Provable Policy Gradient Methods for Average-Reward Markov Potential Games

Min Cheng
Texas A&M University

Ruida Zhou
Univeristy of California, LA

P. R. Kumar
Texas A&M University

Chao Tian
Texas A&M University

Abstract

We study Markov potential games under the infinite horizon average reward criterion. Most previous studies have been for discounted rewards. We prove that both algorithms based on independent policy gradient and independent natural policy gradient converge globally to a Nash equilibrium for the average reward criterion. To set the stage for gradient-based methods, we first establish that the average reward is a smooth function of policies and provide sensitivity bounds for the differential value functions, under certain conditions on ergodicity and the second largest eigenvalue of the underlying Markov decision process (MDP). We prove that three algorithms, policy gradient, proximal-Q, and natural policy gradient (NPG), converge to an ϵ-Nash equilibrium with time complexity $O(\frac{1}{\epsilon^2})$, given a gradient/differential Q function oracle. When policy gradients have to be estimated, we propose an algorithm with $\tilde{O}(\frac{1}{\min_{s,a}\pi(a|s)\delta})$ sample complexity to achieve δ approximation error w.r.t the ℓ_2 norm. Equipped with the estimator, we derive the first sample complexity analysis for a policy gradient ascent algorithm, featuring a sample complexity of $\tilde{O}(1/\epsilon^5)$. Simulation studies are presented.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) (Busoniu et al., 2008; Zhang et al., 2021a) features interactions among multiple agents, with each agent having its own objective and decision-making process. It finds applications in various domains, such as video games (Vinyals et al., 2019; Samvelyan et al., 2019); robotics (Yang and Gu, 2004; Perrusquía et al., 2021); economics (Zheng et al., 2022); and networked system control (Chu et al., 2020). Unlike single-agent RL, interactions among agents create a dynamic and non-stationary environment, making the learning process more challenging. Under the criterion of discounted reward, theoretical investigations have examined Markov general-sum games (Song et al., 2021), zero-sum games (Zhang et al., 2020), and Markov potential games (Leonardos et al., 2021; Zhang et al., 2021b; Ding et al., 2022).

In infinite-horizon tasks, it is more natural to use the average reward over the entire life-span for continuing tasks where optimizing stable, long-term performance becomes crucial, e.g., resource allocation in data centers, congestion games, and control problems (Xu et al., 2014). As shown in (Zhang and Ross, 2020), algorithms designed for discounted reward criterion can lead to unsatisfactory performance under the long-term average cost criterion. However, the average reward criterion which suits long-term strategic games remains largely unexplored. This paper delves into the challenges of employing the average reward criterion in the realm of MARL, specifically for Markov potential games.

Among different approaches to reinforcement learning, policy-based methods are appealing due to the ease of applying function approximation for large state and action spaces. However, most of the existing literature on average reward RL is either based on model-based methods (Auer et al., 2008; Azar et al., 2017), value-based methods (Wei et al., 2020), or based on reduction to discounted MDPs (Jin and Sidford, 2021), with relatively fewer works dedicated to exploring policy-based methods (Li et al., 2022; Wei et al., 2020). This paper examines policy-based methods in the context of average reward Markov potential games and demonstrates the convergence to a Nash policy, which is the primary goal of theoretical investigations in MARL (Leonardos et al., 2021; Ding et al., 2022; Zhang et al.,

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

2022).

1.1 Contributions

- We address the problem of average reward Markov potential games and analyze three algorithms, policy gradient ascent, proximal-Q, and natural policy gradient. We show that with access to a gradient oracle, they converge to an ϵ-Nash equilibrium with time complexity $O(\frac{1}{\epsilon^2})$.

- When the policy gradient has to be estimated, we propose a single-trajectory policy gradient estimator that estimates the policy gradient with $\tilde{O}(\frac{1}{\pi(a|s)\delta})$ sample complexity and δ approximation error w.r.t. the ℓ_2 norm. We also provide the first sample complexity bound $\tilde{O}(\frac{1}{\epsilon^5})$ for the projected policy gradient ascent algorithm.

- On the technical side, we rigorously show that the average reward is an L-smooth function of the policy under an ergodicity assumption. This is the first theoretical analysis of policy gradient for the average reward. We note that the concurrent work of Bai et al. (2023) assumes L-smoothness without proving it. We also establish sensitivity bounds for differential value functions for general single-agent average reward MDPs, which play an important role in providing regret bound independent of the size of the action set (Section 5). These bounds can potentially be used in smoothness analysis of other parameterized policy parameter classes and function approximation analysis, under a further assumption of (Xu et al., 2020, Assumption 1).

Comparison with Previous Works To obtain a sample complexity bound for a projected policy gradient algorithm, existing work (Leonardos et al., 2021) attempts to establish such a bound for a policy projected from a true deterministic gradient, instead of one projected from a gradient estimated from samples (details in Appendix C.4). The only work that analyzes the estimation error of policy gradient under the average reward setting that we are aware of is a recent work (Bai et al., 2023) which separately estimates two parts of policy gradient. Even though they have the same sample complexity, our algorithm calls the estimation algorithm $O(1/\lfloor\frac{1}{\delta}\rfloor)$ times less than theirs, thus reducing the computational burden.

1.2 Related Works

Markov Potential Games originate from Monderer and Shapley (1996), who proposed static potential games. Later, Dechert and O'Donnell (2006) addressed the problem of stochastic lake water usage modeling it as a Markov potential game with known transition probabilities. With the emergence of reinforcement learning, Leonardos et al. (2021) and Zhang et al. (2021b) extended the Markov potential game to the unknown dynamics setting, where they analyzed the convergence to Nash equilibrium by extending the policy gradient techniques developed by Agarwal et al. (2021) and Mei et al. (2020) to the multi-agent setting. Later Ding et al. (2022) proposed a policy ascent algorithm, projecting from the Q-function instead of direct policy gradient. Cen et al. (2022), Zhang et al. (2022), and Sun et al. (2024) have studied the natural policy gradient algorithm in the static and Markov settings. Recently, variants of networked Markov potential games and α-Markov potential games have been studied by Zhou et al. (2023) and Guo et al. (2023). However, all these results are restricted to the discounted reward setting.

Average Reward MDPs date back to the classic results of Howard (1960); Blackwell (1962); Puterman (2004); Kakade (2001); Sutton et al. (1999). When system dynamics are known, average reward MDPs can be solved by linear programming, value iteration, or policy iteration (Howard, 1960; Puterman, 2004). A survey can be found in Dewanto et al. (2020). When the dynamics need to be learned, model-based methods like UCRL2 and UCBVI were proposed in Auer et al. (2008) and Azar et al. (2017), and the reward-biased method originally proposed by Kumar and Becker (1982) has been recently reexamined in Mete et al. (2021). For model-free algorithms, Wei et al. (2020) proposed optimistic mirror ascent, Zhang et al. (2021c) studied TD(λ) and Q learning, and Li et al. (2022) analyzed the policy gradient methods under the mirror descent framework. Zhang and Xie (2023) and Jin and Sidford (2021) proposed algorithms utilizing a reduction to the discounted reward setting.

2 PRELIMINARIES

In this section, we introduce the average reward Markov potential game (AMPG), Nash equilibria, and differential value functions for the average reward MDP.

2.1 Average Reward Markov Potential Games

An N-agent infinite-horizon tabular average-reward Markov game (AMG) is represented by a tuple AMG$(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{r_i\}_{i=1}^N)$. $\mathcal{S}$ denotes a finite state space, $\mathcal{A}_i$ is a finite action space for agent i, and $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times ... \times \mathcal{A}_N$ is the joint action space for all agents. We denote by S, A_1, ...,A_N their respective cardinalities. P denotes the transition probabilities, i.e., $P(\cdot|s, \mathbf{a}) \in \triangle(\mathcal{S})$ is the probability distribution of the next state under a joint action profile $\mathbf{a} = (a_1, a_2, ..., a_N) \in \mathcal{A}$ when the current state is s, and $\triangle(\mathcal{S})$ denotes the probability simplex over set $\mathcal{S}$. $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the one-step reward function for agent i.

A randomized stationary policy for an agent is defined by a map $\pi_i : \mathcal{S} \to \triangle(\mathcal{A}_i)$, i.e., $\pi_i(\cdot|s) \in \triangle(\mathcal{A}_i)$. Denote by $\Pi_i := (\triangle(\mathcal{A}_i))^{\mathcal{S}}$ the set of all randomized policies for agent i. We use $\pi = \{\pi_i\}_{i=1}^N$ to represent the joint policy of all agents, and $\pi_{-i} = \{\pi_j\}_{j\neq i}$ to represent all policies but i. $\Pi = \Pi_1 \times \Pi_2 \times \ldots \times \Pi_N$ is the independent joint policy set. Similarly, we denote $a_{-i} := \{a_j\}_{j\neq i}$ and $\Pi_{-i} := \Pi_1 \times \ldots \Pi_{i-1} \times \Pi_{i+1} \ldots \times \Pi_N$. Under a given joint policy π, the long-term average reward for agent i starting from initial state s is

$$\rho_i^\pi(s) := \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} r_i(s^{(t)}, a^{(t)}) | s^{(0)} = s \right]. \quad (1)$$

In this work, we will restrict our attention to an ergodic underlying MDP:

Assumption 1. *For any joint policy $\pi \in \Pi$, the induced Markov chain is irreducible and aperiodic.*

Under Assumption 1, there exists a unique stationary distribution $\nu^\pi \in \triangle(\mathcal{S})$ independent of the initial state s (Puterman, 2004) i.e., $\nu^\pi(s') = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \mathbb{I}(s^{(t)} = s')|s^{(0)} = s\right] = \lim_{T\to\infty} \mathbb{E}_\pi \left[\mathbb{I}(s^{(T)} = s')|s^{(0)} = s\right]$ for any $s \in \mathcal{S}$. As a result, the average reward $\rho_i^\pi(s) = \langle \nu^\pi, r_i^\pi \rangle$ is also independent of the initial state s and we write it as ρ_i^π for simplicity.

Definition 1 (Average reward Markov potential games)**.** *The average reward Markov potential game (ARMPG) is a special case of AMG, where there exists a potential function $\Phi(\pi) : \Pi_1 \times \Pi_2 \times ...\Pi_N \to \mathbb{R}$, such that for any i, $\pi_i, \pi_i' \in \Pi_i$ and $\pi_{-i} \in \Pi_{-i}$,*

$$\Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}) = \rho_i^{\pi_i,\pi_{-i}} - \rho_i^{\pi_i',\pi_{-i}}.$$

Denote by $C_\Phi := max_{\pi,\pi'} |\Phi(\pi) - \Phi(\pi')|$ the span of the potential function Φ. Note that $C_\Phi \leq N$ since for any joint policies π and π', $|\Phi(\pi) - \Phi(\pi')| = |\rho^\pi - \rho^{\pi'_1,\pi_{2:N}} + \rho^{\pi'_1,\pi_{2:N}} - \rho^{\pi'_{1,2},\pi_{3:N}} + \ldots + \rho^{\pi'_{1:N-1},\pi_N} - \rho^{\pi'}| \leq N$.

Definition 2 (Nash and ϵ-Nash equilibrium)**.** *A policy π^* is a Nash equilibrium if for each agent i,*

$$\rho^{\pi_i^*,\pi_{-i}^*} \geq \rho^{\pi_i,\pi_{-i}^*}, \quad \forall \pi_i \in \Pi_i,$$

or an ϵ-Nash equilibrium if

$$\rho^{\pi_i^*,\pi_{-i}^*} \geq \rho^{\pi_i,\pi_{-i}^*} - \epsilon, \quad \forall \pi_i \in \Pi_i.$$

It may be noted that the maximizer of the potential function is a Nash equilibrium, while the converse may not be true since there could be multiple Nash equilibria.

2.2 Value Functions and Their Properties

The differential value function

$$V_i^\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} (r_i(s^{(t)}, \mathbf{a}^{(t)}) - \rho_i^\pi) | s^{(0)} = s \right] \quad (2)$$

captures the accumulated deviation from the stationary performance. The differential Q function and differential advantage function are defined respectively as:

$$Q_i^\pi(s, \mathbf{a}) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} (r_i(s^{(t)}, \mathbf{a}^{(t)}) - \rho_i^\pi) | s^{(0)} = s, \mathbf{a}^{(0)} = \mathbf{a} \right],$$
$$A_i^\pi(s, \mathbf{a}) := Q_i^\pi(s, \mathbf{a}) - V^\pi(s, \mathbf{a}).$$

Taking the expectation with respect to other policies except j, we can describe how agent j affects Q_i^π by:

$$\overline{Q}_{j;i}^\pi(s, a_j) := \sum_{a_{-j}\in\mathcal{A}_{-j}} \pi_{-j}(a_{-j}|s) Q_i^\pi(s, a_j, a_{-j}). \quad (3)$$

To simplify notation, we will use $\overline{Q}_i^\pi$ to represent $\overline{Q}_{i;i}^\pi$, and define $r_i^\pi(s) := E_{\mathbf{a}\sim\pi(\cdot|s)} r_i(s, \mathbf{a})$, $r_i^{\pi_{-j}}(s, a_j) := E_{a_{-j}\sim\pi_{-j}(\cdot|s)} r_i(s, a_j, a_{-j})$, and $\overline{A}_i^\pi(s, a_i) := \sum_{a_{-i}\in\mathcal{A}_{-i}} \pi_{-i}(a_{-i}|s) A_i^\pi(s, a_i, a_{-i})$.

With $P_\pi \in \mathbb{R}^{S\times S}$ denoting the state transition probability matrix induced by policy π, and $P^\pi \in \mathbb{R}^{SA\times SA}$ denoting the induced state-action transition probability matrix, the differential value function $\mathbf{V_i^\pi}$ and differential Q function $\mathbf{Q_i^\pi}$ are solutions of the following equations, up to a constant shift (Puterman, 2004):

$$\begin{aligned} \mathbf{V}^\pi &= \mathbf{r}_i^\pi - \rho_i^\pi \mathbf{1}_S + P_\pi \mathbf{V}^\pi, \\ \mathbf{Q}^\pi &= \mathbf{r}_i - \rho_i^\pi \mathbf{1}_{SA} + P^\pi \mathbf{Q}^\pi. \end{aligned} \quad (4)$$

Lemma 1 (Performance difference lemma)**.** *For any policies π_j, $\pi_j' \in \Pi_j$, and $\pi_{-j} \in \Pi_{-j}$, the difference between the average rewards for each agent i is*

$$\begin{aligned} &\rho_i^{\pi_j,\pi_{-j}} - \rho_i^{\pi_j',\pi_{-j}} \\ =&\mathbb{E}_{s\sim\nu^{\pi_j,\pi_{-j}}} \langle \overline{Q}_{j:i}^{\pi_j',\pi_{-j}}(s, \cdot), \pi_j(\cdot|s) - \pi_j'(\cdot|s) \rangle \\ =&\mathbb{E}_{s\sim\nu^{\pi_j,\pi_{-j}}} \sum_{a_j} \pi_j(a_j|s) \overline{A}_{j:i}^{\pi_j',\pi_{-j}}(s, a_j). \end{aligned}$$

Let $\overrightarrow{\mathbf{e}}(s, a_j) \in \{0, 1\}^{S\times A_j}$ denote the unit vector where the only non-zero term has the index (s, a_j).

With the definition of Fréchet derivative (Dieudonné, 2011), we can verify that there exists a linear operator $\mathbf{A} := \sum_s \nu^\pi(s) \sum_{a_j} \overrightarrow{\mathbf{e}}(s, a_j)\overline{Q}^\pi_{j;i}(s, a_j)$ mapping the set $U = \{u \in \mathbb{R}^{S\times A_j} : \sum_{a\in\mathcal{A}_j} u(s, a_j) = 0,\ \forall s \in \mathcal{S}\}$ to $\mathbb{R}^{S\times A_j}$ s.t. $\lim_{\|u\|_2\to 0} \frac{\|\rho^{\pi_j+u,\pi_{-j}} - \rho_i^{\pi_j,\pi_{-j}} - \langle \mathbf{A},u\rangle\|_2}{\|u\|_2} = 0$. This can be shown by the performance difference lemma.

Lemma 2 (Partial derivative)**.** *For any i, j, and any policies $\pi_j \in \Pi_j$, $\pi_{-j} \in \Pi_{-j}$,*

$$\frac{\partial \rho_i^\pi}{\partial \pi_j(a_j|s)} = \overline{Q}^\pi_{j;i}(s, a_j)\nu^\pi(s),$$
$$\frac{\partial \Phi(\pi)}{\partial \pi_j(a_j|s)} = \overline{Q}^\pi_j(s, a_j)\nu^\pi(s).$$

Let $\{\lambda_i(\mathbf{M})\}_{i=1,\dots,n}$ be the eigenvalues of matrix $\mathbf{M} \in \mathcal{R}^{n\times n}$, where $|\lambda_1(\mathbf{M})| \geq |\lambda_2(\mathbf{M})| \geq \dots \geq |\lambda_n(\mathbf{M})|$. The largest eigenvalue of P^π is 1 corresponding to the eigenvector $\nu^\pi P^\pi = \nu^\pi$ Puterman (2004). The second largest eigenvalue is strictly less than 1 and is related to the mixing time of the Markov chain induced by π (Kale, 2013, Lemma 2.1).

Definition 3 (Li et al. (2022))**.** *Let $1 - \Gamma$ be the probability of the least visited state of the MDP, with $\Gamma := 1 - \min_{\pi\in\Pi, s\in\mathbb{S}} \nu^\pi(s)$. Note that $\Gamma \in (0, 1)$.*

Definition 4. *The mixing coefficient of the MDP is defined as:*

$$\kappa_0 := \max_{\pi\in\Pi} \frac{1}{1 - |\lambda_2(P^\pi)|}.$$

Unlike our definitions of Γ and κ_0, some literature in the field of average reward MDP employs the concepts of hitting time $t_{hit} := \max_\pi \max_s \frac{1}{\nu^\pi(s)}$ and mixing time $t_{mix} := \max_\pi \min\{t \geq 1 | \|(P^\pi)^t(\cdot|s) - \nu^\pi\|_1 \leq \frac{1}{4}, \forall s \in \mathcal{S}\}$ of the underlying MDP, (Wei et al., 2020; Bai et al., 2023). We can establish a close relationship between these two sets of definitions. Specifically Γ is related to the hitting time as $t_{hit} = \frac{1}{1-\Gamma}$, and from Lemma 3 κ_0 is related to the mixing time as $t_{mix} = O(\frac{1}{\log(1-\frac{1}{\kappa_0})})$. It is also evident that κ_0 is finite under Assumption 1 in the tabular setting.

Lemma 3. *Let $C_p := \min\{\sqrt{\frac{S}{1-\Gamma}}, \frac{1}{1-\Gamma}\}$ and $\varrho := 1 - \frac{1}{\kappa_0}$. Then for any policy π,*

$$\sup_s \|(P^\pi)^t(\cdot|s_0 = s) - \nu^\pi\|_1 \leq C_p\varrho^t, \forall\ t > 0.$$

Definition 5. *Define $\kappa := \max_i \max_\pi \min_{b\in\mathbb{R}} \|\overline{Q}^\pi_i + b\mathbf{1}\|_\infty$ for any reward function $r \in [0, 1]$.*

By Lemma 3, the span of the differential Q function can be bounded as $\kappa \leq \|\overline{Q}^\pi_i\|_\infty \leq 1 + C_p \sum_{t=1}^\infty \varrho^t = 1 + \frac{C_p\varrho}{1-\varrho} \leq C_p\kappa_0$. Details are in Appendix Proposition 3.

2.3 Performance Metric

We study "independent" policy optimization. By this we mean that at step t, each agent i updates its policy π_i^t to π_i^{t+1} based on the information it can collect locally without coordinating with other agents. The goal is to find a Nash equilibrium, and we quantitatively measure the closeness of the joint policy π^t to a Nash equilibrium by the Nash-gap(t) := $\max_i \max_{p\in\Pi_i}(\rho_i^{p,\pi^t_{-i}} - \rho_i^{\pi^t_i,\pi^t_{-i}})$. The optimization algorithm is evaluated by the following notions of Nash regret,

$$\text{Nash-Regret}(T) := \frac{1}{T}\sum_{\tau=0}^{T-1} \text{Nash-gap}(t),$$
$$\text{Nash-Regret}^*(T) := \frac{1}{T}\sum_{\tau=0}^{T-1} \text{Nash-gap}(t)^2.$$

It is clear that the Nash gap is positive and policy π^t is an ϵ-Nash equilibrium if Nash-gap$(t) \leq \epsilon$. Moreover, if the Nash regret (or Nash regret*) is less than ϵ, the policy π^{t^*} with $t^* \in \arg\min_t$ Nash-gap(t), is an ϵ (or $\sqrt{\epsilon}$)-Nash equilibrium.

3 POLICY GRADIENT ALGORITHM

We now analyze the independent projected policy gradient algorithm for average reward Markov potential games. The algorithm adopts a direct policy parameterization. At each step, each agent i updates its policy independently along the gradient direction and projects it back to the policy space via $\text{Proj}_{\Pi_i}(\pi) := \arg\min_{p\in\Pi_i} \|p - \pi\|_2$.

We first consider the oracle-based setting, where the algorithm has access to a gradient-oracle that can exactly calculate the policy gradient of a given policy. We study the convergence performance in this setting and its time complexity. Subsequently, we consider the setting where there is no access to any oracle. We propose a gradient estimator based on trajectory samples, and study its sample complexity.

The key step in the analysis relies on the smoothness of the average value function. Unlike the discounted setting where the gradient has a closed form expression since $I - \gamma P^\pi$ is invertible and its power series can be bounded (Agarwal et al., 2021; Leonardos et al., 2021), in the average case, a similar analysis can not be applied since $\gamma = 1$. In the following Lemma 4, with the help of perturbation theory of stochastic matrices we show that the average reward function is smooth in both single-agent and multi-agent situations. The proof is in Appendix C.2.

Algorithm 1 Independent projected policy gradient ascent

```
Input: learning rate $\beta > 0$
Initialization: $\pi_i^{(0)}(a_i|s) = 1/A_i$ for any $i$, $s$, $a_i$
for $t = 0$ to $T-1$ do
    $\pi_i^{(t+1)} = \mathrm{Proj}_{\Pi_i}(\pi_i^t + \beta \nabla_{\pi_i} \rho_i^{\pi^t}),\ \forall i$
end for
```

Lemma 4 (Smoothness of ρ and Φ)**.** *Denote* $A_{\max} := \max_i A_i$, $L := \kappa_0^2 S^{3/2} A_{\max} + \kappa_0 \sqrt{S} A_{\max}$, *and* $L_\Phi := N(\kappa_0^2 S^{3/2} A_{\max} + \kappa_0 (S A_{\max} + 2A_{\max}) + A_{\max})$.

(a) For any i, *and* $\pi_{-i} \in \Pi_{-i}$, *the average value* ρ_i^π *is* $\kappa_0^2 S^{3/2} A_i + \kappa_0 \sqrt{S} A_i$*-smooth with respect to policy* π_i. *Moreover, for any* i, ρ_i^π *is* L*-smooth with respect to policy* π_i, *i.e.* $\|\nabla_{\pi_i} \rho_i^{\pi_i, \pi_{-i}} - \nabla_{\pi_i} \rho_i^{\pi_i', \pi_{-i}}\|_2 \le L \|\pi_i - \pi_i'\|_2$ *for* $\forall\ i$, *and* $\pi_i, \pi_i' \in \Pi_i$.

(b) The potential function $\Phi(\pi)$ *is* L_Φ*-smooth with respect to joint policy* π, *i.e.* $\|\nabla \Phi(\pi) - \nabla \Phi(\pi')\|_2 \le L_\Phi \|\pi - \pi'\|_2$ *for* $\forall\ \pi, \pi' \in \Pi$.

We note that compared to the smoothness factor in discounted reward settings ($\frac{2\gamma A_{\max}}{(1-\gamma)^3}$ for single-agent, $\frac{2N\gamma A_{\max}}{(1-\gamma)^3}$ for multi-agent), the smoothness factor for the average reward setting has an extra dependence on state size S. The reason is that the second order linear derivative of ρ^π depends on the ℓ_1 norm of the linear derivative of ν^π, while Definition 4 can only guarantee an ℓ_2 bound. The exchange between ℓ_1 and ℓ_2 norms introduces the factor S.

3.1 Policy Gradient Algorithm with Gradient Oracle

We first introduce the distribution mismatch coefficient, which also appears in the convergence behavior of the policy gradient algorithm in discounted single-agent MDPs (Agarwal et al., 2021) and discounted Markov potential games (Ding et al., 2022; Leonardos et al., 2021).

Definition 6 (Distribution mismatch coefficient)**.** $D := \max_{\pi, \pi' \in \Pi} \|\frac{\nu^\pi}{\nu^{\pi'}}\|_\infty$.

In the average reward setting, the coefficient D can be upper bounded by $\frac{1}{1-\Gamma}$. The independent projected policy gradient ascent algorithm is given in Algorithm 1, the regret of which is bounded as follows, with its proof given in Appendix C.2.

Theorem 1. *Choose learning rate* $\beta := \frac{1}{L_\Phi}$. *Then the Nash-regret* of Algorithm 1 is bounded by*

$$\textit{Nash-regret}^*(T) = O\left(\frac{D^2 L_\Phi C_\Phi S}{T}\right).$$

Therefore, by setting time $T := O\left(\frac{N C_\Phi D^2 S^{5/2} A_{\max} \kappa_0^2}{\epsilon^2}\right)$, it yields an ϵ-Nash equilibrium.

Algorithm 2 Policy gradient ascent with estimation

```
Input: learning rate $\beta > 0$, $K$, $N_1$, $N_2$, $N_3$
Initialization: $\pi_i^{(0)}(a_i|s) = 1/A_i$ for any $i$, $s$, $a_i$
for $t = 0$ to $T-1$ do
    agents take action independently and synchronously for $N_1 + KN_2$ time steps to collect trajectories $\{\mathcal{T}_i^t\}$
    for agent $i$ do
        $\hat{g}_i^t \leftarrow$ gradient estimation$(\mathcal{T}_i^t, \pi_i^t, K, N_1, N_2, N_3)$
        $\pi_i^{(t+1)} = \mathrm{Proj}_{\Pi_{i,\alpha}}(\pi_i^t + \beta \hat{g}_i^t)$
    end for
end for
```

The analysis has two parts. Based on the smoothness estimated in Lemma 4, the algorithm will converge to a stationary point by the optimization theory of gradient ascent. We then show that the stationary policy is a Nash equilibrium in the average reward Markov potential game, which establishes the convergence of the algorithm.

3.2 Sample-Based Policy Gradient Estimate

In practice, we usually do not have access to a gradient oracle. To apply the policy gradient algorithm, we need to estimate the gradient from trajectory samples. We propose a gradient estimator (Algorithm 3) which only relies on a single trajectory and is thus more practical in real applications since it does not require resetting the Markov process or a generative model. The sample-based policy gradient ascent algorithm is described in Algorithm 2.

The estimator $\hat{g}_i$ in Algorithm 3 is based on Lemma 2, which relates the policy gradient with $\overline{Q}_i^\pi(s, a_i)$, and we approximate it by $R_i(t) = \sum_{k=0}^N (r_i(s^{t+k}, \mathbf{a}^{t+k}) - \hat{\rho}_i)$. It is generally hard to estimate the policy gradient in the average reward case. Unlike the discounted reward criterion, the policy gradient might be unbounded under average reward setting. We resolve this issue by adapting the sample length N based on the mixing rate in Lemma 3. However, $\hat{g}_i$ may have large variance when $\pi_i(a_i'|s')$ is small. To overcome this, we restrict the policy class to $\Pi_\alpha := \Pi_{1,\alpha} \times \ldots \times \Pi_{N,\alpha}$, where $\Pi_{i,\alpha} := \{(1-\alpha)\pi_i + \alpha u_i | \forall \pi_i \in \Pi_i\}$ with $u_i := (\mathrm{Unif}_{\mathcal{A}_i})^S \in \Pi_i$ being the uniform policy. Such restriction has been considered in Leonardos et al. (2021); Ding et al. (2022) previously. We can balance the representation power of the policy class and the variance of the gradient estimator by adjusting α.

Lemma 5. *For any agent* i, *consider the gradient es-*

Algorithm 3 Gradient estimation

```
Input: trajectory $\mathcal{T} = (s^0, a_i^0, r_i^0, ..., s^t, a_i^t, r_i^t)$, policy $\pi_i$, $K$, $N_1$, $N_2$. $(t = N_1 + KN_2 - 1)$
$\hat{\rho}_i \leftarrow \frac{2}{N_1} \sum_{t=\frac{N_1}{2}}^{N_1 - 1} r_i^t$
$g \leftarrow 0$
for $k = 0$ to $K - 1$ do
    $t_k \leftarrow N_1 + kN_2$
    $R(k) \leftarrow \sum_{\tau=t_k}^{t_k + N_2 - 1} (r_i^\tau - \hat{\rho}_i)$
    $g \leftarrow g + R(k) \nabla_{\pi_i} \log \pi_i(a_i^{t_k} | s^{t_k})$
end for
$\hat{g}_i \leftarrow \frac{g}{K}$
Output:$\hat{g}_i$
```

timate $\hat{g}_i$ defined in Algorithm 3. Given the (s, a, r)-trajectory of length $KN_2 + N_1$ and the policy $\pi_i \in \Pi_{i,\alpha}$ that generated it, the estimated gradient has ℓ_2 error bounded as

$$
\begin{aligned}
\mathbb{E}\|\hat{g}_i - \frac{\partial \rho_i^\pi}{\partial \pi_i}\|_2^2 \leq & \left(\frac{1}{\alpha} + 1\right) \frac{2A_{\max} N_2^2}{K} \\
& + \frac{4C_p A_{\max}}{1 - \varrho^{N_2}} \left(\sqrt{\frac{2}{\alpha}} + \sqrt{2}\right) \frac{N_2^2}{K} \varrho^{N_2} \\
& + \frac{16 A_{\max} C_p^2}{(1 - \varrho)^2} \frac{N_2^2}{N_1^2} \varrho^{N_1} + \frac{2A_{\max} C_p^2}{(1 - \varrho)^2} \varrho^{2N_2}.
\end{aligned}
$$

We can guarantee an ℓ_2 error of $O(\delta)$ by choosing $N_1 = N_2 = O(\log \frac{1}{\delta})$, and $K = \tilde{O}(\frac{A_{\max}}{\alpha\delta})$. A detailed proof is in Appendix C.3.

Theorem 2. *If all players independently and synchronously run Algorithm 2 with learning rate $\beta = \frac{1}{L_\Phi}$, the Nash regret is bounded as:*

$$
\begin{aligned}
\mathbb{E}[\text{Nash-regret}^*(T)] = & O\Big(\frac{D^2 S L_\Phi C_\Phi}{T} + \kappa^2 \alpha^2 \\
& + \kappa^2 D A_{\max} L_\Phi^2 \delta\Big).
\end{aligned}
$$

We can therefore determinate an ϵ-Nash equilibrium by choosing $T = O(\frac{N C_\Phi D^2 S^{5/2} A_{max} \kappa_0^2}{\epsilon^2})$, $\alpha = O(\frac{\epsilon}{\kappa})$, $\delta = O(\frac{\epsilon^2}{\kappa^2 D A_{\max} L_\Phi^2})$, $K = \tilde{O}(\frac{A_{\max}}{\alpha\delta})$, and $N_1 = N_2 = \tilde{O}(1)$. Substituting the bound for κ and L_Φ, the sample complexity is then $T(KN_2 + N_1) = \tilde{O}(\frac{N^3 D^3 C_\Phi S^7 A_{\max}^5 \kappa_0^9}{\epsilon^5 (1-\Gamma)^{3/2}})$.

To analyze Algorithm 2, we introduce the shadow policy $\tilde{\pi}^{t+1} := \text{Proj}_{\Pi_\alpha}(\pi^t + \beta \nabla_\pi \Phi(\pi^t))$ projected from the true deterministic policy gradient. The distance between $\tilde{\pi}^t$ and π^t is bounded by the estimation error. Since the shadow policy $\tilde{\pi}^t$ can capture the optimality criterion in the update rule in Algorithm 2, a bound on the Nash gap with policy improvement w.r.t. potential value can be provided for each time step. Together with the bounded distance from true policy π^t, the conclusion follows.

Algorithm 4 Independent proximal-Q

```
Input: learning rate $\beta > 0$
Initialization: $\pi_i^{(0)}(a_i|s) = 1/A_i$ for any $i$, $s$, $a_i$
for $t = 0$ to $T - 1$ do
    $\pi_i^{t+1}(\cdot|s) = \arg\max_{p(\cdot|s) \in \triangle(\mathcal{A}_i)} \{\beta \langle \overline{Q}_i^{\pi^t}(s, \cdot), p(\cdot|s) \rangle_{\mathcal{A}_i} - \frac{1}{2} \|p(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2\}$ $\forall s, i$
end for
```

Remark. We note that there are some mistakes in the analysis of the sample-based policy gradient in (Leonardos et al., 2021, Theorem 4.7) about the minimizer of the Moreau envelope and the update analysis, which are elaborated in Appendix C.4. We address the difficulties in analyzing the sample-based policy gradient algorithm, and as far as we are aware, Theorem 2 is the first sample-based projected policy gradient algorithm for Markov potential games with a rigorous performance guarantee.

4 PROXIMAL-Q ALGORITHM

In this section, we analyze another algorithm, the proximal-Q algorithm, under the assumption of the availability of an oracle for the differential value function. The more general case where there is no oracle, and its sample complexity analysis are addressed in Appendix D.1. Instead of the policy gradient $\frac{\partial \rho_i^\pi}{\partial \pi_i(a_i|s)} = \overline{Q}_i^\pi(s, a_i) \nu^\pi(s)$, Algorithm 4 uses the differential Q function as the ascent direction, which is less sensitive for states s with a small visiting rate $\nu^\pi(s)$. The key part of the regret analysis is to connect the one-step policy update rule with the difference between the respective potential values. The analysis in the discounted setting is based on the performance difference lemma and second order performance difference lemma (Ding et al., 2022, Lemma 21), both relying on the backward induction enabled by the discount factor $\gamma < 1$, which are unfortunately not applicable in the average reward setting. To bound the second order difference for the average reward, we can use its smoothness property established in Lemma 4, or carefully analyze the sensitivity of the two parts Q^π and ν^π in the performance difference Lemma 1. We emphasize that the sensitivity bounds for differential value functions also play an important role in establishing the regret bound independent of the size of the action set, as shown in Section 5.

Define $\|\pi\|_{1,\infty} := \max_s \|\pi(\cdot|s)\|_1$ and $\|\mathbf{M}\|_\infty := \max_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{M}\mathbf{x}\|_\infty$ for matrix $\mathbf{M}$. One may note that $\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty \leq \|\mathbf{M}_1\|_\infty + \|\mathbf{M}_2\|_\infty$, $\|\mathbf{M}_1\mathbf{M}_2\|_\infty \leq \|\mathbf{M}_1\|_\infty \|\mathbf{M}_2\|_\infty$, and $\|\mathbf{M}\|_\infty \leq \max_i \sum_j |\mathbf{M}_{ij}|$.

Let $P^{\pi,\infty} := ((\nu^\pi)^T, ..., (\nu^\pi)^T)^T$, with $\nu^\pi \in (0, 1)^{1 \times S}$, denote the infinite-step state transition matrix. There exists a closed form expression $V^\pi = (I - P^\pi + P^{\pi,\infty})^{-1}(I - P^{\pi,\infty})r^\pi$ (Puterman, 2004).

Definition 7. *For any policy $\pi \in \Pi$, $(I - P^\pi + P^{\pi,\infty})$ is invertible (Puterman (2004)). Define $\kappa_1 := \max_\pi \|(I - P^\pi + P^{\pi,\infty})^{-1}\|_\infty$.*

Note that $(I - P^\pi + P^{\pi,\infty})^{-1} = I + \sum_{t=1}^\infty (P^\pi - P^{\pi,\infty})^t$. By Lemma 3 and $P^\pi P^{\pi,\infty} = P^{\pi,\infty} P^\pi = P^{\pi,\infty}$, $P^{\pi,\infty} P^{\pi,\infty} = P^{\pi,\infty}$, we have $\kappa_1 \le \|I\|_\infty + \sum_{t=1}^\infty \|(P^\pi)^t - P^{\pi,\infty}\|_\infty \le 1 + \sum_{t=1}^\infty C_p \varrho^t = 1 + C_p \frac{\varrho}{1-\varrho}$.

Consider a general average reward MDP with reward function r. ρ_r^π, V_r^π and Q_r^π can be defined similarly as in Section 2.

Lemma 6 (Sensitivity bounds for average reward MDP)**.** *For any reward function taking values in $[0, 1]$, and any policies $\pi, \pi' \in \Pi$, the following bounds hold:*

$$
\begin{aligned}
|\nu^\pi(s) - \nu^{\pi'}(s)| \le& \kappa \|\pi - \pi'\|_{1,\infty}, \ \forall s \in \mathcal{S} \\
|\rho_r^\pi - \rho_r^{\pi'}| \le& \kappa \|\pi - \pi'\|_{1,\infty}, \\
\|V_r^\pi - V_r^{\pi'}\|_\infty \le& \kappa_1 (2 + S(\kappa + \kappa_1) + S\kappa\kappa_1) \|\pi - \pi'\|_{1,\infty}, \\
\|Q_r^\pi - Q_r^{\pi'}\|_\infty \le& (\kappa + 2\kappa_1 + S\kappa_1(\kappa + \kappa_1) + S\kappa\kappa_1^2) \\
& \times \|\pi - \pi'\|_{1,\infty}.
\end{aligned}
$$

Proofs are provided in Appendix D. Lemma 6 is a general result for both single-agent and multi-agent settings. Consider a joint policy $\pi(\mathbf{a}|s) := \pi_1(a_1|s)\pi_2(a_2|s)...\pi_N(a_N|s)$. If $\pi = (\pi_j, \pi_{-j})$, $\pi' = (\pi'_j, \pi_{-j})$, then $||\pi - \pi'||_{1,\infty} = ||\pi_j - \pi'_j||_{1,\infty}$. The following proposition is directly derived from Lemma 6.

Proposition 1. $\|Q_i^{\pi_j,\pi_{-j}} - Q_i^{\pi'_j,\pi_{-j}}\|_\infty \le \kappa_Q \|\pi_j - \pi'_j\|_{1,\infty}$*, where $\kappa_Q := \kappa + 2\kappa_1 + S\kappa_1(\kappa + \kappa_1) + S\kappa\kappa_1^2$.*

Theorem 3. *If the learning rate is chosen as $\beta \le \max\{\frac{1-\Gamma}{(N-1)(\kappa_Q + S\kappa^2)A_{\max}}, \frac{1-\Gamma}{2L_\Phi}\}$, then Algorithm 4 has a bounded Nash regret:*

$$
\textit{Nash-regret}(T) \le \sqrt{D}(\kappa\sqrt{A_{max}} + \frac{2}{\beta})\sqrt{2\beta C_\Phi}\frac{1}{\sqrt{T}}.
$$

If we set the learning rate to $\beta = \frac{1-\Gamma}{2L_\Phi}$, the time complexity for an ϵ-Nash equilibrium is $T = O(\frac{NC_\Phi DS^{3/2}A_{max}\kappa_0^2}{(1-\Gamma)\epsilon^2})$. The proof is given in Appendix D.

5 NATURAL POLICY GRADIENT

Finally, we analyze the natural policy gradient (NPG) algorithm under the average reward setting. NPG is a powerful technique to accelerate the convergence of policy update with Fisher information via preconditioning (Kakade, 2001). We consider the independent NPG Algorithm 5 under the availability of a differential value function oracle.

Algorithm 5 Independent natural policy gradient ascent

```
1: Input: learning rate $\beta > 0$
2: Initialization: $\pi_i^{(0)}(a_i|s) = 1/A_i$ for any $i$, $s$, $a_i$
3: for $t = 0$ to $T - 1$ do
4:     $\pi_i^{t+1}(\cdot|s) = \arg\max_{p(\cdot|s) \in \triangle(\mathcal{A}_i)} \{\beta\langle \overline{Q}_i^{\pi^t}(s,\cdot), p(\cdot|s)\rangle_{\mathcal{A}_i} - D^p_{\pi_i^t}(s)\} \ \forall s, i$
5: end for
```

Under the softmax parameterization, the joint policy $\pi^\theta = (\pi_1^{\theta_1}, \ldots, \pi_N^{\theta_N})$ with $\theta = (\theta_1, \ldots, \theta_N) \in \mathbb{R}^{SA}$ is $\pi_i^{\theta_i}(s, a) = \frac{\exp(\theta_{s,a_i})}{\sum_{a'_i \in \mathcal{A}_i} \theta_{s,a'_i}}$. The gradient of ρ (or Φ) w.r.t. θ is $\frac{\partial \rho_i^\pi}{\partial \theta} = \sum_{s,a} \nu^\pi(s)\pi(a|s)\frac{\partial \log \pi(s,a)}{\partial \theta} Q_i^\pi(s, a)$ (Sutton et al., 1999). With the Fisher information matrix defined as $F_i(\theta) := \mathbb{E}_{s\sim\nu^{\pi^\theta}, a_i\sim\pi^{\theta_i}(\cdot|s)}[(\frac{\partial \log \pi^{\theta_i}(a_i|s)}{\partial \theta_i})(\frac{\partial \log \pi^{\theta_i}(a_i|s)}{\partial \theta_i})^T]$ (Kakade, 2001), the natural policy gradient update is $\theta_i^{t+1} = \theta_i^t + F_i(\theta_i)^\dagger \nabla_{\theta_i} \rho_i^{\pi^\theta}$. It can be shown that the NPG update is equivalent to the update in Algorithm 5 (the proof is provided in Appendix E). In Algorithm 5, $D_q^p(s) := \sum_a p(a|s) \log \frac{p(a|s)}{q(a|s)}$ is the Kullback–Leibler (KL) Divergence between distributions $p(\cdot|s)$ and $q(\cdot|s)$. There is a closed-form expression for the update, $\pi_i^{t+1}(a_i|s) \propto \pi_i^t(a_i|s) \exp\left(\beta \overline{Q}_i^{\pi^t}(s, a_i)\right) \propto \pi_i^t(a_i|s) \exp\left(\beta \overline{A}_i^{\pi^t}(s, a_i)\right)$, which can be verified by the Karush–Kuhn–Tucker (KKT) condition. This does not require the calculation of the inverse of the Fisher information matrix and the gradient oracle, but employs a differential value function oracle instead.

We first show the monotonic improvement property of the NPG one-step update.

Lemma 7. *Let $Z_t^{i,s} := \sum_{a_i} \pi_i^t(a_i|s) \exp\left(\beta \overline{A}_i^t(s, a_i)\right)$. When $\beta \le \max\{\frac{1-\Gamma}{(N-1)(\kappa_Q + S\kappa^2)}, \frac{1-\Gamma}{L_\Phi}\}$,*

$$
\Phi(\pi^{t+1}) - \Phi(\pi^t) \ge \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1}, \pi_i^t}} \log Z_i^{t,s} \ge 0.
$$

In the discounted reward setting, the performance difference lemma provides a bound for the monotone improvement (Zhang et al., 2022, Lemma 20) which is applicable only when the potential function $\Phi(\pi)$ can be decomposed as the discounted summation of some state action reward function $\Phi(\pi) = \mathbb{E}_{a\sim\pi, s\sim\eta} \sum_t \gamma^t \phi(s, a)$. Here we do not need such an additional assumption, but utilize the smoothness (Lemma 4) or the sensitivity bound for differential Q function (Lemma 6) instead.

Let $c(t) := \min_i \min_s \sum_{a_i^* \in argmax_{a_i \in \mathcal{A}_i} \overline{Q}_i^{\pi^t}(s,a_i)} \pi_i^t(a_i|s)$, which is also considered in Zhang et al. (2022). It indicates the exploration power of the policy π_t, i.e., the probability mass covering the optimal actions. Define $c := \inf_t c(t)$. The following lemma shows that c is strictly positive, which is key to bounding the convergence rate. The proofs are given in Appendix E.

Lemma 8. *If all stationary points of the potential function $\Phi(\theta) = \Phi(\pi(\theta))$ are isolated, $\beta \le \min\{\max\{\frac{1-\Gamma}{(N-1)(\kappa_Q+S\kappa^2)}, \frac{1-\Gamma}{L_\Phi}\}, \frac{1}{2\kappa}\}$, Algorithm 5 asymptotically converges to a Nash equilibrium. Then $c > 0$.*

Theorem 4. *If all stationary points of the potential function $\Phi(\theta) = \Phi(\pi(\theta))$ are isolated, $\beta \le \min\{\max\{\frac{1-\Gamma}{(N-1)(\kappa_Q+S\kappa^2)}, \frac{1-\Gamma}{L_\Phi}\}, \frac{1}{2\kappa}\}$, the regret of Algorithm 5 can be bounded as*

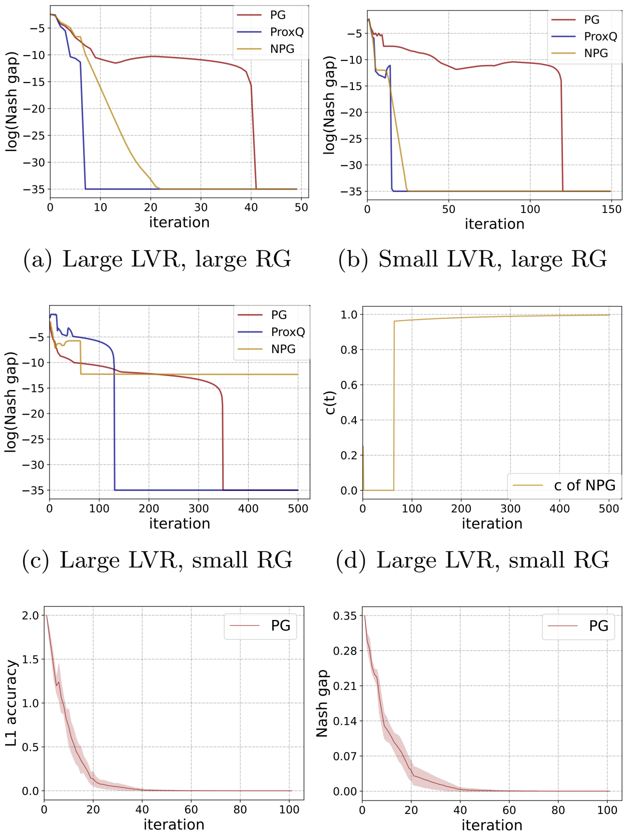
$$\textit{Nash-regret}^*(T) \le \frac{3C_\Phi}{c\beta(1-\Gamma)T}.$$

Substituting the bound for κ and κ_Q, $\beta = \frac{1-\Gamma}{NS^{3/2}\kappa_0^2 \min\{A_{\max}, \frac{S^2\kappa_0}{(1-\Gamma)^{7/2}}\}}$, the time complexity is $T = O(\frac{NC_\Phi S^{3/2}\kappa_0^2 \min\{\frac{S^2\kappa_0}{(1-\Gamma)^{7/2}}, A_{\max}\}}{c(1-\Gamma)^2\epsilon^2})$. When the action set has a large cardinality $A_{\max}$, it gives an $A_{\max}$-independent bound for the chosen learning rate.

6 EXPERIMENTS

We first illustrate the convergence of the algorithms in the oracle setting. We randomly generate a Markov potential game model with $S = 100$ states, and $A_1 = 4$, $A_2 = 3$, $A_3 = 2$ actions for the $N = 3$ agents. We choose the largest learning rate. This yields the fastest convergence for each algorithm. The numerical experiments corroborate our theoretical findings in Theorems 1, 3, and 4.

Recall that in the theorems, a small least visited rate (LVR) $(1-\Gamma)$ has a negative impact on the convergence rate. As shown in Fig. 1(a) and 1(b),the small LVR impedes the convergence of all three algorithms. Comparing the theoretical findings of policy gradient and that of proximal-Q and NPG, we note that the complexity of the former is of order $S^{5/2}$ while the latter is of order $S^{3/2}$. This is also verified in Fig. 1(b) and the effect is more significant when LVR is small. In addition, the convergence of NPG depends on the exploration factor $c(t)$. We generate a reward function with a small reward gap (RG) between the optimal action and the second best optimal one, which increases the exploration difficulty and results in a small value of $c(t)$. Despite $c(t)$ nearing 1, with very small RG, NPG updates can become minor, causing the algorithm to get stuck near a Nash equilibrium (Fig. 1(d)). Further discussion is in Appendix A.



(a) Large LVR, large RG

(b) Small LVR, large RG

(c) Large LVR, small RG

(d) Large LVR, small RG

(e) ℓ_1 accuracy

(f) Nash gap

Figure 1: (a)(b)(c)(d) are results for the oracle setting. Since the Nash gap can be as low as 0.0, we truncate the log(Nash gap) at -35 from below. (d) depicts the change in $c(t)$ of the NPG algorithm for Fig. 1(c). (e) and (f) present the results of Algorithm 2. The solid lines are the means of trajectories over seven random seeds and shaded regions are the standard deviations.

We further demonstrate that the proposed sample-based independent policy gradient Algorithm 2 converges to the desired policy in Fig. 1(e) and 1(f), under both the ℓ_1 accuracy, i.e., $\frac{1}{N}\sum_{i=1}^N \|\pi_i^t - \pi_i^*\|_1$ and Nash gap.

To illustrate the potential of the independent learning scheme in averaged reward MARL, a more complex robot navigation task is conducted. There are two controllers (linear speed controller and angular speed controller) in a robot with each viewed as an agent. The agents gain rewards when the robot is moving toward the target. We implement a practical version of the independent average NPG algorithm with a neural

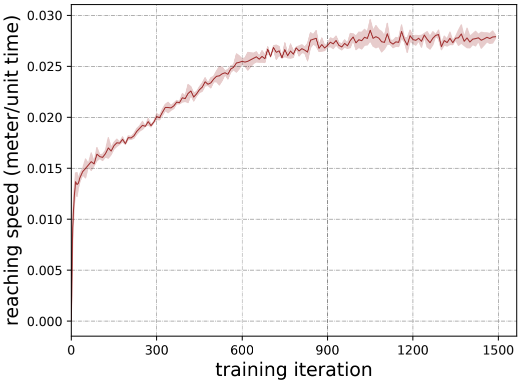


Figure 2: Training process of robot navigation task. The solid line is the mean of trajectories over three random seeds and shaded regions are the standard deviations.

network, which is inspired by Algorithm 1 of Ma et al. (2021), and showcase its performance in Fig. 2.

7 CONCLUSION

In this paper, we study Markov potential games under the average reward criterion and analyze three algorithms, policy gradient ascent, proximal-Q, and NPG under the access to an oracle. We establish time complexity that matches the results in discounted reward settings. We also propose a gradient estimator, which only relies on a single trajectory. The sample-based policy gradient ascent algorithm is shown to converge to a Nash equilibrium, and a sample complexity is provided. We also close several technical gaps in the analysis of policy gradient methods between the discounted and average reward settings.

Acknowledgements

This material is based upon work partially supported by the US Army Contracting Command under W911NF-22-1-0151 and W911NF2120064, US National Science Foundation under CMMI-2038625, and US Office of Naval Research under N00014-21-1-2385.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506.

Auer, P., Jaksch, T., and Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR.

Bai, Q., Mondal, W. U., and Aggarwal, V. (2023). Regret analysis of policy gradient algorithm for infinite horizon average reward Markov decision processes. *arXiv preprint arXiv:2309.01922*.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Blackwell, D. (1962). Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33:719–726.

Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357.

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

Cen, S., Chen, F., and Chi, Y. (2022). Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization. In *2022 61st IEEE Conference on Decision and Control (CDC)*, pages 2833–2838.

Chu, T., Chinchali, S., and Katti, S. (2020). Multi-agent reinforcement learning for networked system control. *arXiv preprint arXiv:2004.01339*.

Davis, D. and Drusvyatskiy, D. (2018). Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*.

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

Dewanto, V., Dunn, G., Eshragh, A., Gallagher, M., and Roosta, F. (2020). Average-reward model-free reinforcement learning: a systematic review and literature mapping. *arXiv preprint arXiv:2010.08920*.

Dieudonné, J. (2011). *Foundations of modern analysis*. Read Books Ltd.

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

Garrett, P. (2005). Hartogs' Theorem: separate analyticity implies joint.

Guo, X., Li, X., Maheshwari, C., Sastry, S., and Wu, M. (2023). Markov α-potential games: Equilibrium approximation and regret analysis. *arXiv preprint arXiv:2305.12553*.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. John Wiley and Sons/MIT Press, New York, NY.

Jin, Y. and Sidford, A. (2021). Towards tight bounds on the sample complexity of average-reward MDPs. In *International Conference on Machine Learning*, pages 5055–5064. PMLR.

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

Kale, S. (2013). Eigenvalues and mixing time. *University of Darthmouth*.

Kazdan, J. L. (1995). Matrices A(t) depending on a parameter t. *unpublished Note*.

Kumar, P. R. and Becker, A. (1982). A new family of optimal adaptive controllers for markov chains. *IEEE Transactions on Automatic Control*, 27(1):137–146.

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

Li, T., Wu, F., and Lan, G. (2022). Stochastic first-order methods for average-reward Markov decision processes. *arXiv preprint arXiv:2205.05800*.

Ma, X., Tang, X., Xia, L., Yang, J., and Zhao, Q. (2021). Average-reward reinforcement learning with trust region methods. *arXiv preprint arXiv:2106.03442*.

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

Mete, A., Singh, R., Liu, X., and Kumar, P. R. (2021). Reward biased maximum likelihood estimation for reinforcement learning. In *Learning for Dynamics and Control*, pages 815–827. PMLR.

Monderer, D. and Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, 14(1):124–143.

Narasimha, D., Lee, K., Kalathil, D., and Shakkottai, S. (2022). Multi-agent learning via Markov potential games in marketplaces for distributed energy resources. In *2022 61st IEEE Conference on Decision and Control (CDC)*, pages 6350–6357.

Perrusquía, A., Yu, W., and Li, X. (2021). Multi-agent reinforcement learning for redundant robot control in task-space. *International Journal of Machine Learning and Cybernetics*, 12:231–241.

Puterman, M. L. (2004). *Markov decision processes: Discrete Stochastic Dynamic Programming.* John Wiley& Sons.

Rengarajan, D., Chaudhary, S., Kim, J., Kalathil, D., and Shakkottai, S. (2022a). Enhanced meta reinforcement learning via demonstrations in sparse reward environments. *Advances in Neural Information Processing Systems*, 35:2737–2749.

Rengarajan, D., Vaidya, G., Sarvesh, A., Kalathil, D., and Shakkottai, S. (2022b). Reinforcement learning with sparse rewards using guidance from offline demonstration. *arXiv preprint arXiv:2202.04628*.

Samvelyan, M., Rashid, T., De Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G., Hung, C.-M., Torr, P. H., Foerster, J., and Whiteson, S. (2019). The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*.

Song, Z., Mei, S., and Bai, Y. (2021). When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*.

Sun, Y., Liu, T., Zhou, R., Kumar, P. R., and Shahrampour, S. (2024). Provably fast convergence of independent natural policy gradient for markov potential games. *Advances in Neural Information Processing Systems*, 36.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12.

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

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10170–10180. PMLR.

Xu, T., Wang, Z., and Liang, Y. (2020). Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369.

Xu, X., Zuo, L., and Huang, Z. (2014). Reinforcement learning algorithms with function approximation: Recent advances and applications. *Information Sciences*, 261:1–31.

Yang, E. and Gu, D. (2004). Multiagent reinforcement learning for multi-robot systems: A survey. Technical report, tech. rep.

Zhang, K., Kakade, S., Basar, T., and Yang, L. (2020). Model-based multi-agent rl in zero-sum Markov games with near-optimal sample complexity. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1166–1178. Curran Associates, Inc.

Zhang, K., Yang, Z., and Başar, T. (2021a). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384.

Zhang, R., Mei, J., Dai, B., Schuurmans, D., and Li, N. (2022). On the global convergence rates of decentralized softmax gradient play in Markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935.

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

Zhang, S., Zhang, Z., and Maguluri, S. T. (2021c). Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242.

Zhang, Y. and Ross, K. W. (2020). Average reward reinforcement learning with monotonic policy improvement. *Handbook of reinforcement learning and control*.

Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward Markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR.

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., and Socher, R. (2022). The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607.

Zhou, R., Liu, T., Cheng, M., Kalathil, D., Kumar, P. R., and Tian, C. (2024). Natural actor-critic for robust reinforcement learning with function approximation. *Advances in neural information processing systems*, 36.

Zhou, Z., Chen, Z., Lin, Y., and Wierman, A. (2023). Convergence rates for localized actor-critic in networked Markov potential games. *arXiv preprint arXiv:2303.04865*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A EXPERIMENTAL DETAILS

A.1 Oracle-based Algorithms

We provide more details regarding the numerical experiments described in Section 6. In the oracle setting, when the state set size is greater than the size of the action set, according to Theorem 1, Theorem 3 and Theorem 4, the time complexities for projected policy gradient ascent, proximal-Q and NPG are

$$O\left(\frac{NC_{\Phi}S^{5/2}A_{\max}\kappa_0^2}{(1-\Gamma)^2\epsilon^2}\right), O(\frac{NC_{\Phi}S^{3/2}A_{max}\kappa_0^2}{(1-\Gamma)^2\epsilon^2}), \text{ and } O(\frac{NC_{\Phi}S^{3/2}A_{\max}\kappa_0^2}{c(1-\Gamma)^2\epsilon^2}),$$

respectively. Notably, projected policy gradient ascent exhibits an additional S dependency, while NPG has an additional dependency on $\frac{1}{c}$. Meanwhile, a small least visited rate (LVR) $1-\Gamma$ has a negative effect on all algorithms. To illustrate these effects, we conducted simulations on Markov potential games with large state spaces, varying LVR (large or small), and different exploration factors (large or small c), as described in the next paragraph.

We randomly generate a cooperative Markov potential game with the following parameters: $S = 100$, $A_1 = 4$, $A_2 = 3$, and $A_3 = 2$. To control the LVR $(1-\Gamma)$, the elements of the transition probability matrix $P \in [0,1]^{SA_1A_2A_3\times S}$ are generated as follows. First, each entry is generated using a uniform distribution Unif$[0,1]$. Then we randomly select half of the states (denoting chosen states by s') and generate the probabilities $P(s'|s,\mathbf{a})$ for every action profile $\mathbf{a}$ according to a uniform distribution Unif$[0,0.01]$ (for small LVR), or Unif$[0,0.1]$ or Unif$[0,1]$ (for large LVR). Subsequently, each row of the matrix is normalized. To reduce the exploration factor c, we generate a reward function with a small reward gap (RG) between the reward of the best action and that of the second best action, thereby increasing exploration difficulty and leading to a smaller c. The reward function is identical for all agents. It is randomly generated and denoted as $R \in [0,1]^{SA_1A_2A_3}$. For scenarios with a small RG, we use a uniform distribution Unif$[0,1]$ for all entries, or set reward $r(s,\mathbf{a}')$ to be 0.001 smaller than $\max_{\mathbf{a}} r(s,\mathbf{a})$. This setup can indeed make $c(t)$ smaller as illustrated in Fig. 1(d). In cases with a large RG, we select an action profile for each state at random and generate rewards according to Unif$[0.4,1]$, with the remaining entities generated according to Unif$[0,0.6]$. We observed that a larger learning rate leads to faster convergence, therefore, we chose the largest learning rate.

A.2 Sample-based Algorithm

In the sample-based setting, we run Algorithm 2 for a manually designed Markov potential game with $S = 2$, $A_1 = A_2 = 2$, an action-independent transition probability matrix $P = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$, and a reward function that is identical for each agent. The rewards for states 1 and 2 are $R_1 = \begin{pmatrix} 1 & 0.2 \\ 0.8 & 0.2 \end{pmatrix}$ and $R_1 = \begin{pmatrix} 0.2 & 1 \\ 0.1 & 0.6 \end{pmatrix}$, where columns indicate actions for agent 1, and rows indicate actions for agent 2. To achieve $\alpha = 0.01$ and $\delta = 0.01$, Theorem 2 suggests a learning rate $\beta \approx \frac{1}{N\kappa_0^2 S^{3/2}A_{\max}} \approx 0.01$ and trajectory length $KN_2+N_1 \approx 10^8$. To reduce the number of samples needed, we choose a trajectory length $KN_2+N_1 = 51000$ with $N_1 = 1000, N_2 = 50, K = 1000$, and reduce the step size every 20 steps, from initially 0.5 to eventually 0.0001, to accommodate inaccurate gradient estimates.

A.3 Robot Navigation Task

We demonstrate the efficacy of average reward MARL by solving a complex robot navigation task for TurtleBot, a two-wheeled mobile robot (Zhou et al., 2024; Rengarajan et al., 2022a,b) in a simulation platform, Gazebo. The navigation task is to guide the robot to any designated target within a 2-meter radius. The state space is defined by a continuous 2-dimensional space representing the distance and relative orientation between the robot and the target. The robot is equipped with two controllers: an angular controller with control effect ranging from -1.5 rad/s to 1.5 rad/s, and a linear velocity controller with control effect ranging from 0 cm/s to 15 cm/s. To model the agents' control strategy, we implemented two independent agents to manage the angular and linear speeds, following an empirical adaptation of Algorithm 5. The reward function is formulated to be proportional to the product of the distance and the scaled orientation between the robot and the target. Hitting the boundary incurs a penalty of -200, while reaching the target yields a reward of 200. Each trajectory terminates upon collision, achievement of the goal, or after 500 simulation steps have elapsed.

All experiments were conducted on a CPU with an 11th Gen Intel(R) Core(TM) i7-11700 @ 2.50GHz.

B PROOFS OF SECTION 2

We first explore a sufficient condition for a Markov game to be a potential game. Previous works have studied the sufficient condition for the discounted reward setting (Leonardos et al., 2021; Zhang et al., 2021a; Narasimha et al., 2022). These conditions encompass some important Markov games.

Proposition 2. *Consider a Markov game that is a static potential game $r_i(s, \mathbf{a})$ at every state $s \in \mathrm{S}$, i.e., there exists a common potential function $\phi(s, \mathbf{a})$ and an utility function $u_i(s, a_{-i})$ for each agent i, such that $r_i(s, \mathbf{a}) = \phi(s, \boldsymbol{a}) + u_i(s, a_{-i})$. If one of the following conditions is satisfied, then the Markov game is an average reward Markov potential game:*

1. *The transition probabilities do not depend on the action $\mathbf{a}$ taken, i.e., $P(s'|s, \mathbf{a}) = P(s'|s)$ for all $\mathbf{a}$.*
2. *For every agent i, there exist a constant c_i such that $\overline{u}_i^{\pi_{-i}}(s) := \sum_{a_{-i}}(\Pi_{j\neq i}\pi_j(a_j|s))u_i(s, a_{-j})$ satisfies $\nabla_{\pi_i(\cdot|s)}\langle \nu^\pi, \overline{u}_i^{\pi_{-i}}\rangle = c_i\mathbf{1}_{A_i}$ for every state s and policy π.*

Proof. Let $\overline{\phi}^\pi(s) := \sum_{\mathbf{a}} \pi(\mathbf{a}|s)\phi(s, \mathbf{a})$. $\rho_i^\pi = \langle \nu^\pi, \overline{\phi}^\pi + \overline{u}_i^{\pi_{-i}}\rangle$. If condition 1 is satisfied, $\nu^\pi \equiv \nu$, then $\langle \nu, \overline{\phi}^\pi\rangle$ is the potential function.

If condition 2 is satisfied, $\rho_i^\pi = \langle \nu^\pi, \overline{\phi}^\pi\rangle + \langle \nu^\pi, \overline{u}_i^{\pi_{-i}}\rangle$. With the interpolation of differential function, for any policies π_i, π_i' and π_{-i}, there exist a constant $a \in [0, 1]$, $\pi_i^* = \pi_i + a(\pi_i' - \pi_i)$, such that $\rho^{\pi_i,\pi_{-i}} - \rho_i^{\pi_i',\pi_{-i}} = \langle \nu^\pi, \overline{\phi}^\pi\rangle - \langle \nu^{\pi_i',\pi_{-i}}, \overline{\phi}^{\pi_i',\pi_{-i}}\rangle + \langle \pi_i - \pi_i', \nabla_{\pi_i(\cdot|s)}\langle \nu^{\pi_i^*,\pi_{-i}}, \overline{u}_i^{\pi_{-i}}\rangle\rangle = \langle \nu^\pi, \overline{\phi}^\pi\rangle - \langle \nu^{\pi_i',\pi_{-i}}, \overline{\phi}^{\pi_i',\pi_{-i}}\rangle + \langle \pi_i - \pi_i', c_i\mathbf{1}_{A_i}\rangle = \langle \nu^\pi, \overline{\phi}^\pi\rangle - \langle \nu^{\pi_i',\pi_{-i}}, \overline{\phi}^{\pi_i',\pi_{-i}}\rangle$. Therefore, $\langle \nu^\pi, \overline{\phi}^\pi\rangle$ is the potential function. □

Remark A fully cooperative game, where all agents have the same reward function $r_i \equiv r$, is an important special case of a Markov potential game. It satisfies Condition 2 in Proposition 2 with $u_i \equiv 0$, while the lake usage problem (Dechert and O'Donnell, 2006) satisfies Condition 1.

Lemma 9 (Restatement of Lemma 1)**.**

$$\begin{aligned}\rho_i^{\hat{\pi}_j,\pi_{-j}} - \rho_i^{\tilde{\pi}_j,\pi_{-j}} &= \mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[\langle \overline{Q}_{j:i}^{\tilde{\pi}_j,\pi_{-j}}(s, \cdot), \hat{\pi}_j(\cdot|s) - \tilde{\pi}_j(\cdot|s)\rangle_{\mathcal{A}_j}] \\ &= \mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}} \sum_{a_j} \hat{\pi}_j(a_j|s)\overline{A}_{j:i}^{\tilde{\pi}_j,\pi_{-j}}(s, a_j).\end{aligned} \tag{5}$$

Proof.

$$\begin{aligned}
&\rho_i^{\hat{\pi}_j,\pi_{-j}} - \rho_i^{\tilde{\pi}_j,\pi_{-j}} \\
=&\rho_i^{\hat{\pi}_j,\pi_{-j}} - \rho_i^{\tilde{\pi}_j,\pi_{-j}} + \mathbb{E}_{s'\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[V_i^{\tilde{\pi}_j,\pi_{-j}}(s')] - \mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[V_i^{\tilde{\pi}_j,\pi_{-j}}(s)] \\
=&\mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}},\mathbf{a}\sim(\hat{\pi}_j,\pi_{-j})}[r_i(s, a) - \rho_i^{\tilde{\pi}_j,\pi_{-j}} + \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_i^{\tilde{\pi}_j,\pi_{-j}}(s')] - V_i^{\tilde{\pi}_j,\pi_{-j}}(s)] \\
=&\mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}},\mathbf{a}\sim(\hat{\pi}_j,\pi_{-j})}[Q_i^{\tilde{\pi}_j,\pi_{-j}}(s, \mathbf{a}) - \langle \overline{Q}_{j;i}^{\tilde{\pi}_j,\pi_{-j}}(\cdot, s), \tilde{\pi}_j(\cdot|s)\rangle_{\mathcal{A}_j}] \\
=&\mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[\mathbb{E}_{a_j\sim\hat{\pi}_i}\overline{Q}_{j;i}^{\tilde{\pi}_j,\pi_{-j}}(s, a_j) - \langle \overline{Q}_{j;i}^{\tilde{\pi}_j,\pi_{-j}}(\cdot, s), \tilde{\pi}_j(\cdot|s)\rangle_{\mathcal{A}_j}] \\
=&\mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[\langle \overline{Q}_{j;i}^{\tilde{\pi}_j,\pi_{-j}}(s, \cdot), \hat{\pi}_j(\cdot|s) - \tilde{\pi}_j(\cdot|s)\rangle_{\mathcal{A}_j}] \\
=&\mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[\sum_{a_j} \overline{Q}_{j;i}^{\tilde{\pi}_j,\pi_{-j}}(s, a_j)\hat{\pi}_j(a_j|s) - V_i^{\tilde{\pi}_j,\pi_{-j}}(s)] \\
=&\mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[\sum_{a_j} \left(\overline{Q}_{j;i}^{\tilde{\pi}_j,\pi_{-j}}(s, a_j) - V_i^{\tilde{\pi}_j,\pi_{-j}}(s)\right) \hat{\pi}_j(a_j|s)] \\
=&\mathbb{E}_{s\sim\nu^{\hat{\pi}_j,\pi_{-j}}}[\sum_{a_j} \hat{\pi}_j(a_j|s)\overline{A}_{j;i}^{\tilde{\pi}_j,\pi_{-j}}(s, a_j)].
\end{aligned}$$

□

Lemma 10 (Restatement of Lemma 2)**.**

$$\frac{\partial \rho_i^\pi}{\partial \pi_j(a_j|s)} = \overline{Q}_{j;i}^\pi(s,a_j)\nu^\pi(s),\ \ \frac{\partial \Phi(\pi)}{\partial \pi_j(a_j|s)} = \overline{Q}_j^\pi(s,a_j)\nu^\pi(s)$$

Proof. Recall that $\overline{Q}_{j;i}^\pi(s,a_j) = \sum\limits_{a_{-j}\in\mathcal{A}_{-j}} \pi_{-j}(a_{-j}|s)Q_i^\pi(s,a_j,a_{-j}) = \sum\limits_{a_{-j}\in\mathcal{A}_{-j}} \pi_{-j}(a_{-j}|s)(r(s,\mathbf{a}) - \rho_i^\pi + \mathbb{E}_{s'\sim P(\cdot|s,a)}V_i^\pi(s'))$. Therefore, $V_i^\pi(s) = \sum_{a_j}\pi_j(a_j|s)\overline{Q}_{j;i}^\pi(s,a_j)$. Differentiating with respect to π_j:

$$\begin{aligned}
\frac{\partial V_i^\pi(s)}{\partial \pi_j} &= \frac{\partial}{\partial \pi_j}\sum_{a_j}\pi_j(a_j|s)\overline{Q}_{j;i}^\pi(s,a_j)\\
&= \sum_{a_j}\frac{\partial \pi_j(a_j|s)}{\partial \pi_j}\overline{Q}_{j;i}^\pi(s,a_j) + \sum_{a_j}\pi_j(a_j|s)\frac{\partial \overline{Q}_{j;i}^\pi(s,a_j)}{\partial \pi_j}\\
&= \sum_{a_j}\frac{\partial \pi_j(a_j|s)}{\partial \pi_j}\overline{Q}_{j;i}^\pi(s,a_j) + \sum_{a_j}\pi_j(a_j|s)\frac{\partial}{\partial \pi_j}(r_i^{\pi_{-j}}(s,a_j) - \rho_i^\pi + \mathbb{E}_{s'\sim P(\cdot|s,a_j,a_{-j}),a_{-j}\sim\pi_{-j}(\cdot|s)}V_i^\pi(s'))\\
&= \sum_{a_j}\frac{\partial \pi_j(a_j|s)}{\partial \pi_j}\overline{Q}_{j;i}^\pi(s,a_j) - \frac{\partial \rho_i^\pi}{\partial \pi_j} + \mathbb{E}_{s'\sim P(\cdot|s,a),\mathbf{a}\sim\pi(\cdot|s)}\frac{\partial V_i^\pi(s')}{\partial \pi_j}.
\end{aligned}$$

Multiplying each side with $\nu^\pi(s)$, taking the summation over s, with the definition of stationary distribution $\mathbb{E}_{s\sim\nu^\pi,s'\sim P(\cdot|s,a),a\sim\pi(\cdot|s)} = \mathbb{E}_{s'\sim\nu^\pi}$, we obtain:

$$\begin{aligned}
\sum_s \nu^\pi(s)\frac{\partial V_i^\pi(s)}{\partial \pi_j} &= \sum_s \nu^\pi(s)\sum_{a_j}\frac{\partial \pi_j(a_j|s)}{\partial \pi_j}\overline{Q}_{j;i}^\pi(s,a_j) - \frac{\partial \rho_i^\pi}{\partial \pi_j} + \mathbb{E}_{s'\sim\nu^\pi}\frac{\partial V_i^\pi(s')}{\partial \pi_j},\\
0 &= \sum_s \nu^\pi(s)\sum_{a_j}\frac{\partial \pi_j(a_j|s)}{\partial \pi_j}\overline{Q}_{j;i}^\pi(s,a_j) - \frac{\partial \rho_i^\pi}{\partial \pi_j},\\
\frac{\partial \rho_i^\pi}{\partial \pi_j} &= \sum_s \nu^\pi(s)\sum_{a_j}\overrightarrow{\mathbf{e}}(s,a_j)\overline{Q}_{j;i}^\pi(s,a_j).
\end{aligned}$$

By the definition of the potential function, it can be noted that $\frac{\partial \Phi(\pi)}{\partial \pi_j(a_j|s)} = \frac{\partial \rho_j^\pi}{\partial \pi_j(a_j|s)} = \overline{Q}_j^\pi(s,a_j)\nu^\pi(s)$. □

Lemma 11 (Kale, 2013, Theorem 3.1, equation (4))**.** *Given a Markov chain with transition probability matrix $P \in \mathbb{R}^{S\times S}$, let ν be its stationary distribution. For any $s, s' \in \mathcal{S}$, we have:*

$$|\frac{P^t(s'|s)}{\nu(s')} - 1| \le \frac{\lambda_2(P)^t}{\sqrt{\nu(s)\nu(s')}}.$$

Lemma 12 (Restatement of Lemma 3)**.** *There exist constants $C_p := \min\{\sqrt{\frac{S}{1-\Gamma}}, \frac{1}{1-\Gamma}\}$ and $\varrho := 1 - \frac{1}{\kappa_0}$ such that for any policy π:*

$$sup_s\|(P^\pi)^t(\cdot|s_0 = s) - \nu^\pi\|_1 \le C_p\varrho^t, \forall\ t > 0.$$

Proof. For any policy $\pi \in \Pi$, $|\frac{(P^\pi)^t(s'|s)}{\nu^\pi(s')} - 1| \le \frac{\lambda_2(P^\pi)^t}{\sqrt{\nu^\pi(s)\nu^\pi(s')}}$ by Lemma 11. Then $\sum_{s'}|(P^\pi)^t(s'|s) - \nu^\pi(s')| \le \lambda_2(P^\pi)^t\sum_{s'}\sqrt{\frac{\nu^\pi(s')}{\nu^\pi(s)}} \le \left(1 - (\frac{1}{1-\lambda_2(P^\pi)})^{-1}\right)^t\sum_{s'}\sqrt{\nu^\pi(s')}\frac{1}{\sqrt{1-\Gamma}} \le (1-\frac{1}{\kappa_0})^t\sqrt{\frac{S}{1-\Gamma}}$.

Similarly, recall $|\frac{(P^\pi)^t(s'|s)}{\nu^\pi(s')} - 1| \le \frac{\lambda_2(P^\pi)^t}{\sqrt{\nu^\pi(s)\nu^\pi(s')}} \le \frac{\lambda_2(P^\pi)^t}{1-\Gamma}$, then $\sum_{s'}|(P^\pi)^t(s'|s) - \nu^\pi(s')| \le \frac{\lambda_2(P^\pi)^t}{1-\Gamma} = (1-\frac{1}{\kappa_0})^2\frac{1}{1-\Gamma}$. □

Proposition 3. *For any policy π and any agent i, the differential Q function has a bounded ℓ_∞ norm:*

$$\|Q_i^\pi\|_\infty \le C_p\kappa_0.$$

Proof. For any $s, \mathbf{a}$ and joint policy π, $|Q_i^\pi(s,\mathbf{a})| = |\mathbb{E}_\pi[\sum_{t=0}^\infty (r_i(s^t,\mathbf{a}^t) - \rho_i^\pi)|s^0 = s, \mathbf{a}^0 = \mathbf{a}]| \leq 1 + |\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})} \sum_{t=1}^\infty \langle (P^\pi)^t(\cdot|s') - \nu^\pi, r^\pi\rangle| \leq 1 + \sum_{t=1}^\infty C_p\varrho^t = 1 + \frac{C_p\varrho}{1-\varrho} = 1 + C_p(\kappa_0 - 1) < C_p\kappa_0$. □

C PROOFS OF SECTION 3

C.1 Auxiliary Lemmas

Lemma 13 (Leonardos et al., 2021, Lemma 4.1)**.** *Let $\pi = (\pi_1, \pi_2, ..., \pi_N)$ be the policy profile for all agents and let $\pi' = \pi + \lambda\nabla_\pi\Phi(\pi)$ be the result from a gradient step on the potential function with learning rate $\lambda > 0$. Then*

$$\mathrm{Proj}_{\Pi_1\times\ldots\times\Pi_N}(\pi') = (\mathrm{Proj}_{\Pi_1}(\pi_1'), ..., \mathrm{Proj}_{\Pi_N}(\pi_N')).$$

Lemma 14 (Agarwal et al., 2021, Proposition B.1)**.** *Let $f(\pi)$ be an l-smooth function. Define the gradient mapping $G(\pi) := \frac{1}{\beta}(\mathrm{Proj}_\Pi(\pi+\beta\nabla_\pi f(\pi)) - \pi)$. Then the update rule for the projected gradient is $\pi^+ = \pi + \beta G(\pi)$. If $\|G(\pi)\|_2 \leq \epsilon$, then*

$$\max_{\pi^+ + \delta\in\Pi, \|\delta\|_2\leq 1} \delta^T \nabla_\pi f(\pi^+) \leq \epsilon(\beta l + 1).$$

There is a typographical mistake in proposition B.1 of Agarwal et al. (2021); the underlying max should be taken on $\pi^+ + \delta \in \Pi$ instead of $\pi + \delta \in \Pi$.

Lemma 15 (Bubeck et al., 2015, Lemma 3.6)**.** *Let f be an l-smooth function over a convex domain C. Let $x \in C$, $x^+ = \mathrm{Proj}_C(x - \beta\nabla f(x))$. Then :*

$$f(x^+) - f(x) \leq (\frac{l}{2} - \frac{1}{\beta})\|x - x^+\|_2^2.$$

Lemma 16. *Assume function $f(\cdot)$ is l-smooth over a convex set C. Let $x^+ := P_C(x - \beta g)$ with $x \in C$. Then*

$$f(x^+) - f(x) \leq (\frac{3l}{4} - \frac{1}{\beta})\|x - x^+\|_2^2 + \frac{1}{l}\|\nabla_x f(x) - g\|_2^2.$$

Proof.

$$\begin{aligned}
f(x^+) - f(x) \leq& \langle\nabla f(x), x^+ - x\rangle + \frac{l}{2}\|x^+ - x\|_2^2 \\
=& \langle g, x^+ - x\rangle + \frac{l}{2}\|x^+ - x\|_2^2 + \langle\nabla f(x) - g, x^+ - x\rangle \\
\overset{(a)}{\leq}& -\frac{1}{\beta}\|x - x^+\|_2^2 + \frac{l}{2}\|x^+ - x\|_2^2 + \frac{1}{2}(\frac{2}{l}\|\nabla_x f(x) - \hat{\nabla}_x f(x)\|_2^2 + \frac{l}{2}\|x^+ - x\|_2^2) \\
=& (\frac{3l}{4} - \frac{1}{\beta})\|x - x^+\|_2^2 + \frac{1}{l}\|\nabla_x f(x) - g\|_2^2.
\end{aligned}$$

We use the property of projection in (a) $\langle x^+ - (x - \beta g), x^+ - x\rangle \leq 0$ and $\langle x, y\rangle \leq \frac{\frac{1}{a}\|x\|_2^2 + a\|y\|_2^2}{2}$ for any positive constant a. □

C.2 Proofs of Lemma 4 and Theorem 1

Lemma 17 (Restatement of Lemma 4)**.** *Denote $A_{\max} := \max_i A_i$, $L := \kappa_0^2 S^{3/2} A_{\max} + \kappa_0\sqrt{S}A_{\max}$, and $L_\Phi := N(\kappa_0^2 S^{3/2} A_{\max} + \kappa_0(SA_{\max} + 2A_{\max}) + A_{\max})$.*

(a) For any i, and $\pi_{-i} \in \Pi_{-i}$, the average value ρ_i^π is $\kappa_0^2 S^{3/2} A_i + \kappa_0\sqrt{S}A_i$-smooth with respect to policy π_i. Moreover, for any i, ρ_i^π is L-smooth with respect to policy π_i, i.e. $\|\nabla_{\pi_i}\rho_i^{\pi_i,\pi_{-i}} - \nabla_{\pi_i}\rho_i^{\pi_i',\pi_{-i}}\|_2 \leq L\|\pi_i - \pi_i'\|_2$ for $\forall$ i, and $\pi_i, \pi_i' \in \Pi_i$.

(b) The potential function $\Phi(\pi)$ is L_Φ-smooth with respect to policy profile π, i.e. $\|\nabla\Phi(\pi) - \nabla\Phi(\pi')\|_2 \leq L_\Phi\|\pi - \pi'\|_2$ for $\forall$ $\pi, \pi' \in \Pi$.

Proof. By the potential function property $\Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}) = \rho_i^{\pi_i,\pi_{-i}} - \rho_i^{\pi_i',\pi_{-i}}$, $\frac{\partial}{\partial \pi_i}\Phi(\pi_i, \pi_{-i}) = \frac{\partial}{\partial \pi_i}\rho_i^{\pi_i,\pi_{-i}}$, and $\nabla_\pi \Phi = (\frac{\partial \rho_1^\pi}{\partial \pi_1}, ..., \frac{\partial \rho_N^\pi}{\partial \pi_N})^T$. To show the smoothness,

$$\begin{aligned}
&\|\nabla_\pi \Phi(\pi) - \nabla_\pi \Phi(\pi')\|_2^2 \\
=&\sum_{i=1}^N \|\nabla_{\pi_i}\Phi(\pi) - \nabla_{\pi_i}\Phi(\pi')\|_2^2 \\
=&\sum_{i=1}^N \|\nabla_{\pi_i}\Phi(\pi) - \nabla_{\pi_i}\Phi(\pi_1', \pi_{2\sim N}) + \nabla_{\pi_i}\Phi(\pi_1', \pi_{2\sim N}) - \nabla_{\pi_i}\Phi(\pi_{1,2}', \pi_{3\sim N}) + \dots \\
&+ \nabla_{\pi_i}\Phi(\pi_{1\sim(N-1)}, \pi_N') - \nabla_{\pi_i}\Phi(\pi')\|_2^2 \\
\leq& \sum_{i=1}^N \sum_{j=1}^N N \|\nabla_{\pi_i}\Phi(\pi_{1\sim j-1}, \pi_{j\sim N}') - \nabla_{\pi_i}\Phi(\pi_{1\sim j}, \pi_{j+1\sim N}')\|_2^2 \\
=&\sum_{i=1}^N \sum_{j=1}^N N \|\frac{\partial \rho_i^{\pi_{1\sim j-1}, \pi_{j\sim N}'}}{\partial \pi_i} - \frac{\partial \rho_i^{\pi_{1\sim j}, \pi_{j+1\sim N}'}}{\partial \pi_i}\|_2^2.
\end{aligned}$$

If we can show $\frac{\partial \rho_i^\pi}{\partial \pi_i}$ is $\frac{L_\Phi}{N}$-lipschitz for any π_j within Π_j, the above inequality implies that $\|\nabla_\pi \Phi(\pi) - \nabla_\pi \Phi(\pi')\|_2^2 \leq \sum_{i=1}^N \sum_{j=1}^N \frac{L_\Phi^2}{N}\|\pi_j - \pi_j'\|_2^2 = L_\Phi^2 \|\pi - \pi'\|_2^2$. In the following proof, we will fix i and denote $\rho = \rho_i$.

Define $\rho(\epsilon) = \rho_i^{\pi_i + \epsilon u_i, \pi_{-i}}$ and $\rho(\tau, \epsilon) = \rho_i^{\pi_i + \epsilon u_i, \pi_j + \tau u_j, \pi_{-ij}}$. We wish to show that $|\frac{d^2\rho}{d\epsilon^2}| \leq L$, $|\frac{d^2 \rho}{d\epsilon d\tau}| \leq \frac{L_\Phi}{N}$ for any u_i, u_j, and any $\epsilon, \tau \in \mathbb{R}$ such that $\pi_i + \epsilon u_i \in \Pi_i$, $\|u_i\|_2 \leq 1$, $\pi_j + \tau u_j \in \Pi_j$, $\|u_j\|_2 \leq 1$ and any $\pi \in \Pi$. Note that $\sum_{a_i} u_i(a_i|s) = 0$, $\sum_{a_j} u_j(a_j|s) = 0$ for any s.

We first note that the unit eigenvector $X(\epsilon)$ of the perturbed square matrix $P^\pi + \epsilon P^{u_i}$ corresponding to the simple eigenvalue 1 is an analytic function Kazdan (1995). Recall that for any ϵ, $X(\epsilon) = c\nu^{\pi_i + \epsilon u_i, \pi_{-i}}$ for some constant $c \neq 0$, thus $\sum_{s\in\mathcal{S}} X_s(\epsilon) \neq 0$. $\nu^{\pi_i+\epsilon u_i, \pi_{-i}} = (\sum_s X_s(\epsilon))^{-1} X(\epsilon)$ is therefore analytic, which guarantees the existence of directional derivatives of both $\nu^{\pi_i + \epsilon u_i, \pi_{-i}}$ and ρ^ϵ. Then by Garrett (2005), $\nu^{\epsilon,\tau} = \nu^{\pi_i + \epsilon u_i, \pi_j + \tau u_j, \pi_{-ij}}$ is jointly analytic, and thus $\frac{d^2 \nu^{\pi_i+\epsilon u_i, \pi_j + \tau u_j, \pi_{-ij}}}{d\epsilon d\tau}$ and $\frac{d^2\rho}{d\epsilon d\tau}$ exist.

We first show $|\frac{d^2\rho}{d\epsilon^2}| \leq L$. Since $\rho(\epsilon) = \sum_{s,\mathbf{a}} \nu^\pi(s)\pi(\mathbf{a}|s) r(s,\mathbf{a}) = \sum_{s,a_i} \nu^{\pi_i + \epsilon u_i, \pi_{-i}}(s)(\pi_i(a_i|s) + \epsilon u_i(a_i|s)) r^{\pi_{-i}}(s, a_i)$, we have

$$\begin{aligned}
\frac{d\rho(\epsilon)}{d\epsilon} &= \sum_{s,a_i} \nu^{\pi_i+\epsilon u_i,\pi_{-i}}(s) u_i(a_i|s) r^{\pi_{-i}}(s,a_i) + \sum_{s,a_i} \frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}(s)}{d\epsilon}(\pi_i(a_i|s) + \epsilon u_i(a_i|s)) r^{\pi_{-i}}(s,a_i), \\
\frac{d^2\rho(\epsilon)}{d\epsilon^2} &= 2\sum_{s,a_i} \frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}(s)}{d\epsilon} u_i(a_i|s) r^{\pi_{-i}}(s,a_i) + \sum_{s,a_i} \frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}(s)}{d\epsilon^2}(\pi_i(a_i|s) + \epsilon u_i(a_i|s)) r^{\pi_{-i}}(s,a_i).
\end{aligned} \tag{6}$$

Since the stationary distribution satisfies $\nu^{\pi_i+\epsilon u_i,\pi_{-i}}\overrightarrow{\mathbf{1}} \equiv 1$, we have $\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\overrightarrow{\mathbf{1}} = 0$ and $\frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2}\overrightarrow{\mathbf{1}} = 0$. In other words, $\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}$ and $\frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2}$ are orthogonal to $\overrightarrow{\mathbf{1}}$. Taking derivatives on both sides of $\nu^{\pi_i+\epsilon u_i,\pi_{-i}} = \nu^{\pi_i+\epsilon u_i,\pi_{-i}} P^{\pi_i+\epsilon u_i,\pi_{-i}}$ gives

$$\begin{aligned}
\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon} &= \nu^{\pi_i+\epsilon u_i,\pi_{-i}} \frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon} + \frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon} P^{\pi_i+\epsilon u_i,\pi_{-i}}, \\
\frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2} &= \nu^{\pi_i+\epsilon u_i,\pi_{-i}} \frac{d^2P^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2} + 2\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon} + \frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2} P^{\pi_i+\epsilon u_i,\pi_{-i}}.
\end{aligned} \tag{7}$$

By Definition 4 and the fact that $\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}$ is orthogonal to the all-1 vector $\mathbf{1}$,

$$\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2 \leq \|\nu^{\pi_i+\epsilon u_i,\pi_{-i}} \frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2 + |\lambda_2(P^{\pi_i+\epsilon u_i,\pi_{-i}})| \|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2,$$

which implies $\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2 \leq \kappa_0 \|\nu^{\pi_i+\epsilon u_i,\pi_{-i}} \frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2$.

Denote $\overline{P}^{\pi_{-i}}(s'|s,a_i) := \sum_{a_{-i}} \pi_{-i}(a_{-i}|s)P(s'|s,a_i,a_{-i})$. We have $\frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}(s'|s) = \sum_{a_i} u_i(a_i|s)\overline{P}^{\pi_{-i}}(s'|s,a_i)$, since $P^{\pi_i+\epsilon u_i,\pi_{-i}}(s'|s) = \sum_{a_i}\sum_{a_{-i}} \pi_{-i}(a_{-i}|s)(\pi_i(a_i|s)+\epsilon u_i(a_i|s))P(s'|s,a)$. For any $\nu \in \Delta(S)$,

$$\begin{aligned}\|\nu \frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2^2 =& \sum_{s'}(\sum_s \nu(s)\sum_{a_i}\overline{P}^{\pi_{-i}}(s'|s,a_i)u(a_i|s))^2 \\ \leq& \sum_{s'}(max_s|\sum_{a_i}\overline{P}^{\pi_{-i}}(s'|s,a_i)u(a_i|s)|)^2 \overset{(a)}{\leq} \sum_{s'}(\frac{1}{2}max_s\|u(\cdot|s)\|_1)^2 \overset{(b)}{\leq} \frac{SA_i}{4}.\end{aligned}$$

(a) follows from $\sum_{a_i} u(a_i|s) = 0$ for any s and the fact that $|\sum_{a_i}\overline{P}(s'|s,a_i)u(a_i|s)|$ is dominated by either the positive part or the negative part of $u(\cdot|s)$. (b) is due to $\|u(\cdot|s)\|_1 \leq \sqrt{A_i}\|u(\cdot|s)\|_2$ and $\|u(\cdot|s)\|_2 \leq 1$. Thus,

$$\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2 \leq \kappa_0 \|\nu^{\pi_i+\epsilon u_i,\pi_{-i}}\frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2 \leq \frac{\kappa_0\sqrt{SA_i}}{2}.$$

Since $\frac{d^2P^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2} = 0$, we have $\|\frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2}\|_2 \leq 2\kappa_0\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2$ by Eq. (7).

$$\begin{aligned}\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\frac{dP^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2 \overset{(a)}{\leq}& \sqrt{\sum_{s'}\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2^2\sum_s(\sum_{a_i}\overline{P}^{\pi_{-i}}(s'|s,a_i)u(a_i|s))^2} \\ \leq& \sqrt{\sum_{s'}\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2^2\sum_s A_i\sum_{a_i}u(a_i|s)^2} \\ \overset{(b)}{\leq}& \sqrt{\sum_{s'}\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2^2 A_i} \leq \frac{\kappa_0 SA_i}{2},\end{aligned} \tag{8}$$

where (a) follows from $\langle x,y\rangle \leq \|x\|_2\|y\|_2$ and (b) is due to $\|u\|_2 \leq 1$. Thus $\|\frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2}\|_2 \leq \kappa_0^2 SA_i$.

Substituting the above inequalities back to $\frac{d^2\rho(\epsilon)}{d\epsilon^2}$ in Eq. (6),

$$|\frac{d^2\rho(\epsilon)}{d\epsilon^2}| \overset{(a)}{\leq} 2\|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2\sqrt{A_i}\|u_i\|_2 + \|\frac{d^2\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon^2}\|_1 \leq \kappa_0\sqrt{S}A_i + \kappa_0^2 S^{3/2}A_i \leq L.$$

(a) is due to $\sum_{s,a_i}\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}(s)}{d\epsilon}u_i(a_i|s)r^{\pi_{-i}}(s,a_i) \leq \|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2\sqrt{\sum_s(\sum_{a_i}u_i(a_i|s)r^{\pi_{-i}}(s,a_i))^2} \leq \|\frac{d\nu^{\pi_i+\epsilon u_i,\pi_{-i}}}{d\epsilon}\|_2\sqrt{\sum_s\|u_i(\cdot|s)\|_1^2}$.

We will next show $|\frac{d^2\rho}{d\epsilon d\tau}| \leq \frac{L_\Phi}{N}$. Use $\nu^{\epsilon,\tau} := \nu^{\pi_i+\epsilon u_i,\pi_j+\tau u_j,\pi_{-ij}}$, $P^{\epsilon,\tau} := P^{\pi_i+\epsilon u_i,\pi_j+\tau u_j,\pi_{-ij}}$. Differentiating twice gives

$$\begin{aligned}\frac{d^2\rho(\epsilon,\tau)}{d\epsilon d\tau} =& \sum_{s,\mathbf{a}}\frac{d^2\nu^{\epsilon,\tau}(s)}{d\epsilon d\tau}(\pi_i(a_i|s)+\epsilon u_i(a_i|s))(\pi_j(a_j|s)+\tau u_j(a_j|s))\pi_{-ij}(a_{-ij}|s)r(s,\mathbf{a}) \\ &+\sum_{s,\mathbf{a}}\frac{d\nu^{\epsilon,\tau}(s)}{d\epsilon}(\pi_i(a_i|s)+\epsilon u_i(a_i|s))u_j(a_j|s)\pi_{-ij}(a_{-ij}|s)r(s,\mathbf{a}) \\ &+\sum_{s,\mathbf{a}}\frac{d\nu^{\epsilon,\tau}(s)}{d\tau}u_i(a_i|s)(\pi_j(a_j|s)+\tau u_j(a_j|s))\pi_{-ij}(a_{-ij}|s)r(s,\mathbf{a}) \\ &+\sum_{s,\mathbf{a}}\nu^{\epsilon,\tau}(s)u_i(a_i|s)u_j(a_j|s)\pi_{-ij}(a_{-ij}|s)r(s,\mathbf{a}),\end{aligned}$$

$$\begin{aligned}|\frac{d^2\rho(\epsilon,\tau)}{d\epsilon d\tau}| \leq& \|\frac{d^2\nu^{\epsilon,\tau}}{d\epsilon d\tau}\|_1 + 2\sqrt{A_j}\|\frac{d\nu^{\epsilon,\tau}}{d\epsilon}\|_2\|u_j\|_2 + 2\sqrt{A_i}\|\frac{d\nu^{\epsilon,\tau}}{d\tau}\|_2\|u_i\|_2 + \|\nu^{\epsilon,\tau}\|_1\max_s\|u_i(\cdot|s)\|_1\max_s\|u_j(\cdot|s)\|_1 \\ \leq& \sqrt{S}\|\frac{d^2\nu^{\epsilon,\tau}}{d\epsilon d\tau}\|_2 + 2\sqrt{A_j}\|\frac{d\nu^{\epsilon,\tau}}{d\epsilon}\|_2 + 2\sqrt{A_i}\|\frac{d\nu^{\epsilon,\tau}}{d\tau}\|_2 + \sqrt{A_iA_j}.\end{aligned}$$

Similarly, as before, $\frac{d^2\nu^{\epsilon,\tau}}{d\epsilon d\tau} = \nu^{\epsilon,\tau}\frac{d^2P^{\epsilon,\tau}}{d\epsilon d\tau} + (\frac{d\nu^{\epsilon,\tau}}{d\tau})(\frac{dP^{\epsilon,\tau}}{d\epsilon}) + (\frac{d\nu^{\epsilon,\tau}}{d\epsilon})(\frac{dP^{\epsilon,\tau}}{d\tau}) + (\frac{d^2\nu^{\epsilon,\tau}}{d\epsilon d\tau})P^{\epsilon,\tau}$, $\|\frac{d\nu^{\epsilon,\tau}}{d\epsilon}\|_2 \leq \frac{\kappa_0\sqrt{SA_i}}{2}$, $\|\frac{d\nu^{\epsilon,\tau}}{d\tau}\|_2 \leq \frac{\kappa_0\sqrt{SA_j}}{2}$, $\frac{d^2P^{\epsilon,\tau}(s'|s)}{d\epsilon d\tau} = \sum_{a_i,a_j}\overline{P}^{\pi_{-ij}}(s'|s,a_i,a_j)u_i(a_i|s)u_j(a_j|s)$. For any $\nu \in \Delta(S)$:

$$\begin{aligned}
\|\nu\frac{d^2P^{\epsilon,\tau}}{d\epsilon d\tau}\|_2^2 &= \sum_{s'}(\sum_{s}\nu(s)\sum_{a_i,a_j}P(s'|s,a_i,a_j)u_i(a_i|s)u_j(a_j|s))^2 \\
&\leq \sum_{s'}(\max_s|\sum_{a_i,a_j}P(s'|s,a_i,a_j)u_i(a_i|s)u_j(a_j|s)|)^2 \\
&\leq \sum_{s'}(\max_s\sum_{a_i,a_j}|u_i(a_i|s)u_j(a_j|s)|)^2 \\
&\leq S(\max_s\|u_i(\cdot|s)\|_1\max_s\|u_j(\cdot|s)\|_1)^2 \leq SA_iA_j.
\end{aligned}$$

For the last inequality, we use $\|x\|_1 \leq \sqrt{n}\|x\|_2$ and $\|u_i\|_2 \leq 1$, $\|u_j\|_2 \leq 1$.

$$\begin{aligned}
\|\frac{d^2\nu^{\epsilon,\tau}}{d\epsilon d\tau}\|_2 &\leq \kappa_0\left(\|\nu^{\epsilon,\tau}\frac{d^2P^{\epsilon,\tau}}{d\epsilon d\tau}\|_2 + \|\frac{d\nu^{\epsilon,\tau}}{d\tau}\frac{dP^{\epsilon,\tau}}{d\epsilon}\|_2 + \|\frac{d\nu^{\epsilon,\tau}}{d\epsilon}\frac{dP^{\epsilon,\tau}}{d\tau}\|_2\right) \\
&\leq \kappa_0\left(\|\nu^{\epsilon,\tau}\frac{d^2P^{\epsilon,\tau}}{d\epsilon d\tau}\|_2 + \frac{\kappa_0SA_i}{2} + \frac{\kappa_0SA_j}{2}\right) \ (\textit{apply Eq. } (8)) \\
&\leq \kappa_0(\sqrt{SA_iA_j} + \frac{\kappa_0S(A_i+A_j)}{2}) \\
|\frac{d^2\rho(\epsilon,\tau)}{d\epsilon d\tau}| &\leq \kappa_0(S\sqrt{A_iA_j} + \frac{\kappa_0S^{3/2}(A_i+A_j)}{2} + 2\sqrt{SA_iA_j}) + \sqrt{A_iA_j} \leq \frac{L_\Phi}{N}.
\end{aligned}$$

Therefore, the L_Φ-smoothness follows. □

Theorem 5 (Restatement of Theorem 1)**.** *Choose learning rate $\beta = \frac{1}{L_\Phi}$. Then Nash-regret* of Algorithm 1 is bounded by*

$$\begin{aligned}
\text{Nash-regret}^*(T) \leq& \frac{32D^2SC_\Phi N(\kappa_0^2S^{5/2}A_{\max} + \kappa_0(SA_{\max} + 2A_{\max}) + A_{\max})}{T} \\
=& O(\frac{D^2\kappa_0^2S^{5/2}A_{\max}NC_\Phi}{T}).
\end{aligned}$$

Proof. First, we bound the Nash-gap for time t:

$$\begin{aligned}
\text{Nash-gap}(t) &\overset{(a)}{=} \rho_i^{\pi_i^{t,*},\pi_{-i}^t} - \rho_i^{\pi^t} \\
&= E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\langle\overline{Q}_i^{\pi_t}(s,\cdot), \pi_i^{t,*}(\cdot|s) - \pi_i^t(\cdot|s)\rangle] \\
&\leq \sum_s \nu^{\pi_i^{t,*},\pi_{-i}^t}(s)\left(\max_{\pi'(\cdot|s)\in\Delta(A_i)}\langle\overline{Q}_i^{\pi^t}(s,\cdot), \pi'(\cdot|s) - \pi_i^t(\cdot|s)\rangle\right) \\
&\overset{(b)}{\leq} \max_s \frac{\nu^{\pi_i^{t,*},\pi_{-i}^t}(s)}{\nu^{\pi^t}(s)}\langle\nabla_{\pi_i}\Phi(\pi^t), \pi' - \pi_i^t\rangle \\
&\overset{(c)}{\leq} 2D\sqrt{S}\max_{\pi_i^t+\delta\in\Pi_i,\|\delta\|_2\leq 1}\delta^T\nabla_{\pi_i}\Phi(\pi^t) \\
&\overset{(d)}{\leq} 2D\sqrt{S}\|\pi_i^t - \pi_i^{t-1}\|_2(L + L_\Phi) \\
&\leq 2(1 + \frac{1}{N})D\sqrt{S}\|\pi^t - \pi^{t-1}\|_2L_\Phi.
\end{aligned}$$

In (a) we use i to denote the agent that achieves the maximum of Nash-gap, i.e., $i \in \arg\max_i \ \max_{p\in\Pi_i}(\rho^{p,\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t})$, and $\pi_i^{t,*} \in \arg\max_{p\in\Pi_i}(\rho^{p,\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t})$. (b) is true since $\max_{\pi'(\cdot|s)\in\Delta(A_i)}\langle\overline{Q}_i^{\pi_t}(s,\cdot), \pi'(\cdot|s) - \pi_t^i(\cdot|s)\rangle \geq 0$

and $\frac{\partial \Phi(\pi^t)}{\pi_i(a_i|s)} = \nu^{\pi^t}(s)\overline{Q}_i^{\pi^t}(s, a_i)$. We also use π' to represent the policy that achieves the maximum. (c) is due to $\|\pi' - \pi_i^t\|_2^2 \leq \sum_s \|\pi'(\cdot|s) - \pi_i^t(\cdot|s)\|_1^2 \leq 4S$. (d) is obtained from Lemma 14 and that $\Phi(\pi_i, \pi_{-i})$ is L-smooth w.r.t. π_i for any $\pi_{-i} \in \Pi_{-i}$.

Since $1 + \frac{1}{N} \leq 2$, we can bound the Nash-regret* as follows

$$\begin{aligned}
\sum_{t=1}^{T} \text{Nash-gap}(t)^2 &\leq 16D^2 L_\Phi^2 S \sum_{t=1}^{T} \|\pi_t - \pi_{t-1}\|_2^2 \\
&\overset{(a)}{\leq} 16D^2 L_\Phi^2 S \sum_{t=1}^{T} \frac{2}{L_\Phi}(\Phi(\pi_{t-1}) - \Phi(\pi_t)) \\
&= 32D^2 L_\Phi S(\Phi(\pi_0) - \Phi(\pi_T)) \\
&\leq 32D^2 L_\Phi C_\Phi S, \\
\text{Nash-regret}^*(T) &\leq \frac{32D^2 L_\Phi C_\Phi S}{T}.
\end{aligned}$$

(a) is due to Lemma 15 by applying $f = -\Phi$. □

C.3 Proof of Lemma 5 and Theorem 2

Lemma 18 (Restatement of Lemma 5)**.** *For any agent i, consider the gradient estimate $\hat{g}_i$ defined in Algorithm 3. Given the (s, a, r)-trajectory of length $KN_2 + N_1$ and the policy $\pi_i \in \Pi_{i,\alpha}$ that generated it, under the assumption that all the other policies π_{-i} are fixed during the generation of the trajectory, the estimated gradient has ℓ_2 error bounded as*

$$\begin{aligned}
&|\mathbb{E}[\hat{\rho}_i] - \rho_i^\pi| \leq \frac{2C_p}{(1-\varrho)N_1}\varrho^{N_1/2}, \\
&\mathbb{E}(\hat{\rho}_i - \rho_i^\pi)^2 \leq (2 + 4C_p\frac{\varrho}{1-\varrho})\frac{1}{N_1}, \\
&\|\mathbb{E}\hat{g}_i - \frac{\partial \rho_i^\pi}{\partial \pi_i}\|_2^2 \leq C_p^2 A_{\max}\left(\frac{N_2^2}{1-\varrho^{2N_2}}\varrho^{2N_1} + \frac{1}{(1-\varrho)^2}\frac{8N_2^2}{N_1^2}\varrho^{N_1} + \frac{2}{(1-\varrho)^2}\varrho^{2N_2}\right), \\
&\mathbb{E}\|\hat{g}_i - \frac{\partial \rho_i^\pi}{\partial \pi_i}\|_2^2 \leq (\frac{1}{\alpha} + 1)\frac{2A_{\max}N_2^2}{K} + \frac{4C_pA_{\max}}{1-\varrho^{N_2}}(\sqrt{\frac{2}{\alpha}} + \sqrt{2})\frac{N_2^2}{K}\varrho^{N_2} + \frac{16A_{\max}C_p^2}{(1-\varrho)^2}\frac{N_2^2}{N_1^2}\varrho^{N_1} + \frac{2A_{\max}C_p^2}{(1-\varrho)^2}\varrho^{2N_2}.
\end{aligned}$$

Proof. Note that $\mathbb{E}[\hat{\rho}_i] = \mathbb{E}[\mathbb{E}[\hat{\rho}_i|s_0]]$, $\mathbb{E}[\hat{\rho}_i] - \rho_i^\pi = \mathbb{E}[\mathbb{E}[\hat{\rho}_i|s_0] - \rho_i^\pi]$,

$$\begin{aligned}
|\mathbb{E}[\hat{\rho}_i|s_0] - \rho_i^\pi| &= |\sum_s(\frac{2}{N_1}\sum_{t=\frac{N_1}{2}}^{N_1-1}(P^\pi)^t(s|s_0) - \nu^\pi(s))\sum_a \pi(a|s)r_i(s,a)| \\
&= |\frac{2}{N_1}\sum_{t=\frac{N_1}{2}}^{N_1-1}(\sum_s((P^\pi)^t(s|s_0) - \nu^\pi(s))\sum_a \pi(a|s)r_i(s,a)| \\
&\leq \frac{2}{N_1}\sum_{t=\frac{N_1}{2}}^{N_1-1}\sum_s|(P^\pi)^t(s|s_0) - \nu^\pi(s)| \\
&\leq \frac{2}{N_1}\sum_{t=\frac{N_1}{2}}^{N_1-1}C_p\varrho^t \\
&\leq \frac{2}{N_1}C_p\varrho^{N_1/2}\frac{1}{1-\varrho}.
\end{aligned}$$

Therefore, $|\mathbb{E}[\hat{\rho}_i] - \rho_i^\pi| \le \frac{2}{N_1} C_p \varrho^{N_1/2} \frac{1}{1-\varrho}$. Use $\mathcal{F}_k$ to denote all the (s, a, r) pairs until episode k.

$$\begin{aligned}
\mathbb{E}(\hat{\rho}_i - \rho_i^\pi)^2 &= \mathbb{E}(\frac{2}{N_1} \sum_{t=\frac{N_1}{2}}^{N_1-1} r_i^t - \rho_i^\pi)^2 \\
&= \frac{4}{N_1^2}(\sum_{t=\frac{N_1}{2}}^{N_1-1} \mathbb{E}(r_i^t - \rho_i^\pi)^2 + 2\sum_{t<\tau} \mathbb{E}[(r_i^t - \rho_i^\pi)(r_i^\tau - \rho_i^\pi)]) \\
&\le \frac{2}{N_1} + \frac{8}{N_1^2} \sum_{t<\tau} \mathbb{E}[(r_i^t - \rho_i^\pi)\mathbb{E}[r_i^\tau - \rho_i^\pi | F_t]] \\
&\le \frac{2}{N_1} + \frac{8}{N_1^2} \sum_{t<\tau} \mathbb{E}[|r_i^t - \rho_i^\pi||\mathbb{E}[r_i^\tau - \rho_i^\pi | F_t]|] \\
&\le \frac{2}{N_1} + \frac{8}{N_1^2} \sum_{t<\tau} \mathbb{E}|\mathbb{E}[r_i^\tau - \rho_i^\pi | F_t]| \\
&\le \frac{2}{N_1} + \frac{8}{N_1^2} \sum_{t<\tau} C_p \varrho^{\tau-t} = \frac{2}{N_1} + \frac{8}{N_1^2} \sum_{\tau=\frac{N_1}{2}}^{N_1-1} \sum_{t=\frac{N_1}{2}}^{\tau-1} C_p \varrho^{\tau-t} \\
&\le \frac{2}{N_1} + C_p \frac{8}{N_1^2} \frac{\varrho}{1-\varrho} \frac{N_1}{2} = (\frac{1}{2} + C_p \frac{\varrho}{1-\varrho}) \frac{4}{N_1}.
\end{aligned}$$

Note that $t_k = N_1 + kN_2$ is the starting time step for the k-th episode, $R(k) = \sum_{\tau=t_k}^{t_k+N_2-1}(r_i^\tau - \hat{\rho}_i)$ is the accumulated bias for the N_2-length interval. Then $\hat{g}_i = \frac{1}{K}\sum_{k=0}^{K-1} R(k)\nabla_{\pi_i} \log \pi_i(a_i^{t_k}|s^{t_k}) \in \mathbb{R}^{S\times A_i}$, where $\nabla_{\pi_i} \log \pi_i(a_i|s) = \frac{1}{\pi_i(a_i|s)} \overrightarrow{\mathbf{e}}_{(s,a_i)} \in \mathbb{R}^{S\times A_i}$ is a unit vector with the only non-zero element corresponding to (s, a_i).

$$\begin{aligned}
&\|\mathbb{E}\hat{g}_i - \frac{\partial \rho_i^\pi}{\partial \pi_i}\|_2^2 \\
=&\|\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[R(k)\nabla_{\pi_i} \log \pi_i(a_i^{t_k}|s^{t_k})] - \mathbb{E}_{s\sim\nu^\pi, a_i\sim\pi_i(\cdot|s)}[\overline{Q}_i^\pi(s, a_i)\nabla_{\pi_i} \log \pi_i(a_i|s)]\|_2^2 \\
\le&\|\frac{1}{K}\sum_{k=0}^{K-1}(\mathbb{E}[R(k)\nabla_{\pi_i} \log \pi_i(a_i^{t_k}|s^{t_k})] - \mathbb{E}_{s^{t_k}\sim\nu^\pi, a_i^{t_k}\sim\pi(\cdot|s)}[R(k)\nabla_{\pi_i} \log \pi_i(a_i^{t_k}|s^{t_k})])\|_2^2 \\
&+\|\frac{1}{K}\sum_{k=0}^{K-1}(\mathbb{E}_{s^{t_k}\sim\nu^\pi, a_i^{t_k}\sim\pi(\cdot|s)}[R(k)\nabla_{\pi_i} \log \pi_i(a_i^{t_k}|s^{t_k})] - \mathbb{E}_{s\sim\nu^\pi, a_i\sim\pi(\cdot|s)}[\overline{Q}_i^\pi(s, a_i)\nabla_{\pi_i} \log \pi_i(a_i|s)])\|_2^2 \\
\le&\|\frac{1}{K}\sum_{k=0}^{K-1}\sum_{\tau=t_k}^{t_k+N_2-1}\sum_{s^{t_k}, a_i^{t_k}, s^\tau, a_i^\tau} \nabla_{\pi_i} \log \pi_i(a_i^{t_k}|s^{t_k})(\mathbb{P}(s^{t_k}, a_i^{t_k}) - \nu^\pi(s^{t_k}, a_i^{t_k}))\mathbb{P}(s^\tau, a_i^\tau|s^{t_k}, a_i^{t_k})(r(s^\tau, a_i^\tau) - \hat{\rho}_i)\|_2^2 \\
&+\frac{1}{K}\sum_{k=0}^{K-1}\|\mathbb{E}_{s^{t_k}\sim\nu^\pi, a_i^{t_k}\sim\pi(\cdot|s)}[(R(k) - \overline{Q}_i^\pi(s^{t_k}, a_i^{t_k}))\nabla_{\pi_i} \log \pi_i(a_i^{t_k}|s^{t_k})]\|_2^2 \\
\le&\|\frac{1}{K}\sum_{k=0}^{K-1}\sum_{\tau=t_k}^{t_k+N_2-1}\sum_{s^{t_k}, a_i^{t_k}} \overrightarrow{\mathbf{e}}_{s^{t_k}, a_i^{t_k}}|(P^\pi)^{t_k}(s^{t_k}|s_0) - \nu^\pi(s^{t_k})|\|_2^2 \\
&+\frac{1}{K}\sum_{k=0}^{K-1}\|\sum_{a_i^{t_k}}\sum_{s^{t_k}} \nu^\pi(s^{t_k})\overrightarrow{\mathbf{e}}_{s^{t_k}, a_i^{t_k}}\mathbb{E}[(R(k) - \overline{Q}_i^\pi(s^{t_k}, a_i^{t_k}))|s^{t_k}, a_i^{t_k}]\|_2^2 \\
\le&\frac{1}{K}\sum_{k=0}^{K-1} N_2^2 \sum_{a_i^{t_k}}\sum_{s^{t_k}} |(P^\pi)^{t_k}(s^{t_k}|s_0) - \nu(s^{t_k})|^2 + \frac{1}{K}\sum_{k=0}^{K-1}\sum_{a_i^{t_k}}\sum_{s^{t_k}} \nu^\pi(s^{t_k})^2|\mathbb{E}[(R(k) - \overline{Q}_i^\pi(s^{t_k}, a_i^{t_k}))|s^{t_k}, a_i^{t_k}]|^2.
\end{aligned}$$

Starting from any $s^0 = s$ and $a_i^0 = a_i$, the difference between $R = \sum_{t=0}^{N_2-1}(r_i^t - \hat{\rho})$ and $\overline{Q}_i^\pi(s, a_i)$ can be bounded as below:

$$\begin{aligned}
&|\mathbb{E}[\sum_{t=0}^{N_2-1}(r_i(s^t,a^t)-\hat{\rho}) - \sum_{t=0}^{\infty}(r_i(s^t,a^t)-\rho_i^\pi)]| \\
=&|\mathbb{E}[\sum_{t=0}^{N_2-1}(\rho_i^\pi - \hat{\rho}_i)] - \mathbb{E}[\sum_{t=N_2}^{\infty}(r_i(s^t,a^t)-\rho_i^\pi)]| \\
\le& N_2\frac{2C_p}{(1-\varrho)N_1}\varrho^{N_1/2} + \sum_{t=N_2}^{\infty}C_p\varrho^t \\
=& N_2\frac{2C_p}{(1-\varrho)N_1}\varrho^{N_1/2} + \frac{C_p}{1-\varrho}\varrho^{N_2}.
\end{aligned}$$

Therefore, we can bound $\|\mathbb{E}\hat{g}_i - \frac{\partial\rho_i^\pi}{\partial\pi_i}\|_2^2$ as:

$$\begin{aligned}
\|\mathbb{E}\hat{g}_i - \frac{\partial\rho_i^\pi}{\partial\pi_i}\|_2^2 \le& \frac{1}{K}\sum_{k=0}^{K-1}N_2^2A_{\max}(C_p\varrho^{t_k})^2 + A_{\max}(N_2\frac{2}{N_1}C_p\varrho^{N_1/2}\frac{1}{1-\varrho} + C_p\varrho^{N_2}\frac{1}{1-\varrho})^2 \\
=& \frac{1}{K}\sum_{k=0}^{K-1}N_2^2A_{\max}(C_p\varrho^{N_1+kN_2})^2 + A_{\max}(N_2\frac{2}{N_1}C_p\varrho^{N_1/2}\frac{1}{1-\varrho} + C_p\varrho^{N_2}\frac{1}{1-\varrho})^2 \\
\le& C_p^2A_{\max}\left(\frac{N_2^2}{1-\varrho^{2N_2}}\varrho^{2N_1} + \frac{1}{(1-\varrho)^2}\frac{8N_2^2}{N_1^2}\varrho^{N_1} + \frac{2}{(1-\varrho)^2}\varrho^{2N_2}\right).
\end{aligned}$$

Denote $\overline{R}(k) := \mathbb{E}_{s^{t_k}\sim\nu^\pi, a^{t_k}\sim\pi(\cdot|s)}[R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})]$. As above, we can bound the variance as:

$$\begin{aligned}
\mathbb{E}\|\hat{g}_i - \frac{\partial\rho_i^\pi}{\partial\pi_i}\|_2^2 \le& 2\mathbb{E}\|\frac{1}{K}\sum_{k=0}^{K-1}R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k}) - \mathbb{E}_{s^{t_k}\sim\nu^\pi, a^{t_k}\sim\pi(\cdot|s)}[R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})]\|_2^2 \\
&+ 2\|\mathbb{E}_{s^{t_k}\sim\nu^\pi, a^{t_k}\sim\pi(\cdot|s)}[R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})] - \mathbb{E}_{s\sim\nu^\pi, a\sim\pi(\cdot|s)}[Q_i^\pi(s,a)\nabla_{\pi_i}\log\pi_i(a_i|s)]\|_2^2 \\
\le& \frac{2}{K^2}\sum_{k=0}^{K-1}\mathbb{E}\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k}) - \overline{R}(k)\|_2^2 \\
&+ \frac{4}{K^2}\sum_{k<\tau}\mathbb{E}\langle R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k}) - \overline{R}(k), R(\tau)\nabla_{\pi_i}\log\pi_i(a_i^{t_\tau}|s^{t_\tau}) - \overline{R}(\tau)\rangle \\
&+ 2A_{\max}(N_2\frac{2}{N_1}C_p\varrho^{N_1/2}\frac{1}{1-\varrho} + C_p\varrho^{N_2}\frac{1}{1-\varrho})^2.
\end{aligned}$$

We bound the first two terms separately.

$$\begin{aligned}
&\mathbb{E}\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k}) - \overline{R}(k)\|_2^2 \\
\le& 2\mathbb{E}\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})\|_2^2 + 2\|\overline{R}(k)\|_2^2 \\
=& 2\sum_{s^{t_k}, a_i^{t_k}}\mathbb{P}(s^{t_k})\pi_i(a_i^{t_k}|s^{t_k})\mathbb{E}[\|\frac{1}{\pi_i(a^{t_k}|s^{t_k})}R(k)\overrightarrow{\mathbf{e}}_{s^{t_k},a_i^{t_k}}\|_2^2|s^{t_k}, a_i^{t_k}] \\
&+ 2\|\sum_{s^{t_k}}\nu^\pi(s^{t_k})\sum_{a_i^{t_k}}\mathbb{E}[R(k)\overrightarrow{\mathbf{e}}_{s^{t_k},a_i^{t_k}}|s^{t_k}, a_i^{t_k}]\|_2^2 \\
=& 2\sum_{s^{t_k}, a_i^{t_k}}\mathbb{P}(s^{t_k})\frac{1}{\pi_i(a_i^{t_k}|s^{t_k})}\mathbb{E}[R(k)^2|s^{t_k}, a_i^{t_k}] + 2\sum_{s^{t_k}}\nu^\pi(s^{t_k})^2\sum_{a_i^{t_k}}\mathbb{E}[R(k)|s^{t_k}, a_i^{t_k}]^2 \\
\le& \frac{2A_{\max}}{\alpha}N_2^2 + 2A_{\max}N_2^2, \ \forall k,
\end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{E}[\langle R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})-\overline{R}(k), R(\tau)\nabla_{\pi_i}\log\pi_i(a_i^{t_\tau}|s^{t_\tau})-\overline{R}(\tau)\rangle]\\
=&\mathbb{E}\langle R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})-\overline{R}(k), \mathbb{E}[R(\tau)\nabla_{\pi_i}\log\pi_i(a_i^{t_\tau}|s^{t_\tau})-\overline{R}(\tau)|F_k]\rangle\\
\leq&\mathbb{E}\Big[\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})-\overline{R}(k)\|_2\|\mathbb{E}[R(\tau)\nabla_{\pi_i}\log\pi_i(a_i^{t_\tau}|s^{t_\tau})-\overline{R}(\tau)\rangle|F_k]\|_2\Big]\\
\leq&\mathbb{E}\Big[\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})-\overline{R}(k)\|_2\|\sum_{s^{t_\tau}}((P^\pi)^{t_\tau-t_k}(s^{t_\tau}|s^{t_k})-\nu^\pi(s^{t_\tau}))\sum_{a_i^{t_\tau}}\mathbb{E}[R(\tau)\overrightarrow{\mathbf{e}}_{s^{t_\tau},a_i^{t_\tau}}|s^{t_\tau},a_i^{t_\tau}]\|_2\Big]\\
\overset{(a)}{\leq}&\mathbb{E}\Big[\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})-\overline{R}(k)\|_2\sum_{s}|(P^\pi)^{t_\tau-t_k}(s|s^{t_k})-\nu^\pi(s)|\sqrt{N_2^2A_{\max}}\Big]\\
\leq&\mathbb{E}\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})-\overline{R}(k)\|_2C_p\varrho^{(\tau-t)N_2}N_2\sqrt{A_{\max}}\\
\overset{(b)}{\leq}&\sqrt{\mathbb{E}\|R(k)\nabla_{\pi_i}\log\pi_i(a_i^{t_k}|s^{t_k})-\overline{R}(t)\|_2^2}C_p\varrho^{(\tau-t)N_2}N_2\sqrt{A_{\max}}\\
\overset{(c)}{\leq}&\left(\sqrt{\frac{2A_{\max}}{\alpha}N_2^2}+\sqrt{2A_{\max}N_2^2}\right)C_p\varrho^{(\tau-t)N_2}N_2\sqrt{A_{\max}}\\
=&\left(\sqrt{\frac{2}{\alpha}}+\sqrt{2}\right)C_p\varrho^{(\tau-t)N_2}A_{\max}N_2^2.
\end{aligned}
$$

In (a) we used the fact $\|x\|_2 \leq \|x\|_1$ in (a). (b) is due to Jensen's inequality, (c) is due to $\sqrt{a^2+b^2} \leq a+b$ for any positive a, b. Therefore, we can bound the variance as

$$
\begin{aligned}
\mathbb{E}\|\hat{g}_i-\frac{\partial\rho_i^\pi}{\partial\pi_i}\|_2^2 \leq 2(\frac{1}{\alpha}+1)A_{\max}N_2^2\frac{1}{K}+\frac{4C_p}{1-\varrho^{N_2}}(\sqrt{\frac{2}{\alpha}}+\sqrt{2})A_{\max}N_2^2\varrho^{N_2}\frac{1}{K}\\
+2A_{\max}(N_2\frac{2}{N_1}C_p\varrho^{N_1/2}\frac{1}{1-\varrho}+C_p\varrho^{N_2}\frac{1}{1-\varrho})^2.
\end{aligned}
$$

□

Theorem 6 (Restatement of Theorem 2)**.** *If all agents independently and synchronously run Algorithm 2 with learning rate $\beta \leq \frac{1}{L_\Phi}$, then Nash regret is bounded by:*

$$
\textit{Nash-regret}^*(T) \leq 48D^2S(L+\frac{1}{\beta})^2\beta\frac{N}{T}+12\kappa^2\alpha^2+\left(\frac{48D^2S(L+\frac{1}{\beta})^2\beta}{L_\Phi}+6\kappa^2A_{\max}\beta^2+24D^2S(2L\beta+1)^2\right)\delta.
$$

Proof. Let $\tilde{\pi}_i^{t+1} = P_{\Pi_{i,\alpha}}(\pi_i^t+\beta\nabla_{\pi_i}\rho_i^{\pi^t})$ be the projection after a exact gradient step. Since $\pi_i^{t+1} = P_{\Pi_{i,\alpha}}(\pi_i^t+\beta\hat{g}_i^t)$, we have $\|\tilde{\pi}_i^{t+1} - \pi_i^{t+1}\|_2 \leq \beta\|\nabla_{\pi_i}\rho_i^{\pi^t} - \hat{g}_i^t\|_2$ by the non-expansion of projection. Then

$$\begin{aligned}
&\text{Nash-gap}(t)\\
\overset{(a)}{=}&\rho_i^{\pi_i^{t,*},\pi_{-i}^t} - \rho_i^{\pi^t}\\
=&E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\langle\overline{Q}_i^{\pi_t}(s,\cdot),\pi_i^{t,*}(\cdot|s) - \pi_i^t(\cdot|s)\rangle]\\
=&E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\langle\overline{Q}_i^{\pi_t}(s,\cdot),\pi_i^{t,*}(\cdot|s) - \tilde{\pi}_i^t(\cdot|s)\rangle] + E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\langle\overline{Q}_i^{\pi_t}(s,\cdot),\tilde{\pi}_i^t(\cdot|s) - \pi_i^t(\cdot|s)\rangle]\\
\leq&E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\langle\overline{Q}_i^{\pi_t}(s,\cdot),\pi_i^{t,*}(\cdot|s) - \tilde{\pi}_i^t(\cdot|s)\rangle] + \kappa\max_s\|\tilde{\pi}_i^t(\cdot|s) - \pi_i^t(\cdot|s)\|_1\\
\leq&E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\langle\overline{Q}_i^{\pi_t}(s,\cdot),(1-\alpha)\pi_i^{t,*}(\cdot|s) + \alpha u_i(\cdot|s) - \tilde{\pi}_i^t(\cdot|s)\rangle] + \alpha E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\langle\overline{Q}_i^{\pi_t}(s,\cdot),\pi_i^{t,*}(\cdot|s) - u_i(\cdot|s)\rangle]\\
&+\kappa\sqrt{A_{\max}}\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
\overset{(b)}{\leq}&E_{s\sim\nu^{\pi_i^{t,*},\pi_{-i}^t}}[\max_{\pi_i'\in\Pi_i}\langle\overline{Q}_i^{\pi_t}(s,\cdot),(1-\alpha)\pi_i'(\cdot|s) + \alpha u_i(\cdot|s) - \tilde{\pi}_i^t(\cdot|s)\rangle] + 2\alpha\|\overline{Q}_i^{\pi_t}\|_\infty + \kappa\sqrt{A_{\max}}\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
\overset{(c)}{\leq}&\max_s\frac{\nu^{\pi_i^{t,*},\pi_t^{-i}}(s)}{\nu^{\pi_t}(s)}\langle\nabla_{\pi_i}\Phi(\pi^t),(1-\alpha)\pi_i' + \alpha u_i - \tilde{\pi}_i^t\rangle + 2\alpha\kappa + \kappa\sqrt{A_{\max}}\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
\leq&2D\sqrt{S}\max_{\tilde{\pi}_i^t+\delta\in\Pi_{i,\alpha},\|\delta\|_2\leq 1}\delta^T\nabla_{\pi_i}\Phi(\pi^t) + 2\alpha\kappa + \kappa\sqrt{A_{\max}}\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
\leq&2D\sqrt{S}\max_{\tilde{\pi}_i^t+\delta\in\Pi_{\alpha},\|\delta\|_2\leq 1}\delta^T\nabla_{\pi_i}\Phi(\tilde{\pi}_i^t,\pi_{-i}^t) + 2D\sqrt{S}\max_{\tilde{\pi}_i^t+\delta\in\Pi_{i,\alpha},\|\delta\|_2\leq 1}\delta^T(\nabla_{\pi_i}\Phi(\pi^t) - \nabla_{\pi_i}\Phi(\tilde{\pi}_i^t,\pi_{-i}^t))\\
&+2\alpha\kappa + \kappa\sqrt{A_{\max}}\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
\overset{(d)}{\leq}&2D\sqrt{S}\|\tilde{\pi}_i^t - \pi_i^{t-1}\|_2(L + \frac{1}{\beta}) + 2D\sqrt{S}L\|\pi_i^t - \tilde{\pi}_i^t\|_2 + 2\alpha\kappa + \kappa\sqrt{A_{\max}}\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
=&2D\sqrt{S}(L + \frac{1}{\beta})\|\tilde{\pi}_i^t - \pi_i^{t-1}\|_2 + 2\alpha\kappa + (\kappa\sqrt{A_{\max}} + 2D\sqrt{S}L)\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
\leq&2D\sqrt{S}(L + \frac{1}{\beta})(\|\tilde{\pi}_i^t - \pi_i^t\|_2 + \|\pi_i^t - \pi_i^{t-1}\|_2) + 2\alpha\kappa + (\kappa\sqrt{A_{\max}} + 2D\sqrt{S}L)\|\tilde{\pi}_i^t - \pi_i^t\|_2\\
\leq&2D\sqrt{S}(L + \frac{1}{\beta})\|\pi_i^t - \pi_i^{t-1}\|_2 + 2\alpha\kappa + (\kappa\sqrt{A_{\max}} + 2D\sqrt{S}(2L + \frac{1}{\beta}))\beta\|\nabla_{\pi_i}\rho_i^{\pi^{t-1}} - \hat{g}_i^{t-1}\|_2.
\end{aligned}$$

In (a) we have used i to denote the agent that achieves the maximum of Nash gap, i.e., $i \in \arg\max_i\ \max_{p\in\Pi_i}(\rho^{p,\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t})$, and $\pi_i^{t,*} \in \arg\max_{p\in\Pi_i}(\rho^{p,\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t})$. We use $\langle x, y\rangle \leq \|x\|_\infty\|y\|_1$ and $\|\pi_i(\cdot|s)\|_1 = 1$ in (b). (c) is true since $\max_{\pi'(\cdot|s)\in\Delta(A_i)}\langle\overline{Q}_i^{\pi_t}(s,\cdot),\pi'(\cdot|s) - \pi_t^i(\cdot|s)\rangle \geq 0$. We also use π_i' to represent the policy that achieves the maximum. (d) comes from Lemma 14 (note that $\Pi_{i,\alpha}$ is convex) and that $\Phi(\pi)$ is L-smooth w.r.t. π_i for any $\pi_{-i} \in \Pi_{-i}$.

Applying Lemma 16 with $f = -\Phi$, $E[\Phi(\pi^{t+1}) - \Phi(\pi^t)] \geq (\frac{1}{\beta} - \frac{3L_\Phi}{4})E\|\pi^t - \pi^{t+1}\|_2^2 - \frac{1}{NL}\sum_i\delta$. Choosing a learning rate $\beta \leq \frac{1}{L_\Phi}$, $4\beta E[\Phi(\pi^{t+1}) - \Phi(\pi^t)] + \frac{4\beta}{L_\Phi}\sum_i\delta \geq E\|\pi^t - \pi^{t+1}\|_2^2$, we arrive at

$$\mathbb{E}[\text{Nash-gap}(t)^2] \leq 12D^2S(L+\frac{1}{\beta})^2\mathbb{E}\|\pi^t-\pi^{t-1}\|_2^2+12\alpha^2\kappa^2+3(\kappa\sqrt{A_{\max}}+2D\sqrt{S}(2L+\frac{1}{\beta}))^2\beta^2\delta,$$

$$\begin{aligned}
\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}\text{Nash-gap}(t)^2] \leq & 12D^2S(L+\frac{1}{\beta})^2\left(\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}4\beta(\Phi(\pi^{t+1})-\Phi(\pi^t)]+\frac{4\beta}{L_\Phi T}\sum_{i,t}\delta\right)+12\alpha^2\kappa^2\\
&+(6\kappa^2A_{\max}\beta^2+24D^2S(2L\beta+1)^2)\delta\\
\leq & 48D^2S(L+\frac{1}{\beta})^2\beta\mathbb{E}[\Phi(\pi^T)-\Phi(\pi^0)]\frac{1}{T}+12\alpha^2\kappa^2\\
&+\left(\frac{48D^2S(L+\frac{1}{\beta})^2\beta C_\Phi}{L_\Phi}+6\kappa^2A_{\max}\beta^2+24D^2S(2L\beta+1)^2\right)\delta\\
\leq & 48D^2S(L+\frac{1}{\beta})^2\beta\frac{C_\Phi}{T}+12\alpha^2\kappa^2\\
&+\left(\frac{48D^2S(L+\frac{1}{\beta})^2\beta C_\Phi}{L_\Phi}+6\kappa^2A_{\max}\beta^2+24D^2S(2L\beta+1)^2\right)\delta.
\end{aligned}$$

□

C.4 Notes of Previous Work

The proof of Theorem 4.7 in Leonardos et al. (2021), where they established a sample complexity result, appears to have two mistakes. In their latest version, the first equation on p. 16 was derived based on:

$$-\Phi_\mu(y^{t+1})+\frac{1}{\lambda}\|\pi^t-y^{t+1}\|_2^2\leq\phi_\lambda(\pi^t), \tag{9}$$

where $y^{t+1}=\text{Proj}_\Pi(\pi^t+\eta\nabla_\pi\Phi_\mu(\pi^t))$, $\phi_\lambda(\pi^t)=\min_{y\in\Pi}(-\Phi_\mu(y)+\frac{1}{\lambda}\|y-\pi^t\|_2^2)$. Note that Eq. (9) is equivalent to y^{t+1} reaching the minimizer of $\phi_\lambda(\pi^t)$. However, it should be noted that the minimizer of Moreau envelope $\arg\min_{y\in\Pi}(-\Phi_\mu(y)+\frac{1}{\lambda}\|y-\pi^t\|_2^2)$ may not be the one step updated policy of the projected gradient ascent algorithm. A counterexample can be found in Beck, 2017, chapter 6.2.3. In Davis and Drusvyatskiy, 2018, Theorem 2.1, which the proof in (Leonardos et al., 2021) mostly follow, y^{t+1} was defined as the minimizer of the Moreau envelope $\arg\min_{y\in\Pi}(-\Phi_\mu(y)+\frac{1}{\lambda}\|y-\pi^t\|_2^2)$, and it was only established that $\|\nabla_\pi\phi_\lambda(\pi^t)\|_2$ is bounded, not that $\|\nabla_\pi\Phi_\mu(\pi^t)\|_2$ is bounded. However, the latter appears critical in establishing the bounded Nash gap in Leonardos et al., 2021, Theorem 4.7.

Furthermore, Eq. (14) in Leonardos et al. (2021) shows that there exists a time t^* s.t. $\|y^{t^*+1}-\pi^{t^*}\|_2$ can be bounded. Subsequently, the authors used their Lemma D.3 and Lemma 4.2 to show a bounded Nash gap. However, Lemma D.3 only guarantees that y^{t^*+1}, instead of π^{t^*}, is a ϵ-stationary point, accordingly an ϵ-Nash equilibrium. It should further be noted that y is not tractable with a finite number of samples.

D PROOFS OF SECTION 4

Lemma 19 (Restatement of Lemma 6). *For any reward function $r\in[0,1]$, any policies $\pi,\pi'\in\Pi$, the following bounds hold, where we replace V_i with V_r to indicate a general reward function:*

$$\begin{aligned}
&|\nu^\pi(s)-\nu^{\pi'}(s)|\leq\kappa\|\pi-\pi'\|_{1,\infty},\forall s\in\mathcal{S}\\
&|\rho_r^\pi-\rho_r^{\pi'}|\leq\kappa\|\pi-\pi'\|_{1,\infty},\\
&\|V_r^\pi-V_r^{\pi'}\|_\infty\leq\kappa_1(2+S(\kappa+\kappa_1)+S\kappa\kappa_1)\|\pi-\pi'\|_{1,\infty},\\
&\|Q_r^\pi-Q_r^{\pi'}\|_\infty\leq(\kappa+2\kappa_1+S\kappa_1(\kappa+\kappa_1)+S\kappa\kappa_1^2)\times\|\pi-\pi'\|_{1,\infty}.
\end{aligned}$$

Proof. First, we provide the sensitivity analysis of the policy-induced reward and the state transition probability matrix:

$$\begin{aligned}|r^\pi(s) - r^{\pi'}(s)| &= |\sum_a (\pi(a|s) - \pi'(a|s))r(s,a)| \le \|\pi(\cdot|s) - \pi'(\cdot|s)\|_1,\ \forall s \in S,\\ |P^\pi(s'|s) - P^{\pi'}(s'|s)| &= |\sum_a (\pi(a|s) - \pi'(a|s))P(s'|s,a)| \le \|\pi(\cdot|s) - \pi'(\cdot|s)\|_1,\ \forall s, s' \in S.\end{aligned}$$

From Definition 5 and Lemma 1, taking $\mathbb{I}(\cdot = s)$ as the reward function, it follows that $\|\nu^\pi - \nu^{\pi'}\|_\infty \le \kappa\|\pi - \pi'\|_{1,\infty}$. Similarly by the performance difference lemma, $|\rho^\pi - \rho^{\pi'}| \le \kappa\|\pi - \pi'\|_{1,\infty}$.

$$\begin{aligned}H^\pi - H^{\pi'} =& (I - P^\pi + P^{\pi,\infty})^{-1}(I - P^{\pi,\infty}) - (I - P^{\pi'} + P^{\pi',\infty})^{-1}(I - P^{\pi',\infty})\\ =& \Big((I - P^\pi + P^{\pi,\infty})^{-1} - (I - P^{\pi'} + P^{\pi',\infty})^{-1}\Big)(I - P^{\pi,\infty}) + (I - P^{\pi'} + P^{\pi',\infty})^{-1}(P^{\pi',\infty} - P^{\pi,\infty})\\ =& \Big((I - P^\pi + P^{\pi,\infty})^{-1}(P^\pi - P^{\pi,\infty} - P^{\pi'} + P^{\pi',\infty})(I - P^{\pi'} + P^{\pi',\infty})^{-1}\Big)(I - P^{\pi,\infty})\\ &+ (I - P^{\pi'} + P^{\pi',\infty})^{-1}(P^{\pi',\infty} - P^{\pi,\infty}),\\ \|V^\pi - V^{\pi'}\|_\infty =& \|H^\pi r^\pi - H^{\pi'} r^{\pi'}\|_\infty\\ =& \|H^\pi(r^\pi - r^{\pi'}) + (H^\pi - H^{\pi'})r^{\pi'}\|_\infty\\ \le& \|H^\pi(r^\pi - r^{\pi'})\|_\infty + \|(I - P^{\pi'} + P^{\pi',\infty})^{-1}(P^{\pi',\infty} - P^{\pi,\infty})r^{\pi'}\|_\infty\\ &+ \|\Big((I - P^\pi + P^{\pi,\infty})^{-1}(P^\pi - P^{\pi,\infty} - P^{\pi'} + P^{\pi',\infty})(I - P^{\pi'} + P^{\pi',\infty})^{-1}\Big)(I - P^{\pi,\infty})r^{\pi'}\|_\infty\\ \le& \|H^\pi\|_\infty\|r^\pi - r^{\pi'}\|_\infty + \|(I - P^{\pi'} + P^{\pi',\infty})^{-1}\|_\infty\|P^{\pi',\infty} - P^{\pi,\infty}\|_\infty\\ &+ \|(I - P^\pi + P^{\pi,\infty})^{-1}(P^\pi - P^{\pi,\infty} - P^{\pi'} + P^{\pi',\infty})(I - P^{\pi'} + P^{\pi',\infty})^{-1}\|_\infty\\ \overset{(a)}{\le}& 2\kappa_1\|r^\pi - r^{\pi'}\|_\infty + \kappa_1\|\nu^\pi - \nu^{\pi'}\|_1 + \kappa_1^2\|P^\pi - P^{\pi'}\|_\infty + \kappa_1^2\|\nu^\pi - \nu^{\pi'}\|_1\\ \le& \kappa_1(2 + S\kappa + \kappa_1 S + \kappa_1 S\kappa)\|\pi - \pi'\|_{1,\infty}.\end{aligned}$$

Above (a) is due to $\|P^{\pi',\infty} - P^{\pi,\infty}\|_\infty \le \|\nu^\pi - \nu^{\pi'}\|_1$, $\|P^\pi - P^{\pi'}\|_\infty \le max_s \sum_{s'} |P^\pi(s'|s) - P^{\pi'}(s'|s)|$ and $\|(I - P^{\pi,\infty})r^{\pi'}\|_\infty \le 1$. Now

$$\begin{aligned}&|Q^\pi(s,a) - Q^{\pi'}(s,a)| = |\rho^\pi - \rho^{\pi'} + \langle P(\cdot|s,a), V^\pi - V^{\pi'}\rangle|,\\ &\|Q^\pi - Q^{\pi'}\|_\infty \le |\rho^\pi - \rho^{\pi'}| + \|V^\pi - V^{\pi'}\|_\infty \le (\kappa + 2\kappa_1 + S\kappa_1(\kappa + \kappa_1) + S\kappa\kappa_1^2)\|\pi - \pi'\|_{1,\infty}.\end{aligned}$$

□

Remark In the case of large state set $\mathcal{S}$ but with finite cardinality $S < \infty$, the general parameterized policy classes (Xu et al., 2020, Assumption 1) is commonly utilized. With the (Xu et al., 2020, Assumption 1(3)), $\|Q^{\pi_w} - Q^{\pi_{w'}}\|_\infty \le \kappa_Q\|\pi - \pi'\|_{1,\infty} \le \kappa_Q C_\pi\|w - w'\|_2$, thus their derivatives in Appendix B still hold. The smoothness of average reward with respect to the general parameterized policy classes can be shown.

Lemma 20 (policy improvement (a)). *Let π^t to be the policy at time t of Algorithm 4,*

$$\Phi(\pi^{t+1}) - \Phi(\pi^t) \ge \frac{1}{\beta}\left(1 - \beta\frac{(N-1)(\kappa_Q + S\kappa^2)A_{\max}}{2(1-\Gamma)}\right)\sum_{i=1}^N \sum_s \nu^{\pi_i^{t+1}, \pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2.$$

Proof. As in Ding et al. (2022), we can derive a bound of $\Phi^{t+1} - \Phi^t$ with the decomposition:

$$\begin{aligned}\Phi^{t+1} - \Phi^t =& \underbrace{\sum_{i=1}^N (\Phi(\pi_i^{t+1}, \pi_{-i}^t) - \Phi(\pi_i^t, \pi_{-i}^t))}_{\text{Diff}_1}\\ &+ \underbrace{\sum_{i=1}^N \sum_{j=i+1}^N (\Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^{t+1}, \pi_j^{t+1}) - \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^t, \pi_j^{t+1}) - \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^{t+1}, \pi_j^t) + \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^t, \pi_j^t)),}_{\text{Diff}_2}\end{aligned} \tag{10}$$

(where $\pi^{t,t+1}_{-(i,j)} := (\pi^t_{1\sim i-1, i+1\sim j-1}, \pi^{t+1}_{j+1\sim N})$).
First we bound each term in Diff_2:

$$\begin{aligned}
&\Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t+1}) - \Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t+1}) - \Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t}) + \Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t})\\
=&\rho_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t+1}} - \rho_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t+1}} - \rho_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t}} + \rho_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t}}\\
=&\mathbb{E}_{s\sim \nu^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t+1}}}[\langle \overline{Q}_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t+1}}(\cdot, s), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}]\\
&- \mathbb{E}_{s\sim \nu^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t}}}[\langle \overline{Q}_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t}}(\cdot, s), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}]\\
=&\mathbb{E}_{s\sim \nu^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t+1}}}[\langle \overline{Q}_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t+1}}(\cdot, s) - \overline{Q}_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t}}(\cdot, s), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}]\\
&+ \sum_s (\nu^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t+1}}(s) - \nu^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t}}(s)) \langle \overline{Q}_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t}}(\cdot, s), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}\\
\geq &- \max_s \| \overline{Q}_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t+1}}(\cdot, s) - \overline{Q}_i^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t}}(\cdot, s)\|_\infty \max_s \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_1\\
&- \kappa S \|\nu^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t}} - \nu^{\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t+1}}\|_\infty \max_s \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_1\\
\geq &- \kappa_Q \|\pi_j^{t+1} - \pi_j^t\|_{1,\infty} \|\pi_i^{t+1} - \pi_i^t\|_{1,\infty} - \kappa^2 S \|\pi_j^{t+1} - \pi_j^t\|_{1,\infty} \|\pi_i^{t+1} - \pi_i^t\|_{1,\infty}\\
= &- (\kappa_Q + S\kappa^2) \|\pi_j^{t+1} - \pi_j^t\|_{1,\infty} \|\pi_i^{t+1} - \pi_i^t\|_{1,\infty}\\
\overset{(a)}{\geq} &- \frac{\kappa_Q + S\kappa^2}{2} (\|\pi_j^{t+1} - \pi_j^t\|^2_{1,\infty} + \|\pi_i^{t+1} - \pi_i^t\|^2_{1,\infty}),
\end{aligned}$$

where (a) is due to $ab \leq \frac{a^2+b^2}{2}$ for any a, b. Therefore, we have

$$\begin{aligned}
\mathrm{Diff}_2 \geq& -\frac{(N-1)(\kappa_Q + S\kappa^2)}{2} \sum_{i=1}^N \|\pi_i^{t+1} - \pi_i^t\|^2_{1,\infty}\\
\geq& -\frac{(N-1)(\kappa_Q + S\kappa^2) A_{\max}}{2(1-\Gamma)} \sum_{i=1}^N \mathbb{E}_{s\sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2.
\end{aligned}$$

To bound Diff_1, we write

$$\begin{aligned}
\mathrm{Diff}_1 =& \sum_{i=1}^N \rho_i^{\pi_i^{t+1}, \pi_{-i}^t} - \rho_i^{\pi_i^{t}, \pi_{-i}^t} = \sum_{i=1}^N \mathbb{E}_{s\sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s, \cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}]\\
\geq& \frac{1}{\beta} \sum_{i=1}^N \mathbb{E}_{s\sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2.
\end{aligned} \tag{11}$$

The last inequality comes from the optimality criterion of the update rule in Algorithm 4. The update $\pi_i^{t+1}(\cdot|s) \in \arg\max_{p(\cdot|s)\in\Delta(\mathcal{A}_i)} \{\beta \langle \overline{Q}_i^{\pi_t}(s, \cdot), p(\cdot|s)\rangle_{A_i} - \frac{1}{2}\|p(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2\}$ is a concave maximization problem. Therefore, $\beta \overline{Q}_i^{\pi_t}(s, \cdot) - \pi_i^{t+1}(\cdot|s) + \pi_i^t(\cdot|s)$ is not an increasing direction:

$$\langle \beta \overline{Q}_i^{\pi_t}(s, \cdot) - \pi_i^{t+1}(\cdot|s) + \pi_i^t(\cdot|s), p(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i} \leq 0, \ \forall p(\cdot|s) \in \Delta(\mathcal{A}_i). \tag{12}$$

The last inequality of Eq. (11) is derived by substituting $p = \pi_i^t$ in the above inequality. □

Lemma 21 (policy improvement (b))**.** *Let π^t to be the policy at time t of Algorithm 4,*

$$\Phi(\pi^{t+1}) - \Phi(\pi^t) \geq \frac{1}{\beta}(1 - \beta \frac{L_\Phi}{1-\Gamma}) \sum_{i=1}^N \mathbb{E}_{s\sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2.$$

Proof. Bound each term in Diff_2 of Eq. (10):

$$\begin{aligned}&\Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t+1}) - \Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t+1}) - \Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1}, \pi_j^{t}) + \Phi(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t}, \pi_j^{t})\\ =&\underbrace{\rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t+1},\pi_j^{t+1}} - \rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t},\pi_j^{t+1}}}_{I_1} - \underbrace{(\rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t+1},\pi_j^{t}} - \rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t},\pi_j^{t}})}_{I_2}.\end{aligned} \tag{13}$$

From the derivative of Lemma 4 $\frac{\partial^2 \rho_i^\pi}{\partial \pi_i}$ is $\frac{L_\Phi}{N}$-Lipschitz w.r.t. π_i or π_j, for any π_i, π_j. We aim to bound $I_1 - I_2$ with l2 norms of $\|\pi_i^t - \pi_i^{t+1}\|_2$ and $\|\pi_j^t - \pi_j^{t+1}\|_2$. Using the interpolation for a differentiable function, there exists $a, b \in [0,1]$, such that

$$\begin{aligned}I_1 =&\rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t+1},\pi_j^{t+1}} - \rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t},\pi_j^{t+1}} = \langle \frac{\partial \rho_i^\pi}{\partial \pi_i}\left(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + a(\pi_i^t - \pi_i^{t+1}), \pi_j^{t+1}\right), \pi_i^{t+1} - \pi_i^t\rangle,\\
I_2 =&\rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t+1},\pi_j^{t}} - \rho_i^{\pi^{t,t+1}_{-(i,j)},\pi_i^{t},\pi_j^{t}} = \langle \frac{\partial \rho_i^\pi}{\partial \pi_i}\left(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + b(\pi_i^t - \pi_i^{t+1}), \pi_j^{t}\right), \pi_i^{t+1} - \pi_i^t\rangle,\\
I_1 - I_2 =&\langle \frac{\partial \rho_i^\pi}{\partial \pi_i}\left(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + a(\pi_i^t - \pi_i^{t+1}), \pi_j^{t+1}\right) - \frac{\partial \rho_i^\pi}{\partial \pi_i}\left(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + a(\pi_i^t - \pi_i^{t+1}), \pi_j^{t}\right), \pi_i^{t+1} - \pi_i^t\rangle\\
&+ \langle \frac{\partial \rho_i^\pi}{\partial \pi_i}\left(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + a(\pi_i^t - \pi_i^{t+1}), \pi_j^{t}\right) - \frac{\partial \rho_i^\pi}{\partial \pi_i}\left(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + b(\pi_i^t - \pi_i^{t+1}), \pi_j^{t}\right), \pi_i^{t+1} - \pi_i^t\rangle\\
\geq& - \|\frac{\partial \rho_i^\pi}{\partial \pi_i}(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + a(\pi_i^t - \pi_i^{t+1}), \pi_j^{t+1}) - \frac{\partial \rho_i^\pi}{\partial \pi_i}(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + a(\pi_i^t - \pi_i^{t+1}), \pi_j^{t})\|_2 \|\pi_i^{t+1} - \pi_i^t\|_2\\
&- \|\frac{\partial \rho_i^\pi}{\partial \pi_i}(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + a(\pi_i^t - \pi_i^{t+1}), \pi_j^{t}) - \frac{\partial \rho_i^\pi}{\partial \pi_i}(\pi^{t,t+1}_{-(i,j)}, \pi_i^{t+1} + b(\pi_i^t - \pi_i^{t+1}), \pi_j^{t})\|_2 \|\pi_i^{t+1} - \pi_i^t\|_2\\
\geq& - \frac{L_\Phi}{N}\|\pi_j^{t+1} - \pi_j^t\|_2 \|\pi_i^{t+1} - \pi_i^t\|_2 - \frac{L_\Phi}{N}|a-b| \|\pi_i^t - \pi_i^{t+1}\|_2 \|\pi_i^{t+1} - \pi_i^t\|_2\\
\geq& - \frac{L_\Phi}{2N}(\|\pi_j^{t+1} - \pi_j^t\|_2^2 + \|\pi_i^{t+1} - \pi_i^t\|_2^2) - \frac{L_\Phi}{N}\|\pi_i^t - \pi_i^{t+1}\|_2^2\\
=& - \frac{L_\Phi}{2N}\|\pi_j^{t+1} - \pi_j^t\|_2^2 - \frac{3L_\Phi}{2N}\|\pi_i^{t+1} - \pi_i^t\|_2^2.\end{aligned}$$

Since the permutation of agents' indexes does not change the above result, we have:

$$\begin{aligned}\text{Diff}_2 &\geq -(N-1)\frac{L_\Phi}{N}\sum_i \|\pi_i^{t+1} - \pi_i^t\|_2^2\\
&\geq -L_\Phi \sum_i \sum_s \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2\\
&\geq -\frac{L_\Phi}{1-\Gamma}\sum_{i=1}^N \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2.\end{aligned}$$

Therefore, by the same bound for Diff_1 in Lemma 20 we can lower bound the policy improvement as

$$\Phi(\pi^{t+1}) - \Phi(\pi^t) \geq (\frac{1}{\beta} - \frac{L_\Phi}{1-\Gamma})\sum_{i=1}^N \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2.$$

□

Theorem 7 (Restatement of Theorem 3)**.** *If $\beta \leq \max\{\frac{1-\Gamma}{(N-1)(\kappa_Q + S\kappa^2)A_{\max}}, \frac{1-\Gamma}{2L_\Phi}\}$, Algorithm 4 has a bounded Nash regret:*

$$\text{Nash-regret}(T) \leq \sqrt{D}(\kappa\sqrt{A_{\max}} + \frac{2}{\beta})\sqrt{2\beta C_\Phi}\frac{1}{\sqrt{T}}. \tag{14}$$

Proof.

$$\begin{aligned}
\text{Nash-gap}(t) &= \max_i(\max_{\pi_i'} \rho_i^{\pi_i',\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t}) \\
&\overset{(1)}{=} \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}[\langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] \\
&= \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}[\underbrace{\langle \overline{Q}^{\pi^t}(s,\cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}}_{I_1} + \underbrace{\langle \overline{Q}^{\pi^t}(s,\cdot), \pi_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i}}_{I_2}].
\end{aligned}$$

In (1), we use π_i to represent the policy that achieves $\arg\max_{\pi_i'}$ in the previous expression and assume that i attains the maximum in $\max_i$. We will bound I_1 and I_2 separately.

Recall Eq. (12) for any p in the feasible policy set, $\langle \beta\overline{Q}_i^{\pi^t}(s,\cdot) - \pi_i^{t+1}(\cdot|s) + \pi_i^t(\cdot|s), p(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i} \leq 0$. We can bound I_1 and I_2 as

$$\begin{aligned}
I_2 &= \langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i} \\
&\leq \frac{1}{\beta}\langle \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s), \pi_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i} \\
&\overset{(a)}{\leq} \frac{1}{\beta}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_\infty \|\pi_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\|_1 \\
&\leq \frac{2}{\beta}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_\infty \\
&\leq \frac{2}{\beta}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2, \\
I_1 &= \langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i} \\
&\overset{(b)}{\leq} \kappa\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_1 \\
&\leq \kappa\sqrt{A_i}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2.
\end{aligned} \tag{15}$$

(a) and (b) result from $\langle x, y\rangle \leq \|x\|_1\|y\|_\infty$.

Therefore we can bound the Nash-gap as:

$$\begin{aligned}
\text{Nash-gap}(t) &\leq \sum_s \nu^{\pi_i,\pi_{-i}^t}(s)(\kappa\sqrt{A_i} + \frac{2}{\beta})\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2 \\
&\leq \sum_s \nu^{\pi_i,\pi_{-i}^t}(s)(\kappa\sqrt{A_{\max}} + \frac{2}{\beta})\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2 \\
&\leq \sqrt{D}(\kappa\sqrt{A_{\max}} + \frac{2}{\beta})\sum_s \sqrt{\nu^{\pi_i,\pi_{-i}^t}(s)}\sqrt{\nu^{\pi_i^{t+1},\pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2}, \\
\sum_{t=0}^{T-1}\text{Nash-gap}(t) &\overset{(a)}{\leq} \sqrt{D}(\kappa\sqrt{A_{\max}} + \frac{2}{\beta})\sqrt{\sum_{t=0}^{T-1}\sum_s \nu^{\pi_i^{t+1},\pi_{-i}^t}(s)}\sqrt{\sum_{t=0}^{T-1}\sum_s\sum_i \nu^{\pi_i^{t+1},\pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2} \\
&= \sqrt{D}(\kappa\sqrt{A_{\max}} + \frac{2}{\beta})\sqrt{T}\sqrt{\sum_{t=0}^{T-1}\sum_s\sum_i \nu^{\pi_i^{t+1},\pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2}.
\end{aligned}$$

(a) is due to the Cauchy-Schwarz inequality.

From Lemma 20 and Lemma 21, we have:

$$\Phi(\pi^{t+1}) - \Phi(\pi^t) \geq \frac{1}{\beta}\left(1 - \beta\min\{\frac{(N-1)(\kappa_Q + S\kappa^2)A_{\max}}{2(1-\Gamma)}, \frac{L_\Phi}{1-\Gamma}\}\right)\sum_{i=1}^N\sum_s \nu^{\pi_i^{t+1},\pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2.$$

Algorithm 6 proximal-Q algorithm with sample estimates

```
Input: learning rate $\beta$, gradient estimation parameters $B$, $N_1$
Initialization: $\pi_i^{(0)}(a_i|s) = 1/A_i$ for any $s \in \mathcal{S}$, $a_i \in \mathcal{A}_i$
for $t = 0$ to $T - 1$ do
  all agents take action independently and synchronously for $B$ time steps to collect trajectories $\{\mathcal{T}_i^t\}$
  for agent $i$ do
    for $s \in \mathcal{S}$ do
      $\hat{q}_i^t(s, \cdot) \leftarrow$ Q estimation$(\mathcal{T}_i^t, s, \pi_i^t, B, N_1)$
    end for
    $\pi_i^{(t+1)}(\cdot|s) = \underset{p(\cdot|s)\in\triangle_\alpha(\mathcal{A}_i)}{\arg\max} \{\beta\langle\hat{q}_i^{\pi^t}(s,\cdot), p(\cdot|s)\rangle_{\mathcal{A}_i} - \frac{1}{2}\|p(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2\},\ \forall s \in \mathcal{S}$
  end for
end for
```

Algorithm 7 Q estimation Wei et al., 2020, Lemma 6

```
Input: trajectory $\mathcal{T} = (s^0, a_i^0, r_i^0, ..., s^B, a_i^B, r_i^B)$, state $s$, policy $\pi_i$, parameters $B$ and $N_1$
$\tau \leftarrow 0$
$k \leftarrow 0$
while $\tau \leq B - N_1$ do
  if $s^\tau = s$ then
    $k \leftarrow k + 1$
    $R \leftarrow \sum_{t=\tau}^{\tau+N_1-1} r_i^t$
    $y_k \leftarrow \frac{R}{\pi_i(a_i^\tau|s)}\mathbf{1}[a = a_i^\tau]$ $(y_k \in \mathbb{R}^{A_i})$
    $\tau \leftarrow \tau + 2N_1$
  else
    $\tau \leftarrow \tau + 1$
  end if
end while
if $k \neq 0$ then
  return $\frac{1}{k}\sum_{j=1}^k y_j$
else
  return $\mathbf{0}$
end if
```

Therefore, by substituting learning rate $\beta \leq \frac{1}{2}\max\{\frac{2(1-\Gamma)}{(N-1)(\kappa_Q+S\kappa^2)A_{\max}}, \frac{1-\Gamma}{L_\Phi}\}$,

$$\begin{aligned}\sum_{t=0}^{T-1}\text{Nash-gap}(t) &\leq \sqrt{D}(\kappa\sqrt{A_{max}} + \frac{2}{\beta})\sqrt{T}\sqrt{\sum_{t=0}^{T-1}2\beta(\Phi(\pi^{t+1}) - \Phi(\pi^t))} \\ &\leq \sqrt{D}(\kappa\sqrt{A_{max}} + \frac{2}{\beta})\sqrt{T}\sqrt{2\beta C_\Phi}.\end{aligned}$$

□

D.1 Sample Complexity of proximal-Q

Lemma 22 (Wei et al., 2020, Lemma 6)**.** *Let $\mathbb{E}_t[x]$ denote the expectation of a random variable x conditioned on all history before episode t (note that π^t is updated at the end of episode $t-1$). If B is large enough, such that there exists $N_2 < B$ with $C_p\varrho^{N_2} \leq \frac{1}{2}(1-\Gamma)$, then for any t, s, a_i, the estimated $\hat{q}$ from Algorithm 7 satisfies:*

$$\mathbb{E}_t[(\hat{q}_i^t(s,a_i) - (\overline{Q}_i^{\pi^t}(s,a_i) + N_1\rho_i^{\pi^t}))^2] \leq 6(1 + \frac{2C_p\varrho^{2N_1}}{1-\Gamma}\frac{B}{2N_1})\left(\frac{N_1^2}{\alpha} + \frac{C_p^2}{(1-\varrho)^2} + \frac{C_p^2\varrho^{2N_1}}{(1-\varrho)^2}\right)\frac{(1-\frac{1-\Gamma}{2})^{n'} + \frac{2N_1+n'N_2}{B}}{1-(1-\frac{1-\Gamma}{2})^{\lfloor\frac{B-N_1}{N_2}\rfloor}}$$
$$+ (1 + \frac{2C_p\varrho^{2N_1}}{1-\Gamma}\frac{B}{2N_1})(\frac{2C_p^2\varrho^{2N_1}}{(1-\varrho)^2} + (1-\frac{1-\Gamma}{2})^{\lfloor\frac{B-N_1}{N_2}\rfloor}\frac{C_p}{1-\varrho}). \tag{16}$$

For completeness, we provide a brief proof below.

Proof. We define some notation first. Let τ_j be the evoked time at line 5 ($s^{\tau_j} = s$), w_j be the waiting time, where $w_j = \tau_j - (\tau_{j-1} + 2N_1)$ for $j > 1$, and $w_1 = \tau_1$ for $j = 1$. Furthermore, let $q_i^\pi(s,a_i) = \overline{Q}_i^\pi(s,a) + N_1\rho^\pi$, $\hat{q}_{i,j}^\pi(s,\cdot) = y_j(\cdot)$ where y_j is defined in line 8 of Algorithm 7, and $\hat{q}_i^\pi(s,\cdot) = \frac{1}{k}\sum_{j=1}^k \hat{q}_{i,j}^\pi(s,\cdot)$ if $k > 0$; otherwise, $\hat{q}_i^\pi = \mathbf{0}$.

The main difficulty of analyzing the bias and variance of $\hat{q}_i^\pi(s,a_i)$ lies in the random number k, which is the times state s is visited (used in line 14 of Algorithm 7), determined by $\{w_1, w_2, ...\}$ only. In the proof of Wei et al. (2020), they first calculate the bias and variance under an "imaginary" world, where the state distribution is reset to ν^π at any time $\tau_j + 2N_1$. Then, they demonstrate that the event $\{\tau_1, \tau_2, ...\}$ has a similar probability measure between the real world and the imaginary world. Since τ_{j+1} and τ_j are independent, it is easier to bound the bias and variance under an imaginary world. We use $\mathbb{E}'$ to denote the expectation in the imaginary world and $\mathbb{E}$ for the real world.

Step 1: Bound the bias and variance under imaginary world

$$\begin{aligned}
\mathbb{E}'[\hat{q}_i^\pi(s,a_i)] =& \mathbb{P}(w_1 \leq B - N_1)\mathbb{E}'[\frac{1}{k}\sum_{j=1}^k \mathbb{E}'[\hat{q}_{i,j}^\pi(s,a_i)|w_1]|w_1 \leq B - N] + \mathbb{P}(w_1 > B - N_1) \times 0 \\
\overset{(a)}{=}& \mathbb{P}(w_1 \leq B - N_1)\mathbb{E}'[\frac{1}{k}(\sum_{j=1}^k q_i^\pi(s,a_i) - \delta(s,a_i))] \\
\overset{(b)}{=}& q_i^\pi(s,a_i) - \delta'(s,a_i).
\end{aligned}$$

In (a), tail is defined as $\delta(s,a_i) := \mathbb{E}_{P,\pi}[\sum_{t=N_1+1}^\infty (r(s,a) - \rho^\pi)|s^0 = s, a_i^0 = a_i]$. It is easy to bound it by $|\delta(s,a_i)| \leq \sum_{t=N_1}^\infty C_p\varrho^t = C_p\varrho^{N_1}\frac{1}{1-\varrho}$. In (b), $\delta'(s,a_i) = (1 - \mathbb{P}(w_1 \leq B - N_1))(q_i^\pi(s,a_i) - \delta(s,a_i)) + \delta(s,a_i)$.

To give a bound for $\delta'(s,a_i)$, let's analyze $|q_i^\pi(s,a_i)|$ and $\mathbb{P}(w_1 \leq B - N_1)$ separately. By Proposition 3 $|q_i^\pi(s,a_i)| \leq C_p\kappa_0 + N_1$. $1 - \mathbb{P}(w_1 \leq B - N_1)$ is the probability of never visiting s during time 0 to time $B - N_1$. If B is large enough, there exists $N_2 < B$ such that $C_p\varrho^{N_2} \leq \frac{1}{2}(1-\Gamma)$, which means $|\mathbb{P}(s^{N_2} = s|s^0 = s') - \nu^\pi(s)| \leq \frac{1}{2}(1-\Gamma)$ and $\mathbb{P}(s^{N_2} = s|s^0 = s') \geq \nu^\pi(s) - \frac{1}{2}(1-\Gamma) \geq \frac{1}{2}(1-\Gamma)$ for any s'. Therefore, $\mathbb{P}(w_1 > B - N_1) \leq (1 - \frac{1-\Gamma}{2})^{\lfloor\frac{B-N_1}{N_2}\rfloor}$.

$$\begin{aligned}
|\mathbb{E}'[\hat{q}_i^\pi(s,a_i)] - q_i^\pi(s,a_i)| =& |\delta'(s,a_i)| \\
\leq& (1 - \frac{1-\Gamma}{2})^{\lfloor\frac{B-N_1}{N_2}\rfloor}(C_p\kappa_0 + N_1) + C_p\varrho^{N_1}\frac{1}{1-\varrho}.
\end{aligned}$$

To bound the variance, denote $\Delta_j = \hat{q}^{\pi}_{i,j}(s,a_i) - q^{\pi}_i(s,a_i) + \delta(s,a_i)$. Then $\mathbb{E}'[\Delta_j|w_j] = 0$.

$$
\begin{aligned}
&\mathbb{E}'[(\hat{q}^{\pi}_i(s,a_i) - q^{\pi}_i(s,a_i))^2] \\
=&\mathbb{P}(w_1 \le B - N_1)\mathbb{E}'[(\hat{q}^{\pi}_i(s,a_i) - q^{\pi}_i(s,a_i))^2 | w_1 \le B - N_1] + \mathbb{P}(w_1 > B - N_1)|q^{\pi}_i(s,a_i)|^2 \\
\le&\mathbb{E}'[(\frac{1}{k}\sum_{j=1}^{k}\Delta_j - \delta(s,a_i))^2 | w_1 \le B - N_1] + (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}(C_p\kappa_0 + N_1)^2 \\
\le&\mathbb{E}'[2(\frac{1}{k}\sum_{j=1}^{k}\Delta_j)^2 + 2\delta(s,a_i))^2 | w_1 \le B - N_1] + (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}(C_p\kappa_0 + N_1)^2 \\
\le&\mathbb{E}'[2(\frac{1}{k}\sum_{j=1}^{k}\Delta_j)^2 | w_1 \le B - N_1] + \frac{2C_p^2\varrho^{2N_1}}{(1-\varrho)^2} + (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}(C_p\kappa_0 + N_1)^2 \\
\le&\mathbb{E}'[\frac{2}{k^2}\sum_{j=1}^{k}\mathbb{E}'[\Delta_j^2|w_1] | w_1 \le B - N_1] + \frac{2C_p^2\varrho^{2N_1}}{(1-\varrho)^2} + (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}(C_p\kappa_0 + N_1)^2 \\
\overset{(a)}{\le}&6\left(\frac{N_1^2}{\pi_i(a_i|s)} + 2C_p^2\kappa_0^2\pi_i(a_i|s) + 2N_1^2\pi_i(a_i|s) + \frac{C_p^2\varrho^{2N_1}\pi_i(a_i|s)}{(1-\varrho)^2}\right)\mathbb{E}'[\frac{1}{k}|w_1 \le B - N_1] \\
&+ \frac{2C_p^2\varrho^{2N_1}}{(1-\varrho)^2} + (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}(C_p\kappa_0 + N_1)^2.
\end{aligned}
$$

In (a) we use $\mathbb{E}'[\Delta_j^2|w_1] \le \pi_i(a_i|s)(3\frac{N_1^2}{\pi_i(a_i|s)^2} + 3(2C_p^2\kappa_0^2 + 2N_1^2) + 3\frac{C_p^2\varrho^{2N_1}}{(1-\varrho)^2})$.

If trajectory length B is large enough, there exists a waiting period length $n'N_2$ such that $k_0 = \lfloor \frac{B}{2N_1 + n'N_2} \rfloor > 1$, $\mathbb{P}(k \le k_0) \le \mathbb{P}(w_1 \ge n'N_2) \le (1 - \frac{1-\Gamma}{2})^{n'}$ is small enough. Then we can bound the average visiting time and variance by:

$$
\begin{aligned}
\mathbb{E}'[\frac{1}{k}|w_1 \le B - N_1] \le& \frac{\mathbb{P}(k \le k_0) \times 1 + \mathbb{P}(k > k_0)\frac{1}{k_0}}{\mathbb{P}(w_1 \le B - N_1)} \\
\le& \frac{(1 - \frac{1-\Gamma}{2})^{n'} + \frac{2N_1 + n'N_2}{B}}{1 - (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}'[(\hat{\beta}^{\pi}_i(s,a) - \beta^{\pi}_i(s,a_i))^2] \le& 6\left(\frac{N_1^2}{\pi_i(a_i|s)} + 2C_p^2\kappa_0^2\pi_i(a_i|s) + 2N_1^2\pi_i(a_i|s) + \frac{C_p^2\varrho^{2N_1}\pi_i(a_i|s)}{(1-\varrho)^2}\right)\frac{(1 - \frac{1-\Gamma}{2})^{n'} + \frac{2N_1 + n'N_2}{B}}{1 - (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}} \\
&+ \frac{2C_p^2\varrho^{2N_1}}{(1-\varrho)^2} + (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B-N_1}{N_2} \rfloor}\frac{C_p}{1-\varrho}.
\end{aligned}
$$

Step2: bound the difference between imaginary world and real world

Since $\hat{q}^{\pi}_i(s,a_i)$ is determined by $X = (k, \tau_1, \mathcal{T}_1, \tau_2, \mathcal{T}_2, ..., \tau_k, \mathcal{T}_k)$, let $\hat{q}^{\pi}_i(s,a_i) = f(X)$. Then $\frac{\mathbb{E}[\hat{q}^{\pi}_i(s,a_i)]}{\mathbb{E}'[\hat{q}^{\pi}_i(s,a_i)]} = \frac{\sum_x f(x)\mathbb{P}(X=x)}{\sum_x f(x)\mathbb{P}'(X=x)} \le \max_x \frac{\mathbb{P}(X=x)}{\mathbb{P}'(X=x)}$. We can bound $\frac{\mathbb{P}(X=x)}{\mathbb{P}'(X=x)}$, $\forall x$ as follow:

$$
\begin{aligned}
\frac{\mathbb{P}(X=x)}{\mathbb{P}'(X=x)} =& \frac{\mathbb{P}(\tau_2|\tau_1, \mathcal{T}_1)...\mathbb{P}(\tau_k|\tau_{k-1}, \mathcal{T}_{k-1})}{\mathbb{P}'(\tau_2|\tau_1, \mathcal{T}_1)...\mathbb{P}'(\tau_k|\tau_{k-1}, \mathcal{T}_{k-1})} \\
\overset{(a)}{\le}&(\max_{s'}\frac{\mathbb{P}(s^{\tau_1 + 2N_1} = s'|\tau_1)}{\nu^{\pi}(s')})\dots(\max_{s'}\frac{\mathbb{P}(s^{\tau_{k-1} + 2N_1} = s'|\tau_{k-1})}{\nu^{\pi}(s')}) \\
\le&(1 + \frac{C_p\varrho^{2N_1}}{1-\Gamma})^{\frac{B}{2N_1}} \le e^{\frac{C_p\varrho^{2N_1}}{1-\Gamma}\frac{B}{2N_1}} \le 1 + \frac{2C_p\varrho^{2N_1}}{1-\Gamma}\frac{B}{2N_1}.
\end{aligned}
$$

Here (a) is derived by that:

$$\mathbb{P}(\tau_{j+1}|\tau_j, \mathcal{T}_j) = \sum_{s' \neq s} \mathbb{P}(\tau_j + 2N_1 = s'|\tau_j, \mathcal{T}_j)\mathbb{P}(s^t \neq s, \forall\ t \in [\tau_j + 2N_1 + 1, \tau_{j+1} - 1], s^{\tau_2} = s|\tau_j + 2N_1 = s'),$$
$$\mathbb{P}'(\tau_{j+1}|\tau_j, \mathcal{T}_j) = \sum_{s' \neq s} \nu^{\pi}(s')\mathbb{P}(s^t \neq s, \forall\ t \in [\tau_j + 2N_1 + 1, \tau_{j+1} - 1], s^{\tau_2} = s|\tau_j + 2N_1 = s'),$$

for $\tau_{j+1} \neq \tau_j + 2N_1$. When $\tau_{j+1} = \tau_j + 2N_1$, we have:

$$\mathbb{P}(\tau_{j+1}|\tau_j, \mathcal{T}_j) = \mathbb{P}(\tau_j + 2N_1 = s|\tau_j, \mathcal{T}_j),\ \mathbb{P}'(\tau_{j+1}|\tau_j, \mathcal{T}_j) = \nu^{\pi}(s).$$

Therefore, the following result can be derived:

$$\begin{aligned}\mathbb{E}[(\hat{q}_i^{\pi}(s, a_i) - q_i^{\pi}(s, a_i))^2] \leq & \mathbb{E}'[(\hat{q}_i^{\pi}(s, a_i) - q_i^{\pi}(s, a_i))^2](1 + \frac{C_p \varrho^{2N_1}}{\nu^{\pi}(s)} \frac{B}{N_1}) \\ \leq & 6(1 + \frac{C_p \varrho^{2N_1}}{1 - \Gamma} \frac{B}{N_1}) \left(\frac{N_1^2}{\alpha} + \frac{2C_p^2}{(1 - \varrho)^2} + 2N_1^2 + \frac{C_p^2 \varrho^{2N_1}}{(1 - \varrho)^2} \right) \frac{(1 - \frac{1-\Gamma}{2})^{n'} + \frac{2N_1 + n' N_2}{B}}{1 - (1 - \frac{1-\Gamma}{2})^{\lfloor \frac{B - N_1}{N_2} \rfloor}} \\ & + (1 + \frac{C_p \varrho^{2N_1}}{1 - \Gamma} \frac{B}{N_1})(\frac{2C_p^2 \varrho^{2N_1}}{(1 - \varrho)^2} + (1 - \frac{1 - \Gamma}{2})^{\lfloor \frac{B - N_1}{N_2} \rfloor} \frac{C_p}{1 - \varrho}).\end{aligned}$$

□

Let $n' = N_1 = O(\log \frac{1}{\alpha\delta})$, $N_2 = O(\log \frac{1}{1-\Gamma})$, $B = \tilde{O}(\frac{1}{\alpha\delta})$. For any agent i, time t, state s, and action a_i, $\mathbb{E}_t[(\hat{q}_i^{\pi}(s, a_i) - q_i^{\pi}(s, a_i))^2] \leq \delta$.

Lemma 23. *(Policy improvement)*

$$\begin{aligned}\Phi(\pi^{t+1}) - \Phi(\pi^t) \geq & \left(\frac{1}{\beta} - \frac{1}{2} - \frac{L_\Phi}{1 - \Gamma} \right) \sum_{i=1}^{N} \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2 \\ & - \frac{1}{2} \sum_{i=1}^{N} \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} [\sum_{a_i} (\hat{q}_i^{\pi^t}(s, a_i) - q_i^{\pi^t}(s, a_i))^2].\end{aligned}$$

Proof. We use the same decomposition as Eq. (10).

Similar as Lemma 21, $\text{Diff}_2 \geq -\frac{L_\Phi}{1-\Gamma} \sum_{i=1}^{N} \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2$.

Note that $\pi_i^{t+1}(\cdot|s) = \underset{p(\cdot|s) \in \triangle_\alpha(\mathcal{A}_i)}{\arg\max} \{\beta \langle \hat{q}_i^{\pi^t}(s, \cdot), p(\cdot|s) \rangle_{\mathcal{A}_i} - \frac{1}{2}\|p(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2)\}$. Deriving from the optimality we have:

$$\langle \beta \hat{q}_i^{\pi_t}(s, \cdot) - \pi_i^{t+1}(\cdot|s) + \pi_i^t(\cdot|s), p(\cdot|s) - \pi_i^{t+1}(\cdot|s) \rangle_{\mathcal{A}_i} \leq 0,\ \forall p(\cdot|s) \in \Delta_\alpha(\mathcal{A}_i). \quad (17)$$

To bound Diff_1,

$$\begin{aligned}
\mathrm{Diff}_1 =& \sum_{i=1}^{N} \rho_i^{\pi_i^{t+1},\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t} \\
=& \sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] \\
=& \sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot) + N_1\rho_i^{\pi^t}, \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] \\
=& \sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} [\langle \hat{q}_i^{\pi^t}(s,\cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] \\
&+ \sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} [\langle q_i^{\pi^t}(s,a_i) - \hat{q}_i^{\pi^t}(s,\cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] \\
\overset{(a)}{\geq}& \frac{1}{\beta}\sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2 - \sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} \frac{\|\hat{q}_i^{\pi^t}(s,\cdot) - q_i^{\pi^t}(s,\cdot)\|_2^2 + \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2}{2} \\
=& (\frac{1}{\beta} - \frac{1}{2})\sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2 - \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} [\sum_{a_i}(\hat{q}_i^{\pi^t}(s,a_i) - q_i^{\pi^t}(s,a_i))^2].
\end{aligned}$$

(a) is derived by applying $p = \pi_i^t$ to equation (27). □

Theorem 8. *If all players run Algorithm 6 independently and synchronously, $\beta \leq (1+\frac{2L_\Phi}{1-\Gamma})^{-1}$, the Nash regret will be bounded by:*

$$\mathbb{E}[\text{Nash-regret}(T)] \leq (\frac{2}{\sqrt{\beta}} + \kappa\sqrt{A_{\max}\beta})\sqrt{\frac{2ND}{T}} + (\frac{2}{\sqrt{\beta}} + \kappa\sqrt{\beta})D\sqrt{NA_{\max}\delta} + 2\sqrt{A_{\max}\delta} + 2\kappa\alpha.$$

Proof.

$$\begin{aligned}
&\text{Nash-gap}(t) \\
=& \max_i(\max_{\pi_i'} \rho_i^{\pi_i',\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t}) \\
\overset{(a)}{=}& \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] \\
=& \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] + \alpha\mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i(\cdot|s) - u_i(\cdot|s)\rangle_{\mathcal{A}_i}] \\
\leq& \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] + 2\alpha\kappa \\
=& \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i}] + \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] \\
&+ 2\alpha\kappa \\
=& \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \hat{q}_i^{\pi^t}(s,\cdot), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i}] \\
&+ \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle q_i^{\pi^t}(s,\cdot) - \hat{q}_i^{\pi^t}(s,\cdot), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i}] \\
&+ \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} [\langle \overline{Q}_i^{\pi^t}(s,\cdot), \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\rangle_{\mathcal{A}_i}] + 2\alpha\kappa.
\end{aligned} \tag{18}$$

In (a), we assume agent i achieved the $\max_i$ and π_i achieved $\max_{\pi_i'}$.

By Eq. (17):

$$\langle \beta\hat{q}_i^t(s,\cdot) - \pi_i^{t+1}(\cdot|s) + \pi_i^t(\cdot|s), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i} \leq 0,$$

we can bound the first term of Eq. (18) as:

$$\begin{aligned}\langle \hat{q}_i^{\pi^t}(s,\cdot), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i} \leq & \frac{1}{\beta}\langle \pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s), (1-\alpha)\pi_i(\cdot|s) + \alpha u_i(\cdot|s) - \pi_i^{t+1}(\cdot|s)\rangle_{\mathcal{A}_i} \\ \leq & \frac{2}{\beta}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2.\end{aligned}$$

The last inequality comes from $\|\pi'(\cdot|s) - \pi(\cdot|s)\|_2 \leq \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \leq 2$.

Therefore, the Nash gap can be bounded as:

$$\begin{aligned}&\text{Nash-gap}(t) \\ \overset{(a)}{\leq} & \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}[\frac{2}{\beta}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2 + 2\|q^{\pi^t}(s,\cdot) - \hat{q}^{\pi^t}(s,\cdot)\|_2 + \kappa\sqrt{A_{\max}}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2 + 2\alpha\kappa] \\ = & (\frac{2}{\beta} + \kappa\sqrt{A_{\max}})\mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2 + 2\mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}\|q^{\pi^t}(s,\cdot) - \hat{q}^{\pi^t}(s,\cdot)\|_2 + 2\alpha\kappa \\ \leq & (\frac{2}{\beta} + \kappa\sqrt{A_{\max}})\sqrt{\|\frac{\nu^{\pi_i,\pi_{-i}^t}}{\nu^{\pi^t}}\|_\infty}\sum_s \sqrt{\nu^{\pi_i,\pi_{-i}^t}(s)\nu^{\pi_i^{t+1},\pi_{-i}^t}(s)}\sqrt{\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2} \\ & + 2\mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}\|q^{\pi^t}(s,\cdot) - \hat{q}^{\pi^t}(s,\cdot)\|_2 + 2\alpha\kappa \\ \leq & (\frac{2}{\beta} + \kappa\sqrt{A_{\max}})\sqrt{D}\sum_s \sqrt{\nu^{\pi_i,\pi_{-i}^t}(s)}\sqrt{\nu^{\pi_i^{t+1},\pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2} \\ & + 2\mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}\|q^{\pi^t}(s,\cdot) - \hat{q}^{\pi^t}(s,\cdot)\|_2 + 2\alpha\kappa.\end{aligned}$$

Applying Lemma 23 and the approximation error bound leads to

$$\begin{aligned}\mathbb{E}[\mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}\|q^{\pi^t}(s,\cdot) - \hat{q}^{\pi^t}(s,\cdot)\|_2] \leq & \sqrt{\mathbb{E}[\mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}}\|q^{\pi^t}(s,\cdot) - \hat{q}^{\pi^t}(s,\cdot)\|_2^2]} \\ \overset{(a)}{=} & \sqrt{\mathbb{E}[\sum_s \nu^{\pi_i,\pi_{-i}^t}(s)\mathbb{E}_t[\sum_{a_i}(q^{\pi^t}(s,a_i) - \hat{q}^{\pi^t}(s,a_i))^2]]} \\ \leq & \sqrt{A_{\max}\delta}.\end{aligned}$$

In (a) we can exchange $\mathbb{E}_t$ and $\sum_s \nu^{\pi_i,\pi_{-i}^t}(s)$ since π_i only depends on π^t, i.e. π_i only depends on $\mathcal{F}_{t-1}$.

Sum over t:

$$
\begin{aligned}
&\mathbb{E}\sum_{t=0}^{T-1}\text{Nash-gap}(t)\\
\leq&(\frac{2}{\beta}+\kappa\sqrt{A_{\max}})\sqrt{D}\mathbb{E}[\sum_{t=0}^{T-1}\sum_{s}\sqrt{\nu^{\pi_i,\pi_{-i}^t}(s)}\sqrt{\nu^{\pi_i^{t+1},\pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\|_2^2}]+2\sqrt{A_{\max}\delta}T+2\alpha\kappa T\\
\overset{(a)}{\leq}&(\frac{2}{\beta}+\kappa\sqrt{A_{\max}})\sqrt{D}\mathbb{E}\sqrt{\sum_{t=0}^{T-1}\sum_{s}\nu^{\pi_i,\pi_{-i}^t}(s)}\sqrt{\sum_{t=0}^{T-1}\sum_{i}\sum_{s}\nu^{\pi_i^{t+1},\pi_{-i}^t}(s)\|\pi_i^{t+1}(\cdot|s)-\pi_i^t(\cdot|s)\|_2^2}\\
&+2\sqrt{A_{\max}\delta}T+2\alpha\kappa T\\
\overset{(b)}{\leq}&(\frac{2}{\beta}+\kappa\sqrt{A_{\max}})\sqrt{DT}\mathbb{E}\sqrt{\left(2\beta\sum_{t=0}^{T-1}(\Phi(\pi^{t+1})-\Phi(\pi^t))+\beta\sum_{t=0}^{T-1}\sum_{i=1}^{N}\sum_{a_i}\mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}}(\hat{q}_i^{\pi^t}(s,a_i)-q_i^{\pi^t}(s,a_i))^2\right)}\\
&+2\sqrt{A_{\max}\delta}T+2\alpha\kappa T\\
\leq&(\frac{2}{\beta}+\kappa\sqrt{A_{\max}})\sqrt{DT}(\sqrt{2\beta C_\Phi}+\mathbb{E}\sqrt{\beta\sum_{t=0}^{T-1}\sum_{i=1}^{N}\sum_{a_i}\mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}}(\hat{q}_i^{\pi^t}(s,a_i)-q_i^{\pi^t}(s,a_i))^2)}\\
&+2\sqrt{A_{\max}\delta}T+2\alpha\kappa T\\
\overset{(c)}{\leq}&(\frac{2}{\beta}+\kappa\sqrt{A_{\max}})\sqrt{DT}(\sqrt{2\beta C_\Phi}+\sqrt{\beta\mathbb{E}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\sum_{a_i}\mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}}(\hat{q}_i^{\pi^t}(s,a_i)-q_i^{\pi^t}(s,a_i))^2)}\\
&+2\sqrt{A_{\max}\delta}T+2\alpha\kappa T\\
\leq&(\frac{2}{\beta}+\kappa\sqrt{A_{\max}})\sqrt{DT}(\sqrt{2\beta C_\Phi}+\sqrt{\beta TNA_{\max}\delta})+2\sqrt{A_{\max}\delta}T+2\alpha\kappa T,
\end{aligned}
$$

where (a) is due to Cauchy-Schwarz inequality. (b) is derived by Lemma 23 and learning rate $\beta\leq(1+\frac{2L_\Phi}{1-\Gamma})^{-1}$. Note that $\Phi(\pi^T)-\Phi(\pi^0)\leq C_\Phi$. (c) is Jensen's inequality.

$$
\mathbb{E}[\text{Nash-regret}(T)]\leq(\frac{2}{\sqrt{\beta}}+\kappa\sqrt{A_{\max}\beta})\sqrt{\frac{2C_\Phi D}{T}}+(\frac{2}{\sqrt{\beta}}+\kappa\sqrt{A_{\max}\beta})\sqrt{NA_{\max}\delta}+2\sqrt{A_{max}\delta}+2\kappa\alpha.
$$

□

If $\beta=(1+\frac{2L_\Phi}{1-\Gamma})^{-1}$, $\mathbb{E}[\text{Nash-regret}(T)]=O(\sqrt{\frac{C_\Phi NDA_{\max}S^{3/2}\kappa_0^2}{(1-\Gamma)T}}+\sqrt{\frac{N^2A_{\max}^2S^{3/2}\kappa_0^2\delta}{1-\Gamma}}+\kappa\alpha)$. To obtain an ϵ-Nash equilibrium, let $T=O(\frac{C_\Phi NDA_{\max}S^{3/2}\kappa_0^2}{(1-\Gamma)\epsilon^2})$, $\delta=O(\frac{(1-\Gamma)\epsilon^2}{N^2A_{\max}^2S^{3/2}\kappa_0^2})$ and $\alpha=O(\frac{\epsilon}{\kappa})$. By Lemma 22, $B=\tilde{O}(\frac{1}{\alpha\delta})=\tilde{O}(\frac{N^2A_{\max}^2S^2\kappa_0^3}{(1-\Gamma)^{3/2}\epsilon^3})$. Therefore, the sample complexity for Algorithm 6 is $TB=\tilde{O}(\frac{C_\Phi N^3DA_{\max}^3S^{7/2}\kappa_0^5}{(1-\Gamma)^{5/2}\epsilon^5})$. Comparing with sample complexity of Algorithm 2 in Theorem 2, the sample complexity of Algorithm 6 is $O(\frac{A_{\max}^2S^{7/2}\kappa_0^4}{1-\Gamma})$ smaller.

E PROOFS OF SECTION 5

Lemma 24. $\frac{\partial\rho_i^{\pi_\theta}}{\partial\theta_{s,a_i}}=\nu^{\pi_\theta}(s)\pi_{\theta_i}(a_i|s)\overline{A}_i^{\pi_\theta}(s,a_i)$.

Proof. From the policy gradient theorem (Sutton et al., 1999), $\frac{\partial\rho_i^{\pi_\theta}}{\partial\theta_{s,a_i}}=\sum_{s',\mathbf{a}'}\nu^{\pi_\theta}(s)\pi_\theta(\mathbf{a}|s)\frac{\partial\log\pi_\theta(\mathbf{a}'|s')}{\partial\theta_{s,a_i}}Q_i^{\pi_\theta}(s,\mathbf{a})$, while $\pi_\theta(\mathbf{a}|s)=\Pi_{i=1}^N\pi_{\theta_i}(a_i|s)$ and $\frac{\partial\log\pi_\theta(\mathbf{a}'|s)}{\partial\theta_{s,a_i}}=\mathbf{1}\{a_i'=a_i,s'=s\}-\mathbf{1}\{s'=s\}\pi_{\theta_i}(a_i|s)$. Therefore, $\frac{\partial\rho_i^{\pi_\theta}}{\partial\theta_{s,a_i}}=\nu^{\pi_\theta}(s)\pi_{\theta_i}(a_i|s)\overline{A}_i^{\pi_\theta}(s,a_i)$. □

Lemma 25 (Lemma 9-12 Zhang et al. (2022))**.** *The update rule* $\theta_i^{t+1} = \theta_i^t + \beta F_i(\theta^t)^\dagger \nabla_{\theta_i} \rho_i^{\pi^t}$ *is equivalent to* $\pi_i^{t+1}(a_i|s) \propto \pi_i^t(a_i|s) \exp\left(\beta \overline{A}_i^{\pi^t}(s, a_i)\right)$*, where* $F_i(\theta) = \mathbb{E}_{s \sim \nu^{\pi_\theta}} \mathbb{E}_{a_i \sim \pi_{\theta_i}(\cdot|s)} [\nabla_{\theta_i} \log \pi_{\theta_i}(a_i|s) \nabla_{\theta_i} \log \pi_{\theta_i}(a_i|s)^T]$ *and* $\pi_i^t = \pi_{\theta_i^t}$

Lemma 26 (Policy improvement, Lemma 7)**.** *If* $\beta \le \max\{\frac{1-\Gamma}{(N-1)(\kappa_Q + S\kappa^2)}, \frac{1-\Gamma}{2L_\Phi}\}$*, the policies updated by Algorithm 5 have:*

$$\Phi(\pi^{t+1}) - \Phi(\pi^t) \ge \frac{1}{\beta} \sum_i \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_i^t}} \log Z_i^{t,s}.$$

Proof.

$$\begin{aligned}\Phi^{t+1} - \Phi^t = & \underbrace{\sum_{i=1}^N (\Phi(\pi_i^{t+1}, \pi_{-i}^t) - \Phi(\pi_i^t, \pi_{-i}^t))}_{\text{Diff}_1} \\ & + \underbrace{\sum_{i=1}^N \sum_{j=i+1}^N (\Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^{t+1}, \pi_j^{t+1}) - \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^t, \pi_j^{t+1}) - \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^{t+1}, \pi_j^t) + \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^t, \pi_j^t)),}_{\text{Diff}_2}\end{aligned}$$

(where $\pi_{-(i,j)}^{t,t+1} := (\pi_{1:(i-1)}^t, \pi_{(i+1):(j-1)}^t, \pi_{(j+1):N}^{t+1})$).

Similar to Lemma 20,

$$\begin{aligned} & \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^{t+1}, \pi_j^{t+1}) - \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^t, \pi_j^{t+1}) - \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^{t+1}, \pi_j^t) + \Phi(\pi_{-(i,j)}^{t,t+1}, \pi_i^t, \pi_j^t) \\ \ge & - \frac{\kappa_Q + S\kappa^2}{2} (\|\pi_j^{t+1} - \pi_j^t\|_{1,\infty}^2 + \|\pi_i^{t+1} - \pi_i^t\|_{1,\infty}^2). \end{aligned} \tag{19}$$

Let $s_i \in \arg\max_{s \in \mathcal{S}} \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_1$. By Pinsker's inequality, we get:

$$\text{Diff}_2 \ge - \frac{(N-1)(\kappa_Q + S\kappa^2)}{2} \sum_{i=1}^N \|\pi_i^{t+1} - \pi_i^t\|_{1,\infty}^2 \ge -(N-1)(\kappa_Q + S\kappa^2) \sum_{i=1}^N D_{\pi_i^t}^{\pi_i^{t+1}}(s_i). \tag{20}$$

Define $Z_t^{i,s} = \sum_{a_i} \pi_i^t(a_i|s) \exp\left(\beta \overline{A}_i^t(s, a_i)\right)$. By the update rule $\pi_i^{t+1}(a_i|s) = \frac{\pi_i^t(a_i|s) \exp\left(\beta \overline{A}_i^t(s, a_i)\right)}{Z_t^{i,s}}$, $\overline{A}_i^t(s, a_i) = \frac{1}{\beta}(\log(\frac{\pi_i^{t+1}(a_i|s)}{\pi_i^t(a_i|s)}) + \log Z_i^{t,s})$. Hence

$$\begin{aligned}\Phi(\pi_i^{t+1}, \pi_{-i}^t) - \Phi(\pi_i^t, \pi_{-i}^t) = & \rho_i^{\pi_i^{t+1}, \pi_i^t} - \rho_i^{\pi_i^t, \pi_{-i}^t} \\ = & \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \sum_{a_i} \pi_i^{t+1}(a_i|s) \overline{A}_i^{\pi^t}(s, a_i) \\ = & \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \sum_{a_i} \pi_i^{t+1}(a_i|s) \frac{1}{\beta} (\log \frac{\pi_i^{t+1}(a_i|s)}{\pi_i^t(a_i|s)} + \log Z_i^{t,s}) \\ = & \frac{1}{\beta} \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} (D_{\pi_i^t}^{\pi_i^{t+1}}(s) + \sum_{a_i} \pi_i^{t+1}(a_i|s) \log Z_i^{t,s}) \\ \overset{(a)}{\ge} & \frac{1 - \Gamma}{\beta} D_{\pi_i^t}^{\pi_i^{t+1}}(s_i) + \frac{1}{\beta} \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \log Z_i^{t,s}.\end{aligned}$$

Therefore,

$$\text{Diff}_1 \ge \frac{1 - \Gamma}{\beta} \sum_i D_{\pi_i^t}^{\pi_i^{t+1}}(s_i) + \frac{1}{\beta} \sum_i \mathbb{E}_{s \sim \nu^{\pi_i^{t+1}, \pi_{-i}^t}} \log Z_i^{t,s},$$

where (a) comes from $1-\Gamma \le \nu^{\pi_i^{t+1},\pi_{-i}^t}(s_i) \le \Gamma$ for any i.

If the learning rate is chosen as $\beta \le \frac{1-\Gamma}{(N-1)(\kappa_Q+S\kappa^2)}$, then combining the above results:

$$\begin{aligned}\Phi(\pi^{t+1}) - \Phi(\pi^t) \ge& (\frac{1-\Gamma}{\beta} - (N-1)(\kappa_Q + S\kappa^2)) \sum_{i=1}^{N} D_{\pi_i^{t+1}}^{\pi_i^t}(s_i) + \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_i^t}} \log Z_i^{t,s} \\ \ge& \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_i^t}} \log Z_i^{t,s}.\end{aligned} \tag{21}$$

If we use Lemma 21 to bound each term in Diff_2, then by $\|x\|_2^2 \le \|x\|_1^2$ and Pinsker's inequality:

$$\mathrm{Diff}_2 \ge -L_\Phi \sum_i \sum_s \|\pi_i^{t+1}(\cdot|s) - \pi_i^t(\cdot|s)\|_2^2 \ge -\frac{2L_\Phi}{1-\Gamma}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} D_{\pi_i^t}^{\pi_i^{t+1}}(s).$$

A similar result for Eq. (21) can be derived:

$$\begin{aligned}\Phi(\pi^{t+1}) - \Phi(\pi^t) =& \mathrm{Diff}_1 + \mathrm{Diff}_2 \\ \ge& \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} (D_{\pi_i^t}^{\pi_i^{t+1}}(s) + \log Z_i^{t,s}) - \frac{2L_\Phi}{1-\Gamma}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} D_{\pi_i^t}^{\pi_i^{t+1}}(s) \\ \ge& \left(\frac{1}{\beta} - \frac{2L_\Phi}{1-\Gamma}\right)\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} D_{\pi_i^t}^{\pi_i^{t+1}}(s) + \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{-i}^t}} \log Z_i^{t,s} \\ \ge& \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_i^t}} \log Z_i^{t,s}.\end{aligned}$$

Therefore, if $\beta \le \max\{\frac{1-\Gamma}{(N-1)(\kappa_Q+S\kappa^2)}, \frac{1-\Gamma}{2L_\Phi}\}$, $\Phi(\pi^{t+1}) - \Phi(\pi^t) \ge \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_i^t}} \log Z_i^{t,s}$. □

With the above monotone improvement, the sufficient condition for asymptotic convergence established by Zhang et al. (2022) is satisfied.

Lemma 27 (Section 12.0.2 Zhang et al. (2022))**.** *If all stationary points of the potential function $\Phi(\theta) = \Phi(\pi(\theta))$ are isolated, $\beta\|A^\pi\|_\infty \le 1$ for any π, and non-negative improvement $\Phi(\pi^{t+1}) - \Phi(\pi^t) \ge \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_i^t}} \log Z_i^{t,s}$ exists, algorithm 5 asymptotically converges to a Nash equilibrium. Also $c > 0$.*

Note that $\|A^\pi\|_\infty \le \|Q^\pi\|_\infty + \|V^\pi\|_\infty \le 2\kappa$ for any $\pi \in \Pi$. With Lemma 26 and Lemma 27 the following lemma holds.

Lemma 28 (Restatement of Lemma 8)**.** *If all stationary points of the potential function $\Phi(\theta) = \Phi(\pi(\theta))$ are isolated, $\beta \le \max\{\frac{1-\Gamma}{(N-1)(\kappa_Q+S\kappa^2)}, \frac{1-\Gamma}{2L_\Phi}\}$ and $\beta \le \frac{1}{2\kappa}$, Algorithm 5 asymptotically converges to a Nash equilibrium. Also $c > 0$.*

Lemma 29 (Lemma 21 Zhang et al. (2022))**.** *When $\beta\|A^\pi\|_\infty \le 1$ for any π, $\log Z_t^{i,s} \ge \frac{c}{3}(\beta \max_{a_i} \overline{A}_i^{\pi_t}(s,a_i))^2$.*

Theorem 9 (Restatement of Theorem 4)**.** *If all stationary points of the potential function $\Phi(\theta) = \Phi(\pi(\theta))$ are isolated, when $\beta \le \max\{\frac{1-\Gamma}{(N-1)(\kappa_Q+S\kappa^2)}, \frac{1-\Gamma}{2L_\Phi}\}$ and $\beta \le \frac{1}{2\kappa}$, the regret of Algorithm 5 is bounded*

$$\text{Nash-regret}^*(T) \le \frac{3C_\Phi}{c\beta(1-\Gamma)T}.$$

Proof.

$$\begin{aligned}\text{Nash-gap}(t) =& \max_i(\max_{\pi_i'} \rho_i^{\pi_i',\pi_{-i}^t} - \rho_i^{\pi_i^t,\pi_{-i}^t}) \\ \overset{(a)}{=}& \mathbb{E}_{s\sim\nu^{\pi_i,\pi_{-i}^t}} \sum_{a_i} \pi_i(a_i|s)\overline{A}_i^{\pi^t}(s,a_i) \\ \le& \max_{s,a_i} \overline{A}_i^{\pi^t}(s,a_i).\end{aligned}$$

In (a), assume i attains the maximum in $\max_i$ and π_i belongs $\arg\max_{\pi_i'}$. Apply the result in Lemma 26:

$$\begin{aligned}\Phi(\pi^{t+1}) - \Phi(\pi^t) \geq& \frac{1}{\beta}\sum_i \mathbb{E}_{s\sim\nu^{\pi_i^{t+1},\pi_{\text{-}i}^t}} \log Z_i^{t,s} \\ \geq& \frac{1-\Gamma}{\beta}\max_s \frac{c}{3}(\beta \max_{a_i} \overline{A}_i^{\pi^t}(s,a_i))^2 \geq \frac{c\beta(1-\Gamma)}{3}\text{Nash-gap}(t)^2.\end{aligned}$$

Therefore, we have

$$\frac{\Phi(\pi^T) - \Phi(\pi^0)}{T} \geq \frac{c\beta(1-\Gamma)}{3}\text{Nash-regret}^*(T).$$

Note that $\Phi(\pi^T) - \Phi(\pi^0) \leq C_\Phi$, the desired result can be shown. □