $r(s,a) = (1-\gamma)\mathbf{1}_{\{s=s^\sharp\}}$, we can view $d_\mu^\pi(s^\sharp)$ as a value function under the policy π and the initial distribution μ. With a slight abuse of notation, we denote such a value function by $V^\pi(\mu; s^\sharp) = d_\mu^\pi(s^\sharp)$. Similarly, we can define $V^\pi(s; s^\sharp)$ and $Q^\pi(s,a; s^\sharp)$, using the same reward function.

By the performance difference lemma (a single-player version of Lemma 1),

$$d_\mu^\pi(s^\sharp) - d_\mu^{\pi'}(s^\sharp) \;=\; V^\pi(\mu; s^\sharp) - V^{\pi'}(\mu; s^\sharp) \;=\; \sum_{s,\,a} d_\mu^\pi(s)\,(\pi(a\,|\,s) - \pi'(a\,|\,s))\,Q^{\pi'}(s,a; s^\sharp).$$

Therefore,

$$\sum_{s^\sharp} \left| d_\mu^\pi(s^\sharp) - d_\mu^{\pi'}(s^\sharp) \right| \;\le\; \sum_{s^\sharp} \sum_{s,a} d_\mu^\pi(s)\,|\pi(a\,|\,s) - \pi'(a\,|\,s)|\,Q^{\pi'}(s,a; s^\sharp). \tag{37}$$

We also note that $Q^{\pi'}(\cdot,\cdot; s^\sharp)$ is the action value function associated with the reward function $r(s,a) = (1-\gamma)\mathbf{1}_{\{s=s^\sharp\}}$. Thus,

$$\begin{aligned}
\sum_{s^\sharp} Q^{\pi'}(s,a; s^\sharp) &= \sum_{s^\sharp} \mathbb{E}\left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbf{1}_{\{s_t = s^\sharp\}} \;\middle|\; (s_0,a_0) = (s,a),\ \pi' \right] \\
&= \mathbb{E}\left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t \;\middle|\; (s_0,a_0) = (s,a),\ \pi' \right] \\
&= 1.
\end{aligned}$$

Therefore, we can arrange (37) as follows,

$$\begin{aligned}
\sum_{s^\sharp} \left| d_\mu^\pi(s^\sharp) - d_\mu^{\pi'}(s^\sharp) \right| &\le \sum_{s,\,a} d_\mu^\pi(s)\,|\pi(a\,|\,s) - \pi'(a\,|\,s)| \\
&= \sum_{s} d_\mu^\pi(s)\,\|\pi(\cdot\,|\,s) - \pi'(\cdot\,|\,s)\|_1 \\
&\le \max_s \|\pi(\cdot\,|\,s) - \pi'(\cdot\,|\,s)\|_1 .
\end{aligned}$$

□

E.3. Auxiliary lemmas for multi-player MDPs

We first extend Lemma 1 in the 1st-order form to the 2nd-order performance difference, which is useful to measure the joint policy improvement from multiple players.

Lemma 21 (The 2nd-order performance difference)**.** *Consider a two-player common-payoff Markov game with state space $\mathcal{S}$ and action sets $\mathcal{A}_1$, $\mathcal{A}_2$. Let $r : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to [0,1]$ be the reward function, and $p : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \Delta(\mathcal{S})$ be the transition function. Let $\Pi_1 = (\Delta(\mathcal{A}_1))^{\mathcal{S}}$ and $\Pi_2 = (\Delta(\mathcal{A}_2))^{\mathcal{S}}$ be player 1 and player 2's policy sets, respectively. Then, for any $x, x' \in \Pi_1$ and $y, y' \in \Pi_2$,*

$$\begin{aligned}
&V^{x,y}(\mu) \;-\; V^{x',y}(\mu) \;-\; V^{x,y'}(\mu) \;+\; V^{x',y'}(\mu) \\
&\le\; \frac{2\kappa_\mu^2 A}{(1-\gamma)^4} \sum_s d_\mu^{x',y'}(s) \left(\|x(\cdot\,|\,s) - x'(\cdot\,|\,s)\|^2 + \|y(\cdot\,|\,s) - y'(\cdot\,|\,s)\|^2 \right)
\end{aligned}$$

where κ_μ is the distribution mismatch coefficient relative to μ (see κ_μ in Definition 1).

Proof of Lemma 21. We define the following non-stationary policies:

$\bar{x}_i$: a Player 1's policy where in steps from 0 to $i-1$, x' is executed; in steps from i to ∞, x is executed.

With this definition, $\bar{x}_0 = x$ and $\bar{x}_\infty = x'$. We define $\bar{y}_i$ similarly. Since $\bar{x}_i$ is non-stationary, we specify its action distribution as $\bar{x}_i(\cdot\,|\,s,h)$ where h is the step index. The joint value function for these non-stationary policies can be defined as usual:

$$V^{\bar{x}_i,\bar{y}_j}(\mu) \;:=\; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t,b_t) \;\middle|\; s_0 \sim \mu,\ a_t \sim \bar{x}_i(\cdot\,|\,s_t,t),\ b_t \sim \bar{y}_j(\cdot\,|\,s_t,t) \right].$$

For simplicity, we omit the initial distribution μ in writing the value function. We first show that for any $H \in \mathbb{N}$,

$$V^{\bar{x}_0,\bar{y}_0} - V^{\bar{x}_H,\bar{y}_0} - V^{\bar{x}_0,\bar{y}_H} + V^{\bar{x}_H,\bar{y}_H} = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1}\left(V^{\bar{x}_i,\bar{y}_j} - V^{\bar{x}_{i+1},\bar{y}_j} - V^{\bar{x}_i,\bar{y}_{j+1}} + V^{\bar{x}_{i+1},\bar{y}_{j+1}}\right)$$

In fact, the right-hand side above is equal to

$$\begin{aligned}
&\sum_{j=0}^{H-1}\sum_{i=0}^{H-1}\left(V^{\bar{x}_i,\bar{y}_j} - V^{\bar{x}_{i+1},\bar{y}_j}\right) + \sum_{j=0}^{H-1}\sum_{i=0}^{H-1}\left(-V^{\bar{x}_i,\bar{y}_{j+1}} + V^{\bar{x}_{i+1},\bar{y}_{j+1}}\right)\\
&= \sum_{j=0}^{H-1}\left(V^{\bar{x}_0,\bar{y}_j} - V^{\bar{x}_H,\bar{y}_j}\right) + \sum_{j=0}^{H-1}\left(-V^{\bar{x}_0,\bar{y}_{j+1}} + V^{\bar{x}_H,\bar{y}_{j+1}}\right)\\
&= \sum_{j=0}^{H-1}\left(V^{\bar{x}_0,\bar{y}_j} - V^{\bar{x}_0,\bar{y}_{j+1}}\right) + \sum_{j=0}^{H-1}\left(-V^{\bar{x}_H,\bar{y}_j} + V^{\bar{x}_H,\bar{y}_{j+1}}\right)\\
&= V^{\bar{x}_0,\bar{y}_0} - V^{\bar{x}_0,\bar{y}_H} - V^{\bar{x}_H,\bar{y}_0} + V^{\bar{x}_H,\bar{y}_H}
\end{aligned}$$

Sending H to infinity and recalling that $\bar{x}_0 = x$, $\bar{x}_\infty = x'$, $\bar{y}_0 = y$, $\bar{y}_\infty = y'$ lead to

$$V^{x,y} - V^{x',y} - V^{x,y'} + V^{x',y'} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left(V^{\bar{x}_i,\bar{y}_j} - V^{\bar{x}_{i+1},\bar{y}_j} - V^{\bar{x}_i,\bar{y}_{j+1}} + V^{\bar{x}_{i+1},\bar{y}_{j+1}}\right).$$

We next focus on the particular summand above with index (i,j) and discuss three cases.

Case 1: $i < j$**.** We first re-write $V^{\bar{x}_i,\bar{y}_j} - V^{\bar{x}_{i+1},\bar{y}_j}$. Notice that the value difference between the policy pairs $(\bar{x}_i,\bar{y}_j)$ and $(\bar{x}_{i+1},\bar{y}_j)$ starts at step i, since both policy pairs are equal to (x',y') from step 0 to step $i-1$. At the ith step, $\bar{x}_i$ changes to x while $\bar{x}_{i+1}$ remains as x'. Therefore,

$$\begin{aligned}
&V^{\bar{x}_i,\bar{y}_j} - V^{\bar{x}_{i+1},\bar{y}_j}\\
&= \frac{1}{1-\gamma}\sum_{s,a,b} d_\mu^{x',y'}(s;i)x(a\,|\,s)y'(b\,|\,s)\left(r(s,a,b) + \mathbb{E}\left[\sum_{t=i+1}^{\infty}\gamma^{t-i}r(s_t,a_t,b_t)\,\middle|\, s_{i+1}\sim p(\cdot\,|\,s,a,b),\bar{x}_i,\bar{y}_j\right]\right)\\
&\quad - \frac{1}{1-\gamma}\sum_{s,a,b} d_\mu^{x',y'}(s;i)x'(a\,|\,s)y'(b\,|\,s)\left(r(s,a,b) + \mathbb{E}\left[\sum_{t=i+1}^{\infty}\gamma^{t-i}r(s_t,a_t,b_t)\,\middle|\, s_{i+1}\sim p(\cdot\,|\,s,a,b),\bar{x}_i,\bar{y}_j\right]\right)\\
&= \frac{1}{1-\gamma}\sum_{s,a,b} d_\mu^{x',y'}(s;i)\Big(x(a\,|\,s) - x'(a\,|\,s)\Big)y'(b\,|\,s)\Bigg(r(s,a,b)\\
&\qquad\qquad + \mathbb{E}\left[\sum_{t=i+1}^{\infty}\gamma^{t-i}r(s_t,a_t,b_t)\,\middle|\, s_{i+1}\sim p(\cdot\,|\,s,a,b),\bar{x}_i,\bar{y}_j\right]\Bigg)
\end{aligned}$$

where we define

$$d_\mu^{x,y}(s;i) = (1-\gamma)\mathbb{E}\left[\gamma^i\mathbf{1}[s_i = s]\mid s_0\sim\mu\right].$$

(note that $d_\mu^{x,y}(s) = \sum_{i=0}^{\infty} d_\mu^{x,y}(s;i)$).
Similarly,

$$\begin{aligned}
&V^{\bar{x}_i,\bar{y}_{j+1}} - V^{\bar{x}_{i+1},\bar{y}_{j+1}}\\
&= \frac{1}{1-\gamma}\sum_{s,a,b} d_\mu^{x',y'}(s;i)\Big(x(a\,|\,s) - x'(a\,|\,s)\Big)y'(b|s)\Bigg(r(s,a,b)\\
&\qquad\qquad + \mathbb{E}\left[\sum_{t=i+1}^{\infty}\gamma^{t-i}r(s_t,a_t,b_t)\,\middle|\, s_{i+1}\sim p(\cdot\,|\,s,a,b),\bar{x}_i,\bar{y}_{j+1}\right]\Bigg)
\end{aligned}$$

We notice that the following difference:

$$\mathbb{E}\left[\sum_{t=i+1}^{\infty}\gamma^{t-i}r(s_t,a_t,b_t)\,\middle|\, s_{i+1}\sim p(\cdot\,|\,s,a,b),\bar{x}_i,\bar{y}_j\right]-\mathbb{E}\left[\sum_{t=i+1}^{\infty}\gamma^{t-i}r(s_t,a_t,b_t)\,\middle|\, s_{i+1}\sim p(\cdot\,|\,s,a,b),\bar{x}_i,\bar{y}_{j+1}\right]$$

is equivalent to

$$\frac{\gamma}{1-\gamma}\sum_{\widetilde{s},\widetilde{a},\widetilde{b}} d^{x,y'}_{p(\cdot\,|\,s,a,b)}(\widetilde{s};j-i-1)x(\widetilde{a}\,|\,\widetilde{s})\Big(y(\widetilde{b}\,|\,\widetilde{s})-y'(\widetilde{b}\,|\,\widetilde{s})\Big)Q^{x,y}(\widetilde{s},\widetilde{a},\widetilde{b}).$$

Hence,

$$\begin{aligned}
&V^{\bar{x}_i,\bar{y}_j}-V^{\bar{x}_{i+1},\bar{y}_j}-V^{\bar{x}_i,\bar{y}_{j+1}}+V^{\bar{x}_{i+1},\bar{y}_{j+1}}\\
&=\frac{\gamma}{(1-\gamma)^2}\sum_{s,a,b}\sum_{\widetilde{s},\widetilde{a},\widetilde{b}} d_\mu^{x',y'}(s;i)d^{x,y'}_{p(\cdot\,|\,s,a,b)}(\widetilde{s};j-i-1)\Big(x(a\,|\,s)-x'(a\,|\,s)\Big)y'(b\,|\,s)x(\widetilde{a}\,|\,\widetilde{s})\\
&\qquad\qquad\times\Big(y(\widetilde{b}\,|\,\widetilde{s})-y'(\widetilde{b}\,|\,\widetilde{s})\Big)Q^{x,y}(\widetilde{s},\widetilde{a},\widetilde{b})\\
&\le\frac{\gamma}{2(1-\gamma)^3}\sum_{s,a,b}\sum_{\widetilde{s},\widetilde{a},\widetilde{b}} d_\mu^{x',y'}(s;i)d^{x,y'}_{p(\cdot\,|\,s,a,b)}(\widetilde{s};j-i-1)y'(b\,|\,s)x(\widetilde{a}\,|\,\widetilde{s})\Big(x(a\,|\,s)-x'(a\,|\,s)\Big)^2\\
&\quad+\frac{\gamma}{2(1-\gamma)^3}\sum_{s,a,b}\sum_{\widetilde{s},\widetilde{a},\widetilde{b}} d_\mu^{x',y'}(s;i)d^{x,y'}_{p(\cdot\,|\,s,a,b)}(\widetilde{s};j-i-1)y'(b\,|\,s)x(\widetilde{a}\,|\,\widetilde{s})\Big(y(\widetilde{b}\,|\,\widetilde{s})-y'(\widetilde{b}\,|\,\widetilde{s})\Big)^2\\
&\qquad\qquad\text{(bounding } |Q^{x,y}(\cdot,\cdot,\cdot)| \text{ by } \tfrac{1}{1-\gamma} \text{ and using AM-GM)}\\
&=\frac{\gamma A}{2(1-\gamma)^3}\sum_{s,a}\sum_{\widetilde{s}} d_\mu^{x',y'}(s;i)d^{x,y'}_{p(\cdot\,|\,s,a,y')}(\widetilde{s};j-i-1)\Big(x(a\,|\,s)-x'(a\,|\,s)\Big)^2\\
&\qquad\qquad\text{(define } p(\cdot\,|\,s,a,y)=\textstyle\sum_b p(\cdot\,|\,s,a,b)y(b|s))\\
&\quad+\frac{\gamma A}{2(1-\gamma)^3}\sum_{s}\sum_{\widetilde{s},\widetilde{b}} d_\mu^{x',y'}(s;i)d^{x,y'}_{p(\cdot\,|\,s,\text{unif},y')}(\widetilde{s};j-i-1)\Big(y(\widetilde{b}\,|\,\widetilde{s})-y'(\widetilde{b}\,|\,\widetilde{s})\Big)^2\\
&\qquad\qquad\text{(define uniform distribution unif} = \tfrac{1}{A}\mathbf{1})
\end{aligned}$$

Summing the inequality above over $i<j$ yields

$$\begin{aligned}
&\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}\left(V^{\bar{x}_i,\bar{y}_j}-V^{\bar{x}_{i+1},\bar{y}_j}-V^{\bar{x}_i,\bar{y}_{j+1}}+V^{\bar{x}_{i+1},\bar{y}_{j+1}}\right)\\
&\le\frac{\gamma A}{2(1-\gamma)^3}\sum_{i=0}^{\infty}\sum_{s,a} d_\mu^{x',y'}(s;i)\Big(x(a\,|\,s)-x'(a\,|\,s)\Big)^2\left(\sum_{\widetilde{s}}\sum_{j=i+1}^{\infty} d^{x,y'}_{p(\cdot\,|\,s,a,y')}(\widetilde{s};j-i-1)\right)\\
&\quad+\frac{\gamma A}{2(1-\gamma)^3}\sum_{i=0}^{\infty}\sum_{\widetilde{s},\widetilde{b}}\sum_{s} d_\mu^{x',y'}(s;i)\Big(y(\widetilde{b}\,|\,\widetilde{s})-y'(\widetilde{b}\,|\,\widetilde{s})\Big)^2\left(\sum_{j=i+1}^{\infty} d^{x,y'}_{p(\cdot\,|\,s,\text{unif},y')}(\widetilde{s};j-i-1)\right)\\
&=\frac{\gamma A}{2(1-\gamma)^3}\sum_{s,a} d_\mu^{x',y'}(s)\Big(x(a\,|\,s)-x'(a\,|\,s)\Big)^2\\
&\quad+\frac{\gamma A}{2(1-\gamma)^3}\sum_{\widetilde{s},\widetilde{b}}\sum_{s} d_\mu^{x',y'}(s)d^{x,y'}_{p(\cdot\,|\,s,\text{unif},y')}(\widetilde{s})\Big(y(\widetilde{b}\,|\,\widetilde{s})-y'(\widetilde{b}\,|\,\widetilde{s})\Big)^2\\
&\qquad\qquad\text{(using the property: } \textstyle\sum_{i=0}^{\infty} d_\mu^{x,y}(s;i)=d_\mu^{x,y}(s))\\
&=\frac{\gamma A}{2(1-\gamma)^3}\sum_{s} d_\mu^{x',y'}(s)\left\|x(\cdot\,|\,s)-x'(\cdot\,|\,s)\right\|^2+\frac{\gamma A}{2(1-\gamma)^3}\sum_{s} d_{\mu'}^{x,y'}(s)\left\|y(\cdot\,|\,s)-y'(\cdot\,|\,s)\right\|^2.
\end{aligned}$$

where μ' is a state distribution that generates the state by the following procedure: first sample a state s_0 according to $d_\mu^{x',y'}(\cdot)$, then execute $(\text{unif},y')=(\frac{1}{A}\mathbf{1},y')$ for one step, and then output the next state.

By Lemma 22 (with $\pi = (x', y')$, $\pi' = (x, y')$, and $\bar{\pi} = (\text{unif}, y')$), we have $\frac{d_\mu^{x,y'}(s)}{d_\mu^{x',y'}(s)} \le \frac{d_{\mu'}^{x,y'}(s)}{\mu(1-\gamma)} \le \frac{\kappa_\mu^2}{\gamma(1-\gamma)}$. Therefore,

$$\begin{aligned}
&\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}\left(V^{\bar{x}_i,\bar{y}_j} - V^{\bar{x}_{i+1},\bar{y}_j} - V^{\bar{x}_i,\bar{y}_{j+1}} + V^{\bar{x}_{i+1},\bar{y}_{j+1}}\right)\\
&\le \frac{\kappa_\mu^2 A}{2(1-\gamma)^4}\sum_s d_\mu^{x',y'}(s)\left(\|x(\cdot\,|\,s) - x'(\cdot\,|\,s)\|^2 + \|y(\cdot\,|\,s) - y'(\cdot\,|\,s)\|^2\right).
\end{aligned}$$

Case 2: $i > j$. This case is symmetric to the case of $i < j$, and can be handled similarly.

Case 3: $i = j$. In this case,

$$\begin{aligned}
&\sum_{i=0}^{\infty}\left(V^{\bar{x}_i,\bar{y}_i} - V^{\bar{x}_{i+1},\bar{y}_i} - V^{\bar{x}_i,\bar{y}_{i+1}} + V^{\bar{x}_{i+1},\bar{y}_{i+1}}\right)\\
&= \frac{1}{1-\gamma}\sum_{i=0}^{\infty}\sum_{s,a,b} d_\mu^{x',y'}(s;i)\Big(x'(a\,|\,s) - x(a\,|\,s)\Big)\Big(y'(a\,|\,s) - y(a\,|\,s)\Big)Q^{x,y}(s,a,b)\\
&= \frac{1}{1-\gamma}\sum_{s,a,b} d_\mu^{x',y'}(s)\Big(x'(a|s) - x(a|s)\Big)\Big(y'(a|s) - y(a|s)\Big)Q^{x,y}(s,a,b)\\
&\le \frac{1}{2(1-\gamma)^2}\sum_{s,a,b} d_\mu^{x',y'}(s)\Big(x'(a\,|\,s) - x(a\,|\,s)\Big)^2 + \frac{1}{2(1-\gamma)^2}\sum_{s,a,b} d_\mu^{x',y'}(s)\Big(y'(a\,|\,s) - y(a\,|\,s)\Big)^2\\
&= \frac{A}{2(1-\gamma)^2}\sum_s d_\mu^{x',y'}(s)\,\|x'(\cdot\,|\,s) - x(\cdot\,|\,s)\|^2 + \frac{A}{2(1-\gamma)^2}\sum_s d_\mu^{x',y'}(s)\,\|y'(\cdot\,|\,s) - y(\cdot\,|\,s)\|^2.
\end{aligned}$$

Summing the bounds in all three cases above completes the proof. □

Lemma 22. *Let π, π' and $\bar{\pi}$ be three policies, and μ be some initial distribution. Let μ' be a state distribution that generates a state according to the following: first sample an s_0 from $d_\mu^\pi(\cdot)$, then execute $\bar{\pi}$ for one step, and then output the next state. Then,*

$$\left\|\frac{d_{\mu'}^{\pi'}}{\mu}\right\|_\infty \le \frac{\kappa_\mu^2}{\gamma} \triangleq \frac{1}{\gamma}\left(\sup_{\tilde{\pi}}\left\|\frac{d_\mu^{\tilde{\pi}}}{\mu}\right\|_\infty\right)^2$$

Proof of Lemma 22. For a particular state $s^\sharp$, we view the supremum $\sup_{\tilde{\pi}} \frac{d_\mu^{\tilde{\pi}}(s^\sharp)}{\mu(s^\sharp)}$ as the optimal value of an MDP whose reward function is $r(s,a) = \frac{1-\gamma}{\mu(s^\sharp)}\mathbf{1}[s = s^\sharp]$ and initial state is generated by μ. The optimal value of this MDP is upper bounded by κ_μ by Definition 1. We next consider the following non-stationary policy for this MDP: first execute $\bar{\pi}$ for one step, and then execute π' in the rest of the steps. The discounted value of this non-stationary policy is lower bounded by

$$\gamma\sum_s \Pr\left(s_1 = s \mid s_0 \sim \mu,\, a_0 \sim \bar{\pi}(\cdot\,|\,s_0)\right) \times \frac{d_s^{\pi'}(s^\sharp)}{\mu(s^\sharp)} = \gamma\sum_{s_0,a_0,s}\mu(s_0)\bar{\pi}(a_0|s_0)p(s\,|s_0,a_0) \times \frac{d_s^{\pi'}(s^\sharp)}{\mu(s^\sharp)}.$$

We can upper and lower bound the discounted sum above as the following:

$$\frac{\gamma}{\kappa_\mu}\sum_{s_0,a_0,s} d_\mu^\pi(s_0)\bar{\pi}(a_0|s_0)p(s\,|s_0,a_0) \times \frac{d_s^{\pi'}(s^\sharp)}{\mu(s^\sharp)} \le \gamma\sum_{s_0,a_0,s}\mu(s_0)\bar{\pi}(a_0|s_0)p(s\,|s_0,a_0) \times \frac{d_s^{\pi'}(s^\sharp)}{\mu(s^\sharp)} \le \kappa_\mu.$$

where the right inequality is due to that this discounted value must be upper bounded by the optimal value of this MDP, which has an upper bound κ_μ, and the left inequality is by the definition of κ_μ. Now notice that

$$\mu'(s) = \sum_{s_0,a_0} d_\mu^\pi(s_0)\bar{\pi}(a_0|s_0)p(s|s_0,a_0)$$

Algorithm 4 Stochastic projected gradient descent with weighted averaging

Parameters: W, $\lambda^{(k)}$, and $\beta_k^{(K)}$.
Input: Stepsize α, total number of iterations $K > 0$.
Initialization: $w^{(0)} = 0$.
for step $k = 1, \ldots, K$ **do**
  Draw $\nabla^{(k)}$ form a distribution such that $\mathbb{E}[\nabla^{(k)} \mid w^{(k)}] \in \partial f(w^{(k)})$.
  Update $w^{(k+1)} = \mathcal{P}_{\|w\| \leq W}\left(w^{(k)} - \lambda^{(k)} \nabla^{(k)}\right)$
end for
Output: $\sum_{k=0}^{K} \beta_k^{(K)} w^{(k)}$

by the definition of μ'. Plugging this into the previous inequality, we get

$$\frac{\gamma}{\kappa_\mu} \times \frac{d_{\mu'}^{\pi'}(s^\sharp)}{\mu(s^\sharp)} \;\leq\; \kappa_\mu.$$

Since this holds for any $s^\sharp$, this gives

$$\left\| \frac{d_{\mu'}^{\pi'}}{\mu} \right\|_\infty \;\leq\; \frac{\kappa_\mu^2}{\gamma}.$$

□

F. Auxiliary Lemmas for Stochastic Projected Gradient Descent

Algorithm 2 serves a sample-based algorithm if we solve the empirical risk minimization problem (7) via a stochastic projected gradient descent,

$$w_i^{(k+1)} \;=\; \mathcal{P}_{\|w\| \leq W}\left(w_i^{(k)} \;-\; \lambda^{(k)}\, \widehat{\nabla}_i^{(t)}(s^{(k)}, a_i^{(k)}))\right) \tag{38}$$

where $\widehat{\nabla}_i^{(t)} := 2(\langle \phi_i, w_i^{(k)} \rangle - R_i^{(k)})\phi_i$ is the kth gradient of (7) and $\lambda^{(k)} > 0$ is the stepsize. We assume that the smallest eigenvalue of correlation matrix $\mathbb{E}_{s,a_i}\left[\phi_i(s,a_i)\phi_i(s,a_i)^\top\right]$ is positive.

For a constrained convex optimization, $\text{minimize}_{w \in \{w \mid \|w\| \leq W\}} f(w)$, where $f(w)$ is a convex function and $W > 0$, we consider a basic method for solving this problem: the stochastic projected gradient descent in Algorithm 4, where $\mathcal{P}_{\|w\| \leq W}$ is a Euclidean projection in $\mathbb{R}^d$ to the constraint set $\|w\| \leq W$.

Lemma 23. *Let* $w^\star := \text{argmin}_{w \in \{w \mid \|w\| \leq W\}} f(w)$. *Suppose* $\text{Var}(\nabla^{(k)}) \leq \sigma^2$. *If we run Algorithm 4 with stepsize* $\lambda^{(k)} = O(\frac{1}{1+k})$ *and* $\beta_k^{(K)} = \frac{1/\lambda^{(k)}}{\sum_{r=0}^{K} 1/\lambda^{(r)}}$, *then,*

$$\mathbb{E}\left[f\left(\sum_{k=0}^{K} \beta_k^{(K)} w^{(k)}\right)\right] - f(w^\star) \;\lesssim\; \frac{\sigma^2 W^2 d}{K}.$$

Proof of Lemma 23. See the proof of Theorem 1 in (Cohen et al., 2017b). □

G. Additional Experiments

We provide details about our experiments as follows.

For illustration, we consider the state space $\mathcal{S} = \{\textit{safe}, \textit{distancing}\}$ and action space $\mathcal{A}_i = \{A, B, C, D\}$, and the number of players $N = 8$. In each state $s \in \mathcal{S}$, the reward for player i taking an action $a \in \mathcal{A}_i$ is the w_s^a-weighted number of players using the action a, where w_s^a specifies the action preference $w_s^A < w_s^B < w_s^C < w_s^D$. The reward in state *distancing* is less than that in state *safe* by a large amount $c > 0$. For state transition, if more than half of players find themselves using the same action, then the state transits to the state *distancing*; transition back to the state *safe* whenever no more than half of players take the same action.

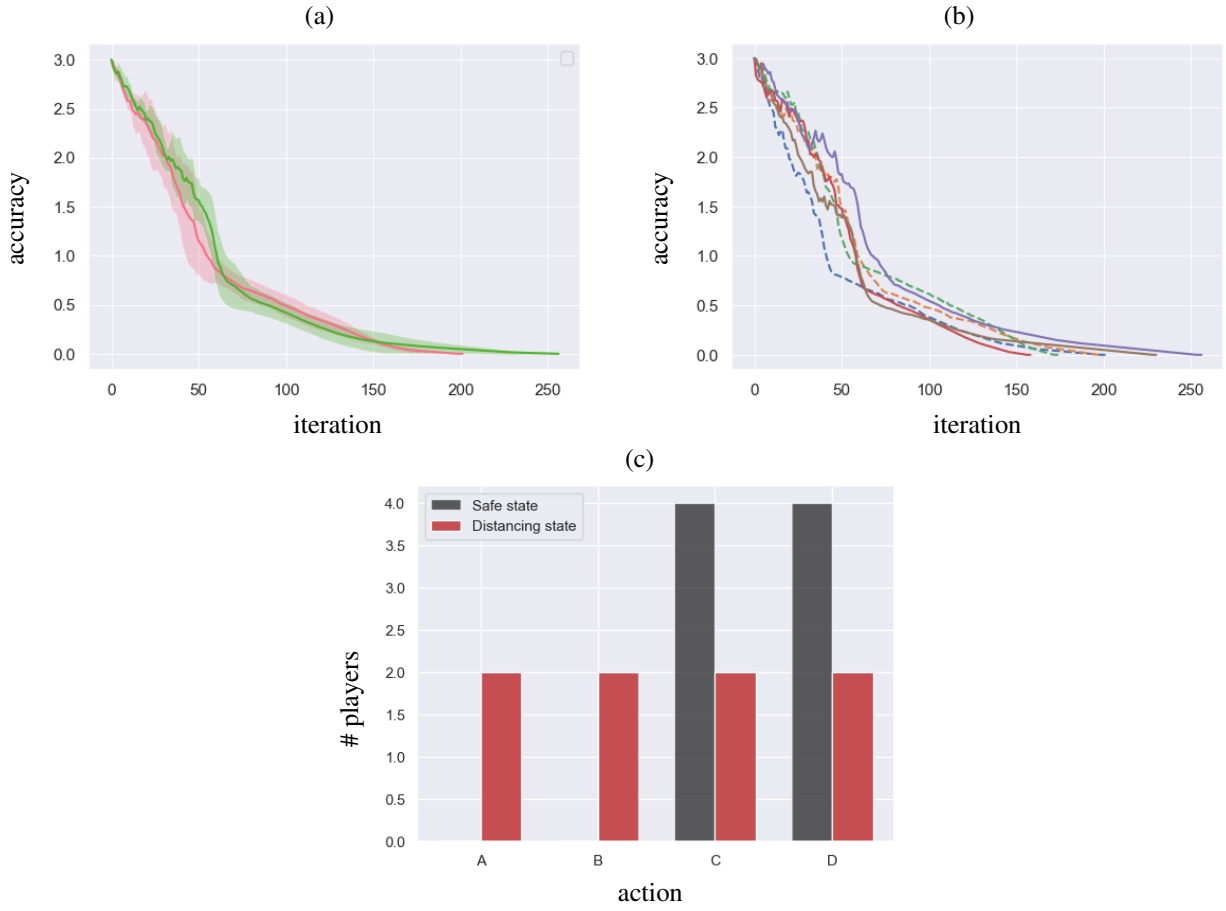

Figure 2. Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.001$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ (Leonardos et al., 2022). Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N}\sum_{i=1}^{N}\|\pi_i^{(t)} - \pi_i^{\text{Nash}}\|_1$. In our computational experiments, the initial distribution ρ is uniform.

In our experiments, we implement our independent policy gradient method based on the code for the projected stochastic gradient ascent (Leonardos et al., 2022). At each iteration, we collect a batch of 20 trajectories to estimate the action-value function and (or) the stationary state distribution under current policy. We choose the discount factor $\gamma = 0.99$, and different the stepsize η, and initial state distributions as we report next.

Continuing Section 7, we further report our computational results using stepsize $\eta = 0.001$ in Figure 2, larger stepsize $\eta = 0.002$ in Figure 3 and stepsize $\eta = 0.005$ in Figure 4. We notice that the stepsize $\eta = 0.001$ for the projected stochastic gradient ascent (Leonardos et al., 2022) does not yield convergence while our independent policy gradient converges as shown in Figure 2. As demonstrated in Section 7, our independent policy gradient permits larger stepsizes with fast convergence, e.g., $\eta = 0.002$ in Figure 3 and $\eta = 0.005$ in Figure 4. Compared Figure 3 with Figure 4, we see an improved convergence of our independent policy gradient using a larger stepsize. We also remark that the learnt policies for all these experiments can generate the same Nash policy that matches the result in Leonardos et al. (2022).

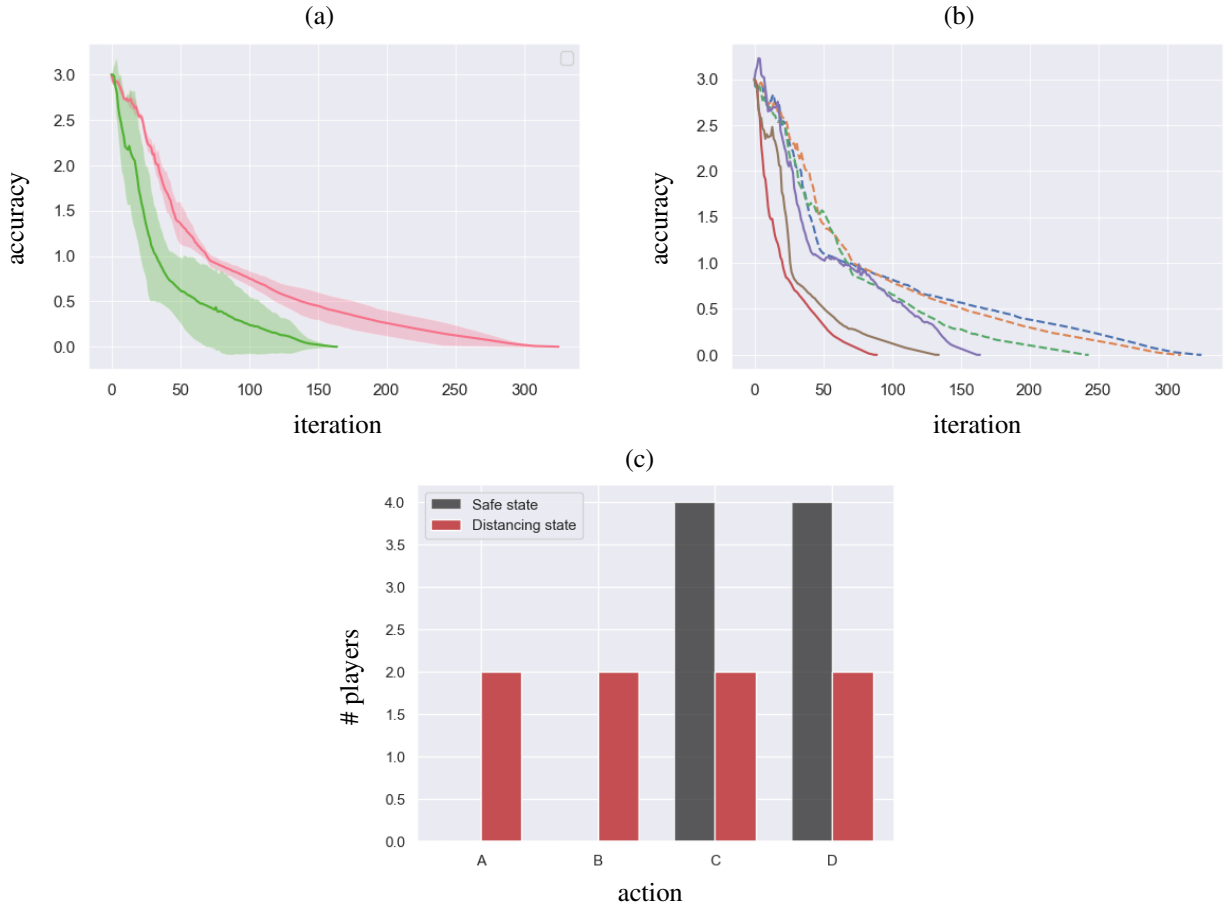

Figure 3. Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.002$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ (Leonardos et al., 2022). Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N}\sum_{i=1}^{N} \|\pi_i^{(t)} - \pi_i^{\text{Nash}}\|_1$. In our computational experiments, the initial distribution ρ is uniform.

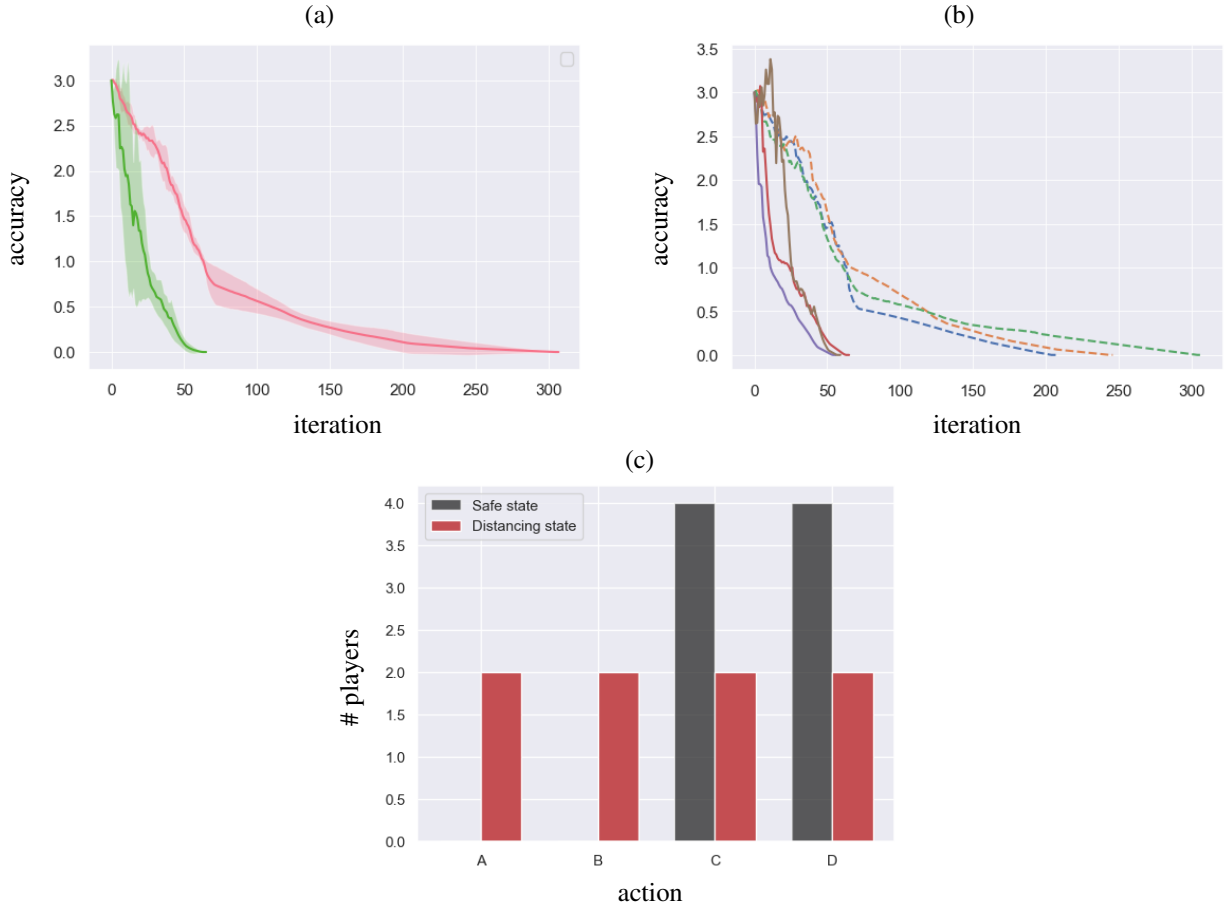

Figure 4. Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.005$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ (Leonardos et al., 2022). Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N}\sum_{i=1}^{N} \|\pi_i^{(t)} - \pi_i^{\text{Nash}}\|_1$. In our computational experiments, the initial distribution ρ is uniform.

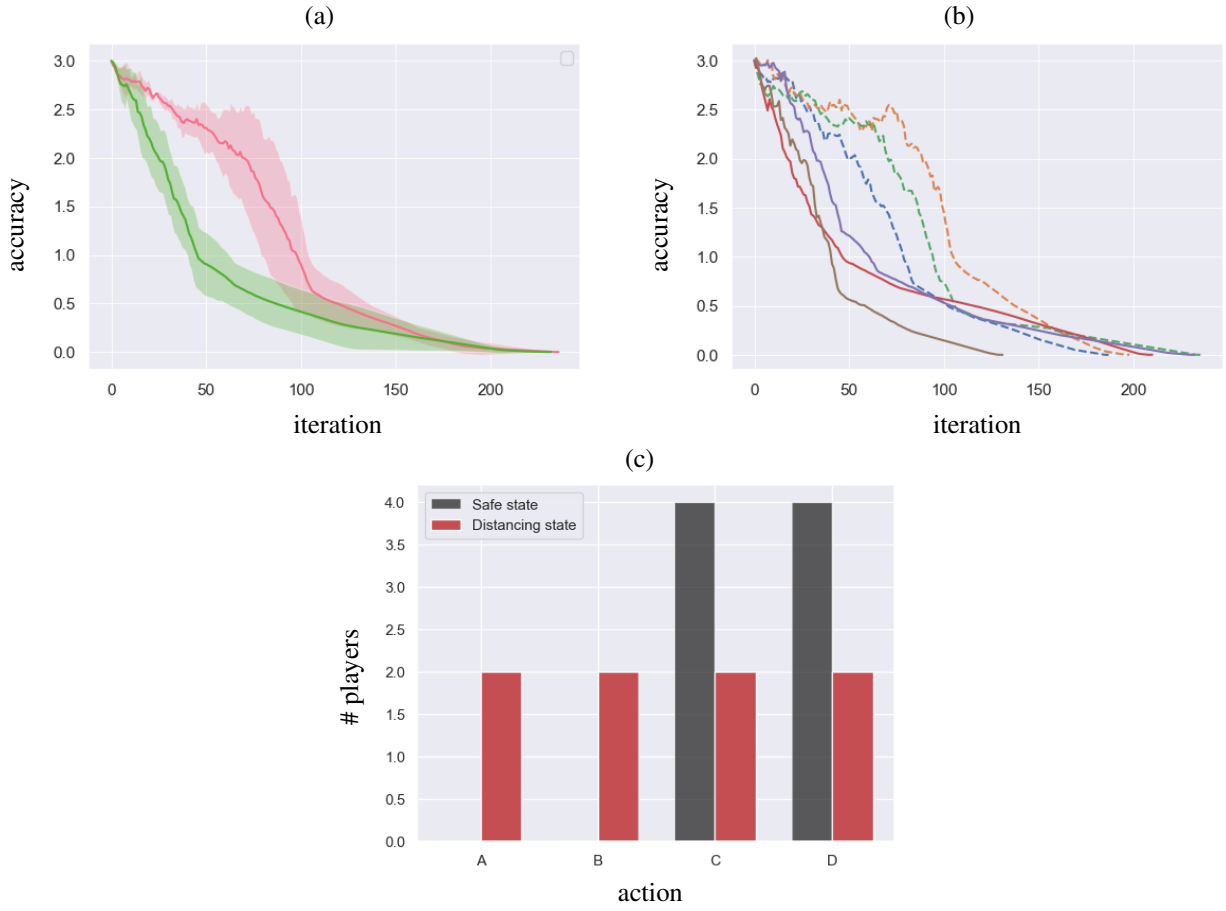

Figure 5. Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.001$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ (Leonardos et al., 2022). Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N}\sum_{i=1}^{N} \|\pi_i^{(t)} - \pi_i^{\text{Nash}}\|_1$. In our computational experiments, the initial distribution is nearly degenerate $\rho = (0.9999, 0.0001)$.

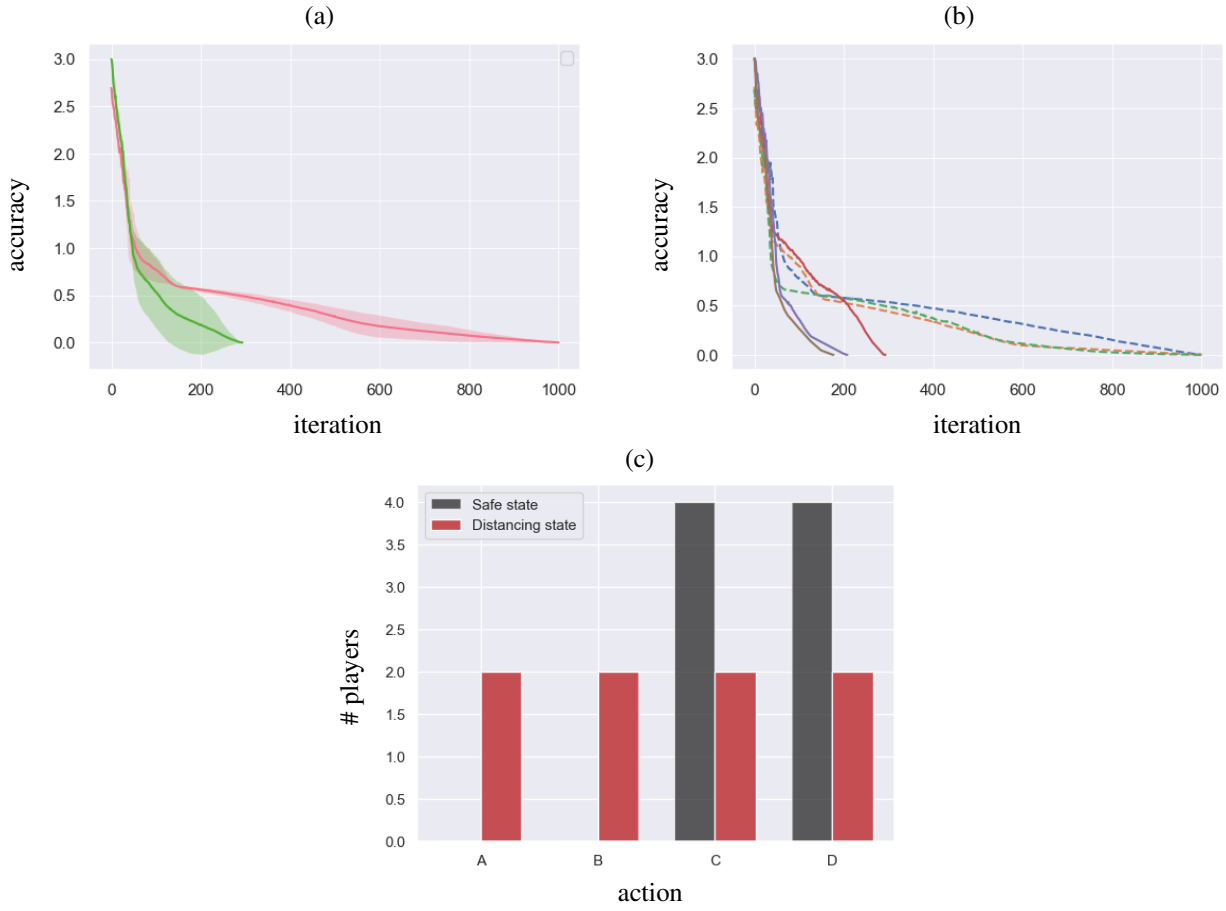

Figure 6. Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.001$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ (Leonardos et al., 2022). Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N}\sum_{i=1}^{N} \|\pi_i^{(t)} - \pi_i^{\text{Nash}}\|_1$. In our computational experiments, the initial distribution is nearly degenerate $\rho = (0.0001, 0.9999)$.

We also examine how sensitive the performance of algorithms depends on initial state distributions. As discussed in Section 4, our independent policy gradient method (4) is different from the projected policy gradient (3) by removing the dependence on the initial state distribution. In the policy gradient theory (Agarwal et al., 2021), convergence of projected policy gradient methods is often restricted by how explorative the initial state distribution is. To be fair, we choose stepsize $\eta = 0.001$ for our algorithm since it achieves a similar performance as the projected stochastic gradient ascent (Leonardos et al., 2022) in Figure 2. We choose two different initial state distributions $\rho = (0.9999, 0.0001)$ and $\rho = (0.0001, 0.9999)$ and report our computational results in Figure 5 and Figure 6, respectively. Compared Figure 5 with Figure 2, both algorithms become a bit slower, but our algorithm is relatively insusceptible to the change of ρ. This becomes more clearer in Figure 6 for another $\rho = (0.0001, 0.9999)$. This demonstrates that practical performance of our independent policy gradient method (4) indeed is invariant to the initial distribution ρ.