Convergence and Price of Anarchy Guarantees of the Softmax Policy Gradient in Markov Potential Games

Dingyang Chen [1] **Qi Zhang** [1] **Thinh T. Doan** [2]

Abstract

We study the performance of policy gradient methods for the subclass of Markov games known as Markov potential games (MPGs), which extends the notion of normal-form potential games to the stateful setting and includes the important special case of the fully cooperative setting where the agents share an identical reward function. Our focus in this paper is to study the convergence of the policy gradient method for solving MPGs under softmax policy parameterization, both tabular and parameterized with general function approximators such as neural networks. We first show the asymptotic convergence of this method to a Nash equilibrium of MPGs for tabular softmax policies. Second, we derive the finite-time performance of the policy gradient in two settings: 1) using the log-barrier regularization, and 2) using the natural policy gradient under the best-response dynamics (NPG-BR). Finally, extending the notion of price of anarchy (POA) and smoothness in normal-form games, we introduce the POA for MPGs and provide a POA bound for NPG-BR. To our knowledge, this is the first POA bound for solving MPGs. To support our theoretical results, we empirically compare the convergence rates and POA of policy gradient variants for both tabular and neural softmax policies.

1. Introduction

The framework of multi-agent sequential decision making is often formulated as (variants of) Markov games (MGs) (Shapley, 1953), which finds a wide range of real-world applications such as coordination of multi-robot systems (Corke et al., 2005), traffic control (Chu et al., 2019), power grid management (Callaway & Hiskens, 2010), etc. Perhaps the most well-known solution concept for MGs is the Nash policy, which is also known as the Nash equilibrium in the special case of stateless Markov games (i.e., normal-form games). In a Nash policy, every agent selects its actions independently of any other agent given the state and plays a best response to all other agents. In the special case of single-agent Markov games, aka Markov decision processes (MDPs), Nash policies reduce to the agent's optimal policies. Most existing algorithms seeking to find Nash policies are value-based (i.e., computing only value functions related to the MG), with examples including Nash Q-learning (Hu & Wellman, 2003), Hyper-Q Learning (Tesauro, 2003), and Nash-VI for the special case of zero-sum MGs (Zhang et al., 2020). Policy-based algorithms, including multi-agent actor-critic algorithms, have recently gained attention with impressive empirical success (Lowe et al., 2017; Foerster et al., 2017) as well as provable guarantees (Zhang et al., 2018; Leonardos et al., 2021; Zhang et al., 2021).

This paper focuses on the MG subclass of *Markov potential games* (MPGs) (Macua et al., 2018; Leonardos et al., 2021; Zhang et al., 2021), which is extended from the notion of (normal-form) potential game and also incorporates as a special case the fully cooperative MGs where all agents share the same reward to optimize. The MPG structure allows for exploiting recent advances in single-agent policy gradient methods (e.g., (Agarwal et al., 2019)) to establish the convergence of policy gradient to (near-)Nash policies in MPGs. Specifically, existing work has established finite-time convergence guarantees under the *direct* policy parameterization. In this paper, we are interested in the alternative *softmax* policy parameterization, both tabularly and with neural networks for learnable state representations. For tabular softmax, we establish several convergence guarantees to (near-)Nash policies in MPGs in Section 3, extending their counterpart from the single-agent setting (Agarwal et al., 2019). We then empirically compare tabular softmax with neural network-based softmax parameterization in terms of their convergence rates.

MPGs can model many problems where outcomes of high social welfare, measured by the sum of all agents' values,

[1]Artificial Intelligence Institute, University of South Carolina [2]Department of Electrical and Computer Engineering, Virginia Tech. Correspondence to: Dingyang Chen <dingyang@email.sc.edu>, Qi Zhang <qz5@cse.sc.edu>.

Decision Awareness in Reinforcement Learning Workshop at the 39th International Conference on Machine Learning (ICML), Baltimore, Maryland, USA, 2022.

are most desirable. In these scenarios, the solution concept of the Nash policy is inadequate. The price of anarchy (POA) of a policy, firstly studied in normal-form games (Roughgarden, 2015), is accordingly defined as the ratio between the sum of all agents' value under this policy and the maximum-possible value sum. In this sense, the POA further measures the quality of a Nash policy. In Section 4, we extend the notion of POA to the stateful MGs and provide first POA bounds for near-Nash policies in MGs and for an approximate best-response dynamics in MPGs. We empirically compare the POA of Nash policies achieved by variants of softmax policy gradient dynamics.

1.1. Related work

Single-agent policy gradient convergence. Agarwal et al. firstly established the policy gradient convergence of to global optima in the single-agent setting under tabular softmax parameterization, specifically, asymptotic convergence of policy gradient ascent, finite-time convergence with log barrier regularization, and finite-time convergence with natural policy gradient. Agarwal et al. also established finite-time convergence for direct policy parameterization (Agarwal et al., 2019). Mei et al. later established finite-time convergence of (regularized) policy gradient ascent under tabular softmax parameterization, with a convergence rate depending on a problem-specific variable (Mei et al., 2020). This problem-specific variable in some sense is necessary, as Li et al. have shown that softmax policy gradient can take exponential time to converge (Li et al., 2021).

Policy gradient convergence in MPGs. Extending the work by Agarwal et al. (Agarwal et al., 2019) from the single-agent setting, Leonardos et al. (Leonardos et al., 2021) and Zhang et al. (Zhang et al., 2021) both established finite-time convergence of projected gradient ascent under tabular softmax parameterization to near-Nash policies in MPGs. Fox et al. (Fox et al., 2022) established the asymptotic convergence of natural policy gradient to Nash policies in MPGs.

POA bounds in normal-form games. Mirrokni and Vetta (Mirrokni & Vetta, 2004) initiated the discussion on the importance of POA bounds beyond Nash equilibria. Roughgarden (Roughgarden, 2015) defined the smoothness of (normal-form) games and then established the first POA bounds of on near-Nash equilibria in smooth games. Roughgarden (Roughgarden, 2015) provided POA bounds for the maximum-gain best-response dynamics in smooth (normal-form) potential games.

2. Preliminaries

Markov game. We consider a Markov game (MG) $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \vec{r} \rangle$ with N agents indexed by $i \in \mathcal{N} = \{1, ..., N\}$, state space $\mathcal{S}$, action space $\mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward functions $\vec{r} = \{r^i\}_{i \in \mathcal{N}}$ with $r^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for each $i \in \mathcal{N}$, and initial state distribution $\mu \in \Delta(\mathcal{S})$. We assume full observability for simplicity, i.e., each agent observes the state $s \in \mathcal{S}$. Under full observability, we consider *product policies*, $\pi : \mathcal{S} \to \times_{i \in \mathcal{N}} \Delta(\mathcal{A}^i)$, that is factored as the product of individual policies $\pi^i : \mathcal{S} \to \Delta(\mathcal{A}^i)$, $\pi(a|s) = \prod_{i \in \mathcal{N}} \pi^i(a^i|s)$. Define the discounted return for agent i from time step t as $G_t^i = \sum_{l=0}^{\infty} \gamma^l r_{t+l}^i$, where $r_t^i := r^i(s_t, a_t)$ is the reward at time step t for agent i. For agent i, product policy $\pi = (\pi^1, ..., \pi^N)$ induces a value function defined as $V_\pi^i(s_t) = \mathbb{E}_{s_{t+1:\infty}, a_{t:\infty} \sim \pi}[G_t^i | s_t]$, and action-value function $Q_\pi^i(s_t, a_t) = \mathbb{E}_{s_{t+1:\infty}, a_{t+1:\infty} \sim \pi}[G_t^i | s_t, a_t]$. Following policy π, agent i's cumulative reward starting from $s_0 \sim \mu$ is denoted as $V_\pi^i(\mu) := \mathbb{E}_{s_0 \sim \mu}[V_\pi^i(s_0)]$.

It will be useful to define the (unnormalized) *discounted state visitation measure* by following policy π after starting at $s_0 \sim \mu$:

$$d_\mu^\pi(s) := \mathbb{E}_{s_0 \sim \mu}\left[\sum_{t=0}^{\infty} \gamma^t \Pr^\pi(s_t = s | s_0)\right]$$

where $\Pr^\pi(s_t = s|s_0)$ is the probability that $s_t = s$ after starting at state s_0 and following π thereafter. We make a standard assumption for the discounted state visitation distribution to be positive for every state under any policy, as formally stated in Assumption 2.1.

Assumption 2.1. For any π and any state s of the Markov game, $d_\mu^\pi(s) > 0$.

Markov potential game.

Definition 2.2 (Markov potential game)**.** A Markov game is called a *Markov potential game* (MPG) if there exists a potential function $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that for any agent i, any pair of product policies (π^i, π^{-i}), $(\bar{\pi}^i, \pi^{-i})$, and any state s:

$$\begin{aligned}
&\mathbb{E}_{s_{t+1:\infty}, a_{t:\infty} \sim (\bar{\pi}^i, \pi^{-i})}\left[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, a_t) | s_0 = s\right] \\
&- \mathbb{E}_{s_{t+1:\infty}, a_{t:\infty} \sim (\pi^i, \pi^{-i})}\left[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, a_t) | s_0 = s\right] \\
=&\mathbb{E}_{s_{t+1:\infty}, a_{t:\infty} \sim (\bar{\pi}^i, \pi^{-i})}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | s_0 = s\right] \\
&- \mathbb{E}_{s_{t+1:\infty}, a_{t:\infty} \sim (\pi^i, \pi^{-i})}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | s_0 = s\right].
\end{aligned}$$

Given a product policy π, we define the *total potential function* as $\Phi_\pi(s) := \mathbb{E}_{s_{t+1:\infty}, a_{t:\infty} \sim \pi}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | s_0 = s]$, and we

can obtain that, for any agent i,

$$V^i_{\bar{\pi}^i,\pi^{-i}}(s) - V^i_{\pi^i,\pi^{-i}}(s) = \Phi_{\bar{\pi}^i,\pi^{-i}}(s) - \Phi_{\pi^i,\pi^{-i}}(s) \quad (1)$$
$$\text{giving} \quad \nabla_{\theta^i} V^i_\theta(s) = \nabla_{\theta^i}\Phi_\theta(s).$$

We also similarly define $\Phi_\pi(\mu) := \mathbb{E}_{s_0\sim\mu}[\Phi_\pi(s_0)]$.

As formally stated in Assumption 2.3, we assume that ϕ, and therefore Φ, are bounded.

Assumption 2.3 (Potential function is bounded)**.** The potential function ϕ is bounded, such that the total potential function Φ is bounded as $\Phi_{\min} \le \Phi_\pi(s) \le \Phi_{\max}\ \forall s, \pi$.

Nash policy. We focus on the solution concept of (ϵ-)Nash policy, as formally defined below.

Definition 2.4 (ϵ-Nash policy)**.** The *Nash-gap* of a policy π is defined as

$$\text{Nash-gap}(\pi) := \max_i \left(\max_{\bar{\pi}^i} V^i_{\bar{\pi}^i,\pi^{-i}}(\mu) - V^i_\pi(\mu) \right)$$

A product policy $\pi = (\pi_1, ..., \pi_N)$ is an *ϵ-Nash policy* if Nash-gap$(\pi) \le \epsilon$.

3. Convergence of the tabular softmax policy gradient in MPGs

In this section, we consider individual policies $(\pi_1, ..., \pi_N)$ to be independently parameterized in the softmax tabular manner from the global state, i.e., we have, for each agent i, its policy parameter $\theta^i = \{\theta^i_{s,a^i} : s \in \mathcal{S}, a^i \in \mathcal{A}^i\}$ and policy

$$\pi^i_{\theta^i}(a^i|s) = \frac{\exp(\theta^i_{s,a^i})}{\sum_{\bar{a}^i\in\mathcal{A}^i}\exp(\theta^i_{s,\bar{a}^i})}.$$

For the rest of this paper, we will abbreviate Φ_{π_θ}, $V^i_{\pi_\theta}$, $A^i_{\pi_\theta}$ as Φ_θ, V^i_θ, A^i_θ, respectively. Lemmas 3.1 and 3.2 formally states the policy gradient form and the smoothness under the tabular softmax parameterization, respectively, which will be used to establish the convergence results in this section.

Lemma 3.1 (Multi-agent tabular softmax policy gradient form, proof in Appendix A)**.** *For the state-based tabular softmax multi-agent policy parameterization, we have:*

$$\frac{\partial\Phi_\theta(\mu)}{\partial\theta^i_{s,a^i}} = \frac{\partial V^i_\theta(\mu)}{\partial\theta^i_{s,a^i}} = d^{\pi_\theta}_\mu(s)\pi^i_{\theta^i}(a^i|s)A^i_\theta(s,a^i) \quad (2)$$

where $A^i_\theta(s,a^i) := \mathbb{E}_{a^{-i}\sim\pi^{-i}_{\theta^{-i}}(\cdot|s)}[A^i_\theta(s,a^i,a^{-i})]$.

Lemma 3.2 (Smoothness of Φ under tabular softmax, proof in Appendix B)**.** *Under tabular softmax π_θ, $\Phi_\theta(s)$ is $\frac{41N}{4(1-\gamma)^3}$-smooth for any state s (hence for any initial state distribution μ).*

We next present our convergence results for the standard policy gradient dynamics without and with log barrier regularization in Sections 3.1 and 3.2, respectively, where Assumptions 2.1 and 2.3 hold.

3.1. Asymptotic convergence of the policy gradient dynamics

In Theorem 3.4, we establish, under the tabular softmax policy parameterization, the asymptotic convergence to a Nash policy in a MPG of the standard policy gradient dynamics:

$$\theta^i_{t+1} = \theta^i_t + \eta\nabla_{\theta^i}V^i_{\theta_t}(\mu) = \theta^i_t + \eta\nabla_{\theta^i}\Phi_{\theta_t}(\mu) \quad (3)$$

where η is the fixed stepsize and the update is performed by every agent $i \in \mathcal{N}$. Theorem 3.4 relies on the assumption on the asymptotic convergence of the policy parameters, formally stated as follows.

Assumption 3.3. Following the policy gradient dynamics (3), the policy parameter of every agent i converges asymptotically, i.e., $\theta^i_t \to \theta^i_*$ as $t \to \infty,\ \forall i$.

We remark here that the assumption that θ^i converges is made to ensure the convergence of $\{Q^i(s,a^i)\}_i$, which is then used to prove the theorem in a similar manner to (Agarwal et al., 2019). Note that, since the gradient is as Equation (2), the gradient converging to zero cannot directly imply the parameters converging to zero. A sufficient condition for Assumption 3.3 to hold is that the stationary points of are *isolated*, which is originally assumed in Fox et al. (Fox et al., 2022) to establish the asymptotic convergence of natural policy gradient to Nash policies.

Theorem 3.4 (Asymptotic convergence of policy gradient, proof in Appendix C)**.** *Suppose every agent $i \in \mathcal{N}$ follows the policy gradient dynamics (3) with $\eta \le min(\frac{1-\gamma}{Nmax(5,\sqrt{N})(\Phi_{max}-\Phi_{min})}, \frac{4(1-\gamma)^3}{41N})$ and Assumption 3.3 holds such that $\theta^i_t \to \theta^i_*$ for every agent i, then the product policy defined by $\theta_* = \{\theta^i_*\}_{i\in\mathcal{N}}$ is a Nash policy.*

3.2. Policy gradient dynamics with log-barrier regularization

Inspired by (Agarwal et al., 2019) for the single-agent setting, we consider the log barrier regularized objective as defined below to establish finite-time convergence guarantees for the policy gradient dynamics:

$$L_\lambda(\theta) := \Phi_\theta(\mu) - \lambda\textstyle\sum_{i=1}^N \mathbb{E}_{s\sim\text{Unif}_\mathcal{S}}\left[\text{KL}(\text{Unif}_{\mathcal{A}^i}, \pi_\theta(\cdot|s))\right]$$
$$= \Phi_\theta(\mu) + \lambda\textstyle\sum_{i=1}^N \left(\frac{\sum_{s,a^i}\log\pi^i_{\theta^i}(a^i|s)}{|\mathcal{S}||\mathcal{A}^i|} + \log|\mathcal{A}^i| \right)$$

where the log barrier regularization, i.e., the KL divergence with respect to the uniform action-selection distribution, is applied to each agent's policy independently. Lemma 3.5

extends the results in (Agarwal et al., 2019) to the multi-agent setting, stating that, with the log barrier regularization, approximate first-order stationary points are near-Nash.

Lemma 3.5 (Log barrier regularization's approximate first-order stationary points are near-Nash, proof in Appendix D.1)**.** *Suppose θ is such that $\|\nabla_\theta L_\lambda(\theta)\|_2 \leq \lambda/(2|\mathcal{S}|\max_i|\mathcal{A}^i|)$. Then the product policy $\pi_\theta = (\pi^1_{\theta^1}, ..., \pi^N_{\theta^N})$ is a $2\lambda M$-Nash policy where $M := \max_{\pi,\pi'} \left\| \frac{d^\pi_\mu}{d^{\pi'}_\mu} \right\|_\infty$, which is well-defined by Assumption 2.1.*

With Lemma 3.5, we establish the convergence rate as stated in Theorem 3.6.

Theorem 3.6 (Convergence rate of the policy gradient with log barrier regularization, proof in Appendix D.2)**.** *Letting $\beta_\lambda := \frac{41N}{4(1-\gamma)^3} + \frac{2\lambda N}{|\mathcal{S}|}$, then β_λ is an upper bound on the smoothness of $L_\lambda(\theta)$. Starting from $\theta_0 = 0$, consider the updates $\theta_{t+1} = \theta_t + \eta\nabla_\theta L_\lambda(\theta_t)$ with $\lambda = \epsilon/2M$ and $\eta = 1/\beta_\lambda$. Then, for any initial distribution μ, we have $\min_{t<T}$ Nash-gap$_t \leq \epsilon$ whenever*

$$T \geq \frac{328NM^2|\mathcal{S}|^2\max_i|\mathcal{A}^i|^2(\Phi_{\max}-\Phi_{\min})}{(1-\gamma)^3\epsilon^2} + \frac{32NM|\mathcal{S}|\max_i|\mathcal{A}^i|^2(\Phi_{\max}-\Phi_{\min})}{\epsilon}.$$

3.3. Approximate best-response natural policy gradient dynamics

In this subsection, we consider the natural policy gradient (NPG) dynamics extended from the single-agent setting to Markov potential games. The NPG dynamics is defined as

$$\theta^i_{t+1} = \theta^i_t + \eta(F^i_{\theta_t})^\dagger\nabla_{\theta^i}V^i_{\theta_t}(\mu) = \theta^i_t + \eta(F^i_{\theta_t})^\dagger\nabla_{\theta^i}\Phi_{\theta_t}(\mu), \tag{4}$$

where $A^\dagger$ denotes the Moore–Penrose inverse of a matrix A and F^i_θ is the Fisher information matrix for agent i under product policy π_θ:

$$F^i_\theta = \mathbb{E}_{s\sim d^{\pi_\theta}_\mu, a^i\sim\pi^i(s)}\left[\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s)\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s)^\top\right].$$

Lemma 3.7 (NPG is effectively soft policy iteration, proof in Appendix E.1)**.** *For any agent i, the NPG update (4) is effectively:*

$$\theta^i_{t+1} = \theta^i_t + \eta A^i_{\theta_t} \quad \text{and}$$
$$\pi^i_{\theta^i_{t+1}}(a^i|s) = \pi^i_{\theta^i_t}(a^i|s)\frac{\exp\left(\eta A^i_{\theta_t}(s,a^i)\right)}{Z^i_t(s)}$$

where $Z^i_t(s) = \sum_{a^i}\pi^i_{\theta^i_t}(a^i|s)\exp\left(\eta A^i_{\theta_t}(s,a^i)\right)$ is the normalization constant for the softmax.

In the single-agent setting with the tabular softmax parameterization, we know that the NPG update (soft policy iteration) can achieve $O(1/\epsilon)$ convergence rate with ϵ being the single-agent optimality gap, compared with the $O(1/\epsilon^2)$ convergence rates achieved by (projected) gradient ascent methods for the direct parameterization and for the tabular softmax parameterization with the log barrier regularization (Agarwal et al., 2019). For MPGs, Fox et al. (Fox et al., 2022) established the asymptotic convergence of natural policy gradient. However, deriving the finite-time convergence with the tabular softmax parameterization when the agents concurrently perform the soft policy iteration is challenging, primarily due to the technical difficulty of relating the potential function value and the Nash-gap. Here, we take a step back and consider the non-concurrent soft policy iteration, where an agent will perform a number of soft policy iterations with fixing other agents' policies: letting $\theta^i_{t,0} = \theta^i_t$, for $k = 1, ..., K$:

$$\theta^i_{t,k} = \theta^i_{t,k} + A^i_{t,k-1} \text{ with} \tag{5}$$
$$A^i_{t,k-1}(s,a^i) = \mathbb{E}_{a^{-i}\sim\pi^{-i}_{\theta^{-i}_t}(\cdot|s)}\left[A^i_{\theta^i_{t,k-1},\theta^{-i}_t}(s,a^i,a^{-i})\right],$$

where $A^i_{t,k}$ is agent i's local advantage of its policy currently parameterized by θ^{-i}_t with respect to the other agents' policies parameterized by θ^{-i}_t, and K is a hyperparameter that controls how close agent i will get to its best response to θ^{-i}_t.

The above update is performed independently for all agents, and for the next iteration $t+1$ we only keep the change of the agent that induces the maximum gain in its own value and, equivalently, in the total potential function:

$$i^*_t = \arg\max_i \Phi_{\theta^i_{t,K},\theta^{-i}_t}(\mu) - \Phi_{\theta_t}(\mu),$$
$$\theta^{i^*_t}_{t+1} = \theta^{i^*_t}_{t,K} \quad \text{and} \quad \theta^i_{t+1} = \theta^i_t \text{ for } i \neq i^*_t \tag{6}$$

which ensembles the standard maximum-gain best-response dynamics for normal-form games (Roughgarden, 2016).

Suppose we aim to converge to a ϵ-Nash policy. We can set K large enough (specifically $K \geq \frac{4}{(1-\gamma)^2\epsilon}$ (Agarwal et al., 2019)), such that every agent's inner-loop update (indexed by k (5)) achieves at least $\epsilon/2$-near-best-response. Therefore, if no agent's improvement in their local value or, equivalently, in the total potential function as computed in (6) is no larger than $\epsilon/2$, then the product policy is already a ϵ-Nash policy; otherwise, we can significantly improve the total potential function such that the total number of outer-loop updates, indexed by t in (6), can be bounded. This establishes the convergence rate of our approximate-best-response NPG dynamics (5,6), as formally stated in Theorem 3.8.

Theorem 3.8 (Convergence of the approximate-best-response NPG, proof in Appendix E.2)**.** *Setting $K \geq \frac{4}{(1-\gamma)^2\epsilon}$ for as the iteration complexity of the inner-loop (5), then the approximate-best-response NPG dynamics (5,6) converges to a ϵ-Nash policy within $O(\frac{\Phi_{\max}-\Phi_{\min}}{(1-\gamma)^2\epsilon^2})$ inner-loop steps.*

4. Bounding the price of anarchy in smooth Markov (potential) games

In Definition 4.1, we formally define the price of anarchy in Markov games, which directly extends the notion in normal-from games that measures the quality of a product policy in terms of maximizing the sum of all agents' values.

Definition 4.1 (Price of anarchy in Markov games)**.** The *price of anarchy* (POA) of a product policy π is defined as $\frac{\sum_i V_\pi^i(\mu)}{\max_{\bar{\pi}} \sum_i V_{\bar{\pi}}^i(\mu)}$, i.e., the ratio between the values summed over all agents and the largest summed values achieved by any product policy $\bar{\pi}$.

For the rest of this section, we formally extend the notion of smoothness from normal-form games (Roughgarden, 2015) to Markov games in Section 4.1, and present our POA bounds in smooth Markov (potential) games in Sections 4.2 and 4.3.

4.1. Definition and sufficient conditions of smooth Markov game.

Definition 4.2 extends the notion of smooth normal-form game to its counterpart in Markov games.

Definition 4.2 (Smooth Markov game)**.** A Markov game is (α, β)-smooth if

$$\sum_i V^i_{\pi_\prime^i, \pi^{-i}}(s) \geq \alpha V_{\pi_\prime}(s) - \beta V_\pi(s)$$

for any s and any pair of product policies $\pi, \pi_\prime$, where $V_\pi(s) := \sum_i V_\pi^i(s)$.

Intuitively, in a smooth Markov game, the externality imposed by one agent on the value of the others is limited. Therefore, we conjecture that a sufficient condition is that both the transition and reward functions of the Markov game are "smooth". Proposition 4.3 verifies this conjecture, which formally defines the smoothness of the transition and reward functions and establishes it as a sufficient condition for the smoothness of the Markov game.

Proposition 4.3 (Transition and reward smoothness as a sufficient condition for Markov game smoothness)**.** *The reward functions $\{r^i\}_{i\in\mathcal{N}}$ of a Markov game is said to be* (λ, μ)-smooth *if*

$$\lambda r_{\pi_\prime}(s) \leq \sum_i r^i_{\pi_\prime^i, \pi^{-i}}(s) \leq \mu r_\pi(s)$$

for any state s and any pair of product policies $\pi, \pi_\prime$, where $r_\pi^i(s) := \mathbb{E}_{a\sim\pi(s)}[r^i(s,a)]$ and $r_\pi(s) := \sum_i r_\pi^i(s)$. Letting $M_\pi := (I - \gamma P_\pi)^{-1}$, the transition function P of a Markov game is said to be (κ, ν)-smooth *if*

$$M_{\pi_\prime^i, \pi^{-i}} r \geq \kappa M_{\pi_\prime} r - \nu M_\pi r$$

for any $r \in \mathbb{R}^{|\mathcal{S}|}$ and any pair of product policies $\pi, \pi_\prime$. For a Markov game, if its reward functions are (λ, μ)-smooth and its transition function is (κ, ν)-smooth, then the Markov game is $(\alpha = \kappa\lambda, \beta = \mu\nu)$-smooth.

Proof. We can establish

$$\begin{aligned}\sum_i V^i_{\pi_\prime^i, \pi^{-i}}(s) =& \sum_i M_{\pi_\prime^i, \pi^{-i}} r^i_{\pi_\prime^i, \pi^{-i}} \\ \geq& \sum_i \kappa M_{\pi_\prime} r^i_{\pi_\prime^i, \pi^{-i}} - \nu M_\pi r^i_{\pi_\prime^i, \pi^{-i}} \\ =& \kappa M_{\pi_\prime} \sum_i r^i_{\pi_\prime^i, \pi^{-i}} - \mu M_\pi \sum_i r^i_{\pi_\prime^i, \pi^{-i}} \\ \geq& \kappa M_{\pi_\prime} \lambda r_{\pi_\prime} - \mu M_\pi \mu r_\pi = \kappa\lambda V_{\pi_\prime} - \mu\nu V_\pi\end{aligned}$$

where the two inequalities are due to the smoothness of the transition function and the reward functions, respectively, which completes the proof. □

4.2. POA bound for near-Nash policies

We here derive our POA bound in Theorem 4.5 for near-Nash policies in smooth Markov games, generalizing from smooth normal-form games (Roughgarden, 2016) to smooth Markov games. Similar to the normal-form game counterpart, we describe the result for ϵ-*ratio*-Nash policies as defined in Definition 4.4 to ease presentation.

Definition 4.4 (ϵ-ratio-Nash policy)**.** A product policy $\pi = (\pi_1, ..., \pi_N)$ is an ϵ-*ratio-Nash policy* if, for any agent i, $\max_{\pi_\prime^i} V^i_{\pi_\prime^i, \pi^{-i}}(\mu) \leq (1-\epsilon) V_\pi^i(\mu)$.

Theorem 4.5 (POA of ϵ-ratio-Nash in smooth Markov games)**.** *In any (α, β)-smooth Markov game, the POA of any ϵ-ratio-Nash policy is at least $\frac{(1-\epsilon)\alpha}{1+(1-\epsilon)\beta}$.*

Proof. Consider setting $\pi_\prime = \pi_*$ in Definition 4.2 where π_* is a policy that achieves the optimal joint value, we have

$$\begin{aligned}V_\pi(s) = \sum_i V_\pi^i(s) \geq& \sum_i (1-\epsilon) V^i_{\pi_*^i, \pi^{-i}}(s) \\ \geq& (1-\epsilon)\left(\alpha V_{\pi_*}(s) - \beta V_\pi(s)\right)\end{aligned}$$

where the first inequality is by the definition of π being ϵ-ratio-Nash and the second inequality is by the definition of smooth Markov game. Rearranging the terms completes the proof. □

4.3. POA bound for the approximate best-response dynamics

Inspired by the POA bounds for the best-response dynamics in smooth (normal-form) potential games (Roughgarden, 2015), we here derive the counterpart for smooth MPGs in Theorem 4.7, which bounds the number of policies generated from the *maximum-gain ϵ-ratio-best-response* dynamics: Until product policy π is ϵ-ratio-Nash, update the maximum-gain agent to its best response, where the maximum-gain agent is the agent that induces the maximum increase in value after its best response.

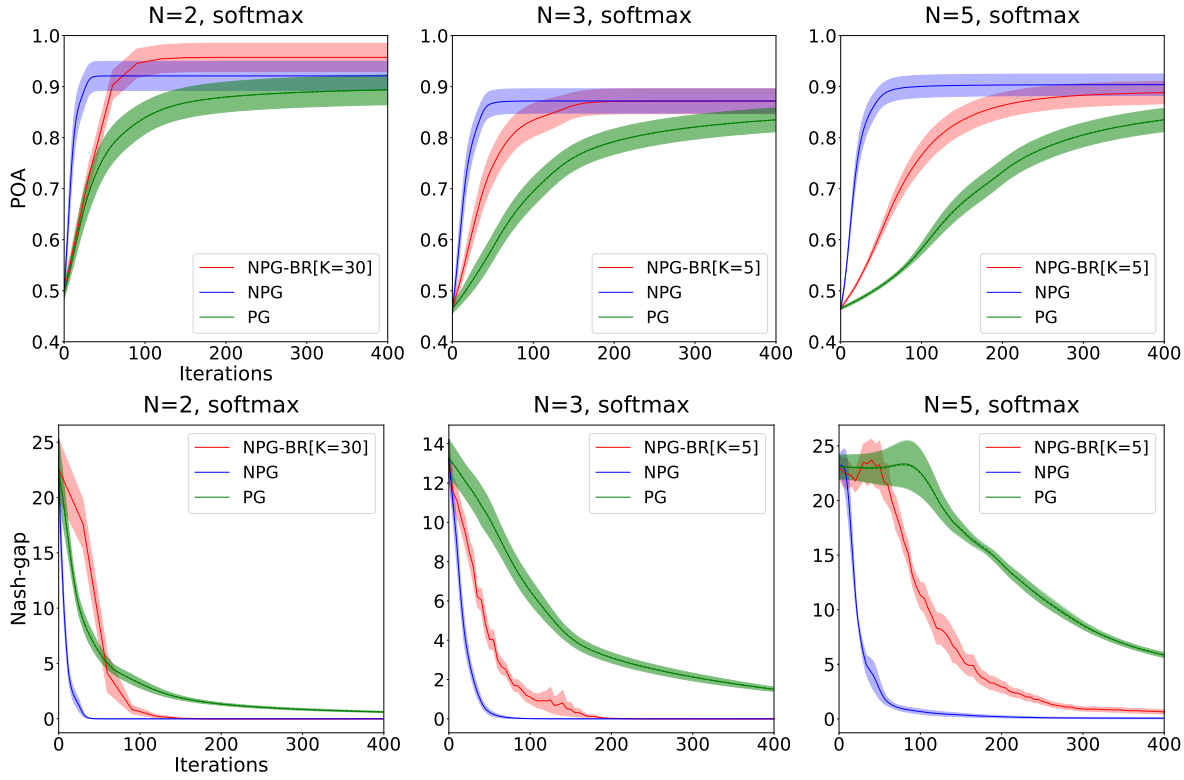

Figure 1. POA (top) and Nash-gap (bottom) under the tabular softmax parameterization (means and standard errors over 10 random initializations).The dashed lines are the curves of the log barrier regularized version of the algorithms with the same color.

We make Assumption 4.6 that also appears in the normal-form game setting (Roughgarden, 2015).

Assumption 4.6. We have $0 < \Phi_\pi(s) \leq V_\pi(s)$ for any product policy π and any state s.

Theorem 4.7 (POA bound of maximum-gain ϵ-ratio-best-response in smooth MPGs, proof in Appendix F.1)**.** *Consider a (α, β)-smooth MPG where Assumption 4.6 holds. Let $\pi_* = \arg\max_\pi V_\pi(\mu)$ be a globally optimal policy and $\sigma > 0$ be a constant for analysis. Consider the sequence of maximum-gain ϵ-ratio-best-response policies $\pi_0, ..., \pi_T$. Then, all but at most*

$$\log_\rho \frac{\Phi_{\max}}{\Phi_0} - T \log_\rho \frac{1}{1-\epsilon} \tag{7}$$

policies π_t in the sequence satisfy

$$V_{\pi_t}(\mu) \geq \frac{\alpha}{(1+\beta)(1+\sigma)} V_{\pi_*}(\mu) \tag{8}$$

where $\rho = (1-\epsilon)(1+\frac{\sigma(1+\beta)}{N})$ and $\Phi_t := \Phi_{\pi_t}(\mu)$.

Since our NPG dynamics (5,6) described in Section 3.3 is an instance of maximum-gain approximate-best-response dynamics, we have Corollary 4.8 directly induced by Theorem 4.7.

Corollary 4.8 (POA bound of the approximate-best-response NPG dynamics (5,6) in smooth MPGs, proof in Appendix F.2)**.** *Consider a (α, β)-smooth MPG where Assumption 4.6 holds. Let $\pi_* = \arg\max_\pi V_\pi(\mu)$ be a globally optimal policy and $\sigma > 0$ be a constant for analysis. Consider the sequence of policies $\pi_0, ..., \pi_T$ generated from the approximate-best-response NPG dynamics (5,6) with $K \geq \frac{4}{(1-\gamma)^2\epsilon}$. Then, all but at most*

$$\log_\rho \frac{\Phi_{\max}}{\Phi_0} - T \log_\rho \left(1 + \frac{\epsilon}{2(1-\gamma)}\right) \tag{9}$$

policies π_t in the sequence satisfy (8), where $\rho = (1 + \frac{\sigma(1+\beta)}{N})/(1 + \frac{\epsilon}{2(1-\gamma)})$.

5. Experiments

Environment. We evaluate the algorithms on Coordination Game, which extends the two players version in (Zhang et al., 2021) to multiple players $N = 2, 3, 5$. The state space and action space are $\mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^N, \mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, respectively, where $\forall i \leq N, \mathcal{S}^i \in \{0, 1\}, \mathcal{A}^i \in \{0, 1\}$. The reward is shared by all the agents (cooperative setting, a special case of Markov Potential Games), and it

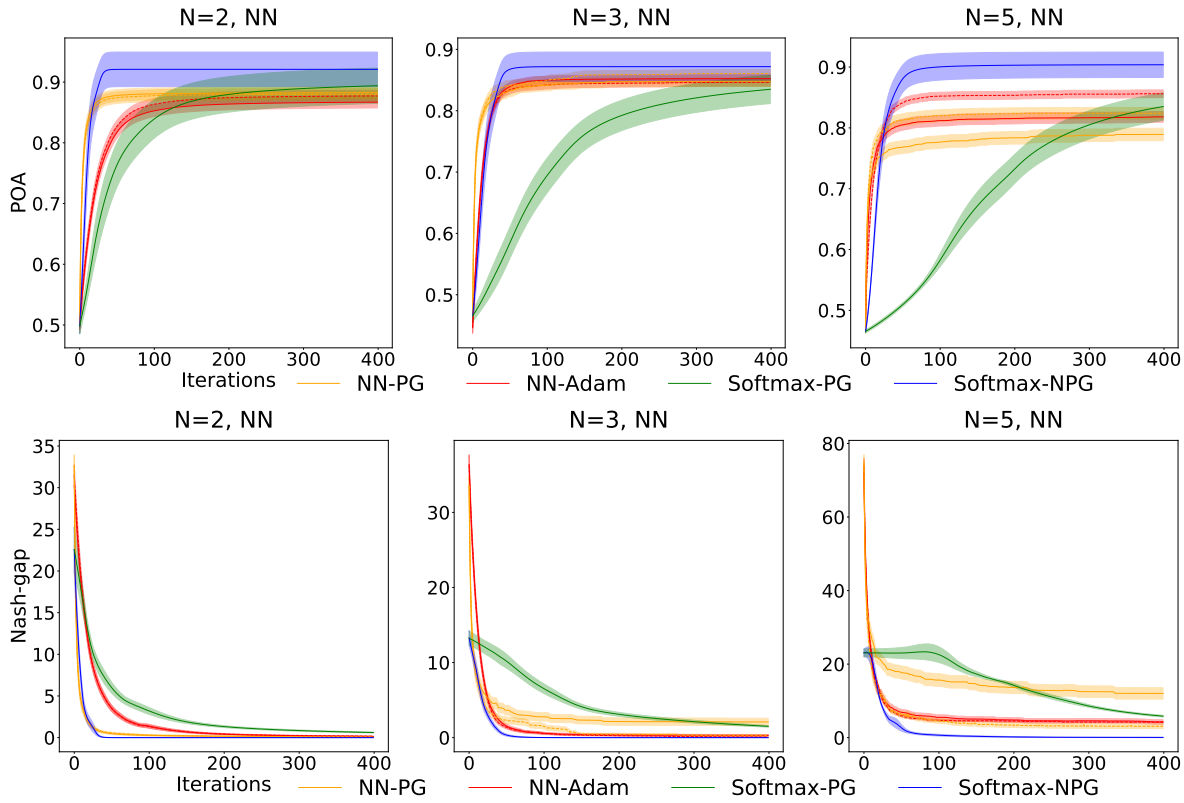

Figure 2. POA (top) and Nash-gap (bottom) under the NN parameterization (means and standard errors over 20 random initializations).The dashed lines are the curves of the log barrier regularized version of the algorithms with the same color.

encourages agents to be in the same local state. To have rewards with more different levels, we design the reward in the way that when the number of agents occupy local state 0 or 1, whichever the maximum, to be the same, the state with more local states of 0s is larger than the one with more 1s. The transition function for each agent i's local state is $P(s^i = 0|a^i = 0) = 1 - \epsilon$, $P(s^i = 0|a^i = 1) = \epsilon$, where $\epsilon = 0.1$.

Algorithms. We exhaustively evaluate the performance of policy gradient **PG**, natural policy gradient **NPG**, and best response natural policy gradient **NPG-BR** with softmax parameterization under the tabular setting, w/wo log barrier regularizer. Besides softmax parameterization for PG, we also consider the neural network parameterization with softmax activation in the last layer **NN-PG**. Precisely, the policy gradient update rule for agent i's neural network policy $\pi^i_{\theta^i} : \mathcal{S} \to \Delta(\mathcal{A}^i)$ is

$$\nabla_{\theta^i}\Phi_\theta(\mu) = \sum_s \sum_{a^i} d^{\pi_\theta}_\mu(s)\pi^i_{\theta^i}(a^i|s)A^i_\theta(s,a^i)\nabla_{\theta^i}\pi^i_{\theta^i}(a^i|s).$$

We run each algorithm in Coordination Game with $N = 2, 3, 5$ agents and plot the Nash-gap and POA as the evaluation metrics. The algorithms, both the tabular softmax and the neural network parameterizations, share the same initial policy parameters, which are sampled from the normal distribution of mean 0 and standard deviation 1. For each log barrier regularized algorithm, we performed a grid search for its coefficient $\lambda \in \{0.01, 0.1, 1.0, 10.0, 100.0\}$ and picked the one with the best POA. Additional details of our experiment are presented in Appendix G.

5.1. Results under the tabular softmax parameterization

Figure 1 presents the POA and the Nash-gap of the algorithms under the tabular softmax parameterization. The results help address the following questions:

How fast do the algorithms converge? In terms of both the POA and the Nash-gap, NPG converges fastest, with NPG-BR the second and PG the slowest. This result demonstrates the improvement in the convergence rate of using the natural policy gradient over the policy gradient.

What is the effect of K for NPG-BR? We did a grid search of $K \in \{1, 5, 10, 20, 50\}$ for NPG-BR (details in Appendix H.1), we show the results for the best-performing K in terms of the POA for $N = 2, 3, 5$ separately in Figure 1. We observe that $K = 5$ is the best for $N = 3, 5$ and $K = 50$, the largest value we searched, is the best for $N = 2$.

How do the algorithms compare in terms of the POA? Consistent with the converge rate, NPG enjoys the overall highest POA, with NPG-BR the second and PG the lowest.

5.2. Results under the neural network parameterization

Figure 2 presents the POA and the Nash-gap of the algorithms under the neural network (NN) parameterization. The results help address the following questions:

Does NN help improve the convergence/POA from tabular softmax? With the NN parameterization, the PG algorithm ("NN-PG") significantly outperforms its tabular softmax counterpart ("Softmax-PG") in terms of both the convergence rate and the POA. NN-PG even outperforms Softmax-NPG in terms of POA at the beginning of the training, although and eventually the POA of Softmax-NPG is the highest among all. This demonstrates the significant improvement of the NN parameterization over the tabular softmax.

What is the effect of the NN regularization? Compared with the results under tabular softmax, the log barrier regularization under NN has a significantly larger impact: it both improves the POA and reduces the Nash-gap at convergence, especially when N is large (e.g., $N = 5$).

What is the effect of the NN optimizer? Among all NN variants, NN-PG is the best in terms of POA when N is small, and the regularized NN-Adam is the best when N is large. When $N = 5$, the POA of the best NN variant, the regularized NN-Adam, is still significantly smaller than Softmax-NPG.

6. Conclusion and discussion

To conclude, we have established in Section 3 convergence to (near-)Nash policies in Markov potential games of several policy gradient-based dynamics under tabular softmax parameterization, including asymptotic convergence of the standard policy gradient dynamics (Section 3.1), its finite-time convergence with log-barrier regularization (Section 3.2), and finite-time convergence of the approximate best-response natural policy gradient dynamics (Section 3.3). In Section 4, we have extended the notion of smoothness in normal-form games to Markov games and established the price-of-anarchy bounds of near-Nash policies in smooth Markov games and of the approximate maximum-gain best-response dynamics in smooth Markov potential games.

Future work. (i) Our theoretical guarantee for the NPG dynamics is limited to the (approximate) best-response variant, although our empirical results imply that the standard NPG dynamics where all agents get updated per iteration should also converge. This suggests that a future direction is to establish the convergence of the standard NPG dynamics. (ii) Our POA bound is also limited to the (approximate) best-response NPG dynamics, and a future direction is to provide POA bounds for other learning dynamics. (iii) Both the theoretical and the empirical parts of this paper are limited to exact gradient computation, and therefore an immediate future direction is to explore sample-based learning dynamics.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*, 2019.

Callaway, D. S. and Hiskens, I. A. Achieving controllability of electric loads. *Proceedings of the IEEE*, 99(1):184–199, 2010.

Chu, T., Wang, J., Codecà, L., and Li, Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1086–1095, 2019.

Corke, P., Peterson, R., and Rus, D. Networked robots: Flying robot navigation using a sensor net. In *Robotics research. The eleventh international symposium*, pp. 234–243. Springer, 2005.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. *arXiv preprint arXiv:1705.08926*, 2017.

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.

Hu, J. and Wellman, M. P. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pp. 3107–3110. PMLR, 2021.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pp. 6379–6390, 2017.

Macua, S. V., Zazo, J., and Zazo, S. Learning parametric closed-loop policies for markov potential games. *arXiv preprint arXiv:1802.00899*, 2018.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Mirrokni, V. S. and Vetta, A. Convergence issues in competitive games. In *Approximation, randomization, and combinatorial optimization. algorithms and techniques*, pp. 183–194. Springer, 2004.

Roughgarden, T. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):1–42, 2015.

Roughgarden, T. *Twenty lectures on algorithmic game theory*. Cambridge University Press, 2016.

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

Tesauro, G. Extending q-learning to general adaptive multi-agent systems. *Advances in neural information processing systems*, 16, 2003.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar, T. Fully decentralized multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1802.08757*, 2018.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020.

Zhang, R., Ren, Z., and Li, N. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021.

A. Proof of Lemma 3.1

Note that

$$\frac{\partial \log \pi^i_{\theta^i}(a^i_\prime|s')}{\partial \theta^i_{s,a^i}} = \mathbb{1}[s=s'](\mathbb{1}[a^i=a^i_\prime] - \pi^i_{\theta^i}(a^i|s)).$$

Plugging it and by similar derivations in the proof of Lemma C.1 in (Agarwal et al., 2019), we have:

$$\begin{aligned}\frac{\partial V^i_\theta(\mu)}{\partial \theta^i_{s,a^i}} =& \mathbb{E}_{s'\sim d^{\pi_\theta}_\mu}\, \mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\left[\mathbb{1}[(s',a^i_\prime)=(s,a^i)]A^i_\theta(s',a')\right]\\ =& d^{\pi_\theta}_\mu(s)\pi^i_{\theta^i}(a^i|s)\,\mathbb{E}_{a^{-i}\sim\pi^{-i}_{\theta^{-i}}(\cdot|s)}\left[A^i_\theta(s,a^i,a^{-i})\right].\end{aligned}$$

This concludes the proof.

B. Proof of Lemma 3.2

Since $\Phi_\theta(s_0)$, abbreviated as Φ_θ in this proof, is (assumed to be) twice-differentialble, as an equivalent condition for smoothness, we will bound the spectral norm of its Hessian $\nabla^2_\theta\Phi_\theta$. Similar to the proof of Lemma 4.4 in (Leonardos et al., 2021), we view Hessian

$$\nabla^2_\theta\Phi_\theta = \left[\frac{\partial^2\Phi_\theta}{\partial\theta^i_{s,a^i}\partial\theta^j_{s',a^j}}\right]_{i,s,a^i,j,s',a^j}$$

as a symmetric $N\times N$ block matrix with submatrices

$$\nabla^2_{\theta^i\theta^j}\Phi_\theta = \left[\frac{\partial^2\Phi_\theta}{\partial\theta^i_{s,a^i}\partial\theta^j_{s',a^j}}\right]_{s,a^i,s',a^j}$$

for all $i,j\in\mathcal{N}$. Claim C.2 in (Leonardos et al., 2021) shows that if we can bound the spectral norm of any submatrix as $\left\|\nabla^2_{\theta^i\theta^j}\Phi_\theta\right\|_2\le L$, then the spectral norm of the block matrix is bounded as $\left\|\nabla^2_\theta\Phi_\theta\right\|_2\le NL$. We then next bound the spectural norm (i.e., the largest absolute eigenvalue) of matrix $\nabla^2_{\theta^j\theta^i}\Phi_\theta$. Noting $\nabla^2_{\theta^j\theta^i}\Phi_\theta=\nabla^2_{\theta^j\theta^i}V^j_\theta=\nabla^2_{\theta^j\theta^i}V^i_\theta$ due to (1), it suffices to define $U(t):=V_{\theta^i+t\cdot u,\theta^{-i}}$ and $W(t,s):=V_{\theta^i+t\cdot u,\theta^j+s\cdot v,\theta^{-i,-j}}$ for scalars $t,s\ge 0$ and unit vectors u,v, and to show

$$\max_{\|u\|_2=1}\left|\frac{d^2U(t)}{dt^2}\right|_{t=0}\le\frac{41}{4(1-\gamma)^3}\ \text{and}\ \max_{\|u\|_2=\|v\|_2=1}\left|\frac{d^2W(t,s)}{dtds}\right|_{t=0,s=0}\le\frac{41}{4(1-\gamma)^3}.$$

For $U(t)$, we decompose it as $U(t)=\sum_{a^i}\sum_{a^{-i}}\pi^i_{\theta^i+t\cdot u}(a^i|s_0)\cdot\pi^{-i}_{\theta^{-i}}(a^{-i}|s_0)\cdot Q_{\theta^i+t\cdot u,\theta^{-i}}(s_0,a^i,a^{-i})$. Abbreviating $\pi^i_{\theta^i+t\cdot u}$ as π^i_t, $\pi^{-i}_{\theta^{-i}}$ as π^{-i}, and $Q_{\theta^i+t\cdot u,\theta^{-i}}$ as Q_t, we have

$$\begin{aligned}\frac{d^2U(t)}{dt^2}=\sum_{a^i}\sum_{a^{-i}}\Bigg(&\frac{d^2\pi^i_t(a^i|s_0)}{dt^2}\cdot\pi^{-i}(a^{-i}|s_0)\cdot Q_t(s_0,a^i,a^{-i})\\ &+2\frac{d\pi^i_t(a^i|s_0)}{dt}\cdot\pi^{-i}(a^{-i}|s_0)\cdot\frac{dQ_t(s_0,a^i,a^{-i})}{dt}\\ &+\pi^i_t(a^i|s_0)\cdot\pi^{-i}(a^{-i}|s_0)\cdot\frac{d^2Q_t(s_0,a^i,a^{-i})}{dt^2}\Bigg)\end{aligned}$$

We then bound $\left|\frac{d^2U(t)}{dt^2}\right|_{t=0}\Big|$ for any unit vector u by bounding the three terms, respectively. For the first term, we have $\sum_{a^i}\left|\frac{d^2\pi^i_t(a^i|s_0)}{dt^2}\right|_{t=0}\Big|\le 6=:C_2$ as proved in Lemma D.4 in (Agarwal et al., 2019), $0\le Q_t(s_0,a^i,a^{-i})\le\frac{1}{1-\gamma}$ assuming the reward is bounded in $[0,1]$, and $\sum_{a^{-i}}\pi^{-i}(a^{-i}|s_0)=1$. For the second term, we have $\sum_{a^i}\left|\frac{d\pi^i_t(a^i|s_0)}{dt}\right|_{t=0}\Big|\le 2=:C_1$

as proved in Lemma D.4 in (Agarwal et al., 2019), and $\left|\frac{dQ_t(s_0,a^i,a^{-i})}{dt}\Big|_{t=0}\right| \leq \frac{\gamma C_1}{(1-\gamma)^2}$ as proved in Lemma D.2 in (Agarwal et al., 2019) and Lemma 4.4 in (Leonardos et al., 2021). For the third term, we have $\left|\frac{d^2Q_t(s_0,a^i,a^{-i})}{dt^2}\Big|_{t=0}\right| \leq \frac{2\gamma^2 C_1}{(1-\gamma)^3} + \frac{\gamma C_2}{(1-\gamma)^2}$ as proved in Lemma D.2 in (Agarwal et al., 2019). We hence derive the bound:

$$\begin{aligned}\max_{\|u\|_2=1}\left|\frac{d^2U(t)}{dt^2}\Big|_{t=0}\right| \leq& \frac{C_2}{1-\gamma} + \frac{2\gamma C_1^2}{(1-\gamma)^2} + \frac{2\gamma^2 C_1}{(1-\gamma)^3} + \frac{\gamma C_2}{(1-\gamma)^2}\\ =& \frac{C_2}{(1-\gamma)^2} + \frac{2\gamma C_1^2}{(1-\gamma)^3} = \frac{6+2\gamma}{(1-\gamma)^3} \qquad (C_1=2, C_2=6)\\ \leq& \frac{8}{(1-\gamma)^3} \leq \frac{41}{4(1-\gamma)^3}\end{aligned}$$

For $W(t,s)$, similarly, we decompose it as $W(t,s) = \sum_{a^i}\sum_{a^j}\sum_{a^{-i,-j}} \pi^i_{\theta^i+t\cdot u}(a^i|s_0) \cdot \pi^j_{\theta^j+s\cdot v}(a^j|s_0) \cdot \pi^{-i,-j}_{\theta^{-i,-j}}(a^{-i,-j}|s_0) \cdot Q_{\theta^i+t\cdot u,\theta^j+s\cdot v,\theta^{-i,-j}}(s_0,a^i,a^j,a^{-i,-j})$. With similar abbreviations, we have

$$\begin{aligned}\frac{d^2W(t,s)}{dtds} = \sum_{a^i}\sum_{a^j}\sum_{a^{-i,-j}}\Bigg(&\frac{d\pi^i_t(a^i|s_0)}{dt}\cdot\frac{d\pi^j_s(a^j|s_0)}{ds}\cdot\pi^{-i,-j}(a^{-i,-j}|s_0)\cdot Q_{t,s}(s_0,a^i,a^j,a^{-i,-j})\\ &+\frac{d\pi^i_t(a^i|s_0)}{dt}\cdot\pi^j_s(a^j|s_0)\cdot\pi^{-i,-j}(a^{-i,-j}|s_0)\cdot\frac{dQ_{t,s}(s_0,a^i,a^j,a^{-i,-j})}{ds}\\ &+\pi^i_t(a^i|s_0)\cdot\frac{d\pi^j_s(a^j|s_0)}{ds}\cdot\pi^{-i,-j}(a^{-i,-j}|s_0)\cdot\frac{dQ_{t,s}(s_0,a^i,a^j,a^{-i,-j})}{dt}\\ &+\pi^i_t(a^i|s_0)\cdot\pi^j_s(a^j|s_0)\cdot\pi^{-i,-j}(a^{-i,-j}|s_0)\cdot\frac{d^2Q_{t,s}(s_0,a^i,a^j,a^{-i,-j})}{dtds}\Bigg).\end{aligned}$$

We then bound $\left|\frac{d^2W(t,s)}{dtds}\Big|_{t=0,s=0}\right|$ for any unit vectors u, v by bounding the four terms, respectively. Similarly, the first term can be bounded by $\frac{C_1^2}{1-\gamma}$, the second term by $\frac{\gamma C_1^2}{(1-\gamma)^2}$, the third term by $\frac{\gamma C_1^2}{(1-\gamma)^2}$, and the fourth term by $\frac{C_2}{(1-\gamma)^2} + \frac{2\gamma C_1^2}{(1-\gamma)^3}$. We hence derive the bound:

$$\begin{aligned}\max_{\|u\|=\|v\|=1}\left|\frac{d^2W(t,s)}{dtds}\Big|_{t=0,s=0}\right| \leq& \frac{C_1^2}{1-\gamma} + \frac{\gamma C_1^2}{(1-\gamma)^2} + \frac{\gamma C_1^2}{(1-\gamma)^2} + \frac{C_2}{(1-\gamma)^2} + \frac{2\gamma C_1^2}{(1-\gamma)^3}\\ =& \frac{-4\gamma^2+2\gamma+10}{(1-\gamma)^3} \qquad (C_1=2, C_2=6)\\ \leq& \frac{41}{4(1-\gamma)^3}.\end{aligned}$$

This concludes the proof.

C. Proof of Theorem 3.4

C.1. Notation

Define

$$V_\phi^{\pi_\theta}(s) = \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)|\pi_\theta, s_0=s]$$
$$\Phi^{\pi_\theta}(\mu) = \mathbb{E}_{s_0\sim\mu}[V_\phi^{\pi_\theta}(s_0)]$$
$$Q_\phi^{\pi_\theta}(s,a) = \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)|\pi_\theta, s_0=s, a_0=a]$$
$$A_\phi^{\pi_\theta}(s,a) = Q_\phi^{\pi_\theta}(s,a) - V_\phi^{\pi_\theta}(s)$$

Suppose $\Phi_{\min} \leq Q_\phi^{\pi_\theta}(s,a) \leq \Phi_{\max}$.

C.2. Smoothness of F

Lemma C.1 (Smoothness of F under tabular softmax)**.** *Fix a state s. Let $\theta_s = [(\theta_s^1)^\top, ..., (\theta_s^N)^\top]^\top \in \mathbb{R}^{\sum_i |\mathcal{A}^i|}$ be the column vector of parameters for state s, with $\theta_s^i \in \mathbb{R}^{|\mathcal{A}^i|}$ for $i \in \mathcal{N}$. For some fixed vector $c_s \in \mathbb{R}^{|\mathcal{A}|}$, define $F_s(\theta_s) := \sum_{a\in\mathcal{A}} \pi_{\theta_s}(a|s)c_{s,a} =: \pi_{\theta_s} \cdot c_s$ with $\pi_{\theta_s} \in \mathbb{R}^{|\mathcal{A}|}$ and $\cdot$ denoting inner product. Then, $F_s(\theta_s)$ is -smooth.*

Proof. We will view Hessian $\nabla^2_{\theta_s} F_s(\theta_s)$ as a $N \times N$ block matrix and bound the spectral norm of each submatrix as $\left\|\nabla^2_{\theta_s^i\theta_s^j} F_s(\theta_s)\right\|_2 \leq L$, which bounds the Hessian's spectral norm as $\left\|\nabla^2_{\theta_s} F_s(\theta_s)\right\|_2 \leq NL$.

We have

$$\nabla_{\theta_s^i} F_s(\theta_s) = \nabla_{\theta_s^i}(\pi_{\theta_s} \cdot c_s) = (\nabla_{\theta_s^i}\pi_{\theta_s})^\top c_s = \nabla_{\theta_s^i}\pi^i_{\theta_s^i}\left(\pi^{-i}_{\theta_s^{-i}} \otimes I_{|\mathcal{A}^i|}\right)^\top M^i c_s$$

where $\nabla_{\theta_s^i} F_s(\theta_s) \in \mathbb{R}^{1\times|\mathcal{A}^i|}$, $M^i \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ is the permutation matrix that permutes all joint actions to be sorted as $a = (a^{-i}, a^i)$, I_n is the $n \times n$ identity matrix, and $\otimes$ is the Kronecker product. For the tabular softmax parameterization, we have

$$\nabla_{\theta_s^i}\pi^i_{\theta_s^i} = \text{diag}\left(\pi^i_{\theta_s^i}\right) - \pi^i_{\theta_s^i}\left(\pi^i_{\theta_s^i}\right)^\top.$$

The submatrix is therefore

$$\nabla^2_{\theta_s^i\theta_s^j} F_s(\theta_s) = \nabla_{\theta_s^j}\left(\nabla_{\theta_s^i}\pi^i_{\theta_s^i}\left(\pi^{-i}_{\theta_s^{-i}} \otimes I_{|\mathcal{A}^i|}\right)^\top M^i c_s\right)$$

If $j = i$:

$$\begin{aligned}\nabla^2_{\theta_s^i\theta_s^i} F_s(\theta_s) =& \nabla_{\theta_s^i}\left(\nabla_{\theta_s^i}\pi^i_{\theta_s^i}\left(\pi^{-i}_{\theta_s^{-i}} \otimes I_{|\mathcal{A}^i|}\right)^\top M^i c_s\right)\\ =& \nabla_{\theta_s^i}(\pi^i_{\theta_s^i} \odot b - (\pi^i_{\theta_s^i} \cdot b)\pi^i_{\theta_s^i})\end{aligned}$$

, where $b = \left(\pi^{-i}_{\theta_s^{-i}} \otimes I_{|\mathcal{A}^i|}\right)^\top M^i c_s$.
For the first term, we get

$$\nabla_{\theta_s^i}(\pi^i_{\theta_s^i} \odot b) = \text{diag}(\pi^i_{\theta_s^i} \odot b) - \pi^i_{\theta_s^i}(\pi^i_{\theta_s^i} \odot b)^\top$$

For the second term we get:

$$\nabla_{\theta_s^i}((\pi^i_{\theta_s^i} \cdot b)\pi^i_{\theta_s^i}) = (\pi^i_{\theta_s^i} \cdot b)\nabla_{\theta_s^i}(\pi^i_{\theta_s^i}) + (\nabla_{\theta_s^i}(\pi^i_{\theta_s^i} \cdot b))(\pi^i_{\theta_s^i})^\top$$

$$\longrightarrow \nabla^2_{\theta_s^i\theta_s^i} F_s(\theta_s) = \text{diag}(\pi^i_{\theta_s^i} \odot b) - \pi^i_{\theta_s^i}(\pi^i_{\theta_s^i} \odot b)^\top - (\pi^i_{\theta_s^i} \cdot b)\nabla_{\theta_s^i}(\pi^i_{\theta_s^i}) - (\nabla_{\theta_s^i}(\pi^i_{\theta_s^i} \cdot b))(\pi^i_{\theta_s^i})^\top$$

Since

$$\begin{aligned}&\max(\left\|\text{diag}(\pi^i_{\theta_s^i} \odot b)\right\|_2, \left\|\pi^i_{\theta_s^i} \odot b\right\|_2, |\pi^i_{\theta_s^i} \cdot b|) \leq \|b\|_\infty = \|c\|_\infty\\ &\left\|\nabla_{\theta_s^i}\pi^i_{\theta_s^i}\right\|_2 = \left\|\text{diag}\left(\pi^i_{\theta_s^i}\right) - \pi^i_{\theta_s^i}\left(\pi^i_{\theta_s^i}\right)^\top\right\|_2 \leq 1\\ &\left\|\nabla_{\theta_s^i}(\pi^i_{\theta_s^i} \cdot b)\right\|_2 \leq \left\|\pi^i_{\theta_s^i} \odot b\right\|_2 + \left\|(\pi^i_{\theta_s^i} \cdot b)\pi^i_{\theta_s^i}\right\|_2 \leq 2\|c\|_\infty,\end{aligned}$$

we know that

$$\left\|\nabla^2_{\theta_s^i\theta_s^i} F_s(\theta_s)\right\|_2 \leq 5\|c\|_\infty$$

If $j \neq i$:

$$\nabla^2_{\theta_s^i\theta_s^j} F_s(\theta_s) = M^j\nabla_{\theta_s^i}\pi^i_{\theta_s^i}\left(\left(\left(\pi^{-i,-j}_{\theta_s^{-i,-j}} \otimes I_{|\mathcal{A}^j|}\right)\nabla_{\theta_s^j}\pi^j_{\theta_s^j}\right) \otimes I_{|\mathcal{A}^i|}\right)^\top M^j M^i c_s$$

Since
$\left\|M^j M^i c_s\right\|_2 \leq \sqrt{N}\left\|c\right\|_\infty$

$$\begin{aligned}\left\|\left(\left(\pi^{-i,-j}_{\theta_s^{-i,-j}} \otimes I_{|\mathcal{A}^j|}\right)\nabla_{\theta_s^j}\pi^j_{\theta_s^j}\right)\otimes I_{|\mathcal{A}^i|}\right\|_2 &= \left\|\left(\pi^{-i,-j}_{\theta_s^{-i,-j}} \otimes I_{|\mathcal{A}^j|}\right)\nabla_{\theta_s^j}\pi^j_{\theta_s^j}\right\|_2 \\ &\leq \left\|\pi^{-i,-j}_{\theta_s^{-i,-j}} \otimes I_{|\mathcal{A}^j|}\right\|_2 \left\|\nabla_{\theta_s^j}\pi^j_{\theta_s^j}\right\|_2 \\ &\leq \left\|\pi^{-i,-j}_{\theta_s^{-i,-j}}\right\|_2 \\ &\leq 1\end{aligned} \tag{10}$$

we know that

$$\nabla^2_{\theta_s^i\theta_s^j} F_s(\theta_s) \leq \sqrt{N}\left\|c\right\|_\infty$$

□

Therefore, we have

$$\left\|\nabla^2_{\theta_s} F_s(\theta_s)\right\|_2 \leq N\max(5,\sqrt{N})\left\|c\right\|_\infty$$

Lemma C.2. *For product policy that can be factorized into the product of individual policies with softmax parameterization, we have:*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta^i_{s,a^i}} = \frac{\partial \Phi^{\pi_\theta}(\mu)}{\partial \theta^i_{s,a^i}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_{\theta^i}(a^i|s) A_\phi^{\pi_\theta,i}(s,a^i)$$

,where $Q_\phi^{\pi_\theta,i}(s,a^i) = \mathbb{E}_{a^{-i}\sim\pi_{\theta^{-i}}(\cdot|s)}\left[Q_\phi^{\pi_\theta}(s,a^i,a^{-i})\right], A_\phi^{\pi_\theta,i}(s,a^i) = Q_\phi^{\pi_\theta,i}(s,a^i) - V_\phi^{\pi_\theta}(s).$

Proof.

$$\begin{aligned}\frac{\partial V_\phi^{\pi_\theta}(\mu)}{\partial \theta^i_{s',a^i}} &= \frac{\partial \Phi^{\pi_\theta}(\mu)}{\partial \theta^i_{s,a^i}} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_\phi^{\pi_\theta}(s,a)\frac{\partial \log \pi_{\theta^i}(a^i|s)}{\partial \theta^i_{s',a^i}}\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A_\phi^{\pi_\theta}(s,a)\mathbb{1}[s=s'](\mathbb{1}[a[m]=a^i] - \pi_{\theta^i}(a^i|s))\right] \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s')\,\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A_\phi^{\pi_\theta}(s',a)(\mathbb{1}[a[m]=a^i] - \pi_{\theta^i}(a^i|s'))\right] \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s')(\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A_\phi^{\pi_\theta}(s',a)\mathbb{1}[a[m]=a^i]\right] - \mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A_\phi^{\pi_\theta}(s',a)\pi_{\theta^i}(a^i|s')\right]) \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s')(\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A_\phi^{\pi_\theta}(s',a)\mathbb{1}[a[m]=a^i]\right] - \pi_{\theta^i}(a^i|s')\,\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A_\phi^{\pi_\theta}(s',a)\right]) \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s')\,\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A_\phi^{\pi_\theta}(s',a)\mathbb{1}[a[m]=a^i]\right] \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s')\sum_a \pi_\theta(a|s')A_\phi^{\pi_\theta}(s',a)\mathbb{1}[a[m]=a^i] \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s')\pi_{\theta^i}(a^i|s')\,\mathbb{E}_{a^{-i}\sim\pi_\theta(\cdot|s')}\left[A_\phi^{\pi_\theta}(s',a^i,a^{-i})\right] \\ &= \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s')\pi_{\theta^i}(a^i|s')A_\phi^{\pi_\theta,i}(s',a^i)\end{aligned}$$

□

Lemma C.3. *For all agents i with a round of parallel update*

$$\theta^i_{t+1} = \theta^i_t + \eta\nabla V^i_{\theta^i_t}(\mu) = \theta^i_t + \eta\nabla\Phi_{\theta^i_t}(\mu)$$

with learning rates $\eta \leq \frac{1-\gamma}{\beta}$, where $\beta = NL(\Phi_{max} - \Phi_{min})$, $L = max(5,\sqrt{N})$, we have

$$V_\phi^{(t+1)}(s) \geq V_\phi^{(t)}(s); Q_\phi^{(t+1)}(s,a) \geq Q_\phi^{(t)}(s,a).$$

Proof. Let us use the notation $\theta_s \in \mathbb{R}^{\sum_i |\mathcal{A}^i_\phi|}$ to refer to the parameters of the product policy on state s. Define

$$F_s(\theta_s) = \sum_a \pi_{\theta_s}(a|s)c(s,a)$$

where $c(s,a)$ is treated as a constant, and is set to be $A^{(t)}_\phi(s,a)$ later in the proof. Thus,

$$\frac{\partial F_s(\theta_s)}{\partial \theta^i_{s,a^i}}\bigg|_{\theta^{t,i}_s} = \sum_{a'} \frac{\partial \pi_{\theta_s}(a'|s)}{\partial \theta^i_{s,a^i}}\bigg|_{\theta^{t,i}_s} c(s,a')$$

$$= \underbrace{\sum_{a'} \mathbb{1}[a'[i]=a^i] \frac{\partial \pi_{\theta_s}(a'|s)}{\partial \theta^i_{s,a^i}}\bigg|_{\theta^{t,i}_s} c(s,a')}_{(1)} + \underbrace{\sum_{a'} \mathbb{1}[a'[i]\neq a^i] \frac{\partial \pi_{\theta_s}(a'|s)}{\partial \theta^i_{s,a^i}}\bigg|_{\theta^{t,i}_s} c(s,a')}_{(2)}$$

$$(1) = \sum_{a'} \mathbb{1}[a'[i]=a^i] \frac{\pi_{\theta_s}(a'|s)}{\pi_{\theta^i_s}(a^i|s)} \Big[\pi_{\theta^i_s}(a^i|s)(1-\pi_{\theta^i_s}(a^i|s))\Big] c(s,a')$$

$$= \sum_{a'} \mathbb{1}[a'[i]=a^i] \pi_{\theta_s}(a'|s)\Big(1-\pi_{\theta^i_s}(a^i|s)\Big)\Big|_{\theta_{s,t}} c(s,a')$$

$$= \sum_{a'} \mathbb{1}[a'[i]=a^i] \pi^t(a'|s)\Big(1-\pi^{t,i}(a^i|s)\Big) c(s,a')$$

$$(2) = \sum_{a'} \mathbb{1}[a'[i]\neq a^i] \frac{\pi_{\theta_s}(a'|s)}{\pi_{\theta^i_s}(a^i|s)}\Big(- \pi_{\theta^i_s}(a^i|s)\pi_{\theta^i_s}(a'[i]|s)\Big)\Big|_{\theta^{t,i}_s} c(s,a')$$

$$= -\sum_{a'} \mathbb{1}[a'[i]\neq a^i] \pi_{\theta_s}(a'|s)\pi_{\theta^i_s}(a^i|s)\Big|_{\theta^{t,i}_s} c(s,a')$$

$$= -\sum_{a'} \mathbb{1}[a'[i]\neq a^i] \pi^t(a'|s)\pi^{t,i}(a^i|s) c(s,a')$$

$(1)+(2) = \sum_{a'} \mathbb{1}[a'[i]=a^i]\pi^t(a'|s)c(s,a') -$

$$\Big(\sum_{a'} \mathbb{1}[a'[i]=a^i]\pi^t(a'|s)\pi^{t,i}(a^i|s)c(s,a') + \sum_{a'} \mathbb{1}[a'[i]\neq a^i]\pi^t(a'|s)\pi^{t,i}(a^i|s)c(s,a')\Big)$$

$$= \sum_{a'} \mathbb{1}[a'[i]=a^i]\pi^t(a'|s)c(s,a') - \sum_{a'} \pi^t(a'|s)\pi^{t,i}(a^i|s)c(s,a')$$

Let $c(s,a') = A_\phi(s,a')$,

$$= \sum_{a'} \mathbb{1}[a'[i]=a^i]\pi^t(a'|s)A_\phi(s,a') - \sum_{a'} \pi^t(a'|s)\pi^{t,i}(a^i|s)A_\phi(s,a')$$

$$= \sum_{a'} \mathbb{1}[a'[i]=a^i]\pi^t(a'|s)A_\phi(s,a')$$

$$= \pi_{\theta^i}(a^i|s)A^{\pi_\theta,i}_\phi(s,a^i)$$

Therefore,

$$\nabla \Phi^t_{\theta^i}(\mu) = \frac{1}{1-\gamma} d^{\pi_\theta}_\mu(s) \frac{\partial F_s(\theta_s)}{\partial \theta^i_{s,a^i}}$$

$$\longrightarrow \theta^{t+1}_s = \theta^t_s + \eta \frac{1}{1-\gamma} d^{\pi_\theta}_\mu(s) \frac{\partial F_s(\theta_s)}{\partial \theta_s}\bigg|_{\theta^t_s}$$

Since $F_s(\theta_s)$ is a β-smooth function for $\beta = N\max(5, \sqrt{N})(\Phi_{\max} - \Phi_{\min})$, then our assumptions that $\eta \leq \frac{1-\gamma}{\beta} = \frac{1-\gamma}{N\max(5,\sqrt{N})(\Phi_{\max}-\Phi_{\min})}$ implies $\eta \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s) \leq \frac{1}{\beta}$, which means

$$F_s(\theta_s^{t+1}) \geq F_s(\theta_s^t)$$

$$\longrightarrow V_\phi^{(t+1)}(s) \geq V_\phi^{(t)}(s); Q_\phi^{(t+1)}(s,a) \geq Q_\phi^{(t)}(s,a).$$

□

Lemma C.4. *For all states s and actions a, there exists values $V_\phi^\infty(s), Q_\phi^\infty(s,a)$ and $Q_\phi^{\infty,i}(s,a)$ such that as $t \to \infty, V_\phi^t(s) \to V_\phi^\infty(s), Q_\phi^t(s,a) \to Q_\phi^\infty(s,a), Q_\phi^{t,i}(s,a) \to Q_\phi^{\infty,i}(s,a)$. Define*

$$\Delta^i = \min_{\{s,a^i | A_\phi^{\infty,i}(s,a^i) \neq 0\}} |A_\phi^{\infty,i}(s,a^i)|.$$

$$\Delta = \min_i \Delta^i.$$

Further, there exists a T_0 such that $\forall t > T_0, s \in \mathcal{S}, a^i \in \mathcal{A}_\phi^i$,

$$Q_\phi^{\infty,i}(s,a^i) - \frac{\Delta}{4} \leq Q_\phi^{t,i}(s,a^i) \leq Q_\phi^{\infty,i}(s,a^i) + \frac{\Delta}{4}$$

Proof. $\{V_\phi^t(s)\}$ is bounded and monotonically increasing, therefore $V_\phi^t(s) \to V_\phi^\infty(s)$. Similarly, we know $Q_\phi^t(s,a) \to Q_\phi^\infty(s,a)$. Since the product policy is assumed to converge, we have that $\{Q_\phi^{t,i}(s,a^i)\}$ is convergent. For agent i, state s, categorize the local action a^i into three groups:

$$I_0^{s,i} = \left\{ a^i | Q_\phi^{\infty,i}(s,a^i) = V_\phi^\infty(s) \right\}$$

$$I_+^{s,i} = \left\{ a^i | Q_\phi^{\infty,i}(s,a^i) > V_\phi^\infty(s) \right\}$$

$$I_-^{s,i} = \left\{ a^i | Q_\phi^{\infty,i}(s,a^i) < V_\phi^\infty(s) \right\}$$

Since $Q_\phi^{t,i}(s,a^i) \to Q_\phi^{\infty,i}(s,a^i)$ as $t \to \infty$, there exists a T_0 such that $\forall t > T_0, s \in \mathcal{S}, a^i \in \mathcal{A}_\phi^i$,

$$Q_\phi^{\infty,i}(s,a^i) - \frac{\Delta}{4} \leq Q_\phi^{t,i}(s,a^i) \leq Q_\phi^{\infty,i}(s,a^i) + \frac{\Delta}{4}$$

□

Lemma C.5. *$\exists T_1$ such that $\forall t > T_1, s \in \mathcal{S}$, we have*

$$A_\phi^{t,i}(s,a^i) < -\frac{\Delta}{4} \text{ for } a^i \in I_-^{s,i}; A_\phi^{t,i}(s,a^i) > \frac{\Delta}{4} \text{ for } a^i \in I_+^{s,i}$$

Proof. Since $V_\phi^t(s) \to V_\phi^\infty(s)$, we have that there exists $T_1 > T_0$ such that for all $t > T_1$,

$$V_\phi^\infty(s) - \frac{\Delta}{4} \leq V_\phi^t(s) \leq V_\phi^\infty(s) + \frac{\Delta}{4}$$

For $a^i \in I_-^{s,i}, t > T_1 > T_0$,

$$\begin{aligned} A_\phi^{t,i}(s,a^i) &= Q_\phi^{t,i}(s,a^i) - V_\phi^t(s) \\ &\leq Q_\phi^{\infty,i}(s,a^i) + \frac{\Delta}{4} - V_\phi^t(s) \\ &\leq Q_\phi^{\infty,i}(s,a^i) + \frac{\Delta}{4} - V_\phi^\infty(s) + \frac{\Delta}{4} \\ &\leq -\Delta + \frac{\Delta}{4} + \frac{\Delta}{4} \\ &\leq -\frac{\Delta}{4} \end{aligned} \tag{11}$$

For $a^i \in I_+^{s,i}, t > T_1 > T_0$,

$$\begin{aligned} A_\phi^{t,i}(s,a^i) &= Q_\phi^{t,i}(s,a^i) - V_\phi^t(s) \\ &\geq Q_\phi^{\infty,i}(s,a^i) - \frac{\Delta}{4} - V_\phi^t(s) \\ &\geq Q_\phi^{\infty,i}(s,a^i) - \frac{\Delta}{4} - V_\phi^\infty(s) \\ &\geq \Delta - \frac{\Delta}{4} \\ &\geq \frac{\Delta}{4} \end{aligned} \tag{12}$$

□

Lemma C.6. *$\frac{\partial \Phi^{\pi_\theta}(\mu)}{\partial \theta^i_{s,a^i}} \to 0$ as $t \to \infty$ for all states s, agents i, actions a^i. This implies that $\forall a^i \in I_-^{s,i} \cup I_+^{s,i}, \pi^{t,i}(a^i|s) \to 0$ and that $\sum_{a^i \in I_0^{s,i}} \pi^{t,i}(a^i|s) \to 1$.*

Proof. Since $\Phi^{\pi_\theta}(\mu)$ is smooth, we know $\frac{\partial \Phi^{\pi_\theta}(\mu)}{\partial \theta^i_{s,a^i}} \to 0$ for all s, i, a^i. From lemma 1 we have

$$\frac{\partial \Phi^t(\mu)}{\partial \theta^i_{s,a^i}} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s) \pi^{t,i}(a^i|s) A_\phi^{\pi^{t,i}}(s,a^i)$$

Since from lemma 4 we know that $|A_\phi^{t,i}(s,a^i)| > \frac{\Delta}{4}$ for all $t > T_1$, for all $a^i \in I_-^{s,i} \cup I_+^{s,i}$, which together with the assumption that μ is strict positive for all state s prove $\pi^{t,i}(a^i|s) \to 0$. Then we also know for all $\sum_{a^i \in I_0^{s,i}} \pi^{t,i}(a^i|s) \to 1$. □

Lemma C.7. *For $t \geq T_1$, θ^i_{s,a^i} is strictly decreasing $\forall a^i \in I_-^{s,i}$ and θ^i_{s,a^i} is strictly increasing $\forall a^i \in I+^{s,i}$.*

Proof. From lemma 1 we have

$$\frac{\partial \Phi^t(\mu)}{\partial \theta^i_{s,a^i}} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s) \pi^{t,i}(a^i|s) A_\phi^{t,i}(s,a^i)$$

From lemma 4, we know for all $t > T_1, a^i \in I_-^{s,i}, A_\phi^{t,i}(s,a^i) \leq -\frac{\Delta}{4}$; For all $a^i \in I_+^{s,i}, A_\phi^{t,i}(s,a^i) \geq \frac{\Delta}{4}$. This implies that after iteration T_1, $\frac{\partial \Phi^t(\mu)}{\partial \theta^i_{s,a^i}} < 0 \forall a^i \in I_-^{s,i}$; $\frac{\partial \Phi^t(\mu)}{\partial \theta^i_{s,a^i}} > 0 \forall a^i \in I_+^{s,i}$. $\longrightarrow$ After iteration T_1, θ^i_{s,a^i} is strictly decreasing $\forall a^i \in I_-^{s,i}$ and θ^i_{s,a^i} is strictly increasing $\forall a^i \in I_+^{s,i}$. □

Lemma C.8. *For all states s where $I_+^{s,i} \neq \emptyset$, we have:*

$$max_{a^i \in I_0^{s,i}} \theta^{t,i}_{s,a^i} \to \infty,\ min_{a^i \in \mathbb{A}^i} \theta^{t,i}_{s,a^i} \to -\infty$$

Proof. Since $I_+^{s,i} \neq \emptyset$, we have some action $a_+^i \in I_+^{s,i}$. From lemma 5, we know

$$\pi^{t,i}(a_+^i|s) \to 0 \text{ as } t \to \infty$$

$$\longrightarrow \frac{\exp(\theta^{t,i}_{s,a_+^i})}{\sum_{a^i \in \mathbb{A}^i} \exp(\theta^{t,i}_{s,a^i})} \to 0 \text{ as } t \to \infty$$

From lemma 6 we know $\theta^{t,i}_{s,a_+^i}$ is monotonically increasing, which implies

$$\sum_{a^i \in \mathbb{A}^i} \exp(\theta^{t,i}_{s,a^i}) \to \infty \text{ as } t \to \infty$$

From lemma 5, we also know

$$\sum_{a^i \in I_0^{s,i}} \pi^{t,i}(a^i|s) \to 1$$

$$\longrightarrow \frac{\sum_{a^i \in I_0^{s,i}} \exp(\theta^{t,i}_{s,a^i})}{\sum_{a^i \in \mathbb{A}^i} \exp(\theta^{t,i}_{s,a^i})} \to 1$$

Since denominator does to ∞, we know

$$\sum_{a^i \in I_0^{s,i}} \exp(\theta^{t,i}_{s,a^i}) \to \infty$$

which implies

$$\max_{a^i \in I_0^{s,i}} \theta^{t,i}_{s,a^i} \to \infty$$

Note this also implies $\max_{a^i \in \mathbb{A}^i} \theta^{t,i}_{s,a^i} \to \infty$. The sum of the gradient is always zero: $\sum_{a^i \in \mathbb{A}^i} \frac{\partial \Phi^t(\mu)}{\partial \theta^i_{s,a^i}} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s) \sum_{a^i \in \mathbb{A}^i} \pi_{\theta^i}(a^i|s) A_\phi^{\pi^{t,i}}(s,a^i) = 0$. Thus, $\sum_{a^i \in \mathbb{A}^i} \theta^{t,i}_{s,a^i} = \sum_{a^i \in \mathbb{A}^i} \theta^{0,i}_{s,a^i}$ which is a constant. Since $\max_{a^i \in \mathbb{A}^i} \theta^{t,i}_{s,a^i} \to \infty$, we know

$$\min_{a^i \in \mathbb{A}^i} \theta^{t,i}_{s,a^i} \to -\infty$$

□

Lemma C.9. *Suppose $a_+^i \in I_+^{s,i}$. $\forall a \in I_0^{s,i}$, if $\exists t \geq T_1$ such that $\pi^{t,i}(a|s) \leq \pi^{t,i}(a_+^i|s)$, then $\forall \tau \geq t, \pi^{\tau,i}(a|s) \leq \pi^{\tau,i}(a_+^i|s)$.*

Proof. Suppose $a_+^i \in I_+^{s,i}, a \in I_0^{s,i}$, if $\pi^{t,i}(a|s) \leq \pi^{t,i}(a_+^i|s)$, then

$$\frac{\partial \Phi^t(\mu)}{\partial \theta^i_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s) \pi^{t,i}(a|s)(Q_\phi^{t,i}(s,a) - V_\Phi^t(s))$$

$$\leq \frac{1}{1-\gamma} d_\mu^{\pi^t}(s) \pi^{t,i}(a_+^i|s)(Q_\phi^{t,i}(s,a_+^i) - V_\Phi^t(s)) = \frac{\partial \Phi^t(\mu)}{\partial \theta^i_{s,a_+^i}}$$

, where the last step holds because $Q_\phi^{t,i}(s,a_+^i) \geq Q_\phi^{\infty,i}(s,a_+^i) - \frac{\Delta}{4} \geq Q_\phi^{\infty,i}(s,a) + \Delta - \frac{\Delta}{4} \geq Q_\phi^{t,i}(s,a) - \frac{\Delta}{4} + \Delta - \frac{\Delta}{4} > Q_\phi^{t,i}(s,a)$ for $t > T_0$.
We can then partition $I_0^{s,i}$ into $B_0^{s,i}(a_+^i)$ and $\bar{B}_0^{s,i}(a_+^i)$ as follows:

$$B_0^{s,i}(a_+^i) : \{a | a \in I_0^{s,i} \text{ and } \forall t \geq T_0, \pi^{t,i}(a_+^i|s) < \pi^{t,i}(a|s)\}$$

$$\bar{B}_0^{s,i}(a_+^i) : I_0^{s,i} \setminus B_0^{s,i}(a_+^i).$$

□

Lemma C.10. *Suppose $I_+^{s,i} \neq \emptyset$. $\forall a_+^i \in I_+^{s,i}$, we have that $B_0^{s,i}(a_+^i) \neq \emptyset$ and that*

$$\sum_{a^i \in B_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s) \to 1, \text{ as } t \to \infty.$$

This implies that:

$$max_{a^i \in B_0^{s,i}(a_+^i)} \theta_s^{t,i} \to \infty.$$

Proof. Let $a_+^i \in I_+^{s,i}$. Consider any $\bar{a}^i \in \bar{B}_0^{s,i}(a_+^i)$. Then by definition of $\bar{B}_0^{s,i}(a_+^i)$, there exists $t' > T_0$ such that $\pi^{t,i}(a_+^i|s) \geq \pi^{t,i}(\bar{a}^i|s)$. From lemma 8, we know $\forall \tau > t, \pi^{\tau,i}(a_+^i|s) \geq \pi^{\tau,i}(\bar{a}^i|s)$. From lemma 5, we know $\pi^{t,i}(a_+^i|s) \to 0$ as $t \to \infty$, which implies

$$\pi^{t,i}(\bar{a}^i|s) \to 0 \text{ as } t \to \infty.$$

Since $B_0^{s,i}(a_+^i) \cup \bar{B}_0^{s,i}(a_+^i) = I_0^{s,i}$ and $\sum_{a^i \in I_0^{s,i}} \pi^{t,i}(a^i|s) \to 1$, we know

$$\sum_{a^i \in B_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s) \to 1$$

$$B_0^{s,i}(a_+^i) \neq \emptyset$$

Using the same techniques in lemma 7, we know

$$\max_{a^i \in B_0^{s,i}(a_+^i)} \theta_s^{t,i} \to \infty$$

□

Lemma C.11. *Consider any s where $I_+^{s,i} \neq \emptyset$. Then, $\forall a_+^i \in I_+^{s,i}, \exists T_{a_+^i}$ such that $\forall t > T_{a_+^i}, \forall a^i \in \bar{B}_0^{s,i}(a_+^i)$,*

$$\pi^{t,i}(a_+^i|s) > \pi^{t,i}(a^i|s)$$

Proof. By the definition of $\bar{B}_0^{s,i}(a_+^i)$ and lemma 8, there exists $t_{a^i} > T_0$ such that $\forall \tau > t_{a^i}, \pi^{\tau,i}(a_+^i|s) > \pi^{\tau,i}(a^i|s)$. We can choose $T_{a_+^i} = \max_{a^i \in B_0^{s,i}(a_+^i)} t_{a^i}$. □

Lemma C.12. *$\forall a_+^i \in I_+^{s,i}$, we have θ^i_{s,a_+^i} is lower bounded as $t \to \infty$. $\forall a_-^i \in I_-^{s,i}$, we have that $\theta^i_{s,a_-^i} \to -\infty$ as $t \to \infty$.*

Proof. From lemma 6, we know that $\forall a_+^i \in I_+^{s,i}$, after T_1, θ^i_{s,a_+^i} is strictly increasing, and is therefore bounded from below. For the second claim, we know from lemma 6 that $\forall a_-^i \in I_-^{s,i}$, after T_1, θ^i_{s,a_-^i} is strictly decreasing. Then, by monotone convergence theorem, we know $\lim_{t\to\infty} \theta^i_{s,a_-^i}$ exists and is either $-\infty$ or some constant θ_0^i. We now prove by contraction that $\lim_{t\to\infty} \theta^i_{s,a_-^i}$ cannot be some constant θ_0^i. Suppose $\lim_{t\to\infty} \theta^i_{s,a_-^i} = \theta_0^i$. We immediately know that $\forall t \geq T_1, \theta^i_{s,a_-^i} > \theta_0^i$. By lemma 7, we know $\exists a^i \in {\mathbb{A}_\phi}^i$ such that

$$\lim \inf_{t\to\infty} \theta^{t,i}_{s,a^i} = -\infty \tag{13}$$

Let us consider some $\delta^i > 0$ such that $\theta^{T_1,i}_{s,a^i} \geq \theta_0^i - \delta^i$. Now for $t \geq T_1$, define $\tau^i(t)$ to be the largest iteration in $[T_1, t]$ such that $\theta^{\tau^i(t),i}_{s,a^i} \geq \theta_0^i - \delta^i$. Define $\mathcal{T}^{t,i}$ to be the subsequence $\{t'\}$ of the interval $(\tau^i(t), t)$ such that $\theta^{t',i}_{s,a^i}$ decreases. Define

$$Z^{t,i} = \sum_{t' \in \mathcal{T}^{t,i}} \frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}}$$

For non-empty $\mathcal{T}^{t,i}$, we have:

$$Z^{t,i} = \sum_{t' \in \mathcal{T}^{t,i}} \frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}} \leq \sum_{t'=\tau^i(t)+1}^{t-1} \frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}} \leq \sum_{t'=\tau^i(t)}^{t-1} \frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}} + \frac{1}{(1-\gamma)}(\Phi_{\max} - \Phi_{\min})$$

$$= \frac{1}{\eta}(\theta^{t,i}_{s,a^i} - \theta^{\tau^i(t),i}_{s,a^i}) + \frac{1}{(1-\gamma)}(\Phi_{\max} - \Phi_{\min})$$

where we have used that $|\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}}| \leq \frac{1}{(1-\gamma)}(\Phi_{\max} - \Phi_{\min})$.
By equation (13), we know

$$\lim \inf_{t\to\infty} Z^{t,i} = -\infty \tag{14}$$

For any $\mathcal{T}^{t,i} \neq \emptyset, \forall t' \in \mathcal{T}^{t,i}$, from lemma 1, we know:

$$\left|\frac{\partial \Phi^{t'}(\mu)/\partial \theta^i_{s,a^i_-}}{\partial \Phi^{t'}(\mu)/\partial \theta^i_{s,a^i}}\right| = \left|\frac{\pi^{t',i}(a^i_-|s)A^{t',i}_\phi(s,a^i_-)}{\pi^{t',i}(a^i|s)A^{t',i}_\phi(s,a^i)}\right| \geq \exp(\theta^i_0 - \theta^{t',i}_{s,a^i})\frac{\Delta}{4(\Phi_{\max} - \Phi_{\min})}$$

$$\geq \exp(\delta^i)\frac{\Delta}{4(\Phi_{\max} - \Phi_{\min})}$$

where we have used that $|A^{t',i}_\phi(s,a^i)| \leq \Phi_{\max} - \Phi_{\min}$ and $\forall t' > T_1, |A^{t',i}_\phi(s,a^i_-)| \geq \frac{\Delta}{4}$. Since both $\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i_-}}$ and $\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}}$ are negative, we can get:

$$\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i_-}} \leq \exp(\delta^i)\frac{\Delta}{4(\Phi_{\max} - \Phi_{\min})}\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}} \tag{15}$$

For non-empty $\mathcal{T}^{t,i}$,

$$\frac{1}{\eta}(\theta^{t,i}_{s,a^i_-} - \theta^{T_1,i}_{s,a^i_-}) = \sum_{t'=T_1}^{t-1}\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i_-}} \leq \sum_{t'\in\mathcal{T}^{t,i}}\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i_-}}$$

By equation (15)

$$\leq \exp(\delta^i)\frac{\Delta}{4(\Phi_{\max} - \Phi_{\min})}\sum_{t'\in\mathcal{T}^{t,i}}\frac{\partial \Phi^{t'}(\mu)}{\partial \theta^i_{s,a^i}}$$

$$= \exp(\delta^i)\frac{\Delta}{4(\Phi_{\max} - \Phi_{\min})}Z^{t,i}$$

which together with the fact that $\theta^{T_1,i}_{s,a^i_-}$ is some finite constant and equation (14) lead to

$$\theta^{t,i}_{s,a^i_-} \to -\infty \text{ as } t \to \infty$$

this contradicts the assumption that $\{\theta^i_{s,a^{t,i}_-}\}_{t\geq T_1}$ is lower bounded by θ^i_0 and complete the proof. □

Lemma C.13. *Consider any s where $I^{s,i}_+ \neq \emptyset$. Then, $\forall a^i_+ \in I^{s,i}_+$,*

$$\sum_{a^i\in B^{s,i}_0(a^i_+)}\theta^{t,i}_{s,a} \to \infty$$

Proof. For any $a^i \in B^{s,i}_0(a^i_+)$. By definition, we know that $\forall t > T_0, \pi^{t,i}(a^i_+|s) < \pi^{t,i}(a|s)$, which implies that $\theta^{t,i}_{s,a^i_+} < \theta^{t,i}_{s,a}$. Since in lemma 11, $\theta^{t,i}_{s,a^i_+}$ is lower bounded as $t \to \infty$, we know that $\theta^{t,i}_{s,a}$ is lower bounded as $t \to \infty$. This together with lemma 9 proves that

$$\sum_{a^i\in B^{s,i}_0(a^i_+)}\theta^{t,i}_{s,a} \to \infty$$

□

Proof of Theorem 3.4. Suppose $I_+^{s,i}$ is non-empty for some s, else the proof is complete. Let $a_+^i \in I_+^{s,i}$. Then, by lemma 12, we know

$$\sum_{a^i \in B_0^{s,i}(a_+^i)} \theta_s^{t,i} \to \infty \tag{16}$$

For $a^i \in I_-^{s,i}$, since $\frac{\pi^{t,i}(a^i|s)}{\pi^{t,i}(a_+^i|s)} = \exp(\theta_s^{t,i} - \theta_{s,a_+^i}^{t,i}) \to 0$ (as $\theta_{s,a_+^i}^{t,i}$ is lower bounded and $\theta_s^{t,i} \to -\infty$ by lemma 11), there exists $T_2 > T_0$ such that

$$\frac{\pi^{t,i}(a^i|s)}{\pi^{t,i}(a_+^i|s)} < \frac{\Delta}{8|\mathcal{A}_\phi^i|(\Phi_{\max} - \Phi_{\min})}$$

$$\longrightarrow -\sum_{a^i \in I_-^{s,i}} \frac{\pi^{t,i}(a^i|s)}{\Phi_{\max} - \Phi_{\min}} > -\pi^{t,i}(a_+^i|s)\frac{\Delta}{8} \tag{17}$$

For $a^i \in \bar{B}_0^{s,i}$, by definition of $\bar{B}_0^{s,i}$, we have $A_\phi^{t,i}(s,a^i) \to 0$ and by lemma 10, $\forall t > T_{a_+^i} 1 < \frac{\pi^{t,i}(a_+^i|s)}{\pi^{t,i}(a^i|s)}$. Then, $\exists T_3 > T_2, T_{a_+^i}$ such that

$$|A_\phi^{t,i}(s,a^i)| < \frac{\pi^{t,i}(a_+^i|s)}{\pi^{t,i}(a^i|s)} \frac{\Delta}{16|\mathcal{A}_\phi^i|}$$

$$\longrightarrow \sum_{a^i \in \bar{B}_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)|A_\phi^{t,i}(s,a^i)| < \pi^{t,i}(a_+^i|s)\frac{\Delta}{16}$$

$$\longrightarrow -\pi^{t,i}(a_+^i|s)\frac{\Delta}{16} < \sum_{a^i \in \bar{B}_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i) < \pi^{t,i}(a_+^i|s)\frac{\Delta}{16} \tag{18}$$

For $t > T_3$,

$$0 = \sum_{a^i \in \mathcal{A}_\phi^i} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i)$$

$$= \sum_{a^i \in I_0^{s,i}} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i) + \sum_{a^i \in I_+^{s,i}} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i) + \sum_{a^i \in I_-^{s,i}} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i)$$

$$\overset{(a)}{\geq} \sum_{a^i \in B_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i) + \sum_{a^i \in \bar{B}_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i)$$

$$+\pi^{t,i}(a_+^i|s)A_\phi^{t,i}(s,a_+^i) + \sum_{a^i \in I_-^{s,i}} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i)$$

$$\overset{(b)}{\geq} \sum_{a^i \in B_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i) + \sum_{a^i \in \bar{B}_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i) + \pi^{t,i}(a_+^i|s)\frac{\Delta}{4} - \sum_{a^i \in I_-^{s,i}} \frac{\pi^{t,i}(a^i|s)}{\Phi_{\max} - \Phi_{\min}}$$

$$\overset{(c)}{\geq} \sum_{a^i \in B_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i) - \pi^{t,i}(a_+^i|s)\frac{\Delta}{16} + \pi^{t,i}(a_+^i|s)\frac{\Delta}{4} - \pi^{t,i}(a_+^i|s)\frac{\Delta}{8}$$

$$> \sum_{a^i \in B_0^{s,i}(a_+^i)} \pi^{t,i}(a^i|s)A_\phi^{t,i}(s,a^i)$$

where (a) uses $\forall a^i \in I_+^{s,i}$ and $t > T_3 > T_1, A_\phi^{t,i}(s,a^i) > 0$ from lemma 3, (b) uses $\forall t > T_3 > T_1, A_\phi^{t,i}(s,a_+^i) > \frac{\Delta}{4}$ from lemma 3 and $A_\phi^{t,i}(s,a^i) \geq -(\Phi_{\max} - \Phi_{\min})$, (c) uses equation (17) and equation (18). This implies that

$$\forall t > T_3, \sum_{a^i \in B_0^{s,i}(a_+^i)} \frac{\partial \Phi^t(\mu)}{\partial \theta_{s,a}^i} < 0$$

which contradicts with equation (16) which leads to

$$\lim_{t\to\infty} \sum_{a^i \in B_0^{s,i}(a_+^i)} (\theta_{s,a^i}^{t,i} - \theta_{s,a^i}^{T_3,i}) = \eta \sum_{t=T_3}^{\infty} \sum_{a^i \in B_0^{s,i}(a_+^i)} \frac{\partial \Phi^t(\mu)}{\partial \theta_{s,a}^i} \to \infty$$

Therefore, the set $I_+^{s,i} = \emptyset$.
Let $\theta = [\theta^{i,\infty}, \theta^{-i,\infty}], \theta' = [\theta^i, \theta^{-i,\infty}]$.

$$V^{\pi_{\theta'}}(\mu) - V^{\pi_\theta}(\mu) = \Phi^{\pi_{\theta'}}(\mu) - \Phi^{\pi_{\theta^i}}(\mu)$$

$$= \mathbb{E}_{s_0\sim\mu}[V_\Phi^{\pi_{\theta'}}(s_0) - V_\Phi^{\pi_\theta}(s_0)]$$

By performance difference lemma,

$$= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta'}}} [\mathbb{E}_{a\sim\pi_{\theta'}(\cdot|s)} A_\Phi^{\pi_\theta}(s,a)]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta'}}} [\mathbb{E}_{a^i\sim\pi^i(\cdot|s)} [\mathbb{E}_{a^{-i}\sim\pi_{\theta^{\infty,-i}}(\cdot|s)} A_\Phi^{\pi_\theta}(s,a)]]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta'}}} [\mathbb{E}_{a^i\sim\pi^i(\cdot|s)} A_\Phi^{\infty,i}(s,a^i)]$$

Since $I_+^{s,i} = \emptyset$,

$$\leq \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta'}}} [\mathbb{E}_{a^i\sim\pi^{\infty,i}(\cdot|s)} A_\Phi^{\infty,i}(s,a^i)]$$

$$= 0$$

which completes the proof.

□

D. Proofs for Section 3.2

D.1. Proof of Lemma 3.5

The proof extends the proof of Theorem 5.2 in (Agarwal et al., 2019) by the usage of the multi-agent performance difference lemma (Lemma C.1 in (Leonardos et al., 2021)).

Fix an arbitrary agent $i \in \mathcal{N}$ and suppose it deviates from $\pi_{\theta^i}^i$ to an optimal policy $\pi_*^i(\theta^{-i})$ w.r.t. the corresponding single-agent MDP specified by θ^{-i}. We will use π_*^i as a shorthand for $\pi_*^i(\theta^{-i})$ and π^{-i} as a shorthand for $\pi_{\theta^{-i}}^{-i}$. By the definition of ϵ-Nash, we need to show that $V_{\pi_*^i,\pi^{-i}}^i(\mu) - V_\theta^i(\mu) \leq 2\lambda M$.

Similar to the proof of Theorem 5.2 in (Agarwal et al., 2019), we can bound $A_\theta^i(s,a^i) \leq$ for any (s,a^i)-pair. It suffices to bound $A_\theta^i(s,a^i)$ for any (s,a^i) where $A_\theta^i(s,a^i) \geq 0$ (else $A_\theta^i(s,a^i) \leq$ is trivially true):

$$\lambda/(2|\mathcal{S}||\mathcal{A}^i|) =: \epsilon_{\text{opt}} \geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a^i}^i} \overset{\text{(i)}}{=} d_\mu^{\pi_\theta}(s)\pi_{\theta^i}^i(a^i|s)A_\theta^i(s,a^i) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}^i|} - \pi_{\theta^i}^i(a^i|s)\right) \geq \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}^i|} - \pi_{\theta^i}^i(a^i|s)\right)$$

where the last inequality is due to $A_\theta^i(s,a^i) \geq 0$, and by rearranging we get $\pi_{\theta^i}^i(a^i|s) \geq 1/2|\mathcal{A}^i|$. Solving (i) for $A_\theta^i(s,a^i)$, we have

$$\begin{aligned}
A_\theta^i(s,a^i) =& \frac{1}{d_\mu^{\pi_\theta}(s)}\left(\frac{1}{\pi_{\theta^i}^i(a^i|s)}\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a^i}^i} + \frac{\lambda}{|\mathcal{S}|}\left(1 - \frac{1}{\pi_{\theta^i}^i(a^i|s)|\mathcal{A}^i|}\right)\right) \\
\leq& \frac{1}{d_\mu^{\pi_\theta}(s)}\left(2|\mathcal{A}^i|\epsilon_{\text{opt}} + \frac{\lambda}{|\mathcal{S}|}\right) \qquad (\pi_{\theta^i}^i(a^i|s) \geq 1/2|\mathcal{A}^i|) \\
\leq& \frac{2\lambda}{d_\mu^{\pi_\theta}(s)|\mathcal{S}|} \qquad (\epsilon_{\text{opt}} = \lambda/(2|\mathcal{S}||\mathcal{A}^i|))
\end{aligned}$$

We are now ready to use the multi-agent performance difference lemma on $\pi_* := (\pi^i_*, \pi^{-i})$ and π_θ:

$$\begin{aligned} V^i_{\pi^i_*,\pi^{-i}}(\mu) - V^i_\theta(\mu) &= \mathbb{E}_{s\sim d^{\pi_*}_\mu}\, \mathbb{E}_{a^i\sim\pi^i_*(s)}\, \mathbb{E}_{a^{-i}\sim\pi^{-i}} \left[A^i_\theta(s,a^i,a^{-i})\right] \\ &= \sum_s d^{\pi_*}_\mu(s) \sum_{a^i} \pi^i_*(a^i|s) A^i_\theta(s,a^i) \\ &\le \sum_s d^{\pi_*}_\mu(s) \frac{2\lambda}{d^{\pi_\theta}_\mu(s)|\mathcal{S}|} \le 2\lambda \max_s \left(\frac{d^{\pi_*}_\mu(s)}{d^{\pi_\theta}_\mu(s)}\right) \le 2\lambda M \end{aligned}$$

which concludes the proof.

D.2. Proof of Theorem 3.6

Lemma 3.2 shows that Φ_θ is $\frac{41N}{4(1-\gamma)^3}$-smooth. Lemma D.4 in (Agarwal et al., 2019) shows that the regularizer for each agent i is $\frac{2\lambda}{|\mathcal{S}|}$-smooth. Thus, β_λ is an upper bound on the smoothness of $L_\lambda(\theta)$. Then, by standard results, we have

$$\min_{t\le T} \left\|\nabla_\theta L_\lambda(\theta^{(t)})\right\|_2^2 \le \frac{2\beta_\lambda(L_\lambda(\theta^*) - L_\lambda(\theta_0))}{T} \le \frac{2\beta_\lambda(\Phi_{\max} - \Phi_{\min})}{T},$$

where the last inequality is because. We need to choose T large enough such that

$$\sqrt{\frac{2\beta_\lambda(\Phi_{\max} - \Phi_{\min})}{T}} \le \lambda/(2|\mathcal{S}| \max_i |\mathcal{A}^i|).$$

Solving the above inequality we obtain $T \ge \frac{8\beta_\lambda|\mathcal{S}|^2 \max_i |\mathcal{A}^i|^2(\Phi_{\max}-\Phi_{\min})}{\lambda^2}$. By Lemma 3.5, we should set $\lambda = \epsilon/2M$ to achieve the specified Nash-gap of ϵ. Plugging in $\lambda = \epsilon/2M$ and $\beta_\lambda := \frac{41N}{4(1-\gamma)^3} + \frac{2\lambda N}{|\mathcal{S}|}$, we have

$$\begin{aligned} T \ge& \frac{32M^2|\mathcal{S}|^2 \max_i |\mathcal{A}^i|^2 \beta_\lambda(\Phi_{\max} - \Phi_{\min})}{\epsilon^2} \\ =& \frac{328NM^2|\mathcal{S}|^2 \max_i |\mathcal{A}^i|^2(\Phi_{\max} - \Phi_{\min})}{(1-\gamma)^3\epsilon^2} + \frac{64\lambda NM^2|\mathcal{S}| \max_i |\mathcal{A}^i|^2(\Phi_{\max} - \Phi_{\min})}{\epsilon^2} \\ =& \frac{328NM^2|\mathcal{S}|^2 \max_i |\mathcal{A}^i|^2(\Phi_{\max} - \Phi_{\min})}{(1-\gamma)^3\epsilon^2} + \frac{32NM|\mathcal{S}| \max_i |\mathcal{A}^i|^2(\Phi_{\max} - \Phi_{\min})}{\epsilon} \end{aligned}$$

which completes the proof.

E. Proofs for Section 3.3

E.1. Proof of Lemma 3.7

The proof is similar to that of the counterpart lemma for the single-agent setting (Lemma 5.1 of (Agarwal et al., 2019)).

For a vector $w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|}$, define the error function

$$L^i_\theta(w) = \mathbb{E}_{s\sim d^{\pi_\theta}_\mu, a^i\sim\pi^i_{\theta^i}(\cdot|s)} \left[w^\top \nabla_{\theta^i} \log \pi^i_{\theta^i}(a^i|s) - A^i_\theta(s,a^i)\right] = \left\|D^i_\theta \left((\nabla_{\theta^i} \log \pi^i_{\theta^i})w - A^i_\theta\right)\right\|_2^2$$

where $D^i_\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|\times|\mathcal{S}||\mathcal{A}^i|}$ is the diagonal matrix with diagonal entries $\{d^{\pi_\theta}_\mu(s)\pi^i_{\theta^i}(a^i|s)\}_{s,a^i}$, and $\nabla_{\theta^i} \log \pi^i_{\theta^i} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|\times|\mathcal{S}||\mathcal{A}^i|}$ is the Jacobian matrix. By the main property of the Moore–Penrose inverse for least squares, i.e., the minimizer of $\|Ax-b\|_2^2$ with the smallest ℓ_2 norm is $A^\dagger b$, we have

$$w^*_\theta = \left(D^i_\theta(\nabla_{\theta^i} \log \pi^i_{\theta^i})\right)^\dagger \left(D^i_\theta A^i_\theta\right)$$

where w^*_θ is the minimizer of $L^i_\theta(w)$ with the smallest ℓ_2 norm. One can verify that $w^*_\theta =:$

$$\begin{aligned} (F^i_\theta)^\dagger \nabla_{\theta^i} V^i_\theta(\mu) &= \left((\nabla_{\theta^i} \log \pi^i_{\theta^i})^\top D^i_\theta \nabla_{\theta^i} \log \pi^i_{\theta^i}\right)^\dagger \left((\nabla_{\theta^i} \log \pi^i_{\theta^i})^\top D^i_\theta A^i_\theta\right) \\ &= \left(D^i_\theta \nabla_{\theta^i} \log \pi^i_{\theta^i}\right)^\dagger \left((\nabla_{\theta^i} \log \pi^i_{\theta^i})^\top\right)^\dagger \left((\nabla_{\theta^i} \log \pi^i_{\theta^i})^\top D^i_\theta A^i_\theta\right) \\ &= \left(D^i_\theta \nabla_{\theta^i} \log \pi^i_{\theta^i}\right)^\dagger \left(D^i_\theta A^i_\theta\right) \\ &= w^*_\theta \end{aligned}$$

We can then follow the same argument in the proof of Lemma 5.1 in (Agarwal et al., 2019) to show the claim of Lemma 3.7.

E.2. Proof of Theorem 3.8

Suppose the inner loop achieves $\frac{\epsilon}{2}$-near-optimal deviation, which require at most $\frac{4}{(1-\gamma)^2\epsilon}$ inner iterations (Agarwal et al., 2019). Then, either the best-response iteration halts, or the total potential function is improved by at least $\frac{\epsilon}{2}$, which implies the number of outer iterations is at most $O(\frac{1}{(1-\gamma)\epsilon})$.

F. Proofs for Section 4

F.1. Proof of Theorem 4.7

We abbreviate $V_\pi^i(\mu)$ as V_π^i and $V_\pi(\mu)$ as V_π. For any policy π_t, define $\delta^i(\pi_t) := V^i_{\pi_*^i,\pi_t^{-i}} - V^i_{\pi_t}$ and $\Delta(\pi_t) := \sum_i \delta^i(\pi_t)$. We now have

$$V_{\pi_t} = \sum_i V^i_{\pi_t} = \sum_i \left(V^i_{\pi_*^i,\pi_t^{-i}} - \delta^i(\pi_t) \right) \geq \alpha V_{\pi_*} - \beta V_{\pi_t} - \Delta(\pi_t)$$

where the inequality is due to the (α, β)-smoothness of the MPG, which implies

$$V_{\pi_t} \geq \frac{\alpha}{1+\beta} V_{\pi_*} - \frac{1}{1+\beta}\Delta(\pi_t). \tag{19}$$

For a "bad" policy π_t that violates (8), we have

$$\Delta(\pi_t) \geq \alpha V_{\pi_*} - (1+\beta)V_{\pi_t} > (1+\beta)(1+\sigma)V_{\pi_t} - (1+\beta)V_{\pi_t} = \sigma(1+\beta)V_{\pi_t} \geq \sigma(1+\beta)\Phi_{\pi_t}$$

where the first inequality is directly from inequality (19), the second inequality due to that π_t is a bad policy, the third due to the assumption that $\Phi_\pi(s) \leq V_\pi(s)$. Therefore, for the maximum-gain agent chosen to update from t to $t+1$, the increase in its local value is at least $\frac{\sigma(1+\beta)}{N}\Phi_{\pi_t}$ since $\Delta(\pi_t) = \sum_i \delta^i(\pi_t)$. Due to the characteristic of Φ in (1), we have $\Phi_{t+1} - \Phi_{\pi_t} \geq \frac{\sigma(1+\beta)}{N}\Phi_{\pi_t}$, i.e.,

$$\Phi_{t+1} \geq \left(1 + \sigma(1+\beta)/N\right)\Phi_{\pi_t}. \tag{20}$$

For a good π_t being updated, Φ can increase by a ratio of at least $\frac{1}{1-\epsilon}$ since

$$\frac{\Phi_{t+1} - \Phi_t}{\Phi_t} = \frac{V^i_{\pi_{t+1}} - V^i_{\pi_t}}{\Phi_t} = \frac{V_{\pi_{t+1}} - V_{\pi_t}}{\Phi_t} \geq \frac{V_{\pi_{t+1}} - V_{\pi_t}}{V_{\pi_t}} > \frac{1}{1-\epsilon} - 1.$$

Let m and $T-m$ be the number of bad and good policies in the sequence, respectively. We then have $\Phi_0(1 + \frac{\sigma(1+\beta)}{N})^m(\frac{1}{1-\epsilon})^{T-m} \leq \Phi_{\max}$, which implies (7) and concludes the proof.

F.2. Proof of Corollary 4.8

Similar to the proof of Theorem 4.7, we can obtain inequality (20) for a bad π_t, and for a good π_t, the $\epsilon/2$ increase per iteration implies

$$\frac{\Phi_{t+1} - \Phi_t}{\Phi_t} = \frac{V^i_{\pi_{t+1}} - V^i_{\pi_t}}{\Phi_t} = \frac{V_{\pi_{t+1}} - V_{\pi_t}}{\Phi_t} \geq \frac{V_{\pi_{t+1}} - V_{\pi_t}}{V_{\pi_t}} > \frac{\epsilon/2}{1-\gamma}.$$

Let m and $T-m$ be the number of bad and good policies in the sequence, respectively. We then have $\Phi_0(1 + \frac{\sigma(1+\beta)}{N})^m(1 + \frac{\epsilon}{2(1-\gamma)})^{T-m} \leq \Phi_{\max}$, which implies (9) and concludes the proof.

G. Experiment details

G.1. Pseudocode for the reward function of our Coordination Game

Algorithm 1 Calculate the team reward for N agents in state s

if $(N = 2)$ or $(N = 3)$ **then**
  `difference_bound=1`
else
  `difference_bound=2`
end if
if abs(s.count("0") − s.count("1")) ≤ difference_bound **then**
  if s.count("0") < s.count("1") **then**
    reward= 1
  else
    reward= 0
  end if
else if s.count('0') > s.count('1') **then**
  reward= 3
else
  reward= 2
end if

G.2. Hyperparameters

Table 1. Hyperparameters

Hyperparameter	Value
γ (discount factor)	0.95
μ (initial state distribution)	Uniform
η (learning rate)	0.1
λ (log barrier coefficient)	searched over $\{0.01, 0.1, 1.0, 10.0, 100.0\}$
K (NPG-BR inner-loop complexity)	searched over $\{1, 5, 10, 20, 50\}$
NN architecture	2^N-FC(2^N)-FC(2^N)-Linear(2)-softmax

*The NN's input is the one-hot representation of the global state s.

G.3. Computing resources

The code is implemented by PyTorch, and a single run of 400 iterations took approximately 30, 50, 200 seconds for 2,3,5 agents version of the coordination game, respectively, using an NVIDIA Tesla V100 GPU and 32 CPU cores.

H. Additional experimental results

H.1. Effect of K for the NPG-BR dynamics

In Figure 1, we plot the best-performing K for $N = 2, 3, 5$, respectively, in terms of the POA, with the results for each individual K shown in Figure 3.

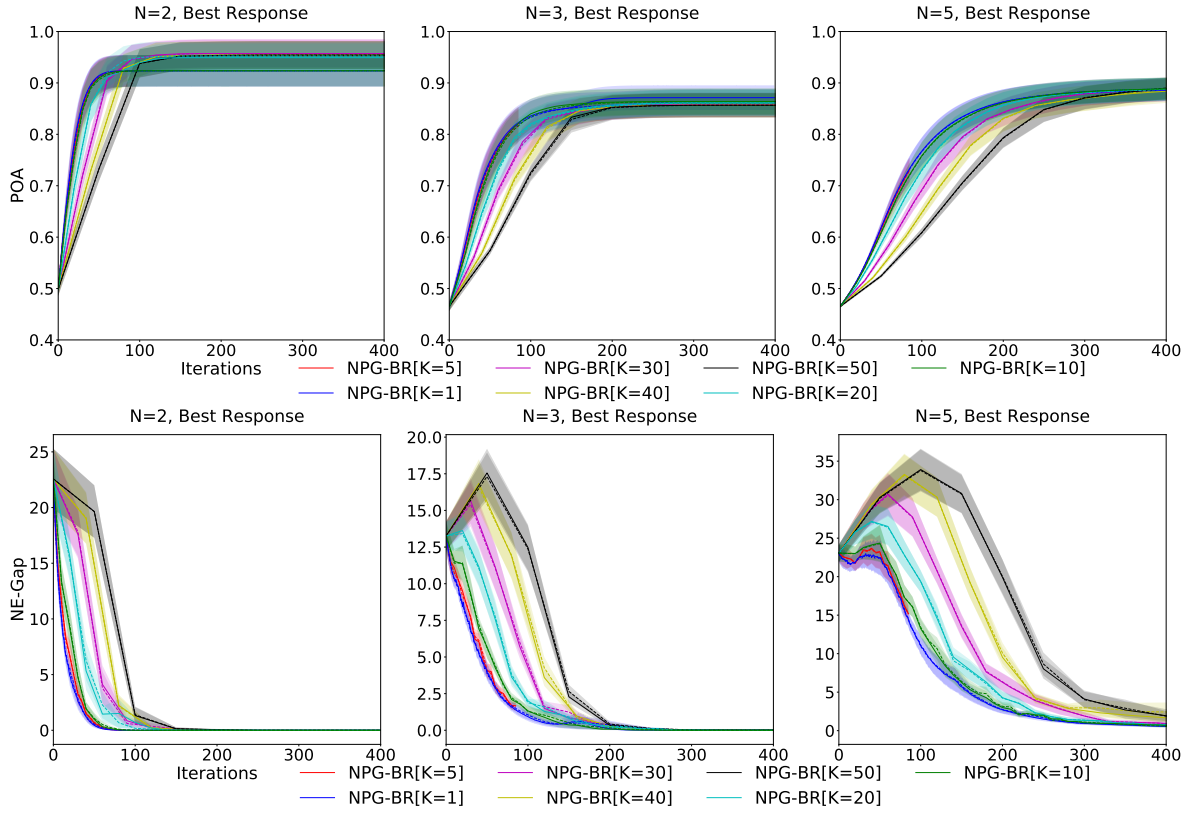

Figure 3. POA (top) and Nash-gap (bottom) under the tabular softmax parameterization (means and standard errors over 10 random initializations).The dashed lines are the curves of the log barrier regularized version of the algorithms with the same color.

H.2. Effect of the log barrier coefficient λ for the PG dynamics under tabular softmax

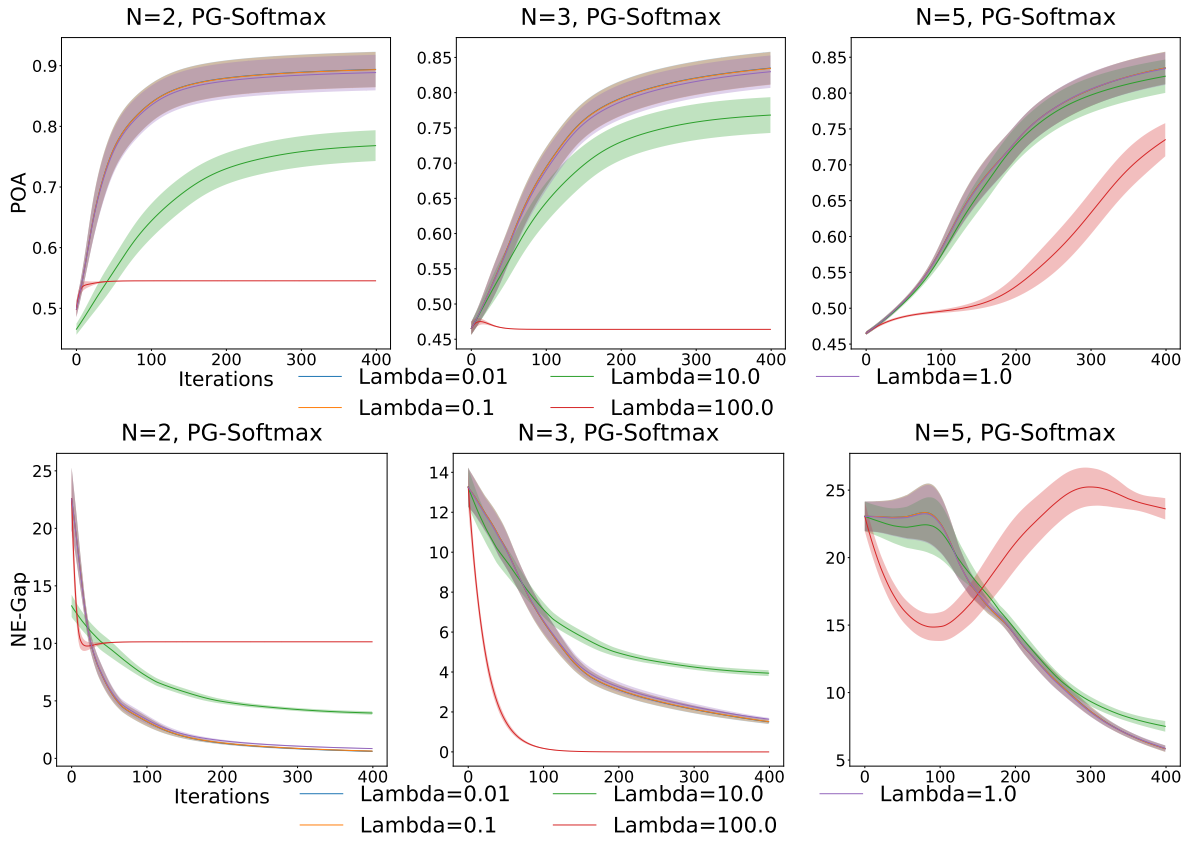

Figure 4. POA (top) and Nash-gap (bottom) for the PG dynamics under the tabular softmax parameterization (means and standard errors over 10 random initializations) with various choices for λ, the log barrier regularization coefficient.

In Figure 1, we plot in the dashed lines the best-performing λ, the log barrier regularization coefficient, for the PG dynamics under tabular softmax. Figure 4 complement the results with the POA and Nash-gap curves with various choices for λ we searched.