Local Optimization Achieves Global Optimality in Multi-Agent Reinforcement Learning

Yulai Zhao [1] **Zhuoran Yang** [2] **Zhaoran Wang** [3] **Jason D. Lee** [1]

Abstract

Policy optimization methods with function approximation are widely used in multi-agent reinforcement learning. However, it remains elusive how to design such algorithms with statistical guarantees. Leveraging a multi-agent performance difference lemma that characterizes the landscape of multi-agent policy optimization, we find that the localized action value function serves as an ideal descent direction for each local policy. Motivated by the observation, we present a multi-agent PPO algorithm in which the local policy of each agent is updated similarly to vanilla PPO. We prove that with standard regularity conditions on the Markov game and problem-dependent quantities, our algorithm converges to the globally optimal policy at a sublinear rate. We extend our algorithm to the off-policy setting and introduce pessimism to policy evaluation, which aligns with experiments. To our knowledge, this is the first provably convergent multi-agent PPO algorithm in cooperative Markov games.

1. Introduction

Recently, multi-agent reinforcement learning (MARL) has demonstrated many empirical successes, e.g., popular strategy games such as Go (Silver et al., 2016), StarCraft II (Vinyals et al., 2019), and poker (Brown & Sandholm, 2018). In contrast to vanilla reinforcement learning (RL), which is only concerned with a single agent seeking to maximize the total reward, MARL studies how multiple agents interact with the shared environment and other agents.

[1]Department of Electrical and Computer Engineering, Princeton University, NJ, USA [2]Department of Statistics and Data Science, Yale University, CT, USA [3]Department of Industrial Engineering and Management Sciences, Northwestern University, IL, USA. Correspondence to: Jason D. Lee <jasonlee@princeton.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Policy optimization methods are widely used in MARL. These algorithms often parameterize policies with a function class and compute the gradients of the cumulative reward using the policy gradient theorem (Sutton et al., 1999) or its variants (e.g., NPG (Kakade, 2001) and PPO (Schulman et al., 2017)) to update the policy parameters.

Despite the empirical successes, theoretical studies of policy optimization in MARL are very limited. Even for the cooperative setting where the agents share a common goal: maximizing the total reward function, numerous challenges arise (Zhang et al., 2021). (1) non-stationarity: each action taken by one agent affects the total reward and the transition of state. Consequently, each learning agent must learn to adapt to the changing environment caused by other agents. From the optimization perspective, the geometry of the multi-agent policy optimization problem becomes unclear. Direct application of traditional single-agent analysis becomes vague due to the lack of stationary Markovian property, which states that evolution in the future only depends on the previous state and individual action. (2) scalability: taking other agents into consideration, each individual agent would face the joint action space, whose dimension increases exponentially with the number of agents. Thus, having numerous agents in the environment problematizes the theoretical analysis of MARL. (3) function approximation: closely related to the scalability issue, the state space and joint action space are often immense in MARL, promoting function approximation to become a necessary component in MARL at the ease of computation and statistical analysis.

In this paper, we aim to answer the following fundamental question:

Can we design a provably convergent multi-agent policy optimization algorithm in the cooperative setting with function approximation?

We answer the above question affirmatively. We propose a multi-agent PPO algorithm in which the local policy of each agent is updated **sequantially** in a similar fashion as vanilla PPO algorithm (Schulman et al., 2017). In particular, we leverage a multi-agent performance difference lemma (cf. Lemma 4.1), assuming the joint policy is decomposed into conditional dependent policies. Such a lemma characterizes

the landscape of policy optimization, showing the superiority of using localized action value functions as the decent direction for each local policy. Such factorized structure essentially bypasses the non-stationarity and scalability concerns. To address large state spaces, we parameterize each local policy using log-linear parametrization and propose to update the policy parameters via KL divergence-regularized mirror descent, where the descent direction is estimated separately. Combining these results, we obtain our multi-agent PPO algorithm. We prove that the multi-agent PPO algorithm converges to globally optimal policy at a sublinear rate. Furthermore, we extend multi-agent PPO to the off-policy setting in which policy is evaluated using samples collected according to *data distribution* μ. We prove similar theoretical guarantees under a coverage assumption of the sampling distribution.

We summarize our contributions below.

Our contributions. First, by focusing on the factorized policies, we prove a multi-agent version of the performance difference lemma showing that the action value functions are ideal descent directions for local policies. Such a geometric characterization functions as a remedy for the non-stationarity concern, motivating our multi-agent PPO algorithm.

Second, we adopt the log-linear function approximation for the policies. We prove that multi-agent PPO converges at a sublinear $\mathcal{O}\left(\frac{N}{1-\gamma}\sqrt{\frac{\log|\mathcal{A}|}{K}}\right)$ rate up to some statistical errors incurred in evaluating/improving policies, where K is the number of iterations, N is the number of agents and $|\mathcal{A}|$ is the action space of each individual agent. The sample complexity depends polynomially on N, thus breaking the curse of scalability.

Third, we propose an off-policy variant of the multi-agent PPO algorithm and introduce pessimism into policy evaluation. The algorithm also converges sublinearly to the globally optimal policy up to the statistical error $\widetilde{\mathcal{O}}(n^{-\frac{1}{3}})$. Here, n is the number of samples used to estimate the critics.[1] A key feature of the sample complexity bound is that it only requires single-policy concentrability.

To our knowledge, this is the first provably convergent multi-agent PPO algorithm in cooperative Markov games with function approximation.

Organization. This paper is organized as follows. In Section 2, we review related literature. In Section 3, we formally describe the problem setup and introduce the necessary definitions. In Section 4, we state the main multi-agent PPO algorithm in detail. We further extend our results to the off-policy setting in Section 5. We conclude in Section 6 and defer the proofs to the Appendix.

[1] $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

2. Related Work

Policy optimization Many empirical works have proven the validity and efficiency of policy optimization methods in games and other applications (Silver et al., 2016; 2017; Guo et al., 2016; Tian et al., 2019). These works usually update the policy parameter in its parametric space using the pioneering policy gradient (PG) theorem by Sutton et al. (1999), or many PG variants invented to improve the empirical performances of vanilla PG methods. In particular, Kakade (2001) introduced the natural policy gradient (NPG) algorithm which searched for the steepest descent direction within the parameter space based on the idea of KL divergence-regularization. Trust region learning-based algorithms are often regarded as advanced policy optimization methods in practice (Lillicrap et al., 2015; Duan et al., 2016), showing superior performances with stable updates. Specifically, TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017) could be seen as KL divergence-constrained variants of NPG. A benign feature of these algorithms is the monotonic improvement guarantees of the expected return.

Despite prosperous empirical findings, the lack of convexity often impedes the development of theories for policy optimization methods. Denote K and T as the number of iterations and samples. Agarwal et al. (2020) showed an iteration complexity of $\mathcal{O}(K^{-\frac{1}{2}})$ and a sample complexity of $\mathcal{O}(T^{-\frac{1}{4}})$ for online NPG with function approximation. Shani et al. (2020) considered a sample-based TRPO and proved a $\tilde{\mathcal{O}}(T^{-\frac{1}{2}})$ rate converging to the global optimum, which could be improved to $\tilde{\mathcal{O}}(1/T)$ when regularized. Making minor modifications to the vanilla PPO algorithm, Liu et al. (2019) presented a convergence rate of $\mathcal{O}(K^{-\frac{1}{2}})$ to global optima when parameterizing both policy and Q functions with neural networks. The key to their analysis is the desirable one-point monotonicity in infinite-dimensional mirror descent that assists in characterizing the policy updates without convexity. We also make use of similar one-point properties in our multi-agent PPO algorithm analysis.

MARL Markov Game (MG) is a commonly used model to characterize the multi-agent decision-making process (Shapley, 1953; Littman, 1994), which can be regarded as a multi-agent extension to the Markov Decision Process (MDP). Policy-based algorithms could generalize to large states through function approximation. There has been growing interest in developing provably efficient algorithms for Markov games (Daskalakis et al., 2020; Cen et al., 2021; Zhao et al., 2022; Ding et al., 2022; Cen et al., 2022). These works often studied competitive RL settings, e.g., zero-sum games. Their convergence rates usually depended on vari-

ous notions of concentrability coefficient and may not scale tightly under the worst scenario.

Policy optimization for MARL Applying policy optimization methods in the MARL setting is more complicated than in the single-agent setting because of the non-stationary environment faced by each agent (Zhang et al., 2021). A learning paradigm called centralized training with decentralized execution (CTDE) is often used in practice (Kraemer & Banerjee, 2016; Lowe et al., 2017; Foerster et al., 2018; Yang et al., 2018; Wen et al., 2019; Zhang et al., 2020). In CTDE, a joint centralized value function helps to address the non-stationarity issue caused by other agents. Each agent has access to the global state and actions of other agents during training, thus allowing them to adjust their policy parameters individually. For instance, Lowe et al. (2017) proposed a multi-agent policy gradient algorithm in which agents learned a centralized critic based on the observations and actions of all agents.

Trust region learning (Schulman et al., 2015) has recently been combined with the CTDE paradigm to ensure monotonic improvements. In particular, IPPO (de Witt et al., 2020) and MAPPO (Yu et al., 2021) showed strong performances of PPO-based methods in the cooperative setting. The practical efficacy of these methods is usually restricted by the *homogeneity* assumption, where the agents share a common action space and policy parameter. Theoretically, providing statistical guarantees for policy optimization algorithms in MARL is more complicated than single-agent scenario (Zhang et al., 2021). In Markov games, the non-stationary environment faced by each agent precludes direct application of the single-agent convergence analysis. A recent attempt by Kuba et al. (2022) proposed the first set of trust region learning algorithms in MARL that enjoyed monotonic improvement guarantees assuming neither homogeneity of agents nor value function decomposition rule. The critical observation leading to their results is the multi-agent advantage function decomposition rule that builds the sequential policy update structure. However, they did not show rates of convergence. In this work, we design a new, provably convergent PPO algorithm for fully cooperative Markov games that converges to globally optimal at policy at sublinear rates by taking advantage of this conditional dependency structure.

Pessimism-based RL methods Though being able to account for large state/action spaces, function approximation also has its own drawbacks. A significant issue arising in using function approximators is the usual occurrence of a positive bias in value function (Thrun & Schwartz, 1993). The learner may not receive an accurate assessment. Numerous empirical works leverage the principle of *pessimism* to correct such overestimation (Fujimoto et al., 2018; Laskin et al., 2020; Lee et al., 2020; Moskovitz et al., 2021). For example, to reduce the evaluation bias brought by function approximation, Fujimoto et al. (2018) constructed the Bellman target by choosing the minimum of two value estimates as an intuitive estimate lower bound. Their approach took a pessimistic view of the value function.

On the theoretical side, a growing body of literature in offline reinforcement learning has also focused pessimism to account for datasets lacking data coverage (Liu et al., 2020; Jin et al., 2021; Uehara & Sun, 2021; Rashidinejad et al., 2021; Zhan et al., 2022). Technically, these works aimed at maximizing the worst-case rewards that a trained agent could obtain. Instead of relying on coverage assumptions on dataset (Munos, 2003; Munos & Szepesvári, 2008; Chen & Jiang, 2019), these methods provided dataset-dependent performance bounds, thus providing robust results for datasets lacking exploration for which traditional methods do not apply. We focus on the off-policy setting in Section 5 where we leverage the Bellman-consistent pessimism (Xie et al., 2021). We show concrete bounds under linear function approximation by assuming a sampling oracle that provides rewards and transition estimates that are used in approximating action value functions.

3. Preliminaries

In this section, we introduce necessary notations, problem setup, and some useful quantities that will be frequently used in this work.

3.1. Setup and Notations

Setup We consider a *fully-cooperative* Markov game (Shapley, 1953; Littman, 1994), which is defined by a tuple $(\mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{P}, r, \gamma)$. Here, $\mathcal{N} = \{1, \ldots, N\}$ denotes the set of agents, $\mathcal{S}$ is the finite state space, $\boldsymbol{\mathcal{A}} = \mathcal{A}^N$ is the product of finite action spaces of all agents(i.e., joint action space), $\mathcal{P} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0, 1]$ decides the transition scheme, a reward function $r : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to [0, 1]$, and $\gamma \in [0, 1)$ is the discount factor.[2] The agents interact with the environment according to the following protocol: at time step t, the agents are at state $s_t \in \mathcal{S}$; every agent i takes action $a_t^i \in \mathcal{A}$, drawn from its policy $\pi^i(\cdot|s_t)$, which together with actions of other agents gives a joint action $\mathbf{a}_t = (a_t^1, \ldots, a_t^N) \in \boldsymbol{\mathcal{A}}$, drawn from the joint policy $\boldsymbol{\pi}(\cdot|s_t) = \prod_{i=1}^N \pi^i(\cdot|s_t)$; the agents receive a joint reward $r_t = r(s_t, \mathbf{a}_t) \in \mathbb{R}$, and move to $s_{t+1} \sim \mathcal{P}(\cdot|s_t, \mathbf{a}_t)$. Given the joint policy $\boldsymbol{\pi}$, the transition probability function $\mathcal{P}$, and the initial state distribution ρ, we define the discounted occupancy state-action distribu-

[2]For clarity, we assume N agents share the same set of actions. It is straightforward to generalize our results to the setting where action sets are different. See Section 4.

tion as

$$d_{\boldsymbol{\pi}}(s,\mathbf{a}) = (1-\gamma)\mathbb{E}\sum_{t=0}^{\infty}\Pr^{\boldsymbol{\pi}}(s_t = s, a_t = \mathbf{a}|s_0 \sim \rho).$$

The standard value function and action value function are defined as

$$V_{\boldsymbol{\pi}}(s) \triangleq \mathop{\mathbb{E}}_{\mathbf{a}_{0:\infty}\sim\boldsymbol{\pi}, s_{1:\infty}\sim\mathcal{P}}\left[\sum_{t=0}^{\infty}\gamma^t r_t \middle| s_0 = s\right],$$

$$Q_{\boldsymbol{\pi}}(s,\boldsymbol{a}) \triangleq \mathop{\mathbb{E}}_{s_{1:\infty}\sim\mathcal{P}, \mathbf{a}_{1:\infty}\sim\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r_t \middle| s_0 = s,\ \mathbf{a}_0 = \boldsymbol{a}\right].$$

The standard advantage function considering all agents is written as $A_{\boldsymbol{\pi}}(s,\boldsymbol{a}) \triangleq Q_{\boldsymbol{\pi}}(s,\boldsymbol{a}) - V_{\boldsymbol{\pi}}(s)$. Later, we shall introduce the agents-specific advantage functions.

Let $\nu_{\boldsymbol{\pi}}(s)$ and $\sigma_{\boldsymbol{\pi}}(s,\mathbf{a}) = \boldsymbol{\pi}(\mathbf{a}|s)\cdot\nu_{\boldsymbol{\pi}}(s)$ denote the stationary state distribution and the stationary state-action distribution associated with a joint policy $\boldsymbol{\pi}$, respectively. Define the underlying optimal policy as $\boldsymbol{\pi}_*$. We use ν_* and σ_* in this paper to indicate $\nu_{\boldsymbol{\pi}_*}$ and $\sigma_{\boldsymbol{\pi}_*}$ for simplicity.

Throughout this paper, we pay close attention to the contribution of different subsets of agents to the performance of the whole team. We introduce the following multi-agent notations before proceeding to multi-agent definitions.

Notations In this work, we index the N agents with integers from 1 to N and use set $\mathcal{N} = \{i | i = 1, \cdots, N\}$ to represent all agents. We use $m \in \mathcal{N}$ to indicate the specific m-th agent. In particular, the set notation on the superscript of a term represents the quantities associated with agents in that set. For example, $\mathbf{a}^{\{1,2,3\}}$ represents the joint action of agents $1, 2$ and 3. We may write index k on superscript when we refer to the specific k-th agent. When bold symbols are used without any superscript (e.g., $\mathbf{a}$), they consider all agents. For simplicity, let $(m : m')$ be shorthand for set: $\{i | m \le i \le m', i \in \mathcal{N}\}$. An example is $\boldsymbol{\pi}^{1:m}(\cdot|s)$ which represents the joint policy considering agents $1, 2\cdots, m$.

We now introduce the multi-agent action value functions and advantage functions that characterize contributions from specific sub-agents.

Definition 3.1. Let P be a subset in $\mathcal{N}$. The multi-agent action value function associated with agents in P is

$$Q_{\boldsymbol{\pi}}^{P}\left(s,\mathbf{a}^{P}\right) \triangleq \mathbb{E}_{\tilde{\mathbf{a}}\sim\tilde{\boldsymbol{\pi}}}\left[Q_{\boldsymbol{\pi}}\left(s,\mathbf{a}^{P},\tilde{\mathbf{a}}\right)\right],$$

here we use a tilde over symbols to refer to the complement agents, namely $\tilde{\mathbf{a}} = \{a^i | i \notin P, i \in \mathcal{N}\}$.

Let $P, P' \subseteq \mathcal{N}$ be two disjoint subsets of agents. The multi-agent advantage function is defined below. Essentially, it accounts for the improvements of setting agents $\mathbf{a}^{P'}$ upon setting agents $\mathbf{a}^{P}$, while all other agents follow $\boldsymbol{\pi}$.

$$A_{\boldsymbol{\pi}}^{P'}\left(s,\mathbf{a}^{P},\mathbf{a}^{P'}\right) \triangleq Q_{\boldsymbol{\pi}}^{P\cup P'}\left(s,\mathbf{a}^{P},\mathbf{a}^{P'}\right) - Q_{\boldsymbol{\pi}}^{P}\left(s,\mathbf{a}^{P}\right).$$

The multi-agent Bellman operators are defined by generalizing the classic versions.

Definition 3.2. For $m \in \mathcal{N}$ and any function $f : \mathcal{S}\times\mathcal{A}^m \to \mathbb{R}$ we define *multi-agent Bellman operator* $\mathcal{T}_{\boldsymbol{\pi}}^{1:m} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}^m} \mapsto \mathbb{R}^{\mathcal{S}\times\mathcal{A}^m}$ as

$$\mathcal{T}_{\boldsymbol{\pi}}^{1:m} f(s,\mathbf{a}^{1:m}) := \mathop{\mathbb{E}}_{\tilde{\mathbf{a}}\sim\tilde{\boldsymbol{\pi}}} r(s,\mathbf{a}^{1:m},\tilde{\mathbf{a}}) + \gamma \mathop{\mathbb{E}}_{\substack{\tilde{\mathbf{a}}\sim\tilde{\boldsymbol{\pi}} \\ s'\sim\mathcal{P}(\cdot|s,\mathbf{a}^{1:m},\tilde{\mathbf{a}})}} f(s',\boldsymbol{\pi}^{1:m})$$

where $f(s',\boldsymbol{\pi}^{1:m})$ is shorthand for $\mathbb{E}_{\mathbf{a}'\sim\boldsymbol{\pi}^{1:m}(\cdot|s')} f(s',\mathbf{a}')$.

It is straightforward to see that $Q_{\boldsymbol{\pi}}^{1:m}$ is the unique fixed point for $\mathcal{T}_{\boldsymbol{\pi}}^{1:m}$, which corresponds to the classic single-agent Bellman operator.

3.2. KL divergence-regularized mirror descent

We review the mirror-decent formulation in provable single-agent PPO algorithm (Liu et al., 2019). At the k-th iteration, the policy parameter θ is updated via

$$\theta_{k+1} \leftarrow \arg\max_{\theta} \hat{\mathbb{E}}\big[\langle A_k(s,\cdot),\pi_\theta(\cdot|s)\rangle - \beta_k KL\left(\pi_\theta(\cdot\|s)\|\pi_{\theta_k}(\cdot|s)\right)\big]. \tag{1}$$

Hereafter we shall use $\langle\cdot,\cdot\rangle$ to represent the inner product over $\mathcal{A}$. The expectation is taken over $\hat{\mathbb{E}}$, which is an empirical estimate of stationary state-action distribution $\nu_{\pi_{\theta_k}}$, and A_k is estimate of advantage function $A^{\pi_{\theta_k}}$.

Adopting the KL-divergence, (1) is closely related to the NPG (Kakade, 2001) update. As a variant, this formulation is slightly different from the vanilla PPO (Schulman et al., 2017): here $KL(\pi_\theta(\cdot\|s)\|\pi_{\theta_k}(\cdot|s))$ is used instead of $KL(\pi_{\theta_k}(\cdot\|s)\|\pi_\theta(\cdot|s))$. Such variation is essential for presenting provable guarantees, which will be shown in the next section.

4. Multi-Agent PPO

Recall that ν_* is the stationary state distribution for $\boldsymbol{\pi}_*$. In this section, we desire to maximize the expected value function under distribution ν_*: $J(\boldsymbol{\pi}) \triangleq \mathbb{E}_{s\sim\nu_*} V_{\boldsymbol{\pi}}(s)$.

This paper aims to present a trust region learning multi-agent algorithm that enjoys a rigorous convergence theory. As we have mentioned, policy optimization for cooperative MARL is challenging because the policy optimization problem becomes a joint optimization involving all the agents. It remains unclear: (a) what the landscape of the total rewards as a multivariate function of the joint policy is and (b) what would be proper policy descent directions for each agent. We come up with a solution to characterize the landscape by taking advantage of a serial decomposition of the performance difference lemma in MARL described below.

Lemma 4.1. *For any joint policy $\boldsymbol{\pi}$ we have*

$$J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}) = \frac{1}{1-\gamma}\sum_{m=1}^{N}\mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}}$$
$$\left\langle Q_{\boldsymbol{\pi}}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot), \pi_*^m(\cdot|s,\mathbf{a}^{1:m-1}) - \pi^m(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle$$

where the inner product is over $a^m \in \mathcal{A}$.

With this geometric characterization, we can justify that using $Q_{\boldsymbol{\pi}}^m(s,\mathbf{a}^{1:m-1},a^m)$ as the descent direction and running KL divergence-regularized mirror descent for **each** agent $m \in \mathcal{N}$ can lead to a better **total** reward, which enables a serial optimization procedure. Below we describe the algorithm in detail.

To represent conditional policies, we adopt log-linear parametrization.

Parametrization For the m-th agent ($m \in \mathcal{N}$), its conditional policy depends on all prior ordered agents $\mathbf{a}^{1:m-1}$. Given a coefficient vector $\theta^m \in \Theta$, where $\Theta = \{\|\theta\| \leq R|\theta \in \mathbb{R}^d\}$ is a convex, norm-constrained set. The probability of choosing action a^m under state s is

$$\pi_{\theta^m}(a^m|s,\mathbf{a}^{1:m-1}) = \frac{\exp\left(\phi^\top(s,\mathbf{a}^{1:m-1},a^m)\theta^m\right)}{\sum_{a^m\in\mathcal{A}}\exp\left(\phi^\top(s,\mathbf{a}^{1:m-1},a^m)\theta^m\right)} \tag{2}$$

where ϕ is a set of feature vector representations. Without loss of generality, we impose a regularity condition such that every $\|\phi\|_2 \leq 1$. This parametrization has been widely used in RL literature (Branavan et al., 2009; Gimpel & Smith, 2010; Heess et al., 2013; Agarwal et al., 2020; Zhao et al., 2022).[3]

4.1. Policy Improvement and Evaluation

At the k-th iteration, we have the current policy $\boldsymbol{\pi}_{\theta_k}$, and we need to: (1) perform **policy evaluation** to obtain the action value function estimates $\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}$ for determining the quality of $\boldsymbol{\pi}_{\theta_k}$. (2) perform **policy improvement** to update policy to $\boldsymbol{\pi}_{\theta_{k+1}}$.

For notational simplicity, we use ν_k and σ_k to represent stationary state distribution $\nu_{\boldsymbol{\pi}_\theta^k}$ and the stationary state-action distribution $\sigma_{\boldsymbol{\pi}_\theta^k}$, which are induced by $\boldsymbol{\pi}_{\theta_k}$.

Policy Improvement At the k-th iteration, we define $\hat{\pi}_{k+1}^m$ as the ideal update based on $\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$ (for agent $m \in \mathcal{N}$), which is an estimator of $Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$. The ideal update is obtained via the following update

$$\hat{\pi}_{k+1}^m \leftarrow \arg\max_{\pi^m} \hat{F}(\pi^m) \tag{3}$$
$$\hat{F}(\pi^m) = \mathbb{E}_{\sigma_k}\Big[\langle \hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot), \pi^m(\cdot|s,\mathbf{a}^{1:m-1})\rangle$$
$$- \beta_k KL\left(\pi^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right)\Big]$$

where θ_k^m is the parameter of the current conditional policy of the m-th agent. In above equation, the distribution is taken over $(s,\mathbf{a}^{1:m-1}) \sim \nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1}$, we write σ_k for simplicity. Under log-linear parametrization: $\pi_{\theta_k^m} \propto \exp\{\phi^\top\theta_k^m\}$, we have the following closed-form ideal policy update.

Proposition 4.2. *Given an estimator $\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$, the KL divergence-regularized update* (3) *has explicit solution*

$$\hat{\pi}_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1}) \propto$$
$$\exp\left\{\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot) + \phi^\top(s,\mathbf{a}^{1:m-1},\cdot)\theta_k^m\right\}.$$

The proof is straightforward by adding the constraint: $\sum_{a^m\in\mathcal{A}}\pi^m(\cdot) = 1$ as a Lagrangian multiplier to $\hat{F}(\pi^m)$. See details in Appendix B.

To approximate the ideal $\hat{\pi}_{k+1}^m$ using a parameterized $\pi_{\theta_{k+1}^m} \propto \exp\{\phi^\top\theta_{k+1}^m\}$, we minimize the following mean-squared error (MSE) as a sub-problem

$$\theta_{k+1}^m \leftarrow \arg\min_{\theta^m\in\Theta} L(\theta^m) \tag{4}$$

where $L(\theta^m)$ is defined as

$$\mathop{\mathbb{E}}_{\sigma_k}\left((\theta^m - \theta_k^m)^\top\phi(s,\mathbf{a}^{1:m-1},a^m) - \frac{\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},a^m)}{\beta_k}\right)^2$$

Intuitively, a small $L(\theta)$ indicates that π_{θ^m} is close to the ideal update $\hat{\pi}_{k+1}^m$. Moreover, if $\hat{\pi}_{k+1}^m$ exactly lies in the log-linear function class, i.e., there exists a $\vartheta \in \Theta$ such that $\hat{\pi}_{k+1}^m \propto \exp\{\phi^\top\vartheta\}$. Then we have $L(\vartheta) = 0$.

To solve the MSE minimization problem (4), we use the classic SGD updates. Let stepsize be η, at each step $t = 0, 1, \cdots, T-1$, parameter θ is updated via

$$\theta(t+\frac{1}{2}) \leftarrow \theta(t) - 2\eta\phi\left((\theta(t) - \theta_k^m)^\top\phi - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}\right)$$
$$\theta(t+1) \leftarrow \Pi_\Theta\theta(t+\frac{1}{2})$$

where we omit $(s,\mathbf{a}^{1:m-1},a^m)$ for simplicity, which is sampled from σ_k. See Algorithm 3 for the detailed solver.

Policy Evaluation In this step, we aim to examine the quality of the attained policy. Thereby, a Q-function estimator is required. We make the following assumption.

[3] We assume that all players share the same parameter set only for clarity. We only need minor modifications in the analysis to extend our results to the setting where N agents have different capabilities. Specifically, we only need to treat norm bounds of updates (R), regularity conditions on features, and β separately for each agent.

Assumption 4.3. Assume we can access an estimator of Q function that returns $\hat{Q}$. The returned $\hat{Q}$ satisfies the following condition for all $m \in \mathcal{N}$ at the k-th iteration

$$\mathop{\mathbb{E}}_{\sigma_k}\left(\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m) - Q^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m)\right)^2 \leq (\xi^m_k)^2.$$

We also have a regularity condition for the estimator: there exists a positive constant B, such that for any $m \in \mathcal{N}$ and $(s, \mathbf{a}^{1:m-1}, a^m) \in \mathcal{S} \times \mathcal{A}^{m-1} \times \mathcal{A}$,

$$\left|\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m)\right| \leq B.$$

In RL practice, such an estimator is often instantiated with deep neural networks (DNNs) (Mnih et al., 2015). While there has been recent interest in studying the theoretical guarantees for DNNs as function approximators (Fan et al., 2020), we assume we have access to such an estimator to ensure the generality of our algorithm. We note that policy estimators like episodic sampling oracle that rolls out trajectories (Agarwal et al., 2020) or neural networks (Mnih et al., 2015; Liu et al., 2019) could all be possible options here. As a generalization, we introduce a specific value function approximation setting in Section 5, in which we assume all Q-functions lie in linear class $\mathcal{F}$. We further adopt the principle of pessimism for better exploration.

Algorithm Equipped with the sub-problem solver for policy improvement and the Q-function estimator, we are prepared to present the provable multi-agent PPO algorithm. The pseudo-code is listed in Algorithm 1. The algorithm runs for K iterations. At the k-th iteration, we estimate Q-function for each agent $m \in \mathcal{N}$ via the estimator (cf. Assumption 4.3) to measure the quality of $\boldsymbol{\pi}_{\theta_k}$. The estimates would also serve as the ideal descent direction for policy improvement. Since we use a constrained parametric policy class, the ideal update is approximated with the best policy parameter $\theta \in \Theta$ by minimizing the MSE problem (4), which runs SGD for T iterations (cf. Algorithm 3). Thanks to the geometric characterization (cf. Lemma 4.1), we are guaranteed to reach a globally improved total reward by updating each agent consecutively.

4.2. Theoretical Analysis

Our analysis relies on problem-dependent quantities. We denote weighted L_p-norm of function f on state-space $\mathcal{X}$ as $\|f\|_{p,\rho} = \left(\sum_{x\in\mathcal{X}} \rho(x)|f(x)|^p\right)^{\frac{1}{p}}$

Definition 4.4. At the k-th iteration, for $m \in \mathcal{N}$ we define the following problem-dependent quantity using Radon-Nikodym derivatives

$$\phi^m_k = \left\|\frac{d(\nu_* \boldsymbol{\pi}^{1:m}_*)}{d(\nu_k \boldsymbol{\pi}^{1:m}_{\theta_k})}\right\|_{2,\sigma_k}$$

Algorithm 1 Multi-Agent PPO

Input: Markov game $(\mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{P}, r, \gamma)$, penalty parameter β, stepsize η for sub-problem, number of SGD iterations T, number of iterations K.

Output: Uniformly sample k from $0, 1, \cdots K-1$, return $\bar{\boldsymbol{\pi}} = \boldsymbol{\pi}_{\theta_k}$.

Initialize $\theta^m_0 = 0$ for every $m \in \mathcal{N}$.
for $k = 0, 1, \ldots, K-1$ **do**
  Set parameter $\beta_k \leftarrow \beta\sqrt{K}$
  for $m = 1, \cdots, N$ **do**
    Sample $\{s_t, \mathbf{a}^{1:m-1}_t, a^m_t\}_{t=0}^{T-1}$ from $\sigma_k = \nu_k \boldsymbol{\pi}_{\theta_k}$.
    Obtain $\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m)$ for each sample .
    Feed samples into Algorithm 3, obtain θ^m_{k+1}.
  end for
end for

These conditions are the well-known concentrability coefficients (Munos, 2003; Farahmand et al., 2010; Chen & Jiang, 2019) for the factorized policy. Still, our conditions are structurally simpler and weaker because they are only density ratios between stationary state-action distributions, not requiring trajectories to roll out.

Now we are prepared to present the main theorem that characterizes the global convergence rate.

Theorem 4.5. *Under Assumption 4.3, for the output policy $\bar{\boldsymbol{\pi}}$ attained by Algorithm 1 in the fully cooperative Markov game, set $\eta = \frac{R}{G\sqrt{T}}$ and*

$$\beta = \sqrt{\frac{NB^2/2}{N\log|\mathcal{A}| + \sum_{m=1}^N \sum_{k=0}^{K-1}(\Delta^m_k + \delta^m_k)}}.$$

After K iterations, we have $J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}})$ upper bounded by

$$\mathcal{O}\left(\frac{B\sqrt{N}}{1-\gamma}\sqrt{\frac{N\log|\mathcal{A}| + \sum_{m=1}^N \sum_{k=0}^{K-1}(\Delta^m_k + \delta^m_k)}{K}}\right)$$

where $\Delta^m_k = \sqrt{2}(\phi^m_k + \phi^{m-1}_k)\cdot\left(\epsilon^m_k + \frac{\xi^m_k}{\beta_k}\right)$ and $\delta^m_k = 2\phi^{m-1}_k\epsilon^m_k$. Here ϵ^m_k is the statistical error of a PPO iteration: for agent $m \in \mathcal{N}$,

$$\mathbb{E}_{\sigma_k}\left((\theta^m_{k+1} - \theta^m_k)^\top\phi - \beta^{-1}_k\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}\right)^2 \leq (\epsilon^m_k)^2$$

where we omit $(s, \mathbf{a}^{1:m-1}, a^m)$ for simplicity.

Let ϵ_{approx} be the approximation capability of the log-linear policy class we adopt, then $\epsilon^m_k = \epsilon_{approx} + \mathcal{O}(T^{-\frac{1}{4}})$.

Theorem 4.5 explicitly characterizes the performance of the output $\bar{\boldsymbol{\pi}}$ in terms of the number of iterations and the iteration errors. When PPO updates are ideal, namely, viewing

δ'_k, Δ^m_k to be 0 for any $m \in \mathcal{N}$, and $k < K$, the rate simplifies to $\mathcal{O}\left(\frac{NB}{1-\gamma}\sqrt{\frac{\log|\mathcal{A}|}{K}}\right)$. The dependency on iteration K is $\mathcal{O}(K^{-\frac{1}{2}})$, matching the same rate as the sample-based single-agent NPG analysis (Agarwal et al., 2020; Liu et al., 2019).

The proof of Theorem 4.5 further requires the following parts: mirror-descent update analysis used in (Liu et al., 2019) and Lemma 4.1 that builds sequential dependency structure among the agents. The full proof is deferred to Appendix B.

4.3. Compare with Independent Learning

In MARL, independent learning refers to a class of algorithms that train multiple agents independently. In these methods, each agent has its own policy function that maps the agent's observations to its actions. The policies are optimized using policy gradient methods in a decentralized manner without explicit communication or coordination, and without explicitly modeling the behavior of the other agents. Independent learning methods are widely used in MARL due to its strong performance and efficiency.

In this subsection, we provide detailed comparisons between our algorithm and previous results on independent learning (both experiments and theories). We also performed a simulation study to showcase the superiority of our sequential policy update structure over naive independent policy gradient updates.

Experiments Some empirical attempts showed independent policy gradient learning could achieve surprisingly strong performance in MARL, such as MAPPO (Yu et al., 2021), IPPO (de Witt et al., 2020), and (Papoudakis et al., 2021).

Despite the empirical success, these methods have several drawbacks. IPPO and MAPPO assume homogeneity (agents share the same action space and policy parameters). Thus, parameter sharing is required. Even though the parameter sharing can be turned off, they still suffer from no monotonic improvement guarantees, though being called PPO-based algorithms. Recall that the main virtue of vanilla TRPO (Schulman et al., 2015) is monotonicity. Also, these methods do not come with any convergence guarantees. The converging problem becomes more severe when parameter-sharing is switched off. A counterexample in (Kuba et al., 2022, Proposition 1) shows parameter sharing could lead to an exponentially-worse sub-optimal outcome.

Thanks to the sequential agents' structure and novel multi-agent mirror-decent analyses, we present the first MARL algorithm that converges at a sub-linear rate. Note that our results neither rely on the homogeneity of agents nor the value function decomposition rule.

Theories Several theoretical works have studied convergence guarantees of independent policy optimization algorithms to a Nash equilibrium (NE) policy in MARL mathematically (Daskalakis et al., 2020; Leonardos et al., 2022; Fox et al., 2022; Ding et al., 2022).

Specifically, Daskalakis et al. (2020) studied competitive RL. And others studied convergence to the NE policy in Markov potential games (an extension of fully-cooperative games). However, we argue that a NE policy is not necessarily optimal in terms of the value function.

In contrast to their work, we present the first provable multi-agent policy optimization algorithm that finds a policy with a near globally optimal value function equipped with a sub-linear convergence rate.

Simulation To further validate the theoretical and experimental benefits of our algorithm, we conducted a numerical simulation to showcase the superiority of our algorithm with sequential updates structure over naive independent policy gradient updates. We consider von Neumann's ratio game, a simple stochastic game also used by Daskalakis et al. (2020). Simulation results show that, unlike our algorithm, the independent learning method has significant difficulty escaping the stationary point. Moreover, our algorithm consistently outperforms independent learning in maximizing value function. See Section E for detailed settings and results.

5. Pessimistic MA-PPO with Linear Function Approximation

In this section, we study the off-policy setting, using samples from a data distribution μ to evaluate $Q_{\boldsymbol{\pi}}$. Experimentally, since function approximators often cause a positive bias in value function (Thrun & Schwartz, 1993), many deep off-policy actor-critic algorithms introduce pessimism to reduce such overestimation (Fujimoto et al., 2018; Laskin et al., 2020). We also adopt pessimistic policy evaluation in this setting, aligning with experimental works.

We focus on the setting where value functions and policies are linearly parameterized. Our results can extend to the general function approximation setting, presented in Appendix D.

Definition 5.1 (Linear Function Approximation)**.** Let ϕ be a set of feature mappings built conditionally, the same definition as Section 4. Define the action value function class as $\mathcal{F}^m = \{\phi^\top\omega : \omega \in \mathbb{R}^d, \|\omega\|_2 \le L, \phi^\top\omega \in [0, 1/1-\gamma]\}$. The policy class is still parameterized by log-linear: $\Pi^m = \{\pi \propto \exp(\phi^\top\theta) : \theta \in \mathbb{R}^d, \|\theta\|_2 \le R\}$ (cf. Section 4).

Remark 5.2. Under the definition, for any $m \in \mathcal{N}$ and

policy $\boldsymbol{\pi}$, there must exist a parameter $\omega \in \mathbb{R}^d$ that satisfies

$$Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m}) = \phi(s, \mathbf{a}^{1:m})^\top \omega$$

In this section, we fix the initial state at a certain s_0. Thus the expected reward we aim to maximize is defined as

$$J(\boldsymbol{\pi}) \triangleq V_{\boldsymbol{\pi}}(s_0).$$

Note that, in single-agent offline RL, only one policy affects the action at a particular state so that we can gauge the quality of value function estimates using an offline dataset $\mathcal{D}$ consisting of states, actions, rewards, and transitions. Intuitively, when the following L_0 approaches 0, we can say f is a nice approximator for the Q-function (Xie et al., 2021).

$$L_0 = \frac{1}{n} \sum_{(s,a,r,s') \sim \mathcal{D}} \left(f(s,a) - r - \gamma f(s', \pi)\right)^2$$

where $f(s', \pi)$ is a shorthand for $\sum_{a'} f(s', a')\pi(a'|s')$ which will be frequently used in this section.

However, in the multi-agent environment, the complex dependent structure precludes the application of such an offline dataset. Specifically, for the m-th agent and policy $\boldsymbol{\pi}$, estimating the multi-agent value function $Q_{\boldsymbol{\pi}}^{1:m}$ demands that all agents not in $\{1:m\}$ must follow $\boldsymbol{\pi}$ (cf. Definition 3.1), which could not be guaranteed by an offline dataset.

Therefore, online interactions are unavoidable in the multi-agent setting we study. Below we make clarifications for the sample-generating protocol.

We will collect state-action samples from a fixed *data distribution* $\mu = \mu_s \mu_a \in \Delta(\mathcal{S} \times \boldsymbol{\mathcal{A}})$. In the benign case, a well-covered μ guarantees adequate exploration over the whole state and action spaces. Assume we have access to a standard RL oracle

Definition 5.3 (Sampling Oracle)**.** The oracle can start from $s \sim \mu_s$, take any action $\mathbf{a} \in \boldsymbol{\mathcal{A}}$, and obtain the next state $s' \sim \mathcal{P}(\cdot|s, \mathbf{a})$, and reward $r(s, \mathbf{a})$.

Our query oracle aligns with the classic **online sampling oracle** for MDP (Kakade & Langford, 2002; Du et al., 2019; Agarwal et al., 2020). The difference is that we transit for one step, while the classic online model usually terminates at the end of each episode. We also note that our oracle is weaker than the **generative model** (Kearns & Singh, 2002; Kakade, 2003; Sidford et al., 2018; Li et al., 2020) which assumes that agent can transit to **any** state, thus greatly weakening the need for explicit exploration. Whereas our oracle starts from a fixed μ_s.[4]

[4]In MDPs, such oracle is called μ-reset model (Kakade & Langford, 2002).

We take advantage of the sampler in the following steps to obtain action value functions that preserve a small error under the multi-agent Bellman operator (cf. Definition 5.1). For agent $m \in \mathcal{N}$ and $\boldsymbol{\pi}$, (1) obtain $s \sim \mu_s$; (2) obtain $\mathbf{a} \sim \mu_a$ and $\mathbf{a}' \sim \boldsymbol{\pi}^{m+1:N}(\cdot|s)$; (3) take $(\mathbf{a}^{1:m}, \mathbf{a}')$ as the joint action to query the oracle where $\mathbf{a}^{1:m}$ represents the $\{1:m\}$ subset of $\mathbf{a}$. The oracle returns (r, s'), which are guaranteed to satisfy:

$$r \sim \mathop{\mathbb{E}}_{\tilde{\mathbf{a}} \sim \boldsymbol{\pi}^{m+1:N}} R(s, \mathbf{a}^{1:m}, \tilde{\mathbf{a}}),$$
$$s' \sim \mathop{\mathbb{E}}_{\tilde{\mathbf{a}} \sim \boldsymbol{\pi}^{m+1:N}} \mathcal{P}(\cdot|s, \mathbf{a}^{1:m}, \tilde{\mathbf{a}}).$$

Repeat the steps for n times. Together this gives dataset $\mathcal{D}^m = \{(s_i, \mathbf{a}_i^{1:m}, r_i, s_i')|i = 1, 2, \cdots n\}$. Define

$$L^{1:m}(f', f, \boldsymbol{\pi}) := \frac{1}{n} \sum_{\mathcal{D}^m} \left(f'(s, \mathbf{a}^{1:m}) - r - \gamma f(s', \boldsymbol{\pi}^{1:m})\right)^2$$

where $f \in \mathcal{F}^m$ (cf. Definition 5.1) and the summation is taken over n quadruples of $(s, \mathbf{a}^{1:m}, r, s')$.

We will need the following Bellman error to evaluate the quality of f.

$$\mathcal{E}^{1:m}(f, \boldsymbol{\pi}) = L^{1:m}(f, f, \boldsymbol{\pi}) - \min_{f' \in \mathcal{F}^m} L^{1:m}(f', f, \boldsymbol{\pi}). \quad (5)$$

Intuitively, we consider f as a nice approximation of $Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m})$ when the quantity is small. This formulation also works for general function approximation. See Appendix D for details.

We shall need a concentrability measure accounting for the distributional mismatch.

Definition 5.4 (Concentrability)**.** The following condition characterizes the distribution shift from the $d_{\boldsymbol{\pi}_*}$ to the sampling distribution.

$$\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}} = \sup_{m \in \mathcal{N}, f \in \mathcal{F}^m, \boldsymbol{\pi} \in \Pi^m} \frac{\left\|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\right\|_{2, d_{\boldsymbol{\pi}_*}}}{\left\|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\right\|_{2, \mathcal{D}^m}}.$$

Recall that $\|\cdot\|_{2,\rho}$ is the weighted L_2-norm. In the nominator, the sum is taken over $(s, \mathbf{a}^{1:m}) \sim d_{\boldsymbol{\pi}_*}$. Whereas in the denominator, the sum is taken over $(s, \mathbf{a}^{1:m})$ from $\mathcal{D}^m$ as an empirical version of μ. The notion serves a similar role as concentrability coefficients in the literature (Munos, 2003; Agarwal et al., 2020): it measures the distributional mismatch between the underlying optimal distribution and the distribution of samples we employ.

Policy Evaluation At the k-th iteration, we have the current policy $\boldsymbol{\pi}_{\theta_k}$. We perform pessimistic policy evaluation via regularization to reduce value bias in evaluating $Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$.

$$\omega_k^m \leftarrow \arg\min_{\omega} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_{\theta_k})\right).$$

Here $\mathcal{E}$ is the Bellman error defined in (5). We obtain $f_k^m = \phi^\top \omega_k^m$ as the pessimistic estimate for $Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$. This update has a closed-form solution under linear function approximation (cf. Definition 5.1). Moreover, under linear function approximation, the minimization on the right-hand side can be solved computationally efficiently because of its quadratic dependency on ω. See details in Appendix C

Policy Improvement When both value functions and policies are linear parameterized (cf. Definition 5.1), the mirror descent policy update for any $(s, \mathbf{a}^{1:m}) \in \mathcal{S} \times \mathcal{A}^m$

$$\pi_{k+1}^m(a^m|s, \mathbf{a}^{1:m-1}) \propto \pi_k^m(a^m|s, \mathbf{a}^{1:m-1}) \cdot \exp(\eta f_k^m(s, \mathbf{a}^{1:m})) \quad (6)$$

could be further simplified to parameter updates in $\mathbb{R}^d$

$$\theta_{k+1}^m = \theta_k^m + \eta \omega_k^m.$$

This observation makes policy improvements in this setting significantly more superficial than in Section 4. For the k-th iteration and agent $m \in \mathcal{N}$, we only need to add $\eta\omega_k^m$ to the policy parameter θ_k^m to improve policy.

Algorithm With the pessimistic policy evaluation and intuitive policy improvement, our pessimistic variant of the multi-agent PPO algorithm is presented in Algorithm 2.

Algorithm 2 Pessimistic Multi-Agent PPO with Linear Function Approximation

Input: Regularization coefficient λ.
Output: Uniformly sample k from $0, 1 \cdots K-1$, return $\bar{\boldsymbol{\pi}} = \boldsymbol{\pi}_{\theta_k}$.

1: Initialize $\theta_0^m = 0$ for every $m \in \mathcal{N}$.
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: **for** $m = 1, 2, \cdots, N$ **do**
4: Pessimistic policy evaluation: $\omega_k^m \leftarrow \arg\min_{\omega} \left(f(s_0, \boldsymbol{\pi}_{\theta_k}^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_{\theta_k}) \right)$.
5: Policy improvement: $\theta_{k+1}^m = \theta_k^m + \eta\omega_k^m$.
6: **end for**
7: **end for**

Now we are prepared to present the main theorem for this section.

Theorem 5.5. *For the output policy $\bar{\boldsymbol{\pi}}$ attained by Algorithm 2 in a fully cooperative Markov game, set $\eta = (1-\gamma)\sqrt{\frac{\log|\mathcal{A}|}{2K}}$ and $\lambda = (1-\gamma)^{-1}\left(\frac{d \log \frac{nLR}{\delta}}{n}\right)^{-2/3}$. After K iterations, w.p. at least $1-\delta$ we have $J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}})$ upper bounded by*

$$\mathcal{O}\left(\frac{N}{(1-\gamma)^2}\sqrt{\frac{\log|\mathcal{A}|}{K}} + \frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{(1-\gamma)^2}\sqrt[3]{\frac{d \log \frac{nLR}{\delta}}{n}} \right)$$

To interpret this bound, the first term accounts for the optimization error accumulating from mirror descent updates (6). The first term has an $(1-\gamma)^{-2}$ dependency on the discount factor, which may not be tight, and we leave it as a future work to improve. The second term represents the estimation errors accumulated during training. We use state-action pairs from μ and the sampling oracle for minimizing $\mathcal{E}^{1:m}(f, \boldsymbol{\pi})$, thereby introducing *distribution mismatch* which is expressed by $\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}$. Note that this single-policy concentrability is already weaker than traditional concentrability coefficients (Munos, 2003; Farahmand et al., 2010; Perolat et al., 2015). Intuitively, a small value of concentrability requires the data distribution μ close to $d_{\boldsymbol{\pi}_*}$, which is the unknown occupancy distribution of optimal policy. On the other hand, if $\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}$ is large, then the bound becomes loose. We provide a similar result for general function approximation in the appendix (cf. Theorem D.7).

There is no explicit dependence on state-space $\mathcal{S}$ in the theorem. Hence the online algorithm proves nice guarantees for function approximation even in the infinite-state setting.

In proving Theorem 5.5, the analysis for Bellman-consistent pessimism (Xie et al., 2021) is useful. We obtain statistical and convergence guarantees by taking advantage of the conditional dependency structure of the cooperative Markov games. See Appendix C for details.

6. Conclusion

In this paper, we present a new multi-agent PPO algorithm that converges to the globally optimal policy at a sublinear rate. The key to the algorithm is a multi-agent performance difference lemma which enables sequential local policy updates. As a generalization, we extend the algorithm to the off-policy setting and present similar convergence guarantees. To our knowledge, this is the first multi-agent PPO algorithm in cooperative Markov games that enjoys provable guarantees.

Acknowledgements

JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Branavan, S. R. K., Chen, H., Zettlemoyer, L. S., and Barzilay, R. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pp. 82–90, USA, 2009. Association for Computational Linguistics.

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

Cen, S., Wei, Y., and Chi, Y. Fast policy extragradient methods for competitive games with entropy regularization. In *Advances in Neural Information Processing Systems*, pp. 27952–27964. Curran Associates, Inc., 2021.

Cen, S., Chi, Y., Du, S. S., and Xiao, L. Faster last-iterate convergence of policy optimization in zero-sum Markov games. *arXiv preprint arXiv:2210.01050*, 2022.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

de Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H., Sun, M., and Whiteson, S. Is independent learning all you need in the StarCraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pp. 5166–5220. PMLR, 2022.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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. PMLR, 2022.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Gimpel, K. and Smith, N. A. Softmax-margin CRFs: Training log-linear models with cost functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 733–736, 2010.

Guo, X., Singh, S., Lewis, R., and Lee, H. Deep learning for reward design to improve Monte Carlo tree search in Atari games. *arXiv preprint arXiv:1604.07095*, 2016.

Heess, N., Silver, D., and Teh, Y. W. Actor-critic reinforcement learning with energy-based policies. In *European Workshop on Reinforcement Learning*, pp. 45–58. PMLR, 2013.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732. PMLR, 2017.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kakade, S. M. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Kraemer, L. and Banerjee, B. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190:82–94, 2016.

Kuba, J. G., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J., and Yang, Y. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EcGGFkNTxdJ.

Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., and Srinivas, A. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33: 19884–19895, 2020.

Lee, K.-H., Fischer, I., Liu, A., Guo, Y., Lee, H., Canny, J., and Guadarrama, S. Predictive information accelerates learning in RL. *Advances in Neural Information Processing Systems*, 33:11890–11901, 2020.

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. URL https://openreview.net/forum?id=gfwON7rAm4.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

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

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

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

Lowe, R., Wu, Y. I., 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, 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Moskovitz, T., Parker-Holder, J., Pacchiano, A., Arbel, M., and Jordan, M. Tactical optimism and pessimism for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12849–12863, 2021.

Munos, R. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, pp. 560–567, 2003.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Papoudakis, G., Christianos, F., Schäfer, L., and Albrecht, S. V. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Perolat, J., Scherrer, B., Piot, B., and Pietquin, O. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pp. 1321–1329, 2015.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

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

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5668–5675, 2020.

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

Sidford, A., Wang, M., Wu, X., Yang, L. F., and Ye, Y. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*, 2018.

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

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Thrun, S. and Schwartz, A. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*, volume 255, pp. 263. Hillsdale, NJ, 1993.

Tian, Y., Ma, J., Gong, Q., Sengupta, S., Chen, Z., Pinkerton, J., and Zitnick, L. Elf opengo: An analysis and open reimplementation of AlphaZero. In *International Conference on Machine Learning*, pp. 6244–6253. PMLR, 2019.

Uehara, M. and Sun, W. Pessimistic model-based offline RL: Pac bounds and posterior sampling under partial coverage. *arXiv e-prints*, pp. arXiv–2107, 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.

Wen, Y., Yang, Y., Luo, R., Wang, J., and Pan, W. Probabilistic recursive reasoning for multi-agent reinforcement learning. *arXiv preprint arXiv:1901.09207*, 2019.

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pp. 550–559. PMLR, 2020.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pp. 5571–5580. PMLR, 2018.

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.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhang, H., Chen, W., Huang, Z., Li, M., Yang, Y., Zhang, W., and Wang, J. Bi-level actor-critic for multi-agent coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7325–7332, 2020.

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, 2021.

Zhao, Y., Tian, Y., Lee, J., and Du, S. Provably efficient policy optimization for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pp. 2736–2761. PMLR, 2022.

A. Sub-problem Solver for Section 4

Algorithm 3 Policy Improvement Solver for MA-PPO

Input: MG $(\mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{P}, r, \gamma)$, iterations T, stepsize η, samples $\{s_t, \mathbf{a}_t^{1:m-1}, a_t^m\}_{t=0}^{T-1}$.
Output: Policy update θ.
1: Initialize $\theta_0 = 0$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Let $(s, \mathbf{a}^{1:m-1}, a) \leftarrow (s_t, \mathbf{a}_t^{1:m-1}, a_t^m)$.
4: $\theta(t+\frac{1}{2}) \leftarrow \theta(t) - 2\eta\phi(s, \mathbf{a}^{1:m-1}, a)\left((\theta(t) - \theta_k^m)^\top \phi(s, \mathbf{a}^{1:m-1}, a^m) - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_k}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)\right)$.
5: $\theta(t+1) \leftarrow \Pi_\Theta \theta(t+\frac{1}{2})$
6: **end for**
7: Calculate average: $\bar{\theta} \leftarrow \frac{1}{T}\sum_{t=1}^T \theta_t$.

B. Proofs for Section 4

First, we note that using SGD updates to solve the MSE problem has the following guarantee.

Lemma B.1 (Average policy)**.** *For a convex objective function $F(\theta)$, suppose the gradient is bounded by G, and the output $\bar{\theta}$ converges to the best function in the class at*

$$F(\bar{\theta}) - \min_{\|\theta\| \le R} F(\theta) \le \frac{GR}{\sqrt{T}}$$

where we set $\eta = \frac{R}{G\sqrt{T}}$.

Proof. Please refer to Theorem 14.8 (Shalev-Shwartz & Ben-David, 2014). □

Now we turn to Algorithm 3, in which, we feed samples $\{s_t, \mathbf{a}_t^{1:m-1}, a_t^m\}_{t=0}^{T-1}$ from $\sigma_k = \nu_k \boldsymbol{\pi}_{\theta_k}$ into the algorithm (for $m \in \mathcal{N}$), in order to minimize

$$L(\theta^m) = \mathbb{E}_{\sigma_k}\left((\theta^m)^\top \phi(s, \mathbf{a}^{1:m-1}, a^m) - (\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m) + (\theta_k^m)^\top \phi(s, \mathbf{a}^{1:m-1}, a^m))\right)^2.$$

We have the following theoretical guarantee for the algorithm.

Lemma B.2 (Policy Improvement error)**.** *At the k-th outer loop, the output policy θ_{k+1} from Algorithm 3 satisfies*

$$\sqrt{L(\theta_{k+1}^m)} \le \epsilon_k^m$$

where $\epsilon_k^m = \epsilon_{approx} + \mathcal{O}(T^{-\frac{1}{4}})$.

Proof. For Algorithm 3, we have $\|\phi\|_2 \le 1$ and $\Theta = \{\|\theta\| \le R | \theta \in \mathbb{R}^d\}$.Thereby the gradient of $L(\theta)$ is bounded by

$$G = 2\left(R + \frac{1}{(1-\gamma)\beta_k}\right).$$

From Lemma B.1, we have the following guarantee holds for any outer iteration $k < K$

$$L(\theta_{k+1}^m) \le= \min_\theta L(\theta) + \mathcal{O}(1/\sqrt{T}),$$

when we set $\eta = \frac{R}{G\sqrt{T}}$. Thus

$$\epsilon_k^m = \sqrt{\min_\theta L(\theta)} + \mathcal{O}(T^{-\frac{1}{4}}) = \epsilon_{approx} + \mathcal{O}(T^{-\frac{1}{4}}).$$

The proof is completed. □

Lemma B.3 (Multi-Agent Advantage Decomposition)**.** *In cooperative Markov games, the following decomposition holds for any joint policy π, state s, and agents $1:m$,*

$$A_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m}) = \sum_{i=1}^{m} A_{\boldsymbol{\pi}}^{i}(s, \mathbf{a}^{1:i-1}, a^i)$$

Proof. Please refer to Lemma 1 (Kuba et al., 2022). □

Proof for Lemma 4.1

Proof. From the classical performance difference lemma (Kakade & Langford, 2002, Lemma 6.1) we have

$$J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}) = \frac{1}{1-\gamma}\mathbb{E}_{\sigma_*} A_{\boldsymbol{\pi}}^{1:N}(s, \mathbf{a}^{1:N})$$

Decomposing the all-agents advantage function into individual contributions via the multi-agent advantage decomposition lemma (cf. Lemma B.3), we have

$$\begin{aligned} J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}) &= \frac{1}{1-\gamma}\mathbb{E}_{\sigma_*} A_{\boldsymbol{\pi}}^{1:N}(s, \mathbf{a}^{1:N}) \\ &= \frac{1}{1-\gamma}\sum_{m=1}^{N} \mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}} \left\langle A_{\boldsymbol{\pi}}^{m}(s, \mathbf{a}^{1:m-1}, a^m), \pi_*^m(\cdot|s, \mathbf{a}^{1:m-1}) \right\rangle. \end{aligned}$$

Note that we have $\sum_a \pi^m(a|s, a^{1:m-1}) A_{\boldsymbol{\pi}}^m(s, \mathbf{a}^{1:m-1}, a) = 0$, then

$$\begin{aligned} &\frac{1}{1-\gamma}\sum_{m=1}^{N} \mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}} \left\langle A_{\boldsymbol{\pi}}^{m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi_*^m(\cdot|s, \mathbf{a}^{1:m-1}) - \pi^m(\cdot|s, a^{1:m-1}) \right\rangle \\ =& \frac{1}{1-\gamma}\sum_{m=1}^{N} \mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}} \left\langle Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi_*^m(\cdot|s, \mathbf{a}^{1:m-1}) - \pi^m(\cdot|s, a^{1:m-1}) \right\rangle \end{aligned}$$

where the last line is because $A_{\boldsymbol{\pi}}^m(s, \mathbf{a}^{1:m-1}, a^m) = Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m) - Q_{\boldsymbol{\pi}}^{1:m-1}(s, \mathbf{a}^{1:m-1})$ and $Q_{\boldsymbol{\pi}}^{1:m-1}(s, \mathbf{a}^{1:m-1})$ can be omitted because it does not change with a^m.

□

Proof for Proposition 4.2.

Proof. For any $(s, \mathbf{a}^{1:m-1}) \in \mathcal{S} \times \mathcal{A}^{m-1}$, policy $\hat{\pi}_{k+1}^m(\cdot|s, \mathbf{a}^{1:m-1})$ is obtained via

$$\begin{aligned} \max_{\pi^m} \mathbb{E}_{\nu_k} &\left[\langle \hat{Q}_{\boldsymbol{\pi}_k}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \rangle - \beta_k KL\left(\pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \| \pi_{\theta_k}(\cdot|s, \mathbf{a}^{1:m-1})\right) \right] \\ &\text{s.t.} \quad \sum_{a^m\in\mathcal{A}} \pi^m(a^m|s, \mathbf{a}^{1:m-1}) = 1 \end{aligned}$$

Adding constraint as a Lagrangian multiplier, we have

$$\begin{aligned} &\int_{\mathcal{S}\times\mathcal{A}^{m-1}} \left[\langle \hat{Q}_{\boldsymbol{\pi}_k}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \rangle - \beta_k KL\left(\pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \| \pi_{\theta_k^m}(\cdot|s, \mathbf{a}^{1:m-1})\right) \right] \sigma_k ds d\mathbf{a}^{1:m-1} \\ &\quad + \int_{\mathcal{S}\times\mathcal{A}^{m-1}} \left(\sum_{a^m\in\mathcal{A}} \pi^m(a^m|s, \mathbf{a}^{1:m-1}) - 1 \right) ds d\mathbf{a}^{1:m-1} \end{aligned}$$

Note that $\pi_{\theta_k^m} \propto \exp\{\theta_k^{m\top}\phi\}$, he optimality condition gives

$$\hat{\pi}_{k+1}^m(\cdot|s, \mathbf{a}^{1:m-1}) \propto \exp\{\beta_k^{-1}\hat{Q}_{\pi_k}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot) + \theta_k^{m\top}\phi(s, \mathbf{a}^{1:m-1}, a)\}.$$

□

Lemma B.4. *Suppose for any agent $m \in \mathcal{N}$, policy improvement error and policy evaluation errors satisfy*

$$\mathbb{E}_{\sigma_k}\left(\theta_{k+1}^{m\top}\phi(s, \mathbf{a}^{1:m-1}, a^m) - (\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m) + \theta_k^{m\top}\phi(s, \mathbf{a}^{1:m-1}, a^m))\right)^2 \leq (\epsilon_k^m)^2, \tag{7}$$

$$\mathbb{E}_{\sigma_k}\left(\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m) - Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)\right)^2 \leq (\xi_k^m)^2. \tag{8}$$

Considering the L_∞-norm of $\theta_{k+1}^{m\top}\phi - \theta_k^{m\top}\phi$ we have

$$\mathbb{E}_{s\sim\nu_*, \mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}\left\|(\theta_{k+1}^m - \theta_k^m)^\top\phi(s, \mathbf{a}^{1:m-1}, \cdot) - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot)\right\|_\infty \leq \frac{\delta_k^m}{2}$$

where $\delta_k^m = 2\phi_k^{m-1}\epsilon_k^m$.

Proof. The proof is straightforward,

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu_*, \mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}}\left\|(\theta_{k+1}^m - \theta_k^m)^\top\phi(s, \mathbf{a}^{1:m-1}, \cdot) - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot)\right\|_\infty \\
\leq &\mathop{\mathbb{E}}_{\substack{s\sim\nu_* \\ \mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}}\left\|(\theta_{k+1}^m - \theta_k^m)^\top\phi(s, \mathbf{a}^{1:m-1}, a^m) - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)\right\|.
\end{aligned}$$

We shift from σ_* to σ_k and introduce concentrability coefficients to measure distributional shift

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{s\sim\nu_k \\ \mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_{\theta_k}}}\left\|(\theta_{k+1}^m - \theta_k^m)^\top\phi(s, \mathbf{a}^{1:m-1}, a^m) - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)\right\| \cdot \frac{\nu_*\boldsymbol{\pi}_*^{1:m-1}}{\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1}} \\
\leq &\left[\mathbb{E}_{\sigma_k}\left((\theta_{k+1}^m - \theta_k^m)^\top\phi - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right)^2\right]^{1/2} \cdot \left[\mathbb{E}_{\sigma_k}\left|\frac{d(\nu_*\boldsymbol{\pi}_*^{1:m-1})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1})}\right|^2\right]^{1/2} \\
= &\epsilon_k^m\phi_k^{m-1}
\end{aligned}$$

where we use Cauchy-Schwartz inequality in the second line.

The proof is completed. □

Recall that we define $\hat{\pi}_{k+1}^m$ as the ideal update policy based on $\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$. Correspondingly, we define the ideal update based on the exact value function $Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$ as

$$\begin{aligned}
\pi_{k+1}^m &\leftarrow \arg\max_{\pi^m} F(\pi^m) \\
F(\pi^m) &= \mathbb{E}_{\sigma_k}\Big[\langle Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi^m(\cdot|s, \mathbf{a}^{1:m-1})\rangle - \beta_k KL\left(\pi^m(\cdot|s, \mathbf{a}^{1:m-1})\|\pi_{\theta_k^m}(\cdot|s, \mathbf{a}^{1:m-1})\right)\Big].
\end{aligned}$$

Under log-linear parametrization: $\pi_{\theta_k^m} \propto \exp\{\phi^\top\theta_k^m\}$, analogously we have

$$\pi_{k+1}^m(\cdot|s, \mathbf{a}^{1:m-1}) \propto \exp\left\{\beta_k^{-1}Q_{\pi_k}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot) + \phi^\top(s, \mathbf{a}^{1:m-1}, \cdot)\theta_k^m\right\}.$$

Lemma B.5 (Error Propagation)**.** *Suppose for any agent $m \in \mathcal{N}$ and $(s, \mathbf{a}^{1:m-1}) \in \mathcal{S} \times \mathcal{A}^{m-1}$, policy improvement and policy evaluation errors satisfy,*

$$\mathbb{E}_{\sigma_k}\left((\theta_{k+1}^m - \theta_k^m)^\top\phi - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}\right)^2 \leq (\epsilon_k^m)^2,$$

$$\mathbb{E}_{\sigma_k}\left(\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}} - Q^{1:m}_{\boldsymbol{\pi}_{\theta_k}}\right)^2 \le (\xi_k^m)^2$$

where we omit $(s, \mathbf{a}^{1:m-1}, a^m)$ *for simplicity.*

Compare the statistical error, we have

$$\left|\mathbb{E}_{s\sim\nu_*,\mathbf{a}\sim\boldsymbol{\pi}_*}\left\langle \log\frac{\pi_{\theta^m_{k+1}}(\cdot|s,\mathbf{a}^{1:m-1})}{\pi^m_{k+1}(\cdot|s,\mathbf{a}^{1:m-1})}, \pi^m_*(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_{\theta^m_k}(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle\right| \le \Delta^m_k.$$

where $\Delta^m_k = \sqrt{2}(\phi^m_k + \phi^{m-1}_k)\cdot\left(\epsilon^m_k + \frac{\xi^m_k}{\beta_k}\right)$

Lemma B.5 presents the quantitative differences between the actual parameterized $\pi_{\theta^m_{k+1}}$ based on $\hat{Q}^{1:m}$ and the ideal policy π^m_{k+1} based on the exact value function $Q^{1:m}$.

Proof. First, from definition for any $m \in \mathcal{N}$ and $s, \mathbf{a}^{1:m-1} \in \mathcal{S}\times\mathcal{A}^{m-1}$ we have

$$\pi^m_{k+1}(\cdot|s,\mathbf{a}^{1:m-1}) = \exp\left\{{\beta_k}^{-1}Q_{\boldsymbol{\pi}_{\theta_k}} + {\theta^m_k}^\top\phi(s,\mathbf{a}^{1:m-1},\cdot)\right\}/W(s,\mathbf{a}^{1:m-1}), \tag{9}$$

$$\pi_{\theta^m_{k+1}}(\cdot|s,\mathbf{a}^{1:m-1}) = \exp\left\{{\theta^m_{k+1}}^\top\phi(s,\mathbf{a}^{1:m-1},\cdot)\right\}/M(s,\mathbf{a}^{1:m-1}). \tag{10}$$

Substituting this into the expression, we have

$$\begin{aligned}
\text{LHS} &= \left\langle \log\pi_{\theta^m_{k+1}} - \log\pi^m_{k+1}, \pi^m_* - \pi_{\theta^m_k}\right\rangle \\
&= \left\langle {\theta^m_{k+1}}^\top\phi - (\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}} + {\theta^m_k}^\top\phi), \pi^m_* - \pi_{\theta^m_k}\right\rangle \\
&= \underbrace{\left\langle {\theta^m_{k+1}}^\top\phi - (\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} + {\theta^m_k}^\top\phi), \pi^m_* - \pi_{\theta^m_k}\right\rangle}_{(a)} + \underbrace{\left\langle \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} - \beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k}\right\rangle}_{(b)}
\end{aligned}$$

In the second line, we use the fact that $\left\langle \log\frac{W}{M}, \pi^m_* - \pi_{\theta^m_k}\right\rangle = \log\frac{W}{M}\sum_{a^m}\left(\boldsymbol{\pi}^m_*(\cdot) - \pi_{\theta^m_k}(\cdot)\right) = 0.$

Bounding the two terms separately, we have

- For (a), taking the expectation over $\mathcal{S}\times\mathcal{A}^{m-1}$ we have

$$\begin{aligned}
&\left|\mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}\left\langle {\theta^m_{k+1}}^\top\phi - (\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} + {\theta^m_k}^\top\phi), \pi^m_* - \pi_{\theta^m_k}\right\rangle\right| \\
&= \left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[(\theta^m_{k+1} - \theta^m_k)^\top\phi - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\left(\pi^m_* - \pi_{\theta^m_k}\right)da^m\cdot\boldsymbol{\pi}^{1:m-1}_*d(\mathbf{a}^{1:m-1})\cdot\nu_*(s)ds\right|
\end{aligned}$$

  Here the expectation is taken w.r.t. $\sigma*$, we change it to be expectation over σ_k by introducing concentrability coefficients

$$\begin{aligned}
&\left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[(\theta^m_{k+1} - \theta^m_k)^\top\phi - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\left(\frac{\boldsymbol{\pi}^{1:m}_*}{\boldsymbol{\pi}^{1:m}_{\theta_k}} - \frac{\boldsymbol{\pi}^{1:m-1}_*\pi_{\theta^m_k}}{\boldsymbol{\pi}^{1:m}_{\theta_k}}\right)\boldsymbol{\pi}^{1:m}_{\theta_k}(\mathbf{a}^{1:m}|s)d(\mathbf{a}^{1:m})\cdot\nu_*(s)ds\right| \\
&= \left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[(\theta^m_{k+1} - \theta^m_k)^\top\phi - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\frac{\nu_*(s)}{\nu_k(s)}\left(\frac{\boldsymbol{\pi}^{1:m}_*}{\boldsymbol{\pi}^{1:m}_{\theta_k}} - \frac{\boldsymbol{\pi}^{1:m-1}_*\pi_{\theta^m_k}}{\boldsymbol{\pi}^{1:m}_{\theta_k}}\right)d\sigma_k\right| \\
&\overset{(i)}{\le} \left[\mathbb{E}_{\sigma_k}\left((\theta^m_{k+1} - \theta^m_k)^\top\phi - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right)^2\right]^{1/2}\cdot\left[\mathbb{E}_{\sigma_k}\left|\frac{d(\nu_*\boldsymbol{\pi}^{1:m}_*)}{d(\nu_k\boldsymbol{\pi}^{1:m}_{\theta_k})} - \frac{d(\nu_*\boldsymbol{\pi}^{1:m-1}_*)}{d(\nu_k\boldsymbol{\pi}^{1:m-1}_{\theta_k})}\right|^2\right]^{1/2} \\
&\overset{(ii)}{\le} \sqrt{2}\epsilon^m_k(\phi^m_k + \phi^{m-1}_k)
\end{aligned}$$

  (i) : This is because of the Cauchy-Schwartz inequality.

  (ii) : This is because: $\sqrt{\int|f-g|^2d\sigma} \le \sqrt{\int 2|f|^2d\sigma + \int 2|g|^2d\sigma} \le \sqrt{2}(\|f\|_{2,\sigma} + \|g\|_{2,\sigma})$.

- For (b), taking the expectation over $\mathcal{S} \times \mathcal{A}^{m-1}$ we have

$$\left|\mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}\left\langle \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} - \beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}, \pi_*^m - \pi_{\theta_k^m}\right\rangle\right|$$
$$= \left|\int\limits_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}} \left[\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} - \beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\left(\pi_*^m - \pi_{\theta_k^m}\right) da^m \cdot \boldsymbol{\pi}_*^{1:m-1} d(\mathbf{a}^{1:m-1})\cdot\nu_*(s)ds\right|$$

Analogously we replace the expectation over $\sigma*$ with expectation over σ_k

$$\left|\int\limits_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}} \left[\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} - \beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\frac{\nu_*(s)}{\nu_k(s)}\left(\frac{\boldsymbol{\pi}_*^{1:m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}} - \frac{\boldsymbol{\pi}_*^{1:m-1}\pi_{\theta_k^m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}\right)\boldsymbol{\pi}_{\theta_k}^{1:m}(\mathbf{a}^{1:m}|s)d(\mathbf{a}^{1:m})\cdot\nu_k(s)ds\right|$$
$$= \left|\int\limits_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}} \left[\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} - \beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\frac{\nu_*(s)}{\nu_k(s)}\left(\frac{\boldsymbol{\pi}_*^{1:m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}} - \frac{\boldsymbol{\pi}_*^{1:m-1}}{\boldsymbol{\pi}_{\theta_k}^{1:m-1}}\right)d\sigma_k\right|$$
$$\overset{(i)}{\leq} \left[\mathbb{E}_{\sigma_k}\left(\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}} - \beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right)^2\right]^{1/2}\cdot\left[\mathbb{E}_{\sigma_k}\left|\frac{d(\nu_*\boldsymbol{\pi}_*^{1:m})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m})} - \frac{d(\nu_*\boldsymbol{\pi}_*^{1:m-1})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1})}\right|^2\right]^{1/2}$$
$$\overset{(ii)}{\leq} \frac{\sqrt{2}}{\beta_k}\xi_k^m(\phi_k^m + \phi_k^{m-1})$$

(i) : This is because of the Cauchy-Schwartz inequality.
(ii) : This is because: $\sqrt{\int|f-g|^2d\sigma} \leq \sqrt{\int 2|f|^2d\sigma + \int 2|g|^2d\sigma} \leq \sqrt{2}(\|f\|_{2,\sigma} + \|g\|_{2,\sigma})$.

Combining the bounds, we conclude the proof for Lemma B.5

$$\Delta_k^m = \sqrt{2}(\phi_k^m + \phi_k^{m-1})\cdot\left(\epsilon_k^m + \frac{\xi_k^m}{\beta_k}\right).$$

□

Below we introduce a lemma that is crucial in our multi-agent PPO analysis. The original version is widely found and proven to be useful for mirror descent analysis (Nesterov, 2003).

Lemma B.6 (One-Step Descent)**.** *For the ideal updated policy π_{k+1}^m, the real updated policy $\pi_{\theta_{k+1}^m}$ and current policy $\pi_{\theta_k^m}$, we have that for any $(s, \mathbf{a}^{1:m-1}) \in \mathcal{S}\times\mathcal{A}^{m-1}$,*

$$KL\left(\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})\right) - KL\left(\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right)$$
$$\leq \left\langle \log\frac{\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})}{\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})}, \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle$$
$$+ \frac{1}{\beta_k}\left\langle A_{\pi_{\theta_k}}^m(s,\mathbf{a}^{1:m-1},\cdot), \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle - \frac{1}{2}\left\|\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right\|_1^2$$
$$- \left\langle(\theta_{k+1}^m - \theta_k^m)^\top\phi(s,\mathbf{a}^{1:m-1},\cdot), \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle$$

Proof. In the proof, we simply omit $(s, \mathbf{a}^{1:m-1})$ when making no abuse of notation.

Using the definition, we have

$$KL\left(\pi_*^m\|\pi_{\theta_k^m}\right) - KL\left(\pi_*^m\|\pi_{\theta_{k+1}^m}\right)$$
$$= \left\langle \log\frac{\pi_{\theta_{k+1}^m}}{\pi_{\theta_k^m}}, \pi_*^m\right\rangle$$

$$= \left\langle \log \frac{\pi_{\theta^m_{k+1}}}{\pi_{\theta^m_k}}, \pi^m_* - \pi_{\theta^m_{k+1}} \right\rangle + KL\left(\pi_{\theta^m_{k+1}} \| \pi_{\theta^m_k}\right)$$
$$= \left\langle \log \frac{\pi_{\theta^m_{k+1}}}{\pi_{\theta^m_k}} - \beta_k^{-1} Q_{\pi_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle + \left\langle \beta_k^{-1} Q_{\pi_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle + KL\left(\pi_{\theta^m_{k+1}} \| \pi_{\theta^m_k}\right) + \left\langle \log \frac{\pi_{\theta^m_{k+1}}}{\pi_{\theta^m_k}}, \pi_{\theta^m_k} - \pi_{\theta^m_{k+1}} \right\rangle. \tag{11}$$

Recall the definitions of π^m_{k+1} (9) and $\pi_{\theta^m_{k+1}}$ (10), we have the following two equations

$$\left\langle \log \frac{\pi_{\theta^m_{k+1}}}{\pi_{\theta^m_k}}, \pi_{\theta^m_k} - \pi_{\theta^m_{k+1}} \right\rangle = \left\langle (\theta^m_{k+1} - \theta^m_k)^\top \phi, \pi_{\theta^m_k} - \pi_{\theta^m_{k+1}} \right\rangle, \tag{12}$$
$$\left\langle \beta_k^{-1} Q_{\pi_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle = \left\langle \log \frac{\pi^m_{k+1}}{\pi_{\theta^m_k}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle. \tag{13}$$

Plugging these results (12), (13) into the RHS of (11) we have

$$KL\left(\pi^m_* \| \pi_{\theta^m_k}\right) - KL\left(\pi^m_* \| \pi_{\theta^m_{k+1}}\right)$$
$$= \left\langle \log \frac{\pi_{\theta^m_{k+1}}}{\pi^m_{k+1}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle + \left\langle \beta_k^{-1} Q^{1:m}_{\pi_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle + \left\langle (\theta^m_{k+1} - \theta^m_k)^\top \phi, \pi_{\theta^m_k} - \pi_{\theta^m_{k+1}} \right\rangle + KL\left(\pi_{\theta^m_{k+1}} \| \pi_{\theta^m_k}\right)$$
$$\geq \left\langle \log \frac{\pi_{\theta^m_{k+1}}}{\pi^m_{k+1}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle + \left\langle \beta_k^{-1} A^m_{\pi_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle + \left\langle (\theta^m_{k+1} - \theta^m_k)^\top \phi, \pi_{\theta^m_k} - \pi_{\theta^m_{k+1}} \right\rangle + \frac{1}{2} \left\| \pi_{\theta^m_{k+1}} - \pi_{\theta^m_k} \right\|_1^2$$

In the last line: (1) From the Definition 3.1 of multi-agent advantage functions, we have

$$\langle Q^{1:m}_{\pi_{\theta_k}} - A^m_{\pi_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k} \rangle = 0,$$

and (2) Pinsker's inequality in information theory gives a lower bound of the KL-divergence.

$$KL\left(\pi_{\theta^m_{k+1}} \| \pi_{\theta^m_k}\right) \geq \frac{1}{2} \left\| \pi_{\theta^m_{k+1}} - \pi_{\theta^m_k} \right\|_1^2,$$

plugging these into the expression, which concludes the proof. □

With these results, we are ready to present the proofs for the main theorem.

Proofs for Theorem 4.5.

Proof. With Lemma B.6, take expectation with respect to $s \sim \nu_*$ and $\mathbf{a} \sim \boldsymbol{\pi}_*$, we have

$$\frac{1}{\beta_k} \mathbb{E}_{\sigma_*} \left\langle A^m_{\pi_{\theta_k}}(s, \mathbf{a}^{1:m-1}, \cdot), \pi^m_*(\cdot|s, \mathbf{a}^{1:m-1}) - \pi_{\theta^m_k}(\cdot|s, \mathbf{a}^{1:m-1}) \right\rangle$$
$$\leq \mathbb{E}_{\sigma_*} \left[KL\left(\pi^m_*(\cdot|s, \mathbf{a}^{1:m-1}) \| \pi_{\theta^m_k}(\cdot|s, \mathbf{a}^{1:m-1})\right) - KL\left(\pi^m_*(\cdot|s, \mathbf{a}^{1:m-1}) \| \pi_{\theta^m_{k+1}}(\cdot|s, \mathbf{a}^{1:m-1})\right) \right]$$
$$- \mathbb{E}_{\sigma_*} \left\langle \log \frac{\pi_{\theta^m_{k+1}}(\cdot|s, \mathbf{a}^{1:m-1})}{\pi^m_{k+1}(\cdot|s, \mathbf{a}^{1:m-1})}, \pi^m_*(\cdot|s, \mathbf{a}^{1:m-1}) - \pi_{\theta^m_k}(\cdot|s, \mathbf{a}^{1:m-1}) \right\rangle$$
$$- \frac{1}{2} \mathbb{E}_{\sigma_*} \left\| \pi_{\theta^m_{k+1}}(\cdot|s, \mathbf{a}^{1:m-1}) - \pi_{\theta^m_k}(\cdot|s, \mathbf{a}^{1:m-1}) \right\|_1^2$$
$$- \mathbb{E}_{\sigma_*} \left\langle {\theta^m_{k+1}}^\top \phi(s, \mathbf{a}^{1:m-1}, \cdot) - {\theta^m_k}^\top \phi(s, \mathbf{a}^{1:m-1}, \cdot), \pi_{\theta^m_k}(\cdot|s, \mathbf{a}^{1:m-1}) - \pi_{\theta^m_{k+1}}(\cdot|s, \mathbf{a}^{1:m-1}) \right\rangle.$$

Analogous to the previous section, we omit $(s, \mathbf{a}^{1:m-1})$ below for simplicity when making no abuse of notation. Rearrange the inequality by plugging in Lemma. B.5

$$\frac{1}{\beta_k} \mathbb{E}_{\sigma_*} \left\langle A^m_{\pi_{\theta_k}}, \pi^m_* - \pi_{\theta^m_k} \right\rangle$$

$$
\begin{aligned}
&\leq \mathbb{E}_{\sigma_*}\left[KL\left(\pi_*^m \| \pi_{\theta_k^m}\right) - KL\left(\pi_*^m \| \pi_{\theta_{k+1}^m}\right)\right] \\
&\qquad - \mathbb{E}_{\sigma_*}\left[\left\langle \log \frac{\pi_{\theta_{k+1}^m}}{\pi_{k+1}^m}, \pi_*^m - \pi_{\theta_k^m}\right\rangle + \frac{1}{2}\left\|\pi_{\theta_{k+1}^m} - \pi_{\theta_k^m}\right\|_1^2 + \left\langle (\theta_{k+1}^m - \theta_k^m)^\top \phi, \pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\rangle\right] \\
&\leq \underbrace{\mathbb{E}_{\sigma_*}\left[KL\left(\pi_*^m \| \pi_{\theta_k^m}\right) - KL\left(\pi_*^m \| \pi_{\theta_{k+1}^m}\right)\right]}_{(i)} + \Delta_k^m \underbrace{- \mathbb{E}_{\sigma_*}\left[\frac{1}{2}\left\|\pi_{\theta_{k+1}^m} - \pi_{\theta_k^m}\right\|_1^2 + \left\langle (\theta_{k+1}^m - \theta_k^m)^\top \phi, \pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\rangle\right]}_{(ii)}
\end{aligned}
\tag{14}
$$

Take summation over $m = 1, \cdots, N$ and $k = 0, \cdots, K-1$ for both sides, note that

LHS This equals $(1-\gamma)\sum_{k=0}^{K-1} \frac{1}{\beta_k}(J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}_{\theta_k}))$ by performance difference lemma (cf. Lemma 4.1).

RHS (i) : After taking summation over k and m, this term is upper bounded by $N \log \mathcal{A}$ because for any $m \in \mathcal{N}$, we have $\mathop{\mathbb{E}}_{\sigma_*} KL\left(\pi_*^m \| \pi_{\theta_0}^m\right) \leq \log |\mathcal{A}|$ since the initial policy π_{θ_0} is uniformly distributed over action spaces.

(ii): Using Hölder inequality, we have

$$
-\mathop{\mathbb{E}}_{\sigma_*}\left\langle (\theta_{k+1}^m - \theta_k^m)^\top \phi, \pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\rangle \leq \mathop{\mathbb{E}}_{\sigma_*}\left\|(\theta_{k+1}^m - \theta_k^m)^\top \phi\right\|_\infty \cdot \left\|\pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\|_1
$$

Using the triangle inequality we can upper bound it by

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma_*}\left\|(\theta_{k+1}^m - \theta_k^m)^\top \phi - \beta_k^{-1}\hat{Q}\right\|_\infty \cdot \left\|\pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\|_1 + \mathop{\mathbb{E}}_{\sigma_*}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty \cdot \left\|\pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\|_1 \\
\leq\ & \delta_k^m + \mathop{\mathbb{E}}_{\sigma_*}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty \cdot \left\|\pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\|_1,
\end{aligned}
$$

where we plug in Lemma B.4 and: $\left\|\pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\|_1 \leq \left\|\pi_{\theta_k^m}\right\|_1 + \left\|\pi_{\theta_{k+1}^m}\right\|_1 = 2$.

We have

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{\sigma_*}\left[\frac{1}{2}\left\|\pi_{\theta_{k+1}^m} - \pi_{\theta_k^m}\right\|_1^2\right] + \mathop{\mathbb{E}}_{\sigma_*}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty \cdot \left\|\pi_{\theta_k^m} - \pi_{\theta_{k+1}^m}\right\|_1 \\
\leq\ & \mathop{\mathbb{E}}_{\sigma_*}\frac{1}{2}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty^2, \quad \text{because } \forall x, y \text{ it holds: } -\frac{1}{2}x^2 + yx \leq \frac{1}{2}y^2. \\
\leq\ & \frac{B^2}{2\beta_k^2}
\end{aligned}
$$

Finally, by combining these results, we rearrange (14) and obtain

$$
(1-\gamma)\sum_{k=0}^{K-1}\frac{1}{\beta_k}(J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}_{\theta_k})) \leq N \log |\mathcal{A}| + \sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m + \delta_k^m) + \sum_{m=1}^{N}\sum_{k=0}^{K-1}\frac{B^2}{2\beta_k^2}.
$$

Setting the penalty parameter $\beta_k = \beta\sqrt{K}$ and noting that $\bar{\boldsymbol{\pi}}$ is uniformly sampled from $\boldsymbol{\pi}_{\theta_k}, k = 1, 2 \cdots K-1$, we have

$$
J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}}) \leq \frac{N\beta^2 \log |\mathcal{A}| + NB^2/2 + \beta^2 \sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m + \delta_k^m)}{(1-\gamma)\beta\sqrt{K}}.
$$

Considering policy improvement/evaluation errors, the optimal choice for β is

$$
\beta = \sqrt{\frac{NB^2/2}{N \log |\mathcal{A}| + \sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m + \delta_k^m)}}
$$

then

$$
J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}}) \leq \mathcal{O}\left(\frac{B\sqrt{N}}{1-\gamma}\sqrt{\frac{N \log |\mathcal{A}| + \sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m + \delta_k^m)}{K}}\right).
$$

The proof is completed. □

C. Proofs for Section 5

C.1. Computational efficiency

Observe the pessimistic evaluation

$$\omega_k^m \leftarrow \arg\min_{\omega} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_{\theta_k}) \right).$$

Under linear function approximation, $f(s_0, \boldsymbol{\pi}^{1:m})$ is instantiated as $\phi(s_0, \boldsymbol{\pi}^{1:m})^\top \omega$, then

$$L^{1:m}(f', f, \boldsymbol{\pi}) = \frac{1}{n} \sum_{\mathcal{D}^m} \left(\phi(s, \mathbf{a}^{1:m})^\top \omega' - r - \gamma \phi(s', \boldsymbol{\pi}^{1:m})^\top \omega \right)^2,$$

thus we have

$$\mathcal{E}^{1:m}(f, \boldsymbol{\pi}) = \sum_{\mathcal{D}_m} (\phi(s, \mathbf{a}^{1:m})^\top \omega - r - \gamma \phi(s', \boldsymbol{\pi}^{1:m})^\top \omega)^2 - \min_{\omega'} \sum_{\mathcal{D}_m} (\phi(s, \mathbf{a}^{1:m})^\top \omega' - r - \gamma \phi(s', \boldsymbol{\pi}^{1:m})^\top \omega)^2,$$

summation is taken over $(s, \mathbf{a}^{1:m}, r, s')$ from $\mathcal{D}^m$.

Therefore, the Bellman error has a quadratic-form dependency on value function parameter ω, allowing the application of many efficient numerical solvers.

C.2. Proofs

The linear function approximation directly implies the **Realizability** and **Completeness** conditions: For any $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$,

$$\inf_{f \in \mathcal{F}^m} \sup_{\text{admissable } \nu} \|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\nu}^2 = 0,$$

and

$$\sup_{f \in \mathcal{F}^m} \inf_{f \in \mathcal{F}^m} \|f' - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\mu}^2 = 0.$$

These conditions hold because we assume a linear structure for state-action value functions: $Q_{\boldsymbol{\pi}}^{1:m} \in \mathcal{F}^m$ (cf. Definition 5.1).

First we examine concentration analysis for linear function approximation (Xie et al., 2021).

Lemma C.1. *For any $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$, with probability at least $1 - \delta$ it holds*

$$\mathcal{E}^{1:m}(Q^{\boldsymbol{\pi}}, \pi) \leq \mathcal{O}\left(\frac{d \log \frac{nLR}{\delta}}{(1-\gamma)^2 n} \right) = \varepsilon_r.$$

Lemma C.2. *For any $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$, $f \in \mathcal{F}^m$ (cf. Definition 5.1) , if $\mathcal{E}^{1:m}(f, \boldsymbol{\pi}) \leq \varepsilon$, with probability at least $1 - \delta$ it holds*

$$\left\| f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f \right\|_{2,\mathcal{D}^m} \leq \mathcal{O}\left(\sqrt{\frac{d \log \frac{nLR}{\delta}}{(1-\gamma)^2 n}} \right) + \sqrt{\varepsilon}.$$

For simplicity, below, we shall define

$$R(s, \mathbf{a}^{1:m}) = \mathbb{E}_{\tilde{\mathbf{a}} \sim \tilde{\boldsymbol{\pi}}_k} r(s, \mathbf{a}^{1:m}, \tilde{\mathbf{a}}).$$

In the following lemma, we show that at every iteration k of Algorithm 2, there exists a Markov game $\mathcal{M}_k$ whose multi-agent value function is exactly f_k^m, $m \in \mathcal{N}$. Moreover, the transition dynamics of $\mathcal{M}_k$ are the same as those of the original $\mathcal{M}$. We have the following theoretical guarantees to control the differences between the reward of $\mathcal{M}$ and rewards of $\mathcal{M}_k$.

Lemma C.3. *At each iteration k, there exists a Markov game $\mathcal{M}_k$ that has the same dynamics as original $\mathcal{M}$. Let the reward function of $\mathcal{M}_k$ be R_k, then*

$$\left\| R_k^{1:m}(s, \mathbf{a}^{1:m}) - R(s, \mathbf{a}^{1:m}) \right\|_{2,\mathcal{D}^m}^2 \leq \varepsilon_r.$$

Proof. Set $R_k = (I - \gamma\mathcal{P})f_k$. It directly implies that

$$\begin{aligned} f_k^m &= \mathcal{T}^{1:m}_{\boldsymbol{\pi}_k,\mathcal{M}_k}(s.\mathbf{a}^{1:m}) \\ &= R_k^{1:m}(s,\mathbf{a}^{1:m}) + \gamma\mathbb{E}f_k^m(s',\mathbf{a}'). \end{aligned}$$

Therefore

$$\left\|R_k^{1:m}(s,\mathbf{a}^{1:m}) - R(s,\mathbf{a}^{1:m})\right\|^2_{2,\mathcal{D}^m} = \|f_k^m - \mathcal{T}^{1:m}_{\boldsymbol{\pi}_k} f_k^m\|^2_{2,\mathcal{D}^m} \le \varepsilon_r$$

□

Lemma C.4. *For any conditional policy $\pi : \mathcal{S}\times\mathcal{A}^{m-1} \to \Delta(\mathcal{A})$, and $s\in\mathcal{S}, \mathbf{a}^{1:m-1}\in\mathcal{A}^{m-1}$,*

$$\begin{aligned} &\sum_{k=1}^K \left\langle \pi_{k+1}(\cdot|s,\mathbf{a}^{1:m-1}), f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle - \ell_{sa}(\pi_1(\cdot|s,\mathbf{a}^{1:m-1})) \\ \ge &\sum_{k=1}^K \left\langle \pi(\cdot|s,\mathbf{a}^{1:m-1}), f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle - \ell_{sa}(\pi) \end{aligned}$$

where we assume $\ell_{sa}(\pi) = \frac{1}{\eta}\sum_{a\in\mathcal{A}}\pi(a|s,\mathbf{a}^{1:m-1})\cdot\log\pi(a|s,\mathbf{a}^{1:m-1})$

Proof. Proofs are straightforward by noticing the fact that $\theta_{k+1}^m = \theta_k^m + \eta w_k^m$. □

Lemma C.5. *For any conditional policy $\pi : \mathcal{S}\times\mathcal{A}^{m-1} \to \Delta(\mathcal{A})$, and $s\in\mathcal{S}, \mathbf{a}^{1:m-1}\in\mathcal{A}^{m-1}$,*

$$\begin{aligned} &\sum_{k=1}^K \left\langle \pi(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_k(\cdot|s,\mathbf{a}^{1:m-1}), f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle \\ \le &\sum_{k=1}^K \left\langle \pi_{k+1}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_k(\cdot|s,\mathbf{a}^{1:m-1}), f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle - \ell_{sa}(\pi_1(\cdot|s,\mathbf{a}^{1:m-1})) \end{aligned}$$

Proof. The results could be obtained by applying Lemma C.4. □

Lemma C.6. *For any conditional policy $\pi : \mathcal{S}\times\mathcal{A}^{m-1} \to \Delta(\mathcal{A})$, and $s\in\mathcal{S}, \mathbf{a}^{1:m-1}\in\mathcal{A}^{m-1}$, if set stepsize $\eta = \sqrt{\frac{\log|\mathcal{A}|}{2K}}$*

$$\sum_{k=1}^K \left\langle \pi(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_k(\cdot|s,\mathbf{a}^{1:m-1}), f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle \le 2\sqrt{2\log|\mathcal{A}|K}$$

Proof. Define $\mathcal{L}^m_{sa,k} = \sum_{k'=1}^k \langle \pi(\cdot|s,\mathbf{a}^{1:m-1}), f_{k'}(s,\mathbf{a}^{1:m-1},\cdot)\rangle - \ell_{sa}(\pi)$. Let $B_{\mathcal{L}^m_{sa,k}}(\cdot\|\cdot)$ and $B_{\ell_{sa}}(\cdot\|\cdot)$ be the Bregman divergences w.r.t. losses $\mathcal{L}^m_{sa,k}$ and ℓ_{sa}. Using the property of divergence we have

$$\begin{aligned} \mathcal{L}^m_{sa,k}(\pi_k^m) &\le \mathcal{L}^m_{sa,k}(\pi_{k+1}^m) + B_{\mathcal{L}^m_{sa,k}}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})) \\ &= \mathcal{L}^m_{sa,k}(\pi_{k+1}^m) - B_{\ell_{sa}}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})). \end{aligned}$$

Reordering it we have

$$B_{\ell_{sa}}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})) \le \mathcal{L}^m_{sa,k}(\pi_{k+1}^m) - \mathcal{L}^m_{sa,k}(\pi_k^m).$$

RHS of the expression above is not greater than

$$\left\langle \pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_k^m(\cdot|s,\mathbf{a}^{1:m-1}), f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle.$$

Then

$$\left\langle \pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_k^m(\cdot|s,\mathbf{a}^{1:m-1}), f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle$$

$$\leq \sqrt{2\eta B_{\ell_{sa}}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1}))}\cdot\left\|f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\|_\infty$$
$$\leq \sqrt{2\eta}\sqrt{\langle\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\rangle}\frac{1}{1-\gamma}.$$

Thus we have $\left\langle\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle\leq\frac{2\eta}{(1-\gamma)^2}$. Substituting it into Lemma C.5 we have

$$\begin{aligned}&\sum_{k=1}^K\left\langle\pi(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle\\ \leq&\frac{2\eta K}{(1-\gamma)^2}+\frac{\log|\mathcal{A}|}{\eta}\\ \leq&\frac{2\sqrt{2\log|\mathcal{A}|K}}{1-\gamma}\end{aligned}$$

where the last line is by setting $\eta=(1-\gamma)\sqrt{\frac{\log|\mathcal{A}|}{2K}}$. □

Lemma C.7. *For any conditional policy $\pi:\mathcal{S}\times\mathcal{A}^{m-1}\to\Delta(\mathcal{A})$,*

$$Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})\geq\min_{f\in\mathcal{F}^m}\left(f(s_0,\boldsymbol{\pi}_k^{1:m})+\lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi}_k)\right)-\lambda\varepsilon_r.$$

Proof. For any conditional policy $\boldsymbol{\pi}^{1:m}$, and $\forall m\in\mathcal{N}$. With **realizability** assumption we have $Q_{\boldsymbol{\pi}}^{1:m}=\arg\min_f\sup_\nu\|f-\mathcal{T}_{\boldsymbol{\pi}}^{1:m}f\|_{2,\nu}^2$ for any admissible ν

$$\begin{aligned}J(\boldsymbol{\pi})&=Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})\\&=Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})-\left(Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})+\lambda\mathcal{E}^{1:m}(Q_{\boldsymbol{\pi}}^{1:m},\boldsymbol{\pi})\right)+\left(Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})+\lambda\mathcal{E}^{1:m}(Q_{\boldsymbol{\pi}}^{1:m},\boldsymbol{\pi})\right)\\&\geq\min_{f\in\mathcal{F}^m}\left(f(s_0,\boldsymbol{\pi}^{1:m})+\lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi})\right)-\lambda\varepsilon_r,\end{aligned}$$

where in the last line we use Lemma C.1. □

Proofs for Theorem 2.

Proof. Use Lemma C.7, at the k-th iteration we have

$$\begin{aligned}J(\boldsymbol{\pi}_k)&\geq\min_{f\in\mathcal{F}^m}\left(f(s_0,\boldsymbol{\pi}_k^{1:m})+\lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi}_k)\right)-\lambda\varepsilon_r\\&\geq f_k^m(s_0,\boldsymbol{\pi}_k^{1:m})-\lambda\varepsilon_r\\&=J_{\mathcal{M}_k}(\boldsymbol{\pi}_k)-\lambda\varepsilon_r\end{aligned}$$

where f_k^m is the multi-agent value function of Markov game $\mathcal{M}_k$ (cf. Lemma C.3).

Therefore we have,

$$\begin{aligned}J(\boldsymbol{\pi}_*)-J(\bar{\boldsymbol{\pi}})&=\frac{1}{K}\sum_{k=1}^K(J(\boldsymbol{\pi}_*)-J(\boldsymbol{\pi}_k))\\&\leq\frac{1}{K}\sum_{k=1}^K(J(\boldsymbol{\pi}_*)-J_{\mathcal{M}_k}(\boldsymbol{\pi}_k))+\lambda\varepsilon_r\\&\leq\underbrace{\frac{1}{K}\sum_{k=1}^K(J_{\mathcal{M}_k}(\boldsymbol{\pi}_*)-J_{\mathcal{M}_k}(\boldsymbol{\pi}_k))}_{\text{I}}+\underbrace{\frac{1}{K}\sum_{k=1}^K(J(\boldsymbol{\pi}_*)-J_{\mathcal{M}_k}(\boldsymbol{\pi}_*))}_{\text{II}}+\lambda\varepsilon_r\end{aligned}$$

Term I. Apply performance difference lemma we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}(J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_k)) \\
&= \frac{1}{1-\gamma}\frac{1}{K}\sum_{m=1}^{N}\mathbb{E}_{s\sim d_{\boldsymbol{\pi}_*}}\mathbb{E}_{\mathbf{a}^{1:m-1}}\sum_{k=1}^{K}\left(Q^{\boldsymbol{\pi}_k}_{\mathcal{M}_k}(s,\mathbf{a}^{1:m-1},\boldsymbol{\pi}^m_*) - Q^{\boldsymbol{\pi}_k}_{\mathcal{M}_k}(s,\mathbf{a}^{1:m-1},\boldsymbol{\pi}^m_k)\right) \\
&= \frac{1}{1-\gamma}\frac{1}{K}\sum_{m=1}^{N}\mathbb{E}_{s\sim d_{\boldsymbol{\pi}_*},\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}\sum_{k=1}^{K}\left\langle \boldsymbol{\pi}^m_*(\cdot|s,\mathbf{a}^{1:m-1}) - \boldsymbol{\pi}^m_k(\cdot|s,\mathbf{a}^{1:m-1}), f^m_k(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle
\end{aligned}$$

where the last line is because f^m_k is the multi-agent state-action value function for $\mathcal{M}_k$. Then, if $\eta = (1-\gamma)\sqrt{\frac{\log|\mathcal{A}|}{2K}}$, Lemma C.6 gives

$$\text{Term I} \le \frac{2N}{(1-\gamma)^2}\sqrt{\frac{2\log|\mathcal{A}|}{K}}.$$

Term II. The following analysis holds for any $m \in \mathcal{N}$ so we omit m for clarity. For this term, again, we use Lemma 1 (Xie & Jiang, 2020) to transform it into norm over state-action distributions

$$\begin{aligned}
J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) &= Q_{\boldsymbol{\pi}_*}(s,\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) \\
&\le \frac{1}{1-\gamma}\left\|Q_{\boldsymbol{\pi}_*} - \mathcal{T}_{\boldsymbol{\pi}_*,\mathcal{M}_k}Q_{\boldsymbol{\pi}_*}\right\|_{2,d_{\boldsymbol{\pi}_*}}
\end{aligned}$$

Note the definition of the auxiliary Markov game for which f^k is its value function(cf. Lemma C.3) we have

$$\begin{aligned}
&\frac{1}{1-\gamma}\left\|Q_{\boldsymbol{\pi}_*} - R_k - \gamma P_{\boldsymbol{\pi}_k}Q_{\boldsymbol{\pi}_*}\right\|_{2,d_{\boldsymbol{\pi}_*}} \qquad R_k = (I-\gamma\mathcal{P})f_k \\
&= \frac{1}{1-\gamma}\left\|f_k - \mathcal{T}_{\boldsymbol{\pi}_k}f_k\right\|_{2,d_{\boldsymbol{\pi}_*}} \\
&\le \frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma}\left\|f_k - \mathcal{T}_{\boldsymbol{\pi}_k}f_k\right\|_{2,\mathcal{D}}
\end{aligned}$$

which is no greater than $\frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma}\left(\sqrt{\varepsilon_r} + \sqrt{\frac{1}{\lambda(1-\gamma)}}\right)$, because $\mathcal{E}(f_k,\boldsymbol{\pi}_k) \le \varepsilon_r + \frac{1}{(1-\gamma)\lambda}$:

$$\begin{aligned}
f_k(s_0,\boldsymbol{\pi}_k) + \lambda\mathcal{E}(f_k,\boldsymbol{\pi}_k) &= \min_f\left(f(s_0,\boldsymbol{\pi}_k) + \lambda\mathcal{E}(f,\boldsymbol{\pi}_k)\right) \\
&\le Q_{\boldsymbol{\pi}_k}(s_0,\boldsymbol{\pi}_k) + \lambda\mathcal{E}(Q_{\boldsymbol{\pi}_k},\boldsymbol{\pi}_k), \quad \text{Lemma C.2} \\
&\le \frac{1}{1-\gamma} + \lambda\varepsilon_r.
\end{aligned}$$

The proof is completed by substituting λ to the original expression. □

D. Pessimistic MA-PPO with General Function Approximation

In this section we extend the results from linear function approximation to general function approximation (cf. Section 5).

In this setting, the value function is searched over a finite set $\mathcal{F}^m$. We impose the following regularity conditions on the general function class

Assumption D.1. For any $m \in \mathcal{N}$, $f \in \mathcal{F}^m$ and $(s,\mathbf{a}^{1:m}) \in \mathcal{S}\times\mathcal{A}^m$, it holds

$$|f(s,\mathbf{a}^{1:m-1},a^m)| \le \frac{1}{1-\gamma}, \tag{15}$$

Algorithm 4 Pessimistic Multi-Agent PPO with General Function Approximation

Input: Regularization coefficient λ.
Output: Uniformly sample k from $0, 1 \cdots K-1$, return $\bar{\boldsymbol{\pi}} = \boldsymbol{\pi}_k$.

1: Initialize uniformly: $\theta_0^m = 0$ for every $m \in \mathcal{N}$.
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: **for** $m = 1, 2, \cdots, N$ **do**
4: Obtain the pessimistic estimate:
 $f_k^m \leftarrow \arg\min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_k) \right)$.
5: Policy improvement: for any $(s, \mathbf{a}^{1:m}) \in \mathcal{S} \times \mathcal{A}^m$,

$$\pi_{k+1}^m(a^m | s, \mathbf{a}^{1:m-1}) \propto \pi_{k+1}^m(a^m | s, \mathbf{a}^{1:m-1}) \cdot \exp\left(\eta f_k^m(s, \mathbf{a}^{1:m-1}, a^m)\right).$$

6: **end for**
7: **end for**

$$|\mathcal{F}^m| \leq |\mathcal{F}| \tag{16}$$

where $|\mathcal{F}|$ is a certain positive number.

Instead of a pre-defined fixed policy class Π, now policy improvement is made upon $\mathcal{F}^m$, formally, for $m \in \mathcal{N}$ and $(s, \mathbf{a}^{1:m-1}) \in \mathcal{S} \times \mathcal{A}^{m-1}$

$$\Pi^m = \left\{ \pi^m(\cdot | s, \mathbf{a}^{1:m-1}) \propto \exp\left(\eta \sum_{j=0}^{k} f_j(s, \mathbf{a}^{1:m-1}, \cdot) \right) : f_j \in \mathcal{F}^m, 0 \leq k \leq K-1 \right\}. \tag{17}$$

Also note that under general function approximation, policy improvement has to be specific for each $(s, \mathbf{a}^{1:m})$, which might become troublesome when the state space is enormous.

For general function approximation, two common expressivity assumptions on $\mathcal{F}$ are required (Antos et al., 2008; Xie et al., 2021).

Assumption D.2 (Realizability)**.** For any $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$,

$$\inf_{f \in \mathcal{F}^m} \sup_{\text{admissable } \nu} \|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\nu}^2 = \zeta_{\mathcal{F}}.$$

where ν can be any admissible distribution over $\mathcal{S} \times \mathcal{A}^{1:m}$, and

Assumption D.3 (Completeness)**.** For any $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$,

$$\sup_{f \in \mathcal{F}^m} \inf_{f \in \mathcal{F}^m} \|f' - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\mu}^2 = \zeta'_{\mathcal{F}}.$$

We have the following concentration guarantees for general function approximation (Xie et al., 2021).

Lemma D.4. *For any $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$, let*

$$f_{\boldsymbol{\pi}}^{1:m} = \arg\min_{f \in \mathcal{F}^m} \sup_{\text{admissable } \nu} \|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\nu}^2,$$

with probability at least $1 - \delta$, it holds

$$\begin{aligned} \mathcal{E}^{1:m}(f_{\boldsymbol{\pi}}^{1:m}, \pi) &\leq \mathcal{O}\left(\frac{\log \frac{|\mathcal{F}^m||\Pi^m|}{\delta}}{n(1-\gamma)^2} + \zeta_{\mathcal{F}} \right) \\ &= \mathcal{O}\left(\frac{K \log \frac{|\mathcal{F}^m|}{\delta}}{n(1-\gamma)^2} + \zeta_{\mathcal{F}} \right) = \varepsilon_r, \end{aligned} \tag{18}$$

where we note that $|\Pi^m| \leq |\mathcal{F}^m|^K$ from (17).

Lemma D.5. *For any $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$, $f \in \mathcal{F}^m$, if $\mathcal{E}^{1:m}(f, \boldsymbol{\pi}) \le \varepsilon$, with probability at least $1-\delta$ it holds*

$$\left\|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\right\|_{2,\mathcal{D}^m} \le \mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{\frac{K \log \frac{|\mathcal{F}^m|}{\delta}}{n}}\right) + \sqrt{\zeta'_{\mathcal{F}}} + \sqrt{\zeta'_{\mathcal{F}} + \varepsilon}.$$

Lemma D.6. *For any conditional policy $\pi : \mathcal{S} \times \mathcal{A}^{m-1} \to \Delta(\mathcal{A})$,*

$$Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) \ge \min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_k)\right) - \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma} - \lambda \varepsilon_r.$$

Proof. For any conditional policy $\boldsymbol{\pi}^{1:m}$, and $\forall m \in \mathcal{N}$. Let $f_{\boldsymbol{\pi}}^{1:m} = \arg\min_f \sup_\nu \|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\nu}^2$ for any admissible ν

$$\begin{aligned}
J(\boldsymbol{\pi}) &= Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) \\
&= Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) - \left(f_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) + \lambda \mathcal{E}^{1:m}(Q_{\boldsymbol{\pi}}^{1:m}, \boldsymbol{\pi})\right) + \left(f_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) + \lambda \mathcal{E}^{1:m}(f_{\boldsymbol{\pi}}^{1:m}, \boldsymbol{\pi})\right) \\
&\ge \min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi})\right) - \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma} - \lambda \varepsilon_r,
\end{aligned}$$

where in the last line we use Lemma C.1 and the realizability assumption

$$f_{\boldsymbol{\pi}}^{1:m} - Q_{\boldsymbol{\pi}}^{1:m} \le \frac{\|f_{\boldsymbol{\pi}}^{1:m} - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f_{\boldsymbol{\pi}}^{1:m}\|_{2,d_{\boldsymbol{\pi}}}}{1-\gamma} \le \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}.$$

□

Equipped with these useful lemmas, we are prepared to proceed to the main theorem for the general function approximation setting

Theorem D.7. *Recall the definition of ε_r (18), for the output policy $\bar{\boldsymbol{\pi}}$ attained by Algorithm 4 in a fully cooperative Markov game, set $\eta = (1-\gamma)\sqrt{\frac{\log|\mathcal{A}|}{2K}}$ and $\lambda = (1-\gamma)^{-1}\varepsilon_r^{-\frac{2}{3}}$. After K iterations, w.p. at least $1-\delta$ we have*

$$J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}}) \le \mathcal{O}\left(\frac{N}{(1-\gamma)^2}\sqrt{\frac{\log|\mathcal{A}|}{K}} + \frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma} \cdot \left(\frac{1}{1-\gamma}\sqrt[3]{\frac{K \log \frac{|\mathcal{F}|}{n}}{n}} + \sqrt{\zeta_{\mathcal{F}} + \zeta'_{\mathcal{F}}} + \sqrt[3]{\zeta_{\mathcal{F}}}\right)\right)$$

Proof. Use Lemma D.6, at the k-th iteration we have

$$\begin{aligned}
J(\boldsymbol{\pi}_k) &\ge \min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_k)\right) - \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma} - \lambda \varepsilon_r \\
&\ge f_k^m(s_0, \boldsymbol{\pi}_k^{1:m}) - \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma} - \lambda \varepsilon_r \\
&= J_{\mathcal{M}_k}(\boldsymbol{\pi}_k) - \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma} - \lambda \varepsilon_r
\end{aligned}$$

Analogous to Appendix C, we have

$$\begin{aligned}
J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}}) &= \frac{1}{K}\sum_{k=1}^K (J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}_k)) \\
&\le \frac{1}{K}\sum_{k=1}^K (J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_k)) + \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma} + \lambda \varepsilon_r \\
&\le \underbrace{\frac{1}{K}\sum_{k=1}^K (J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_k))}_{\text{I}} + \underbrace{\frac{1}{K}\sum_{k=1}^K (J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*))}_{\text{II}} + \frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma} + \lambda \varepsilon_r
\end{aligned}$$

Term I. The analysis for the optimization term is the same as that in Appendix C. If $\eta = (1-\gamma)\sqrt{\frac{\log|\mathcal{A}|}{2K}}$, Lemma C.6 gives

$$\text{Term I} \leq \frac{2N}{(1-\gamma)^2}\sqrt{\frac{2\log|\mathcal{A}|}{K}}.$$

Term II. Similar with Appendix C

$$\begin{aligned} J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) &= Q_{\boldsymbol{\pi}_*}(s, \boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) \\ &\leq \frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma}\left\|f_k - \mathcal{T}_{\boldsymbol{\pi}_k} f_k\right\|_{2,\mathcal{D}} \end{aligned}$$

which is no greater than

$$\frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma}\left(\mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{\frac{K\log\frac{|\mathcal{F}^m|}{\delta}}{n}}\right) + \sqrt{\zeta'_{\mathcal{F}}} + \sqrt{\zeta'_{\mathcal{F}} + \varepsilon_r} + \sqrt{\frac{1}{(1-\gamma)\lambda}}\right)$$

, because $\mathcal{E}(f_k, \boldsymbol{\pi}_k) \leq \varepsilon_r + \frac{1}{(1-\gamma)\lambda}$:

$$\begin{aligned} f_k(s_0, \boldsymbol{\pi}_k) + \lambda\mathcal{E}(f_k, \boldsymbol{\pi}_k) &= \min_f \left(f(s_0, \boldsymbol{\pi}_k) + \lambda\mathcal{E}(f, \boldsymbol{\pi}_k)\right) \\ &\leq Q_{\boldsymbol{\pi}_k}(s_0, \boldsymbol{\pi}_k) + \lambda\mathcal{E}(Q_{\boldsymbol{\pi}_k}, \boldsymbol{\pi}_k), \quad \text{Lemma C.2} \\ &\leq \frac{1}{1-\gamma} + \lambda\varepsilon_r. \end{aligned}$$

The proof is completed by substituting λ to the original expression. □

E. Simulation

In this section, we perform a toy example to showcase the superiority of our sequential update structure over naive independent policy gradient updates. We consider von Neumann's ratio game, a simple stochastic game also used by (Daskalakis et al., 2020).

In the game, there are only two agents, and each has an action space of 2. There is only one state, i.e., no state transition exists. The immediate reward for selecting actions (a, b) is $R(a, b)$ the probability of stopping in each round is $S(a, b)$. The value function $V(\pi_x, \pi_y)$ for this game is given by

$$V = \frac{\pi_x^\top R \pi_y}{\pi_x^\top S \pi_y}.$$

The two agents cooperate with each other to maximize the value function. From now on, we shall use $(x, 1-x)$ and $(y, 1-y)$ to represent both policies. We set parameters as

$$R = \begin{bmatrix} 1 & 0.5 \\ -0.5 & 1 \end{bmatrix}, \quad \text{and} \quad R = \begin{bmatrix} 1 & 1 \\ 0.1 & 0.1 \end{bmatrix}.$$

Consider the value function as a function of variables x and y , then the stationary point is near $(x, y) = (0.5, 0)$, at which the value function is $V \approx 0.46$, which is smaller than the global maximum $V = 1$.

To solve the problem, we adopt two algorithms: (1) our algorithm with sequential updates and (2) the independent (policy gradient) learning method. In both algorithms, we use softmax parametrization for policies. In particular, our log-linear parameterization (2) becomes softmax parametrization by setting ϕ as one-hot representations, i.e., for action a , $\phi(a)^\top\theta = \theta_a$ where θ_a represents the specific entry of θ that corresponds to a .

Results We test our algorithm with sequential gradient updates and the independent learning method in different settings. The results are shown in Figure 1. [5] Below we discuss the empirical findings from this simulation study.

First, we find that the independent policy optimization method often struggles around the stationary point (see (a)-(c)) that is not necessarily globally optimal. In this example, a big stepsize would help alleviate the issue (e.g., in (c), independent PG escapes the stationary point after 3000 iterations). However, the convergence to global optima is still slower than our method. We note that noise might help to escape the stationary points (Jin et al., 2017). Our findings align with the theoretical comparisons we made aforementioned. Even if the independent PG method is not trapped by a stationary point, from (d)-(f), our algorithm consistently outperforms in terms of maximizing the value function.

In this toy example, the optimization landscape is quite simple: only two agents participate, each with only two possible actions. No state transition is allowed, which is the main difficulty in performing sequential decision-making. We point out that, globally, there is only one stationary point. In such an effortless case, our algorithm consistently outperforms independent PG in mainly two folds. First, our algorithm does not struggle like independent PG when the current policy is near the stationary point where gradient information is few. Second, our algorithm demonstrates a fast convergence rate to the global maximum value function. Therefore, when the complexity of the environment increases significantly, for instance: (1) multiple heterogeneous agents interact with each other and the unknown environment, (2) complex function approximators are adopted (e.g., deep neural networks), utilizing independent PG would be more problematic in terms of locating the global optimum because there will be more stationary points in the landscape.

Our findings showcase the necessity and usefulness of the conditional dependency structure, which helps us find a policy that enjoys a globally sub-optimal value function.

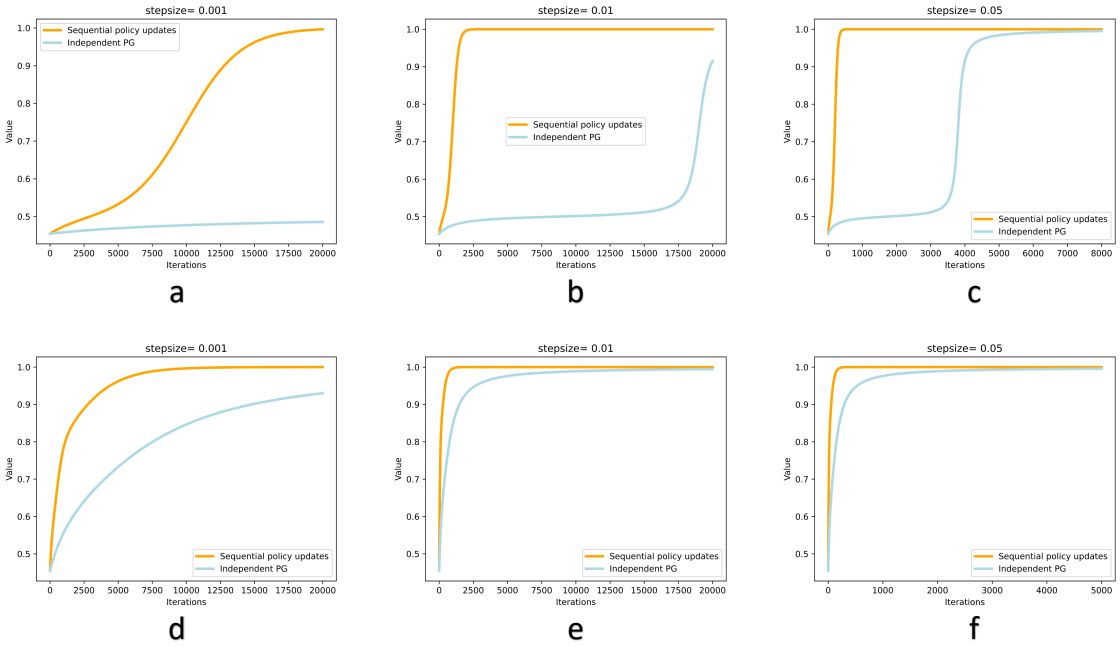

Figure 1. Performances of our algorithm and the independent learning method. In (a)-(c): policies are initialized close to the stationary point. In (d)-(f): both policies are uniformly initialized.

[5] Implementation can be found at https://github.com/zhaoyl18/ratio_game.