Markov α-Potential Games: Equilibrium Approximation and Regret Analysis

Xin Guo
IEOR
University of California Berkeley
Berkeley, CA 94720
xinguo@berkeley.edu

Xinyu Li*
IEOR
University of California Berkeley
Berkeley, CA 94720
xinyu_li@berkeley.edu

Chinmay Maheshwari*
EECS
University of California Berkeley
Berkeley, CA 94720
chinmay_maheshwari@berkeley.edu

Shankar Sastry
EECS
University of California Berkeley
Berkeley, CA 94720
shankar_sastry@berkeley.edu

Manxi Wu
ORIE
Cornell University
Ithaca, NY 14853
manxiwu@cornell.edu

Abstract

This paper proposes a new framework to study multi-agent interaction in Markov games: Markov α-potential games. Markov potential games are special cases of Markov α-potential games, so are two important and practically significant classes of games: Markov congestion games and perturbed Markov team games. In this paper, α-potential functions for both games are provided and the gap α is characterized with respect to game parameters. Two algorithms – the projected gradient-ascent algorithm and the sequential maximum improvement smoothed best response dynamics – are introduced for approximating the stationary Nash equilibrium in Markov α-potential games. The Nash-regret for each algorithm is shown to scale sub-linearly in time horizon. Our analysis and numerical experiments demonstrates that simple algorithms are capable of finding approximate equilibrium in Markov α-potential games.

1 Introduction

Markov potential games (MPG). It is a class of Markov game, originated in Macua et. al. [1] and further generalized by [2] as a framework to study multi-agent interaction in Markov games beyond zero sum games and common interest games. In this game, the change in utility of any player who unilaterally deviates from her policy can be evaluated by the change in the value of a potential function. Consequently, finding the Nash equilibrium of game can be reduced to solving for the global optimum of the potential function. In fact, MPG is the only class of game besides the zero-sum and common interest game for which provably convergent multi-agent reinforcement

*Corresponding authors. Authors ordered alphabetically.

Preprint. Under review.

learning (MARL) algorithms exist. Since its inception, there is an extensive body of research on MPG for the approximation and computation of the Nash equilibria [3–10, 1, 11].

Despite the promise and the potential of MPG, there is a lack of prominent applications for this class of games in the existing literature. One major reason behind this is the difficulty in verifying the existence or constructing the potential function for the game, which also is the main criticism of MPG [11]. This is especially true when the game does not satisfy certain restrictive assumptions, such as the state transition matrix being independent of players' actions, or all players' payoff functions being identical. Furthermore, recent study by [2] has shown that even for games with each stage being a static potential game and with a simple state transition structure may not be MPG.

Markov α-potential games. In this paper, we propose a new framework to study multi-agent interaction in Markov games. We introduce a less restrictive and new notion of Markov games: *Markov α-potential games* with *α-potential function* Φ, where the difference between the change of any player's long-run utility induced by a unilateral policy deviation and the change of the α-potential function value is bounded by some positive number α. MPGs are special classes of Markov α-potential games, with $\alpha = 0$.

We provide a detailed analysis of two important classes of Markov α-potential games that hold practical significance. The first is the *Markov congestion game* (MCG), which is popular in dynamic traffic routing, robotics systems, ride-hailing markets, and cloud computing. It extends the well-known static congestion games studied in [12] to a dynamic setting. The second is the *perturbed Markov team game* (PMTG) whose special case – *Markov team game* – has been extensively studied in the MARL community [13–17]. For both classes of game, we explicitly characterize their α and α-potential functions in Propositions 3.3 and 3.4. We then show that any stationary Nash equilibrium of the MPG is a α-stationary Nash equilibrium of the Markov α-potential game (Proposition 3.7).

Additionally, we introduce and analyze two algorithms for computing an approximate stationary Nash equilibrium in Markov α-potential games, along with their corresponding Nash-regret analyses. The first is the projected gradient-ascent algorithm, proposed and studied in [7] (Algorithm 1). The second algorithm, termed the *sequential maximum improvement smoothed best response*, is a new algorithm presented in this work (Algorithm 2). In the latter algorithm, players employ a smoothed sequential best response dynamics, where at each stage a player with the maximum improvement in their Q-function value is selected to update their policy as the smoothed best response. The Nash-regret for both of these algorithms is analyzed, and the dependence of the regret on the gap parameter α is presented in Theorems 4.1 and 4.2, respectively. A crucial technical step for the latter algorithm's Nash-regret analysis involves studying the *path length* of policy updates, which is bounded in terms of the potential function's change, in addition to the error arising from the gap between the Markov α-potential game and the MPG. It is worth noting that this new algorithm and its Nash-regret analysis can also be applied to compute equilibria in MPG. To the best of our knowledge, this is the first regret analysis of smoothed best response dynamics, even for MPGs. Lastly, we conduct a numerical study to illustrate and compare the convergence speed of the two proposed algorithms for the MCG and PMTG.

Related works. The closest related work to ours is on static near-potential games, as introduced in [18, 19], which is a relaxation of (static) potential games. Our work extends this static framework to a Markov setting. Furthermore, previous studies on equilibrium approximation algorithms for static near-potential games [18, 20–22] do not extend to Markov games. Furthermore, it is worth noting that the MCGs studied in our work are closely related to *Markov state-wise potential games*, where each state corresponds to a static potential game. However, existing research on Markov state-wise potential games is limited, with only a few exceptions such as [10, 11], and cannot be directly applied to the study of MCGs. In [10], the players are assumed to be myopic, whereas our work considers non-myopic players. On the other hand, [11] propose certain conditions for a Markov state-wise potential game to be a MPG, but these conditions impose restrictions such as action independence or state independence in the state transition matrix, as well as separability of players' rewards in state and action. In our analysis of MCGs, we do not impose any such restrictions on either the state transition or the reward structure. Additionally, a recent work [23] introduces an approximation algorithm for MCGs and investigates Nash-regret. However, their approach specifically considers Markov games with a finite time horizon and independent state transitions for each facility. This is fundamentally different from the infinite time horizon game examined in our paper, where we do not make any assumptions regarding the structure of state evolution. Furthermore, the results in [23] are

tailored exclusively for congestion games, whereas our work focuses on a much broader framework of Markov α-potential games. Some prior works have also focused on Markov team game in the MARL literature, as evidenced by works such as [16, 17], among others. However, the prevailing paradigm in prior research in this direction assumes that players share the same utility function. Our work on PMTGs naturally subsume Markov team games while also allowing flexibility to incorporate additional heterogeneity in the reward structure of players.

2 Setup

Consider a Markov game $\mathcal{G} = \langle I, S, (A_i)_{i\in I}, (u_i)_{i\in I}, P, \delta\rangle$, where I is a finite set of agents, S is a finite set of states s; A_i is a finite set of actions with generic member a_i for each player $i \in I$, and $a = (a_i)_{i\in I} \in A = \times_{i\in I} A_i$ is the action profile of all players; $u_i : S \times A \to \mathbb{R}$ such that $u_i(s, a)$ is the one-stage payoff for player i with state $s \in S$ and action profile $a \in A$; $P = (P(s'|s, a))_{s,s'\in S,a\in A}$ is the probability transition matrix, where $P(s'|s, a)$ is the one-step probability that the state changes from s to s' with action profile a; and $\delta \in (0, 1)$ is the discount factor. We consider *a stationary Markov policy* $\pi_i = (\pi_i(s, a_i))_{s\in S,a_i\in A_i} \in \Pi_i = \Delta(A_i)^{|S|}$, where $\pi_i(s, a_i)$ is the probability that player i chooses action a_i given state s. Let $\pi_i(s) = (\pi_i(s, a_i))_{a_i\in A_i}$ for each $i \in I$ and each $s \in S$. Additionally, denote the joint policy profile as $\pi = (\pi_i)_{i\in I} \in \Pi = \times_{i\in I}\Pi_i$, and the joint policy of all players except player i as $\pi_{-i} = (\pi_j)_{j\in I\setminus\{i\}} \in \Pi_{-i} = \times_{j\in I\setminus\{i\}}\Pi_j$. Let $\bar{A} = \max_{i\in I} |A_i|$ and $C = \max_{i\in I,s\in S,a\in A} u_i(s, a)$.

The game proceeds in discrete-time step indexed by $k = \{0, 1, ...\}$. At $k = 0$, the initial state s^0 is sampled from a probability distribution μ. At every step k, given the state s^k, each player i's action $a_i^k \in A_i$ is realized from the policy $\pi_i(s^k)$, and the realized action profile is $a^k = (a_i^k)_{i\in I}$. The state of the next step s^{k+1} is realized according to the probability transition matrix $P(\cdot|s^k, a^k)$ based on the current state s^k and action profile a^k. Given an initial state distribution μ, and a stationary policy profile π, the expected total discounted payoff for each player $i \in I$ is given by $V_i(\mu, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k u_i(s^k, a^k)\right]$ where $s^0 \sim \mu$, $a^k \sim \pi(s^k)$, and $s^k \sim P(\cdot|s^{k-1}, a^{k-1})$. For the rest of the article, with slight abuse of notation, $V_i(s, \pi)$ is used to denote the expected total payoff for player i when the initial state is a fixed state $s \in S$, and $P^\pi(\cdot|\cdot)$ denotes the transition probability given a policy π, i.e., $P^{(\pi_i,\pi_{-i})}(s'|s) = \sum_{a_{-i}\in A_{-i}} \sum_{a_i\in A_i} \pi_{-i}(s, a_{-i})\pi_i(s, a_i)P(s'|s, a)$. Given a policy $\pi \in \Pi$ and initial state distribution $\mu \in \Delta(S)$, the discounted state visitation distribution is defined as $d_\mu^\pi(s) = (1 - \delta) \sum_{k=0}^{\infty} \delta^k P(s^k = s|s^0 \sim \mu)$.

3 Markov α-potential games

This section introduces the notion of Markov α-potential games with a few examples, and presents the properties of stationary Nash equilibrium in this game framework. To begin with, recall the original definition of Markov potential games proposed by [2].

Definition 3.1 (Markov potential games (MPG) [2])**.** *A Markov game $\mathcal{G}$ is a Markov potential game if there exists a state-dependent potential function $\Phi : S \times \Pi \to \mathbb{R}$ such that for every $s \in S$,*

$$\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i}), \tag{1}$$

for any $i \in I$, any $\pi_i, \pi_i' \in \Pi_i$, and any $\pi_{-i} \in \Pi_{-i}$.

Intuitively, a game is a MPG if there exists a potential function such that when a player unilaterally deviates from her policy, the change of the potential function equals to the change of all players' total expected payoff. As pointed earlier, a key shortcoming of MPG framework is the difficulty ([11]) to check the existence or to compute Φ, thus limiting the scope of its real-world applications. Let us now relax the equality (1) so that the difference between the change of any player's utility from unilateral deviation and the change of the potential function value is bounded by some $\alpha > 0$, hence the notion of *Markov α-potential games*.

Definition 3.2 (Markov α-potential game)**.** *A Markov game $\mathcal{G}$ is a Markov α-potential game for some $\alpha > 0$, if there exists a state-dependent potential function $\Phi : S \times \Pi \to \mathbb{R}$ such that for every $s \in S$,*

$$\left|\left(\Phi\left(s, \pi_i', \pi_{-i}\right) - \Phi\left(s, \pi_i, \pi_{-i}\right)\right) - \left(V_i\left(s, \pi_i', \pi_{-i}\right) - V_i\left(s, \pi_i, \pi_{-i}\right)\right)\right| \leq \alpha, \tag{2}$$

for any $i \in I$, any $\pi_i, \pi_i' \in \Pi_i$, and any $\pi_{-i} \in \Pi_{-i}$. We refer Φ as a α-potential function.

3.1 Examples of Markov α-potential game

Markov potential game. It is clearly a Markov α-potential game with $\alpha = 0$. Furthermore, there exist two important classes of games that are widely recognized as not being Markov potential games. We will demonstrate that these classes of games can be categorized as Markov α-potential games, with explicit construction of their respective α-potential functions.

Markov congestion game. It is a natural extension of the static congestion game introduced by [24], where a finite number of players use a set of resources, and each player's reward depends on their aggregated usage of these resources. Each stage of a Markov congestion game, denoted as $\mathcal{G}_{\mathsf{mcg}}$, is a static congestion game with a state-dependent reward function of each resource, and the state transition depends on players' aggregated usage of each resource. Specifically, denote the finite set of resources in the one-stage congestion game as E. The action $a_i \in A_i \subseteq 2^E$ of each player $i \in I$ is the set of resources chosen by player i, and the action set A_i is the set of all resource combinations that are feasible for player i. The total usage demand of all players is $D > 0$, and each player's demand is $D/|I|$, for sake of simplicity and without loss of generality. Given an action profile $a = (a_i)_{i\in I}$, the aggregated usage demand of each resource $e \in E$ is given by

$$w_e(a) = \sum_{i\in I} \mathbb{1}(e \in a_i) \cdot D/|I|. \tag{3}$$

In each state s, the reward for using resource e, denoted as $c_e(s, w_e(a))$, depends on the aggregated usage demand. Thus, the one-stage payoff for player $i \in I$ in state $s \in S$ given the joint action profile $a \in A$ is $u_i(s, a) = \sum_{e\in a_i} c_e(s, w_e(a))$. The state transition probability, denoted as $P(s'|s, w)$, depends on the aggregate usage vector $w = (w_e)_{e\in E}$ induced by players' action profile as in (3). The next proposition shows that $\mathcal{G}_{\mathsf{mcg}}$ is a Markov α-potential game under a suitable Lipchtiz conditions for the state transition probability.

Proposition 3.3. *If there exists some $\zeta > 0$ such that for all $s, s' \in S$ it holds that $|P(s'|s, w) - P(s'|s, w')| \le \zeta\|w - w'\|_1$. Then the congestion game $\mathcal{G}_{\mathsf{mcg}}$ is a Markov α-potential game with $\alpha = \frac{2\zeta|S|D\delta|E|\max_{s,\pi}\Phi(s,\pi)}{|I|(1-\delta)}$, and the corresponding α-potential function Φ as*

$$\Phi(\mu, \pi) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \delta^k \Big(\sum_{e\in E} \sum_{j=1}^{w_e^k |I|/D} c_e\big(s^k, \frac{jD}{|I|}\big)\Big)\Big], \tag{4}$$

$s^0 \sim \mu$, the aggregate usage vector w^k is induced by $a^k \sim \pi(s^k)$, and $s^k \sim P(\cdot|s^{k-1}, w^{k-1})$.

Note that the Markov potential function as in (4) is the expected value of the summation of each stage's static potential function. There are two key steps to derive it: first, each static congestion game at stage k is a potential game with a potential function $\sum_{e\in E}\sum_{j=1}^{w_e^k|I|/D} c_e\left(s^k, jD/|I|\right)$; second, we establish in Lemma 7.1, that the change of state transition probability induced by a single player's policy devation decreases as the number of players increases, thus the gap between $\mathcal{G}_{\mathsf{mcg}}$ and the MPG decreases as the total demand is dispersed across more players. Furthermore, α scales linearly with respect to the Lipchitz constant ζ, the size of state space $|S|$, resource set $|E|$, and decreases as $|I|$ increases. As $|I| \to \infty$, the $\mathcal{G}_{\mathsf{mcg}}$ becomes a MPG.

Perturbed Markov team game. This is a a Markov game denoted as $\mathcal{G}_{\mathsf{pmtg}} = \langle S, (A_i)_{i\in I}, (u_i)_{i\in I}, P, \delta\rangle$, where the payoff function for each player $i \in I$ can be decomposed as $u_i(s, a) = r(s, a) + \xi_i(s, a)$, where $r(s, a)$ represents the common interest for the team, and $\xi_i(s, a)$ the perturbed payoff component which represents each player i's heterogeneous preference. In this game, the perturbed payoff component is assumed to satisfy $\|\xi_i(\cdot)\|_\infty \le \kappa$, where κ is a small positive number measuring each individual player's deviation from the team's common interest. As $\kappa \to 0$, $\mathcal{G}_{\mathsf{pmtg}}$ becomes an Markov team game, which is a MPG [2]. The next proposition shows that a $\mathcal{G}_{\mathsf{pmtg}}$ is a Markov α-potential game, and the gap α decreases in the magnitude of payoff perturbation κ. Moreover, the potential function is the total long-horizon expected team payoff.

Proposition 3.4. *A perturbed Markov team game $\mathcal{G}_{\mathsf{pmtg}}$ is a Markov α-potential game with $\alpha = \frac{2\kappa}{(1-\delta)^2}$, and the corresponding α-potential function $\Phi(\mu, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k r(s^k, a^k)\right]$, where $s^0 \sim \mu$, $a^k \sim \pi(s^k)$, and $s^k \sim P(\cdot|s^{k-1}, w^{k-1})$.*

3.2 Properties of Markov α-potential game

Now we will study the stationary Nash Equilibrium of Markov α-potential games. First, recall

Definition 3.5 (Stationary Nash equilibrium)**.** *A policy profile π^* is a stationary Nash equilibrium of $\mathcal{G}$ if for any $i \in I$, any $\pi_i \in \Pi_i$, and any $\mu \in \Delta(S)$, $V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*)$.*

That is, a stationary policy profile is a stationary Nash equilibrium of the game $\mathcal{G}$ if each player i's policy maximizes her expected total payoff given her opponents' equilibrium policy. A stationary Nash equilibrium always exists in any Markov game with finite states and actions [25]. To analyze the Markov α-potential game, we will also need

Definition 3.6 (ϵ-Stationary Nash equilibrium)**.** *For any $\epsilon \geq 0$, a policy profile π^* is an ϵ-stationary Nash equilibrium of $\mathcal{G}$ if for any $i \in I$, any $\pi_i \in \Pi_i$, and any $\mu \in \Delta(S)$, $V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon$.*

Clearly, as $\epsilon \to 0$, an ϵ-stationary Nash equilibrium becomes a stationary Nash equilibrium. Now we can establish the equilibrium properties of Markov α-potential game.

Proposition 3.7. *Consider a Markov α-potential game $\mathcal{G}$ with an α-potential function $\Phi(\mu, \pi)$. Any $\pi^* \in \Pi$ such that $\pi^* \in \arg\max_{\pi \in \Pi} \Phi(s, \pi)$ for every $s \in S$ is an α-stationary Nash equilibrium policy. Moreover, any ϵ-stationary Nash equilibrium policy in the MPG associated with Φ yields an $(\epsilon + \alpha)$-stationary Nash equilibrium for the Markov α-potential game.*

Proposition 3.7 shows that the maximizer of any α-potential function is an approximate stationary Nash equilibrium in the Markov α-potential game, and the equilibrium approximation gap ϵ depends on the gap of the potential function α. This indicates that when the α-potential function is known, any algorithm that computes the maximizer of the potential function can be used to compute an approximate stationary Nash equilibrium of the Markov α-potential game.

4 Approximation algorithms and Nash-regret analysis

In this section, we present two algorithms for computing an approximate stationary equilibrium in Markov α-potential games. First, we present projected gradient-ascent algorithm, originally proposed in [7] for MPGs. Next, we propose a new algorithm – Sequential Maximum Improvement Smoothed Best Response algorithm. We evaluate the convergence rate of both algorithms in terms of Nash-regret. Analysis in this section demonstrates that simple algorithms are capable of finding approximate equilibrium in Markov α-potential games.

4.1 Approximation algorithms

First recall some notations: Given a joint policy $\pi \in \Pi$, we define the *Q-function* $Q_i(s, a_i; \pi)$ for a player $i \in I$ as the infinite-horizon discounted utility of player i when choosing a_i in the first stage, and choosing policy π starting from the second stage given that the opponents choose π_{-i} in all stages, i.e. $Q_i(s, a_i; \pi) = \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \big(u_i(s, a_i, a_{-i}) + \delta \sum_{s' \in S} P(s'|s, a) V_i(s', \pi)\big)$. We denote the vector of Q-functions for all $a_i \in A_i$ as $Q_i(s; \pi) = (Q_i(s, a_i; \pi))_{a_i \in A_i}$. Moreover, following [3], we introduce the *smoothed* Markov game $\tilde{\mathcal{G}}$, where the one-stage expected payoff of each player i with state s and policy $\pi(s)$ is the original one-stage payoff $\mathbb{E}_{a \sim \pi(s)}[u_i(s, a)]$ along with the entropy regularizer $\nu_i(s, \pi_i) = \sum_{a_i \in A_i} \pi_i(s, a_i) \log(\pi_i(s, a_i))$. That is, $\tilde{u}_i(s, \pi) = \mathbb{E}_{a \sim \pi(s)}[u_i(s, a)] - \tau \nu_i(s, \pi_i)$, where $\tau > 0$. Under the smoothed one-stage payoffs, the expected total discounted infinite horizon payoff of player i is given by $\tilde{V}_i(s, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i)\right) | s^0 = s\right]$, the *smoothed* Q-function is given by $\tilde{Q}_i(s, a_i; \pi) = \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i})(\tilde{u}_i(s, a_i, \pi_i) + \delta \sum_{s' \in S} P(s'|s, a) \tilde{V}_i(s', \pi))$, and the *smoothed* potential function is given by $\tilde{\Phi}(s, \pi) = \Phi(s, \pi) - \tau \mathbb{E}[\sum_{i \in I} \sum_{k=0}^{\infty} \delta^k \nu_i(s^k, \pi_i) | s^0 = s]$.

Projected gradient-ascent (Algorithm 1). The algorithm iterates for T steps. In every step $t \in [T-1]$, each player $i \in I$ updates her policy following a projected gradient-ascent algorithm as in (5). This algorithm was recently proposed and analyzed in [7] for MPGs.

Algorithm 1: Projected Gradient-Ascent Algorithm

Input: Step size η, for every $i \in I, a_i \in A_i, s \in S$ set $\pi_i^{(0)}(s, a_i) = 1/|A_i|$.
for $t = 0, 1, 2, ..., T-1$ **do**
For every $i \in I, s \in S$ update the policies as follows

$$\pi_i^{(t+1)}(s) = \mathcal{P}_{\Pi_i}\left(\pi_i^{(t)}(s) + \eta Q_i^{(t)}(s)\right), \tag{5}$$

where $Q_i^{(t)}(s) = Q_i(s, \pi^{(t)})$ and $\mathcal{P}_{\Pi_i}$ denotes the orthogonal projection on Π_i.
end

Sequential maximum improvement smoothed best response (Algorithm 2). The algorithm iterates for T time steps. In every time step $t \in [T-1]$, based on the current policy profile $\pi^{(t)}$, the smoothed Q-function can be computed as $\tilde{Q}_i^{(t)}(s, \pi_i) = \sum_{a_i \in A_i} \pi_i(s, a_i) \tilde{Q}_i(s, a_i; \pi^{(t)})$ for all $s \in S$ and all $i \in I$. Then, each agent compute their one-stage best response strategy that maximizes the smoothed Q-function value: for every $i \in I, a_i \in A_i, s \in S$

$$BR_i^{(t)}(s, a_i) = \left(\underset{\pi_i' \in \Pi_i}{\arg\max} \left(\tilde{Q}_i^{(t)}(s, \pi_i') - \tau \nu_i(s, \pi_i')\right)\right)_{a_i} = \frac{\exp(\tilde{Q}_i^{(t)}(s, a_i)/\tau)}{\sum_{a_i' \in A_i} \exp(\tilde{Q}_i^{(t)}(s, a_i')/\tau)},$$

and their maximum improvement of smoothed Q-function value in comparison to current policy:

$$\Delta_i^{(t)}(s) = \max_{\pi_i'} \left(\tilde{Q}_i^{(t)}(s, \pi_i') - \tau \nu_i(s, \pi_i')\right) - \left(\tilde{Q}_i^{(t)}(s, \pi_i^{(t)}) - \tau \nu_i(s, \pi_i^{(t)})\right), \quad \forall s \in S. \tag{6}$$

If the maximum improvement $\Delta_i^{(t)}(s) \leq 0$ for all $i \in I$ and all $s \in S$, then the algorithm terminates and returns the current policy profile $\pi^{(t)}$. Otherwise, the algorithm chooses a tuple of player and state $(\bar{i}^{(t)}, \bar{s}^{(t)})$ associated with the maximum improvement value $\Delta_i^{(t)}(s)$, and updates the policy of player $\bar{i}^{(t)}$ in state $\bar{s}^{(t)}$ with the one-stage best response strategy.[2] The policies of all other players and other states remain unchanged.

Algorithm 2: Sequential Maximum Improvement Smoothed Best Response

Input: Smoothness parameter τ, for every $i \in I, a_i \in A_i, s \in S$ set $\pi_i^{(0)}(s, a_i) = 1/|A_i|$.
for $t = 0, 1, 2, ..., T-1$ **do**
Compute the maximum improvement of smoothed Q-function $\{\Delta_i^{(t)}(s)\}_{i \in I, s \in S}$ as in (6).
if $\Delta_i^{(t)}(s) \leq 0$ *for all* $i \in I$ *and all* $s \in S$ **then**
return $\pi^{(t)}$
else
Choose the tuple $(\bar{i}^{(t)}, \bar{s}^{(t)})$ with the maximum improvement

$$(\bar{i}^{(t)}, \bar{s}^{(t)}) \in \underset{i,s}{\arg\max}\, \Delta_i^{(t)}(s), \tag{7}$$

and update policy

$$\pi_{\bar{i}^{(t)}}^{(t+1)}(\bar{s}^{(t)}, a_i) = BR_{\bar{i}^{(t)}}^{(t)}(\bar{s}^{(t)}, a_i) \; \forall \, a_i \in A_{\bar{i}^{(t)}} \tag{8}$$
$$\pi_i^{(t+1)}(s) = \pi_i^{(t)}(s), \quad \forall (i, s) \neq (\bar{i}^{(t)}, \bar{s}^{(t)}).$$

end
end

4.2 Nash-regret Analysis

In this section, we present Nash-regret analysis of Algorithm 1-2 to study global non-asymptotic convergence property. Nash-regret of an algorithm is defined to be the average deviation of iterates of

[2] Any tie breaking rule can be used here if the maximum improvement is achieved by more than one tuple.

the algorithm from Nash equilibrium. Formally,

$$\text{Nash-regret}\ (T) := \frac{1}{T}\sum_{t=1}^{T}\max_{i\in I} R_i^{(t)},\ \text{where}\ R_i^{(t)} := \max_{\pi_i'\in\Pi_i} V_i(\mu,\pi_i',\pi_{-i}^{(t)}) - V_i(\mu,\pi^{(t)}).$$

Note that Nash-regret is always non-negative; and if Nash-regret$(T)\le\epsilon$ for some $\epsilon>0$ then there exists t^* such that $\pi^{(t^*)}$ is an ϵ-Nash equilibrium.

Nash-regret analysis for Algorithm 1.

Theorem 4.1. *Consider a Markov α-potential game with a α-potential function Φ and initial state distribution μ. Then the policy updates generated from Algorithm 1 with $\eta = \frac{(1-\delta)^{2.5}\sqrt{C_\Phi+|I|^2\alpha T}}{2|I||A|\sqrt{T}}$ ensures that*

$$\textit{Nash-regret}(T) \le \mathcal{O}\left(\frac{\sqrt{\tilde{\kappa}_\mu|A||I|}}{(1-\delta)^{\frac{9}{4}}}\left(\frac{C_\Phi}{T}+|I|^2\alpha\right)^{\frac{1}{4}}\right).$$

where $\tilde{\kappa}_\mu := \min_{\nu\in\Delta(\mathcal{S})}\max_{\pi\in\Pi}\|\frac{d_\mu^\pi}{\nu}\|_\infty < +\infty$ is the minmax distribution mismatch coefficient and $C_\Phi>0$ is a constant satisfying $|\Phi(\mu,\pi)-\Phi(\mu,\pi')|\le C_\Phi$ for any π,π',μ.

Proof sketch. First, from [7, Proof of Theorem 1], we note that for any $\nu\in\Delta(S)$ the Nash-regret can be bounded as the path length of policy updates of players

$$\text{Nash-regret}(T) \le \frac{3\sqrt{\tilde{\kappa}_\mu}}{\eta\sqrt{T}(1-\delta)^{\frac{3}{2}}}\times\sqrt{\sum_{t=1}^{T}\sum_{i\in I,s\in S} d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}(s)}\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\|_2^2}, \quad (9)$$

Next, we generalize [7, Lemma 3] from MPG to Markov α-potential game by appropriately accounting for difference between unilateral deviations of players and potential function via (2) and obtain

$$\frac{1}{2\eta(1-\delta)}\sum_{\substack{s\in S\\ i\in I}} d_\mu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\|_2^2 \le \Phi(\mu,\pi^{(t+1)})-\Phi(\mu,\pi^{(t)}) + \frac{4\eta^2|A|^2|I|^2}{(1-\delta)^5}+|I|^2\alpha.$$

By summing over all t and combining with (9), we obtain

$$\text{Nash-regret}(T) \le \frac{3\sqrt{\tilde{\kappa}_\mu}}{(1-\delta)^{\frac{3}{2}}}\times\sqrt{\frac{2(1-\delta)}{\eta T}(C_\Phi+2|I|^2\alpha T)+\frac{8\eta|A|^2|I|^2}{(1-\delta)^4}}$$

The claim of Theorem 4.1 follows by choosing the optimal stepsize η that minimizes the upper bound.

Nash regret analysis for Algorithm 2.

Theorem 4.2. *Consider a Markov α-potential game with a α-potential function Φ and initial state distribution μ such that $\min_{s\in S}\mu(s)=\bar{\mu}>0$. Then the policy updates generated from Algorithm 2 with parameter τ*

$$\tau = \left(\log(\bar{A})+\frac{\sqrt{\frac{2\log(\bar{A})}{1-\delta}}}{\sqrt{\alpha+\frac{C_\Phi}{T}}}+\frac{\log(\bar{A})(1-\delta)\sqrt{\bar{\mu}}}{4C\sqrt{\bar{A}}\sqrt{\alpha+\frac{C_\Phi}{T}}}\right)^{-1}$$

satisfies that

$$\textit{Nash-Regret}(T) \le \mathcal{O}\left(\frac{\sqrt{|I|\bar{A}\log(\bar{A})}}{(1-\delta)^{5/2}\sqrt{\bar{\mu}}}\left(\max\{\sqrt{\alpha},(\alpha)^{1/4}\}+\left(\frac{C_{\tilde{\Phi}}}{T}\right)^{1/4}\right)\right)$$

where $C_\Phi>0$ is a constant satisfying $|\Phi(\mu,\pi)-\Phi(\mu,\pi')|\le C_\Phi$ for any π,π',μ.

Proof sketch. First, we bound the instantaneous regret for any player $i \in I$ at time $t \in [T]$ as follows:

$$R_i^{(t)} \underset{(a)}{=} \frac{1}{1-\delta}\left(\sum_s d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s)\Delta_i^{(t)}(s) + 2\tau\log(|A|)\right) \underset{(b)}{\leq} \frac{1}{1-\delta}\left(\Delta_{\bar{i}^{(t)}}^{(t)}(\bar{s}^{(t)}) + 2\tau\log(|A|)\right),$$

where (a) is due to multi-agent performance difference lemma for smoothed game (Lemma 11(b) in [3]), and (b) is due to the fact that $\sum_s d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s) = 1$ and $\Delta_i^{(t)}(s) \leq \Delta_{\bar{i}^{(t)}}^{(t)}(\bar{s}^{(t)})$ as in (7). We emphasize that (7) is a consequence of the fact that in each stage, our Algorithm 2 selects player $\bar{i}^{(t)}$ and state $\bar{s}^{(t)}$ for policy update such that the $(\bar{i}^{(t)}, \bar{s}^{(t)})$ has the maximum one-stage improvement. Thus,

$$\text{Nash-regret}(T) = \frac{1}{T}\sum_{t\in[T]}\max_{i\in I} R_i^{(t)} \leq \frac{1}{T(1-\delta)}\sum_{t\in[T]}\left(\Delta_{\bar{i}^{(t)}}^{(t)}(\bar{s}^{(t)}) + 2\tau\log(|A|)\right). \tag{10}$$

Next, we prove the following technical lemma:

Lemma 4.3. *(a)* $\Delta_{\bar{i}^{(t)}}^{(t)}(\bar{s}^{(t)}) \leq 4C\sqrt{|A|}\frac{1+\tau\log(|A|)}{(1-\delta)}\left\|\pi_{\bar{i}_t}^{(t+1)}(\bar{s}_t) - \pi_{\bar{i}_t}^{(t)}(\bar{s}_t)\right\|_2 \quad \forall\, t \in [T].$

(b) $\sum_{t=1}^{T-1}\|\pi_{\bar{i}_t}^{(t+1)}(\bar{s}_t) - \pi_{\bar{i}_t}^{(t)}(\bar{s}_t)\|_2^2 \leq \frac{2}{\tau\bar{\mu}}\left(\tilde{\Phi}(\mu,\pi^{(T)}) - \tilde{\Phi}(\mu,\pi^{(0)}) + \alpha T\right).$

In Lemma 4.3, (a) builds on the Cauchy-Schwartz inequality and the upper bound on the total discounted payoff of the smoothed game. Additionally, (b) involves several technical steps that exploit the definition of Markov α-potential game, the performance difference lemma of the smoothed game, and the fact that Algorithm 2 is ensures that only one player is allowed to update its policy that too in one state. Moreover, by combining (10) and Lemma 4.3, we obtain

$$\text{Nash-Regret}(T) \leq \sqrt{|A|}\frac{1+\tau\log(|A|)}{\sqrt{T}(1-\delta)^2}\sqrt{\sum_{t=1}^{T-1}\left\|\pi_{\bar{i}_t}^{(t+1)}(\bar{s}_t) - \pi_{\bar{i}_t}^{(t)}(\bar{s}_t)\right\|_2^2} + \frac{2\tau\log(|A|)}{(1-\delta)}, \tag{11a}$$

$$\leq \sqrt{|A|}\frac{1+\tau\log(|A|)}{(1-\delta)^2\sqrt{\tau\bar{\mu}}}\left(\sqrt{\epsilon + \frac{C_\Phi}{T}} + |I|^{1/2}\sqrt{\frac{\tau\log(|A|)}{T}}\right) + \frac{2\tau\log(|A|)}{(1-\delta)} \tag{11b}$$

where (11a) builds on Lemma 4.3 (a) and the Cauchy-Schwartz inequality, and (11b) builds on Lemma 4.3 (b) and the fact that $\left|\left(\tilde{\Phi}(\mu,\pi) - \tilde{\Phi}(\mu,\pi')\right) - (\Phi(\mu,\pi) - \Phi(\mu,\pi'))\right| \leq \frac{2\tau|I|\log(|A|)}{1-\delta}$. By choosing the optimal τ the minimizes the right-hand-side of (11b), we obtain the regret bound in Theorem 4.2.

Remark 4.4. *As $\alpha \to 0$, Theorem 4.2 provides the Nash regret bound of Algorithm 2 in MPG. This is the first analysis of the sequential best response algorithm in MPG, and thus of independent interest.*

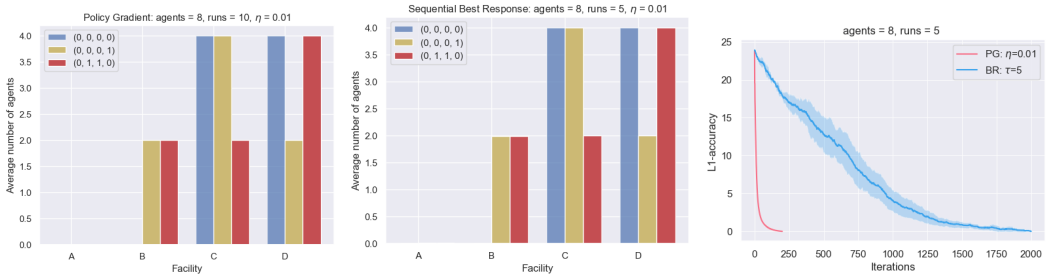
Comparison of regret bounds. In Table 1, we summarize the dependency of Nash regret of each algorithm with respect to the number of stages T, the number of players $|I|$, action set size $|A|$, discount factor δ and gap parameter α. We can see that both algorithms have the same regret dependency on T. Algorithm 1 has better regret dependency on the size of action profiles, α, and the discount factor of the game. Other other hand, Algorithm 2 algorithm has better regret dependency on the size of players.

Algorithm 1	$\mathcal{O}(\frac{1}{T^{1/4}})$	$\mathcal{O}(\|I\|)$	$\mathcal{O}(\sqrt{\|A\|})$	$\mathcal{O}\left(\frac{1}{(1-\delta)^{9/4}}\right)$	$\mathcal{O}(\alpha^{1/4})$
Algorithm 2	$\mathcal{O}(\frac{1}{T^{1/4}})$	$\mathcal{O}(\sqrt{\|I\|})$	$\mathcal{O}(\sqrt{\|A\|}\log(\|A\|))$	$\mathcal{O}\left(\frac{1}{(1-\delta)^{5/2}}\right)$	$\mathcal{O}(\max\{\sqrt{\alpha}, \alpha^{1/4}\})$

Table 1: Comparison of Nash regret of Algorithms 1 and 2.

5 Numerical experiments

This section studies the empirical performance of Algorithm 1 and Algorithm 2 for Markov congestion game (MCG) and perturbed Markov team game (PMTG) discussed in Section 3. Both algorithms



(a) (b) (c)

Figure 1: **Markov congestion game:** (a) Distribution of players taking four actions in representative states. (b) and (c) are mean L1-accuracy with shaded region of one standard deviation over all runs: (b) using policy gradient with stepsize $\eta = 0.01$; (c) using sequential best response with regularizer $\tau_t = 0.999^t \cdot 5$

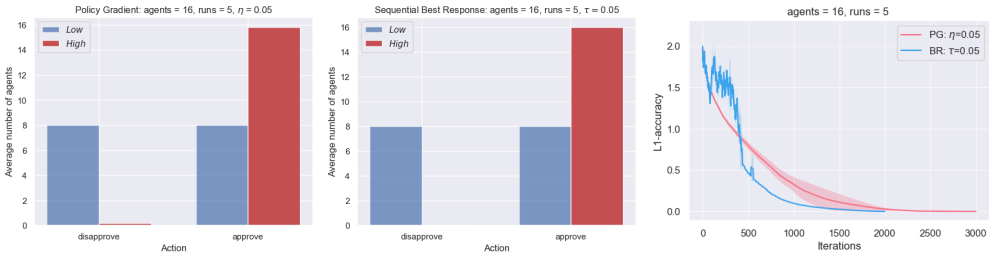
will be shown to converge in both class of games, although Algorithm 1 converges faster for PMTG while Algorithm 2 converges faster for MCG. Below are the details for the setup of the experiments and results.

MCG: We consider MCG with $|I| = 8$ players, where there are $|E| = 4$ facilities that each player can select from, i.e, $|A_i| = 4$. For each facility j, there is an associated state s_j: *normal* ($s_j = 0$) or *congested* ($s_j = 1$), and the state of the game is $s = (s_j)_{j \in E}$. The reward for each player being at facility k is equal to a predefined weight w_k^{safe} times the number of players at $k = A, B, C, D$. The weights are $w_A^{\text{safe}} = 1 < w_B^{\text{safe}} = 2 < w_C^{\text{safe}} = 4 < w_D^{\text{safe}} = 6$, i.e., facility D is most preferable by all players. However, if more than $|I|/2$ players find themselves in the same facility, then this facility transits to the *congested* state, where the reward for each player is reduced by a large constant $c = -100$. To return to the*normal* state, the facility should contain no more than $|I|/4$ players.

PMTG: We consider an experiment with $|I| = 16$ players, and there are $|A_i| = 2$ actions, *approve* ($a_i = 1$) or *disapprove* ($a_i = 0$), where each player can select; there are $S = 2$ states: *high* ($s = 1$) and *low* ($s = 0$) levels of excitement for the project. A project will be conducted if at least $\frac{|I|}{2}$ player approves. If the project is not conducted, each player's reward is 0; otherwise, each player has the common reward equal to 1 plus her individual reward. The individual reward of player i equals to the sum of $w_i \mathbf{1}_{\{a_i = s\}}$ *(not disrupting the atmosphere)* and $-w_i' a_i$ *(the cost of approving the project)*, where $w_i = 10 \cdot (|I| + 1 - i)/|I|$ and $w_i' = (i + 1)/|I|$ are predefined positive weights based on the index of players. The state transits from *high* excitement to *low* if there are less than $|I|/4$ players approving the current project; the state moves from *low* excitement to *high* if there are at least $|I|/2$ players approving the current project. For both games, we perform episodic updates with 20 steps and a discount factor $\delta = 0.99$. We estimate the Q-functions and the value functions using the average of mini-batches of size 10. And for Algorithm 2, we apply a discounting regularizer $\tau_t = \delta_\tau^t \cdot \tau$ to accelerate convergence. Figure 1a, 1b and 2a, 2b show that the players learn the expected Nash profile in selected states in all runs in both MCG and PMTG. Figure 1c and 2c depict the L1-accuracy in the policy space at each iteration which is defined as the average distance between the current policy and the final policy of all 8 players, i.e., L1-accuracy $= (1/|I|) \sum_{i \in I} \left\| \pi_i - \pi_i^{\text{final}} \right\|_1$.

6 Conclusion

We propose a new framework to study multi-agent interaction in Markov games: Markov α-potential games. This framework is demonstrated to encompass two practically relevant classes of games with diverse applications, while ensuring efficient computation of approximate equilibria through simple algorithms. Therefore, Markov α-potential games emerge as a promising game framework to be explored in the field of MARL.



(a) (b) (c)

Figure 2: **Perturbed Markov team game:** (a) and (b) are distributions of players taking actions in all states: (a) using policy gradient with stepsize $\eta = 0.05$; (b) using sequential best response with regularizer $\tau_t = 0.9975^t \cdot 0.05$. (c) is mean L1-accuracy with shaded region of one standard deviation over all runs.

References

[1] S. V. Macua, J. Zazo, and S. Zazo, "Learning parametric closed-loop policies for Markov potential games," *arXiv preprint arXiv:1802.00899*, 2018.

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

[3] C. Maheshwari, M. Wu, D. Pai, and S. Sastry, "Independent and decentralized learning in markov potential games," *arXiv preprint arXiv:2205.14590*, 2022.

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

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

[6] W. Mao, T. Başar, L. F. Yang, and K. Zhang, "Decentralized cooperative multi-agent reinforcement learning with exploration," *arXiv preprint arXiv:2110.05707*, 2021.

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

[8] R. Fox, S. M. Mcaleer, W. Overman, and I. Panageas, "Independent natural policy gradient always converges in Markov potential games," in *AISTATS*, pp. 4414–4425, PMLR, 2022.

[9] R. Zhang, J. Mei, B. Dai, D. Schuurmans, and N. Li, "On the global convergence rates of decentralized softmax gradient play in Markov potential games," in *Advances in Neural Information Processing Systems*, 2022.

[10] J. R. Marden, "State based potential games," *Automatica*, vol. 48, no. 12, pp. 3075–3088, 2012.

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

[12] R. W. Rosenthal, "A class of games possessing pure-strategy nash equilibria," *International Journal of Game Theory*, vol. 2, pp. 65–67, 1973.

[13] L. Baudin and R. Laraki, "Best-response dynamics and fictitious play in identical-interest and zero-sum stochastic games," *arXiv preprint arXiv:2111.04317*, 2021.

[14] G. Arslan and S. Yüksel, "Decentralized Q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2016.

[15] M. O. Sayin and O. Unlu, "Logit-q learning in Markov games," *arXiv preprint arXiv:2205.13266*, 2022.

[16] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal nash equilibrium in team Markov games," *Advances in neural information processing systems*, vol. 15, 2002.

[17] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonomous agents and multi-agent systems*, vol. 11, pp. 387–434, 2005.

[18] O. Candogan, A. Ozdaglar, and P. A. Parrilo, "Near-potential games: Geometry and dynamics," *ACM Transactions on Economics and Computation (TEAC)*, vol. 1, no. 2, pp. 1–32, 2013.

[19] O. Candogan, I. Menache, A. Ozdaglar, and P. A. Parrilo, "Flows and decompositions of games: Harmonic and potential games," *Mathematics of Operations Research*, vol. 36, no. 3, pp. 474–503, 2011.

[20] S. Aydın, S. Arefizadeh, and C. Eksin, "Decentralized fictitious play in near-potential games with time-varying communication networks," *IEEE Control Systems Letters*, vol. 6, pp. 1226–1231, 2021.

[21] S. Aydın and C. Eksin, "Networked policy gradient play in markov potential games," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[22] I. Anagnostides, I. Panageas, G. Farina, and T. Sandholm, "On last-iterate convergence beyond zero-sum games," in *International Conference on Machine Learning*, pp. 536–581, PMLR, 2022.

[23] Q. Cui, Z. Xiong, M. Fazel, and S. S. Du, "Learning in congestion games with bandit feedback," *arXiv preprint arXiv:2206.01880*, 2022.

[24] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124–143, 1996.

[25] D. Fudenberg, F. Drew, D. K. Levine, and D. K. Levine, *The Theory of Learning in Games*, vol. 2. MIT Press, 1998.

[26] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 4431–4506, 2021.

7 Proof of results in Section 3

Before presenting the proofs in this section we define some notation which simplify the exposition. Given a policy $(\pi_i, \pi_{-i}) \in \Pi$, and state $s, s' \in S$ we define

$$P^{\pi}(s'|s) = \sum_{a_{-i} \in A_{-i}} \pi_{-i}(a_{-i}|s) \sum_{a_i \in A_i} \pi_i(a_i|s) P(s'|s,a) \tag{12}$$

7.1 Proof of Proposition 3.3

Before presenting the proof of Proposition 3.3 we present a crucial lemma which is central to the proof.

Lemma 7.1. *If there exists some $\zeta > 0$ such that for all $s, s' \in S$ it holds that $|P(s'|s,w) - P(s'|s,w')| \leq \zeta \|w - w'\|_1$ then for any $i \in I, \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$ it holds that*

$$\|P^{\pi_i, \pi_{-i}} - P^{\pi_i', \pi_{-i}}\|_\infty \leq \frac{2\zeta |S| D \max_{a_i \in A_i} |a_i|}{|I|} \tag{13}$$

Proof. For any $i \in I, \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, $s, s' \in S$, we observe that

$$\begin{aligned}
&P^{\pi_i, \pi_{-i}}(s'|s) - P^{\pi_i', \pi_{-i}}(s'|s) \\
&\overset{(a)}{=} \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \sum_{a_i \in A_i} \pi_i(s, a_i) P(s'|s,a) - \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \sum_{a_i \in A_i} \pi_i'(s, a_i) P(s'|s,a) \\
&\overset{(b)}{=} \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \left(\sum_{a_i \in A_i} \pi_i(s, a_i) P(s'|s, w(a)) - \sum_{a_i \in A_i} \pi_i'(s, a_i) P(s'|s, w(a)) \right) \\
&\leq \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \left(\max_{a_i} P(s'|s, w(a_i, a_{-i})) - \min_{a_i} P(s'|s, w(a_i, a_{-i})) \right) \\
&= \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \left(P(s'|s, w(\bar{a}_i, a_{-i})) - P(s'|s, w(\underline{a}_i, a_{-i})) \right),
\end{aligned} \tag{14}$$

where $\bar{a}_i \in \arg\max_{a_i} P(s'|s, w(a_i, a_{-i}))$ and $\underline{a}_i \in \arg\min_{a_i} P(s'|s, w(a_i, a_{-i}))$. In the above set of equations, (a) is due to (12), (b) is due to the structure of congestion games considered in Section 3 where the transition matrix only depends action through aggregate usage vector w.

Using (14), we observe that for any $s \in S$

$$\begin{aligned}
&\sum_{s' \in S} |P^{\pi_i, \pi_{-i}}(s'|s) - P^{\pi_i', \pi_{-i}}(s'|s)| \\
&\overset{(a)}{\leq} \sum_{s' \in S} \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) | \left(P(s'|s, w(\bar{a}_i, a_{-i})) - P(s'|s, w(\underline{a}_i, a_{-i})) \right) | \\
&= \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \sum_{s' \in S} | \left(P(s'|s, w(\bar{a}_i, a_{-i})) - P(s'|s, w(\underline{a}_i, a_{-i})) \right) | \\
&\overset{(b)}{\leq} \zeta |S| \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \|w(\bar{a}_i, a_{-i}) - w(\underline{a}_i, a_{-i})\|_1 \\
&= \zeta |S| \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \sum_{e \in E} |w_e(\bar{a}_i, a_{-i}) - w_e(\underline{a}_i, a_{-i})| \\
&\overset{(c)}{=} \zeta |S| \frac{D}{|I|} \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \sum_{e \in E} |\mathbb{1}(e \in \bar{a}_i) - \mathbb{1}(e \in \underline{a}_i)| \\
&= \frac{2\zeta |S| D \max_{a_i \in A_i} |a_i|}{|I|},
\end{aligned}$$

where (a) is due to (14) and triangle inequality, (b) is due to the Lipschitz property of transition assumed in statement of Lemma 7.1 and (c) is due to the definition of aggregate usage vector in (3). This concludes the proof of the claim. □

Proof of Proposition 3.3. Recall that for any $s \in S$, the stage game is a potential game with potential function $\varphi(s,a) = \sum_{e\in E}\sum_{j=1}^{w_e(a)|I|/D} c_e\left(s, jD/|I|\right)$. That is, for any $s \in S, i \in I, a_i, a_i' \in A_i, a_{-i} \in A_{-i}$ it holds that

$$\varphi(s,a_i,a_{-i}) - \varphi(s,a_i',a_{-i}) = u_i(s,a_i,a_{-i}) - u_i(s,a_i',a_{-i}). \tag{15}$$

Under this notation we can equivalently write (4) as

$$\Phi(s,\pi) = \mathbb{E}\Big[\sum_{k=0}^{\infty}\delta^k\varphi(s^k,a^k)\Big|s_0 = s\Big] = \varphi(s,\pi) + \delta\sum_{s'\in S}P(s'|s,\pi)\Phi(s',\pi) \tag{16}$$

For any $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, using (16), we can write the difference of potential function at two policies $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi_i', \pi_{-i})$ as

$$\Phi(s,\pi) - \Phi(s,\pi') = \varphi(s,\pi) - \varphi(s,\pi') + \delta\sum_{s'\in S}\left(P(s'|s,\pi)\Phi(s',\pi) - P(s'|s,\pi')\Phi(s',\pi')\right) \tag{17}$$

Additionally, recall that the value function of player $i \in I$ with policy $\pi \in \Pi$ starting from state $s \in S$ is given by

$$V_i(s,\pi) = u_i(s,\pi) + \delta\sum_{s'\in S}P(s'|s,\pi)V_i(s',\pi) \tag{18}$$

The difference in value function of player i at two policies is given by

$$V_i(s,\pi) - V_i(s,\pi') = u_i(s,\pi) - u_i(s,\pi') + \delta\sum_{s'\in S}P(s'|s,\pi)V_i(s',\pi) - P(s'|s,\pi')V_i(s',\pi') \tag{19}$$

Subtracting (17) from (19) we obtain

$$\begin{aligned}
&\left(V_i(s,\pi) - V_i(s,\pi')\right) - \left(\Phi(s,\pi) - \Phi(s,\pi')\right)\\
&= \left(u_i(s,\pi) - u_i(s,\pi')\right) - \left(\varphi(s,\pi) - \varphi(s,\pi')\right)\\
&\quad + \delta\sum_{s'\in S}P(s'|s,\pi)\left(V_i(s',\pi) - \Phi(s',\pi)\right) - \sum_{s'\in S}P(s'|s,\pi')\left(V_i(s',\pi') - \Phi(s',\pi')\right)\\
&\overset{(a)}{=} \delta\sum_{s'\in S}P(s'|s,\pi)\left(V_i(s',\pi) - \Phi(s',\pi)\right) - \sum_{s'\in S}P(s'|s,\pi')\left(V_i(s',\pi') - \Phi(s',\pi')\right)\\
&\overset{(b)}{=} \delta\sum_{s'\in S}P(s'|s,\pi)\left(V_i(s',\pi) - V_i(s',\pi') + \Phi(s',\pi') - \Phi(s',\pi)\right)\\
&\quad - \sum_{s'\in S}\left(P(s'|s,\pi') - P(s'|s,\pi)\right)\left(V_i(s',\pi') - \Phi(s',\pi')\right),
\end{aligned}$$

where (a) is due to (15), (b) is by adding and subtracting the term $\sum_{s'\in S}P(s'|s,\pi)\left(V_i(s',\pi') - \Phi(s',\pi')\right)$.

Thus it follows

$$\begin{aligned}
&\max_{s\in S}\left|V_i(s,\pi) - V_i(s,\pi') - \left(\Phi(s,\pi) - \Phi(s,\pi')\right)\right|\\
&\le \delta\max_{s\in S}\left|V_i(s,\pi) - V_i(s,\pi') - \left(\Phi(s,\pi) - \Phi(s,\pi')\right)\right|\\
&\qquad + \delta\max_{s\in S}\left|\sum_{s'\in S}\left(\left(P(s'|s,\pi) - P(s'|s,\pi')\right)\left(\Phi(s',\pi) - V_i(s',\pi)\right)\right)\right|\\
&\le \delta\max_{s\in S}\left|V_i(s,\pi) - V_i(s,\pi') - \left(\Phi(s,\pi) - \Phi(s,\pi')\right)\right|\\
&\qquad + \delta\max_{s'\in S}\left|\Phi(s',\pi) - V_i(s',\pi)\right|\max_{s\in S}\sum_{s'\in S}\left|P(s'|s,\pi) - P(s'|s,\pi')\right|
\end{aligned}$$

Rearranging terms leads to

$$\begin{aligned}
&\max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - (\Phi(s,\pi) - \Phi(s,\pi'))| \\
&\leq \frac{\delta}{1-\delta} \max_{s'\in S} |\Phi(s',\pi) - V_i(s',\pi)| \max_{s\in S} \sum_{s'\in S} |P(s'|s,\pi) - P(s'|s,\pi')| \\
&= \frac{\delta}{1-\delta} \max_{s'\in S} |\Phi(s',\pi) - V_i(s',\pi)| \|P^{\pi} - P^{\pi'}\|_\infty
\end{aligned}$$

Using Lemma 7.1 we obtain that

$$\max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - (\Phi(s,\pi) - \Phi(s,\pi'))| \leq \frac{2\delta\zeta|S|D\max_{a_i\in A_i}|a_i|}{(1-\delta)|I|} \max_{s'\in S} |\Phi(s',\pi) - V_i(s',\pi)|$$

The claim follows by noting that for and $s' \in S$

$$\begin{aligned}
|\Phi(s',\pi) - V_i(s',\pi)| &= \left| \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \sum_{e\in E} \left(\varphi(s^k,a^k) - u_i(s^k,a^k) \right) \right] \right| \\
&\overset{(a)}{\leq} \left| \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \varphi(s^k,a^k) \right] \right| \\
&\leq \max_{s,\pi} \Phi(s,\pi),
\end{aligned}$$

where (a) follows by noting that

$$u_i(s^k,a^k) = \sum_{e\in E} c_e(s^k,w_e^k)\mathbb{1}(e\in a_i^k) \leq \sum_{e\in E} c_e(s^k,w_e^k) \leq \varphi(s^k,a^k).$$

□

7.2 Proof of Proposition 3.4

Proof. The goal is to show that for every $i \in I, \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$ it holds that

$$\max_{s\in S} |V_i(s,\pi_i,\pi_{-i}) - V_i(s,\pi_i',\pi_{-i}) - (\Phi(s,\pi_i,\pi_{-i}) - \Phi(s,\pi_i',\pi_{-i}))| \leq \frac{2\kappa}{(1-\delta)^2},$$

where

$$\Phi(s,\pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k r(s^k,a^k) | s^0 = s \right] = r(s,\pi) + \delta \sum_{s'\in S} P(s'|s,\pi)\Phi(s',\pi), \tag{20}$$

Note that the potential function Φ satisfies Using (20) we can write the difference of potential function at two policies $\pi = (\pi_i,\pi_{-i}), \pi' = (\pi_i',\pi_{-i})$ as

$$\Phi(s,\pi) - \Phi(s,\pi') = \varphi(s,\pi) - \varphi(s,\pi') + \delta \sum_{s'\in S} (P(s'|s,\pi)\Phi(s',\pi) - P(s'|s,\pi')\Phi(s',\pi')) \tag{21}$$

Similarly, the difference in value function of player i at π, π' is given as

$$V_i(s,\pi) - V_i(s,\pi') = u_i(s,\pi) - u_i(s,\pi') + \delta \sum_{s'\in S} P(s'|s,\pi)V_i(s',\pi) - P(s'|s,\pi')V_i(s',\pi') \tag{22}$$

Subtracting (21) from (22) we obtain

$$
\begin{aligned}
&(V_i(s,\pi) - V_i(s,\pi')) - (\Phi(s,\pi) - \Phi(s,\pi')) \\
&= (u_i(s,\pi) - u_i(s,\pi')) - (r(s,\pi) - r(s,\pi')) \\
&\quad + \delta \sum_{s' \in S} P(s'|s,\pi) \left(V_i(s',\pi) - \Phi(s',\pi)\right) - \sum_{s' \in S} P(s'|s,\pi') \left(V_i(s',\pi') - \Phi(s',\pi')\right) \\
&\overset{(a)}{=} (u_i(s,\pi) - u_i(s,\pi')) - (r(s,\pi) - r(s,\pi')) \\
&\quad + \delta \sum_{s' \in S} P(s'|s,\pi) \left(V_i(s',\pi) - V_i(s',\pi') + \Phi(s',\pi') - \Phi(s',\pi)\right) \\
&\quad - \sum_{s' \in S} \left(P(s'|s,\pi') - P(s'|s,\pi)\right) \left(V_i(s',\pi') - \Phi(s',\pi')\right),
\end{aligned}
$$

where (a) is by adding and subtracting the term $\sum_{s' \in S} P(s'|s,\pi) \left(V_i(s',\pi') - \Phi(s',\pi')\right)$. Thus it follows

$$
\begin{aligned}
&\max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Phi(s,\pi) - \Phi(s,\pi'))| \\
&\leq \max_{s \in S} |\left(u_i(s,\pi) - u_i(s,\pi')\right) - \left(r(s,\pi) - r(s,\pi')\right)| \\
&\quad + \delta \max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Phi(s,\pi) - \Phi(s,\pi'))| \\
&\quad\quad + \delta \max_{s \in S} \left| \sum_{s' \in S} \left(\left(P(s'|s,\pi) - P(s'|s,\pi')\right) \left(\Phi(s',\pi) - V_i(s',\pi)\right) \right) \right| \\
&\overset{(a)}{\leq} 2\kappa + \delta \max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Phi(s,\pi) - \Phi(s,\pi'))| \\
&\quad\quad + \delta \max_{s' \in S} |\Phi(s',\pi) - V_i(s',\pi)| \max_{s \in S} \sum_{s' \in S} |P(s'|s,\pi) - P(s'|s,\pi')|,
\end{aligned} \tag{23}
$$

where (a) is due to the fact

$$
\begin{aligned}
&|\left(u_i(s,\pi) - u_i(s,\pi')\right) - \left(r(s,\pi) - r(s,\pi')\right)| \\
&\leq 2 \max_{\pi \in \Pi} |u_i(s,\pi) - r(s,\pi)| \\
&\leq 2 \max_{s \in S, \pi \in \Pi} |\xi_i(s,\pi)| \leq 2\kappa.
\end{aligned}
$$

Rearranging terms in (23) we obtain

$$
\begin{aligned}
&\max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Phi(s,\pi) - \Phi(s,\pi'))| \\
&\leq \frac{2\kappa}{1-\delta} + \frac{\delta}{1-\delta} \max_{s' \in S} |\Phi(s',\pi) - V_i(s',\pi)| \max_{s \in S} \sum_{s' \in S} |P(s'|s,\pi) - P(s'|s,\pi')| \\
&= \frac{2\kappa}{1-\delta} + \frac{2\delta}{1-\delta} \max_{s' \in S} |\Phi(s',\pi) - V_i(s',\pi)|.
\end{aligned} \tag{24}
$$

Finally, we note that

$$
|\Phi(s',\pi) - V_i(s',\pi)| = |\sum_{k=0}^{\infty} \delta^k \xi_i(s,\pi(s^k))| \leq \frac{\kappa}{1-\delta}. \tag{25}
$$

Combining (24) and (25) we conclude that

$$
\max_{s \in S} |V_i(s,\pi) - V_i(s,\pi') - (\Phi(s,\pi) - \Phi(s,\pi'))| \leq \frac{2\kappa}{1-\delta} + \frac{2\delta\kappa}{(1-\delta)^2} = \frac{2\kappa}{(1-\delta)^2}
$$

□

7.3 Proof of Proposition 3.7

Proof. First, we show that any $\pi^* \in \Pi$ such that $\pi^* \in \arg\max_{\pi\in\Pi} \Phi(s,\pi)$, for every $s \in S$, is α-stationary Nash equilibrium of $\mathcal{G}$. That is, we want to show that for every $i \in I, s \in S, \pi_i \in \Pi_i$

$$V_i(s,\pi^*) - V_i(s,\pi_i,\pi^*_{-i}) \geq -\alpha.$$

To see this, we note that

$$\begin{aligned} V_i(s,\pi^*) - V_i(s,\pi_i,\pi^*_{-i}) &\overset{(a)}{\geq} \Phi(s,\pi^*) - \Phi(s,\pi_i,\pi^*_{-i}) - \alpha, \\ &\overset{(b)}{\geq} -\alpha, \end{aligned}$$

where (a) is due to (2) and (b) is due to the fact that π^* is the maximizer of potenial function Φ.

Next, we show that any ϵ-stationary Nash equilibrium for MPG associated with Φ is an $(\epsilon + \alpha)-$stationary Nash equilibrium for $\mathcal{G}$. Let $\hat{\mathcal{G}}$ be the MPG associate with Φ and let $\hat{V}$ be the value function associated with $\hat{\mathcal{G}}$. That is, for any $i \in I, \pi_i, \pi'_i \in \Pi_i, \pi_{-i}, \pi'_{-i} \in \Pi_{-i}, \mu \in \Delta(S)$ it holds that

$$\hat{V}_i(s,\pi_i,\pi_{-i}) - \hat{V}_i(s,\pi'_i,\pi_{-i}) = \Phi(s,\pi_i,\pi_{-i}) - \Phi(s,\pi'_i,\pi_{-i}). \tag{26}$$

Let $\tilde{\pi}$ be a ϵ-stationary Nash equilibrium of $\hat{\mathcal{G}}$. That is, for every $i \in I, s \in S, \pi_i \in \Pi_i$

$$\hat{V}_i(s,\tilde{\pi}) - \hat{V}_i(s,\pi_i,\tilde{\pi}_{-i}) \geq -\epsilon. \tag{27}$$

Then, the goal is to show that that for every $i \in I, s \in S, \pi_i \in \Pi_i$

$$V_i(s,\tilde{\pi}) - V_i(s,\pi_i,\tilde{\pi}_{-i}) \geq -\alpha - \epsilon.$$

Towards that goal, we note that

$$\begin{aligned} V_i(s,\tilde{\pi}) - V_i(s,\pi_i,\tilde{\pi}_{-i}) &\overset{(a)}{\geq} \Phi(s,\tilde{\pi}) - \Phi(s,\pi_i,\tilde{\pi}_{-i}) - \alpha \\ &\overset{(b)}{=} \hat{V}_i(s,\tilde{\pi}) - \hat{V}_i(s,\pi_i,\tilde{\pi}_{-i}) - \alpha \\ &\overset{(c)}{\geq} -\epsilon - \alpha, \end{aligned}$$

where (a) is due to (2), (b) is due to (26) and (c) is due to (27). □

8 Proof of results in Section 4

8.1 Proof of Theorem 4.1

Lemma 8.1 (Performance difference)**.** *For the ith player, if we fix the policy π_{-i} and any state distribution μ, then for any two policies π'_i and π_i,*

$$V_i(\mu,\pi'_i,\pi_{-i}) - V_i(\mu,\pi_i,\pi_{-i}) = \frac{1}{1-\delta} \sum_{s,a_i} d_\mu^{\pi'_i,\pi_{-i}}(s) \cdot (\pi'_i(s,a_i) - \pi_i(s,a_i))\, Q_i\left(s,a_i;\pi_i,\pi_{-i}\right).$$

Proof. It is a direct application of the performance difference lemma in [26]. □

For $i,j \in \{1,...,|I|\}$ with $i<j$, we denote by "$i \sim j$" the set of indices $\{k \mid i<k<j\}$, "$<i$" the set of indices $\{k \mid k=1,\ldots,i-1\}$, and "$>j$" the set of indices $\{k \mid k=j+1,\ldots,|I|\}$. We use the shorthand $\pi_I := \{\pi_k\}_{k\in I}$ to represent the joint policy for all players $k \in I$. For example, when $I = i \sim j, \pi_I = \{\pi_k\}_{k=i+1}^{j-1}$ is a joint policy for players from $i+1$ to $j-1$; $\pi_{<i,i\sim j}, \pi_{<i}$, and $\pi_{>j}$ can be introduced similarly.

Lemma 8.2 (Lemma 2 in [7])**.** *For any function $f^\pi : \Pi \to \mathbb{R}$, and any two policies $\pi, \pi' \in \Pi$,*

$$\begin{aligned} f^{\pi'} - f^\pi =& \sum_{i=1}^{|I|} \left(f^{\pi'_i,\pi_{-i}} - f^\pi\right) \\ &+ \sum_{i=1}^{|I|} \sum_{j=i+1}^{|I|} \left(f^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi'_j} - f^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi'_j} - f^{\pi_{<i,i\sim j},\pi'_{>j},\pi'_i,\pi_j} + f^{\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi_j}\right). \end{aligned}$$

Lemma 8.3 (Policy improvement: Markov α-potential games)**.** *For Markov α-potential game* (3.2) *with any state distribution ν, the potential function $\Phi(\nu,\pi)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 1 satisfies*

$$\Phi(\nu,\pi^{(t+1)})-\Phi(\nu,\pi^{(t)})\geq\frac{1}{2\eta(1-\delta)}\sum_{i\in I,s\in S}d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}(s)}\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\|^2 -\frac{4\eta^2|A|^2|I|^2}{(1-\delta)^5}-|I|^2\alpha$$

Proof. We let $\pi'=\pi^{(t+1)}$ and $\pi=\pi^{(t)}$ to ease the exposition. By Lemma 8.2 with $f^\pi=\Phi(\nu,\pi)$, it is equivalent to analyze

$$\Phi(\nu,\pi^{(t+1)})-\Phi(\nu,\pi^{(t)})=\mathrm{Diff}_a+\mathrm{Diff}_b$$

where

$$\begin{aligned}\mathrm{Diff}_a&=\sum_{i=1}^{|I|}\Phi(\nu,\pi_i',\pi_{-i})-\Phi(\nu,\pi),\\ \mathrm{Diff}_b&=\sum_{i=1}^{|I|}\sum_{j=i+1}^{|I|}\big(\Phi(\nu,\pi_{<i,i\sim j},\pi'_{>j},\pi_i',\pi_j')-\Phi(\nu,\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi_j')\\ &\qquad-\Phi(\nu,\pi_{<i,i\sim j},\pi'_{>j},\pi_i',\pi_j)+\Phi(\nu,\pi_{<i,i\sim j},\pi'_{>j},\pi_i,\pi_j)\big)\end{aligned}\tag{28}$$

Bounding Diff$_a$. By the property of the potential function $\Phi(\nu,\pi)$,

$$\begin{aligned}\Phi(\nu,\pi_i',\pi_{-i})-\Phi(\nu,\pi)+\alpha&\geq V_i(\nu,\pi_i',\pi_{-i})-V_i(\nu,\pi)\\ &=\frac{1}{1-\delta}\sum_{s,a_i}d_\nu^{\pi_i',\pi_{-i}}(s)\left(\pi_i'\left(s,a_i\right)-\pi_i\left(s,a_i\right)\right)Q_i\left(s,a_i;\pi_i,\pi_{-i}\right)\end{aligned}$$

where the equality follows from Lemma 8.1. The optimality of $\pi_i'=\pi_i^{(t+1)}$ in (5) leads to

$$\left\langle\pi_i'(s),Q_i(s;\pi_i,\pi_{-i})\right\rangle_{A_i}-\frac{1}{2\eta}\left\|\pi_i'(s)-\pi_i(s)\right\|^2\geq\left\langle\pi_i(s),Q_i(s;\pi_i,\pi_{-i})\right\rangle_{A_i}.$$

Combine the above two parts to get

$$\Phi(\nu,\pi_i',\pi_{-i})-\Phi(\nu,\pi)+\alpha\geq\frac{1}{2\eta(1-\delta)}\sum_s d_\nu^{\pi_i',\pi_{-i}}(s)\left\|\pi_i'(s)-\pi_i(s)\right\|^2.$$

By summing up i from 1 to $|I|$, we can get

$$\mathrm{Diff}_a\geq\frac{1}{2\eta(1-\delta)}\sum_{i=1}^{|I|}\sum_s d_\nu^{\pi_i^{(t+1)},\pi_{-i}^{(t)}}(s)\left\|\pi_i^{(t+1)}(s)-\pi_i^{(t)}(s)\right\|^2-|I|\alpha\tag{29}$$

Bounding Diff$_b$. For any $\tilde\pi_{-ij},\pi_i',\pi_i,\pi_j',\pi_j$, it holds that

$$\begin{aligned}&\Phi(\nu,\tilde\pi_{-ij},\pi_i',\pi_j')-\Phi(\nu,\tilde\pi_{-ij},\pi_i,\pi_j')-\Phi(\nu,\tilde\pi_{-ij},\pi_i',\pi_j)+\Phi(\nu,\tilde\pi_{-ij},\pi_i,\pi_j)+2\alpha\\ \geq&V_i(\nu,\tilde\pi_{-ij},\pi_i',\pi_j')-V_i(\nu,\tilde\pi_{-ij},\pi_i,\pi_j')-V_i(\nu,\tilde\pi_{-ij},\pi_i',\pi_j)+V_i(\nu,\tilde\pi_{-ij},\pi_i,\pi_j)\\ \geq&-\frac{8\eta^2|A|^2}{(1-\delta)^5}\end{aligned}$$

where the second inequality is from results in Lemma 3 in [7]. Thus,

$$\mathrm{Diff}_b\geq-\frac{|I|(|I|-1)}{2}\times\left(\frac{8\eta^2|A|^2}{(1-\delta)^5}+2\alpha\right)\geq-\frac{4\eta^2|A|^2|I|^2}{(1-\delta)^5}-|I|(|I|-1)\alpha\tag{30}$$

Combining (29) and (30) finishes the proof. □

8.1.1 Proof of Theorem 4.1

Proof. Observe that (5) is equivalent to

$$\pi_i^{(t+1)}(s) := \overset{\pi_i(s)\in\Delta(A_i)}{\operatorname{argmax}} \left\{ \left\langle \pi_i(s), Q_i^{(t)}(s) \right\rangle_{A_i} - \frac{1}{2\eta} \left\| \pi_i(s) - \pi_i^{(t)}(s) \right\|^2 \right\}. \tag{31}$$

which implies that for any $\pi_i' \in \Pi_i$,

$$\left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \eta Q_i^{(t)}(s) - \pi_i^{(t+1)}(s) + \pi_i^{(t)}(s) \right\rangle_{A_i} \le 0,$$

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

$$\begin{aligned}
&\left\langle \pi_i'(s) - \pi_i^{(t)}(s), Q_i^{(t)}(s) \right\rangle_{A_i} \\
=&\left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), Q_i^{(t)}(s) \right\rangle_{A_i} + \left\langle \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s), Q_i^{(t)}(s) \right\rangle_{A_i} \\
\le&\frac{1}{\eta} \left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\rangle_{A_i} + \left\langle \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s), Q_i^{(t)}(s) \right\rangle_{A_i}
\end{aligned}$$

Note that for any two probability distributions p_1 and p_2, $\|p_1 - p_2\| \le \|p_1 - p_2\|_1 \le 2$. Thus we can continue the above calculations by

$$\begin{aligned}
&\left\langle \pi_i'(s) - \pi_i^{(t)}(s), Q_i^{(t)}(s) \right\rangle_{A_i} \\
\le&\frac{2}{\eta} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| + \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| \left\| Q_i^{(t)}(s) \right\| \\
\le&\frac{3}{\eta} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|
\end{aligned} \tag{32}$$

where the last inequality is due to $\left\| Q_i^{(t)}(s) \right\| \le \frac{\sqrt{A}}{1-\delta}$ and $\eta \le \frac{1-\delta}{\sqrt{A}}$. Hence, by the Performance Difference Lemma 8.1 and (32),

$$\begin{aligned}
T \cdot \text{Nash-regret}(T) &= \sum_{t=1}^{T} \max_i \left(\max_{\pi_i'} V_i(\mu, \pi_i', \pi_{-i}^{(t)}) - V_i(\mu, \pi^{(t)}) \right) \\
&\overset{(a)}{=} \frac{1}{1-\delta} \sum_{t=1}^{T} \max_{\pi_i'} \sum_{s,a_i} d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s) \left(\pi_i'(s, a_i) - \pi_i^{(t)}(s, a_i) \right) Q_i^{(t)}(s, a_i) \\
&\overset{(b)}{\le} \frac{3}{\eta(1-\delta)} \sum_{t=1}^{T} \sum_{s} d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|
\end{aligned}$$

where in the equality (a) we slightly abuse the notation i to represent $\arg\max_i$ and in (b) we slightly abuse the notation π_i' to represent $\arg\max_{\pi_i'}$. Now we proceed the above calculation by choosing an arbitrary $\nu \in \Delta(S)$ and using the following inequality

$$\frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \le \frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{(1-\delta)\nu(s)} \le \frac{\sup_{\pi\in\Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}{1-\delta}$$

to get:

$$\begin{aligned}
&T \cdot \text{Nash-regret}(T) \\
\le&\frac{3\sqrt{\sup_{\pi\in\Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}}{\eta(1-\delta)^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_{s} \sqrt{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s) \times d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| \\
\overset{(a)}{\le}& \frac{3\sqrt{\sup_{\pi\in\Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}}{\eta(1-\delta)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_{s} d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T} \sum_{s} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2} \\
\overset{(b)}{\le}& \frac{3\sqrt{\sup_{\pi\in\Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}}{\eta(1-\delta)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_{s} d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)} \times \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{|I|} \sum_{s} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2}
\end{aligned}$$

where (a) follows from Cauchy-Schwartz inequality and in (b) we replace i ($\arg\max_i$) by the sum over all players.

Now we proceed the above calculation with $\nu^* = \arg\min_{\nu\in\Delta(S)} \max_{\pi\in\Pi} \|\frac{d_\mu^\pi}{\nu}\|_\infty$ and apply Lemma 8.3 to get

$T \cdot \text{Nash-regret}(T)$

$$\leq \frac{3\sqrt{\widetilde{\kappa}_\mu}}{\eta(1-\delta)^{\frac{3}{2}}}\sqrt{T} \times \sqrt{2\eta(1-\delta)\left(\Phi(\nu^*,\pi^{(T)}) - \Phi(\nu^*,\pi^{(0)})\right) + \frac{8\eta^3|A|^2|I|^2}{(1-\delta)^4}T + 2\eta(1-\delta)|I|^2\alpha\cdot T}$$

$$\leq \frac{3\sqrt{\widetilde{\kappa}_\mu}}{\eta(1-\delta)^{\frac{3}{2}}}\sqrt{T} \times \sqrt{2\eta(1-\delta)C_\Phi + \frac{8\eta^3|A|^2|I|^2}{(1-\delta)^4}T + 2\eta(1-\delta)|I|^2\alpha\cdot T}$$

$$= 3\sqrt{\frac{2\widetilde{\kappa}_\mu T(C_\Phi + |I|^2\alpha\cdot T)}{\eta(1-\delta)^2} + \frac{8\widetilde{\kappa}_\mu\eta T^2|A|^2|I|^2}{(1-\delta)^7}}$$

where the second inequality follows from $\|\Phi(\nu^*,\pi) - \Phi(\nu^*,\pi')\| \leq C_\Phi$ for any π,π'. By taking step size $\eta = \frac{(1-\delta)^{2.5}\sqrt{C_\Phi + |I|^2\alpha T}}{2|I|A\sqrt{T}}$, we can complete the proof with

$$\text{Nash-regret}(T) \leq \frac{3\cdot 2^{\frac{3}{2}}\sqrt{\tilde{\kappa}_\rho A|I|}}{(1-\delta)^{\frac{9}{4}}}\left(\frac{C_\Phi}{T} + |I|^2\alpha\right)^{\frac{1}{4}}$$

□

8.2 Proof of Theorem 4.2

Before presenting the proof of Theorem 4.2, we some crucial lemmas.

Lemma 8.4. *If $\mathcal{G}$ is Markov α-potential game with Φ as its α-potential then $\tilde{\mathcal{G}}$ is also a Markov α-potential game with $\tilde{\Phi}$ as its α-potential where for any $s\in S, \pi\in\Pi$*

$$\tilde{\Phi}(s,\pi) = \Phi(s,\pi) - \tau\mathbb{E}[\sum_{i\in I}\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i)]$$

Proof. We want to show that if for all $s\in S, i\in I, \pi_i', \pi_i\in\Pi_i, \pi_{-i}\in\Pi_{-i}$

$$\left|\left(\Phi\left(s,\pi_i',\pi_{-i}\right) - \Phi\left(s,\pi_i,\pi_{-i}\right)\right) - \left(V_i\left(s,\pi_i',\pi_{-i}\right) - V_i\left(s,\pi_i,\pi_{-i}\right)\right)\right| \leq \alpha,$$

then

$$\left|\left(\tilde{\Phi}\left(s,\pi_i',\pi_{-i}\right) - \tilde{\Phi}\left(s,\pi_i,\pi_{-i}\right)\right) - \left(\tilde{V}_i\left(s,\pi_i',\pi_{-i}\right) - \tilde{V}_i\left(s,\pi_i,\pi_{-i}\right)\right)\right| \leq \alpha.$$

It is sufficient to show that for all $s\in S, i\in I, \pi_i', \pi_i\in\Pi_i, \pi_{-i}\in\Pi_{-i}$

$$\begin{aligned}&\left(\tilde{\Phi}\left(s,\pi_i',\pi_{-i}\right) - \tilde{\Phi}\left(s,\pi_i,\pi_{-i}\right)\right) - \left(\tilde{V}_i\left(s,\pi_i',\pi_{-i}\right) - \tilde{V}_i\left(s,\pi_i,\pi_{-i}\right)\right)\\ &= \left(\Phi\left(s,\pi_i',\pi_{-i}\right) - \Phi\left(s,\pi_i,\pi_{-i}\right)\right) - \left(V_i\left(s,\pi_i',\pi_{-i}\right) - V_i\left(s,\pi_i,\pi_{-i}\right)\right).\end{aligned}$$

To see this, recall that for every $i\in I, \pi_i\in\Pi_i, \pi_{-i}\in\Pi_{-i}, \mu\in\Delta(S)$

$$\tilde{V}_i(\mu,\pi_i,\pi_{-i}) = V_i(\mu,\pi_i,\pi_{-i}) - \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i)|s^0\sim\mu\right]. \tag{33}$$

Thus, we observe that for every $s\in S, i\in I, \pi_i', \pi_i\in\Pi_i, \pi_{-i}\in\Pi_{-i}$

$$\begin{aligned}&\left(\tilde{\Phi}\left(s,\pi_i',\pi_{-i}\right) - \tilde{\Phi}\left(s,\pi_i,\pi_{-i}\right)\right) - \left(\tilde{V}_i\left(s,\pi_i',\pi_{-i}\right) - \tilde{V}_i\left(s,\pi_i,\pi_{-i}\right)\right)\\ &= \left(\Phi(s,\pi_i',\pi_{-i}) - \tau\mathbb{E}[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i')] - \Phi(s,\pi_i,\pi_{-i}) + \tau\mathbb{E}[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i)]\right)\\ &- \left(V_i(s,\pi_i',\pi_{-i}) - \tau\mathbb{E}[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i')] - V_i(s,\pi_i,\pi_{-i}) + \tau\mathbb{E}[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i)]\right)\\ &= \left(\Phi\left(s,\pi_i',\pi_{-i}\right) - \Phi\left(s,\pi_i,\pi_{-i}\right)\right) - \left(V_i\left(s,\pi_i',\pi_{-i}\right) - V_i\left(s,\pi_i,\pi_{-i}\right)\right).\end{aligned}$$

This concludes the proof. □

Lemma 8.5. *[[3, Lemma 11]] For any $i \in I, \mu \in \Delta(S)$, $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi_i', \pi_{-i})$ it holds that*

$$\tilde{V}_i(\mu, \pi') - \tilde{V}_i(\mu, \pi) = \frac{1}{1-\delta} \sum_{s' \in S} d_\mu^{\pi'}(s') \left((\pi_i' - \pi_i)^\top \tilde{Q}_i(s; \pi) + \tau \nu_i(s, \pi_i) - \tau \nu_i(s, \pi_i') \right). \quad (34)$$

Lemma 8.6. *For any $i \in I, s \in S, \pi \in \Pi$ the update $\pi_i^+ = \arg\max_{\pi_i' \in \Pi_i} \left(\tilde{Q}_i(s)^\top \pi_i' - \tau \nu_i(s, \pi_i') \right)$ satisfies*

$$\sum_{a_i \in A_i} \tilde{Q}_i(s, a_i) \left(\pi_i^+(s, a_i) - \pi_i'(s, a_i) \right) \geq \tau \sum_{a_i \in A_i} \log \left(\pi_i^+(s, a_i) \right) \left(\pi_i^+(s, a_i) - \pi_i'(s, a_i) \right), \quad \forall\, \pi_i' \in \Pi_i$$

Proof. From the first order conditions of optimality for the update $\pi_i^+ = \arg\max_{\pi_i' \in \Pi_i} \left(\tilde{Q}_i^{(t)}(s, \pi_i') - \tau \nu_i(s, \pi_i') \right)$ it holds that

$$\left(\tilde{Q}_i(s) - \tau \nabla_{\pi_i} \nu_i(s, \pi_i^+) \right)^\top \left(\pi_i^{(+)}(s) - \pi_i'(s) \right) \geq 0, \quad \forall\, \pi_i' \in \Pi_i.$$

Finally, we note that for any $s \in S, i \in I, a_i \in A_i$ it holds that $\nabla_{\pi_i(s,a_i)} \nu_i(s, \pi_i) = 1 + \log(\pi_i(s, a_i))$. Therefore,

$$\begin{aligned} \sum_{a_i} \tilde{Q}_i(s, a_i) \left(\pi_i^+(s, a_i) - \pi_i'(s, a_i) \right) &\geq \tau \sum_{a_i} \left(1 + \log \left(\pi_i^+(s, a_i) \right) \right) \left(\pi_i^+(s, a_i) - \pi_i'(s, a_i) \right), \quad \forall\, \pi_i' \in \Pi_i \\ &\geq \tau \sum_{a_i} \log \left(\pi_i^+(s, a_i) \right) \left(\pi_i^+(s, a_i) - \pi_i'(s, a_i) \right), \quad \forall\, \pi_i' \in \Pi_i, \end{aligned}$$

where in the last inequality we are able to drop 1 because the policies add upto 1. □

Lemma 8.7. *For any $i \in I, s \in S, \pi_i, \pi_i' \in \Pi_i$ it holds that*

$$\nu_i(s, \pi_i) - \nu_i(s, \pi_i') \geq \sum_{a_i \in A_i} \left(\log(\pi_i'(s, a_i)) \right) \left(\pi_i(s, a_i) - \pi_i'(s, a_i) \right) + \frac{1}{2} \|\pi_i(s) - \pi_i'(s)\|^2$$

Proof. To prove the lemma, we first claim that entropy $\pi \mapsto \nu_i(s, \pi)$ is a 1-strongly convex function for every $s \in S$. This can be observed by computing the Hessian which is a $\mathbb{R}^{A_i \times A_i}$ diagonal matrix with (a_i, a_i) entry as $1/\pi(s, a_i)$. Note that since $\pi_i(s, a_i) \leq 1$ it follows that the diagonal entries of Hessian matrix are all greater than 1. Thus, $\nu_i(s, \pi)$ is 1-strongly convex function.

The result in Lemma follows by noting that for any κ-strongly convex function f it holds that

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\kappa}{2} \|y - x\|^2.$$

□

Lemma 8.8. *For any $i \in I, \mu \in \Delta(S), \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$*

$$\left| (V_i(\mu, \pi_i, \pi_{-i}) - V_i(\mu, \pi_i', \pi_{-i})) - \left(\tilde{V}_i(\mu, \pi_i, \pi_{-i}) - \tilde{V}_i(\mu, \pi_i', \pi_{-i}) \right) \right| \leq \frac{2\tau \log |A_i|}{1 - \delta}$$

Proof. The claim follows by expanding the definition of smoothed infinite horizon utility. Indeed, for every $i \in I, \mu \in \Delta(S), \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, we observe that

$$\begin{aligned}
&\left|(V_i(\mu,\pi_i,\pi_{-i}) - V_i(\mu,\pi_i',\pi_{-i})) - \left(\tilde{V}_i(\mu,\pi_i,\pi_{-i}) - \tilde{V}_i(\mu,\pi_i',\pi_{-i})\right)\right| \\
&= \Bigg| (V_i(\mu,\pi_i,\pi_{-i}) - V_i(\mu,\pi_i',\pi_{-i})) \\
&\quad - \left(V_i(\mu,\pi_i,\pi_{-i}) - \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i)\right)\right] - V_i(\mu,\pi_i',\pi_{-i}) + \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i')\right)\right]\right)\Bigg| \\
&= \left|\left(-\tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i)\right)\right] + \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i')\right)\right]\right)\right| \\
&\leq \frac{2\tau\max_{s,\pi_i}\nu_i(s,\pi_i)}{(1-\delta)} = \frac{2\tau\log(|A_i|)}{(1-\delta)}.
\end{aligned}$$

□

Lemma 8.9. *For any $i \in I, t \in [T], a_i \in A_i$*

$$\tau|\log(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i))| \leq 2\|\tilde{Q}^{(t^*)}_{\bar{i}(t)}(s)\|_\infty + \tau\log(|A_i|)$$

Proof. We note that there exists $t^* \leq t$ when player $\bar{i}(t)$ updated its state $\bar{s}^{(t)}$ before time t. Therefore,

$$\begin{aligned}
\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) &= \frac{\exp(\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i))}{\sum_{a_i'\in A_i}\exp(\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i'))} \\
&\geq \frac{\exp(\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},\underline{a}_i)/\tau)}{|A_i|\exp(\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},\bar{a}_i)/\tau)} \\
&= \frac{1}{|A_i|}\exp\left(\left(\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},\underline{a}_i) - \tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},\bar{a}_i)\right)/\tau\right),
\end{aligned}$$

where $\bar{a}_i \in \arg\max_{a_i\in A_i}\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i)$ and $\underline{a}_i \in \arg\min_{a_i\in A_i}\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i)$. Consequently, we obtain that

$$\tau|\log(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i))| \leq 2\|\tilde{Q}^{(t^*)}_{\bar{i}(t)}(s)\|_\infty + \tau\log(|A_i|).$$

□

Lemma 8.10. *For any $t \in [T], i \in I, s \in S$*

$$\|\tilde{Q}^{(t)}_i(s)\|_\infty \leq C\frac{1+\tau\log(|A_i|)}{(1-\delta)},$$

where $C = \max_{s,i\in I,a\in A}|u_i(s,a)|$.

Proof. First, we show that $\max_{s\in S}|\tilde{V}_i(s,\pi)| \leq C\frac{1+\tau\log(|A_i|)}{(1-\delta)}$. Indeed,

$$\begin{aligned}
\max_{s\in S}|\tilde{V}_i(s,\pi)| &= \max_{s\in S}\left|\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\tilde{u}_i(s^k,a^k)\right]\right| \\
&\leq \max_{s\in S}\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k|\tilde{u}_i(s^k,a^k)|\right] \\
&\leq \max_{s\in S}\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(|u_i(s^k,a^k)| + \tau\log(|A_i|)\right)\right] \\
&\leq C\frac{1+\tau\log(|A_i|)}{(1-\delta)}.
\end{aligned}$$

Finally, from the definition of $\tilde{Q}_i$ it follows that

$$\begin{aligned}
|\tilde{Q}_i(s,a_i)| &= \left| \sum_{a_{-i}\in A_{-i}} \pi_{-i}(s,a_{-i})(\tilde{u}_i(s,a_i,\pi_i) + \delta \sum_{s'\in S} P(s'|s,a)\tilde{V}_i(s',\pi)) \right| \\
&\leq \sum_{a_{-i}\in A_{-i}} \pi_{-i}(s,a_{-i}) \left| \tilde{u}_i(s,a_i,\pi_i) + \delta \sum_{s'\in S} P(s'|s,a)\tilde{V}_i(s',\pi) \right| \\
&\leq \sum_{a_{-i}\in A_{-i}} \pi_{-i}(s,a_{-i}) \left(\left| \tilde{u}_i(s,a_i,\pi_i) \right| + \delta \sum_{s'\in S} P(s'|s,a) \left| \tilde{V}_i(s',\pi) \right| \right) \\
&\leq C \sum_{a_{-i}\in A_{-i}} \pi_{-i}(s,a_{-i}) \left((1+\tau\log(|A_i|)) + \frac{\delta}{1-\delta}(1+\tau\log(|A_i|)) \right) \\
&= C\frac{1+\tau\log(|A_i|)}{(1-\delta)}.
\end{aligned}$$

□

8.2.1 Proof of Lemma 4.3

Proof of Lemma 4.3. (a) We note that for any $t \in [T]$

$$\begin{aligned}
\Delta^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}) &= \frac{1}{1-\delta}\Bigg(\sum_{a_i\in A_{\bar{i}(t)}} \left(\left(\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right) \tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right) \\
&\qquad + \tau\nu_{\bar{i}(t)}(\bar{s}^{(t)}, \pi^{(t)}_{\bar{i}(t)}) - \tau\nu_{\bar{i}(t)}(\bar{s}^{(t)}, \pi^{(t+1)}_{\bar{i}(t)}) \Bigg) \\
&\overset{(a)}{\leq} \frac{1}{1-\delta}\Bigg(\sum_{a_i\in A_{\bar{i}(t)}} \left(\left(\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right) \tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right) \\
&\qquad + \tau\Big(\sum_{a_i\in A_{\bar{i}(t)}} \log(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i)) \left(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right) \Big) \Bigg) \\
&= \frac{1}{1-\delta}\Bigg(\sum_{a_i\in A_{\bar{i}(t)}} \left(\left(\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right) \left(\tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \tau\log(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i)) \right) \right) \\
&\leq \frac{1}{1-\delta}\Bigg(\sum_{a_i\in A_{\bar{i}(t)}} \left(\left| \left(\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right) \right| \left| \left(\tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \tau\log(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i)) \right) \right| \right) \\
&\overset{(b)}{\leq} \frac{1}{1-\delta}\sqrt{\bar{A}} \max_{a_i\in A_i} \left| \left(\tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \tau\log(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i)) \right) \right| \left\| \pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right\|_2 \\
&\leq \frac{1}{1-\delta}\sqrt{\bar{A}} \left(\max_{a_i\in A_i} \left| \tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right| + \max_{a_i\in A_i} \tau \left| \log(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i)) \right| \right) \left\| \pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right\|_2 \\
&\overset{(c)}{\leq} \frac{1}{1-\delta}\sqrt{\bar{A}} \left(\|\tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\|_\infty + 2\|\tilde{Q}^{(t^*)}_{\bar{i}(t)}(\bar{s}^{(t)})\|_\infty + \tau\log(\bar{A}) \right) \left\| \pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right\|_2 \\
&\overset{(d)}{\leq} 4C\frac{1+\tau\log(\bar{A})}{1-\delta}\sqrt{\bar{A}} \left\| \pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)},a_i) \right\|_2
\end{aligned}$$

where (a) follows due to convexity of $\nu_i(s,\cdot)$ and (b) follows due to Cauchy-Schwartz inequality, (c) follows due to Lemma 8.9 and (d) is due to Lemma 8.10.

(b) Here, we show that

$$\sum_{t=1}^{T-1} \|\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\|_2^2 \leq \frac{2}{\tau\bar{\mu}} \left(\tilde{\Phi}(\mu,\pi^{(T)}) - \tilde{\Phi}(\mu,\pi^{(0)}) + \alpha T \right).$$

To see this, we note that

$$
\begin{aligned}
&\tilde{\Phi}(\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}) - \tilde{\Phi}(\pi^{(t)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}) \\
&\overset{(a)}{=} \tilde{V}_{\bar{i}(t)}(\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}) - \tilde{V}_{\bar{i}(t)}(\pi^{(t)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}) - \alpha \\
&\overset{(b)}{=} \frac{1}{1-\delta} \sum_{s\in S, a\in A_{\bar{i}(t)}} d^{\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}}(s) \Bigg((\pi^{(t+1)}_{\bar{i}(t)}(s,a) - \pi^{(t)}_{\bar{i}(t)}(s,a)) \tilde{Q}^{(t)}_{\bar{i}(t)}(s, a_i) \\
&\qquad\qquad + \tau \nu_{\bar{i}(t)}(s, \pi^{(t)}_{\bar{i}(t)}) - \tau \nu_{\bar{i}(t)}(s, \pi^{(t+1)}_{\bar{i}(t)}) \Bigg) - \alpha \\
&\overset{(c)}{=} \frac{1}{1-\delta} \sum_{a\in A_{\bar{i}(t)}} d^{\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}}(\bar{s}^{(t)}) \Bigg((\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}, a) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}, a)) \tilde{Q}^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}, a) \\
&\qquad\qquad + \tau \nu_{\bar{i}(t)}(\bar{s}^{(t)}, \pi^{(t)}_{\bar{i}(t)}) - \tau \nu_{\bar{i}(t)}(\bar{s}^{(t)}, \pi^{(t+1)}_{\bar{i}(t)}) \Bigg) - \alpha
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(d)}{\geq} \frac{\tau}{1-\delta} \sum_{a\in A_{\bar{i}(t)}} d^{\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}}(\bar{s}^{(t)}) \log(\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}, a)) (\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}, a) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}, a)) \\
&\quad + \frac{\tau}{1-\delta} \sum_{a\in A_{\bar{i}(t)}} d^{\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}}(\bar{s}^{(t)}) \log(\pi^{(t+1)}_{i}(\bar{s}^{(t)}, a)) \left(\pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}, a) - \pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}, a) \right) \\
&\quad + \frac{\tau}{2(1-\delta)} d^{\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}}(\bar{s}^{(t)}) \|\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\|^2 - \alpha \\
&= \frac{\tau}{2(1-\delta)} d^{\pi^{(t+1)}_{\bar{i}(t)}, \pi^{(t)}_{-\bar{i}(t)}}(\bar{s}^{(t)}) \|\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\|^2 - \alpha \\
&\overset{(e)}{\geq} \frac{\tau \mu(\bar{s}^{(t)})}{2} \|\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\|^2 - \alpha,
\end{aligned}
$$

where (a) follows from Lemma 8.4, (b) follows from Lemma 8.5, (c) follows from the fact that $\pi^{(t+1)}_{\bar{i}(t)}(s) = \pi^{(t)}_{\bar{i}(t)}(s)$ for all $s \neq \bar{s}^{(t)}$, (d) follows from Lemma 8.6 and 8.7, (e) follows by noting that $d^{\pi^{(t+1)}_i, \pi^{(t)}_{-i}}(\bar{s}^{(t)}) \geq (1-\delta)\mu(\bar{s}^{(t)})$.

□

8.2.2 Proof of Theorem 4.2

Proof. Note that for all i it holds that

$$
R^{(t)}_i = \max_{\pi'_i \in \Pi_i} V_i(\mu, \pi'_i, \pi^{(t)}_{-i}) - V_i(\mu, \pi^{(t)}) = V_i(\mu, \pi^{\dagger}_i, \pi^{(t)}_{-i}) - V_i(\mu, \pi^{(t)}),
$$

where $\pi^{\dagger}_i \in \arg\max_{\pi'_i \in \Pi_i} V_i(\mu, \pi'_i, \pi^{(t)}_{-i})$. Using Lemma 8.8 we obtain

$$
R^{(t)}_i \leq \tilde{V}_i(\mu, \pi^{\dagger}_i, \pi^{(t)}_{-i}) - \tilde{V}_i(\mu, \pi^{(t)}) + \frac{2\tau \log(\bar{A})}{(1-\delta)}
$$

Next, we note that for any $i \in I, \mu \in \Delta(S)$ it holds that

$$
\begin{aligned}
&\tilde{V}_i(\mu, \pi_i^\dagger, \pi_{-i}^{(t)}) - \tilde{V}_i(\mu, \pi_i^{(t)}, \pi_{-i}^{(t)}) \\
&\overset{(a)}{=} \frac{1}{1-\delta} \sum_{s,a_i} d_\mu^{\pi_i^\dagger, \pi_{-i}(t)}(s) \left(\left(\pi_i^\dagger(s,a_i) - \pi_i^{(t)}(s,a_i) \right) \tilde{Q}_i^{(t)}(s,a_i) + \tau\nu_i(s,\pi_i^{(t)}) - \tau\nu_i(s,\pi_i') \right) \\
&\overset{(b)}{\leq} \frac{1}{1-\delta} \sum_{s} d_\mu^{\pi_i^\dagger, \pi_{-i}(t)}(s) \left(\max_{\pi_i'} \sum_{a_i \in A_i} \left(\left(\pi_i'(s,a_i) - \pi_i^{(t)}(s,a_i) \right) \tilde{Q}_i^{(t)}(s,a_i) \right) + \tau\nu_i(s,\pi_i^{(t)}) - \tau\nu_i(s,\pi_i') \right), \\
&\overset{(c)}{=} \frac{1}{1-\delta} \sum_{s} d_\mu^{\pi_i^\dagger, \pi_{-i}(t)}(s) \left(\Delta_i^{(t)}(s) \right), \\
&\overset{(d)}{\leq} \frac{1}{1-\delta} \sum_{s} d_\mu^{\pi_i^\dagger, \pi_{-i}(t)}(s) \left(\Delta_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) \right), \\
&\overset{(e)}{=} \frac{1}{1-\delta} \left(\Delta_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) \right),
\end{aligned}
$$

where (a) is due to Lemma 8.1 and (b) is due to the fact that $d_\mu^{\pi_i^\dagger, \pi_{-i}(t)}(s) \geq 0$, (c) is by (6), (d) is because $\Delta_i^{(t)}(s) \leq \Delta_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)})$ for all $i \in I, s \in S$, (e) is because $\sum_s d_\mu^{\pi_i^\dagger, \pi_{-i}(t)}(s) = 1$. To summarize, we obtain

$$
R_i^{(t)} \leq \frac{1}{1-\delta} \left(\Delta_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) + 2\tau \log(\bar{A}) \right).
$$

Therefore,

$$
\begin{aligned}
\text{Nash-Regret}(T) &\leq \frac{1}{T(1-\delta)} \sum_{t \in [T]} \left(\Delta_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) + 2\tau \log(\bar{A}) \right) \\
&\overset{(a)}{\leq} \frac{4C\sqrt{\bar{A}}(1+\tau\log(\bar{A}))}{T(1-\delta)^2} \sum_{t \in [T]} \left\| \pi_{\bar{i}(t)}^{(t+1)}(\bar{s}^{(t)}, a_i) - \pi_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}, a_i) \right\|_2 + \frac{2\tau\log(\bar{A})}{(1-\delta)} \\
&\overset{(b)}{\leq} \frac{4C\sqrt{\bar{A}}(1+\tau\log(\bar{A}))}{\sqrt{T}(1-\delta)^2} \sqrt{\sum_{t \in [T]} \left\| \pi_{\bar{i}(t)}^{(t+1)}(\bar{s}^{(t)}) - \pi_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) \right\|_2^2} + \frac{2\tau\log(\bar{A})}{(1-\delta)} \\
&\overset{(c)}{\leq} \frac{8C\sqrt{\bar{A}}(1+\tau\log(\bar{A}))}{\sqrt{\tau\bar{\mu}}(1-\delta)^2} \sqrt{\alpha + \frac{C_{\tilde{\Phi}}}{T}} + \frac{2\tau\log(\bar{A})}{(1-\delta)} \\
&\overset{(d)}{\leq} \frac{8C\sqrt{\bar{A}}(1+\tau\log(\bar{A}))}{\sqrt{\tau\bar{\mu}}(1-\delta)^2} \sqrt{\alpha + \frac{C_{\Phi}}{T} + \frac{2\tau|I|\log(\bar{A})}{T(1-\delta)}} + \frac{2\tau\log(\bar{A})}{(1-\delta)} \\
&\overset{(e)}{\leq} \frac{8C\sqrt{\bar{A}}(1+\tau\log(\bar{A}))}{\sqrt{\tau\bar{\mu}}(1-\delta)^2} \left(\sqrt{\alpha + \frac{C_{\Phi}}{T}} + \sqrt{\frac{2\tau|I|\log(\bar{A})}{T(1-\delta)}} \right) + \frac{2\tau\log(\bar{A})}{(1-\delta)}
\end{aligned}
$$

where (a) is due to Lemma 4.3(a), (b) is due to Cauchy-Schwartz inequality, (c) is due to Lemma 4.3(b), (d) is the fact that $|C_\Phi - C_{\tilde{\Phi}}| \leq \frac{2\tau|I|\log(|A|)}{1-\delta}$, and (e) is due to the fact that for any two positive scalars x, y it holds that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$.

For ease of exposition, we define

$$
D_1 = \frac{8C\sqrt{\bar{A}}}{\sqrt{\bar{\mu}}(1-\delta)^2}, \quad D_2 = \sqrt{\alpha + \frac{C_{\tilde{\Phi}}}{T}}, \quad D_3 = \sqrt{\frac{2\log(\bar{A})}{(1-\delta)}}
$$

Then, it follows that

$$
\begin{aligned}
\text{Nash-Regret}(T) &\leq \left(\frac{D_1}{\sqrt{\tau}} + \sqrt{\tau} D_1 \log(\bar{A}) \right) D_2 + \left(\frac{D_1}{\sqrt{\tau}} + \sqrt{\tau} D_1 \log(\bar{A}) \right) D_3 \sqrt{|I|} \sqrt{\frac{\tau}{T}} + \tau D_3^2 \\
&= \sqrt{\tau} \left(D_1 D_2 \log(\bar{A}) \right) + \tau \left(D_1 D_3 \log(\bar{A}) \sqrt{\frac{|I|}{T}} + D_3^2 \right) + \frac{D_1 D_2}{\sqrt{\tau}} + \frac{D_1 D_3 \sqrt{|I|}}{\sqrt{T}}
\end{aligned}
$$

Consider $\tau < 1$ then it holds that $\tau \leq \sqrt{\tau}$. Thus, we obtain

$$\text{Nash-Regret}(T) = \sqrt{\tau}\left(D_1 D_2 \log(\bar{A}) + D_1 D_3 \log(\bar{A})\sqrt{\frac{|I|}{T}} + D_3^2\right) + \frac{D_1 D_2}{\sqrt{\tau}} + \frac{D_1 D_3\sqrt{|I|}}{\sqrt{T}}$$

If we select τ to be

$$\tau = \sqrt{\frac{D_1 D_2}{D_1 D_2 \log(\bar{A}) + D_1 D_3 \log(\bar{A})\sqrt{\frac{|I|}{T}} + D_3^2}},$$

which is strictly less than 1, it follows that

$$\begin{aligned}
\text{Nash-Regret}(T) &\leq \sqrt{D_1^2 D_2^2 \log(\bar{A}) + D_1^2 D_2 D_3 \log(|A|)\sqrt{\frac{|I|}{T}} + D_1 D_2 D_3^2} + \frac{D_1 D_3\sqrt{|I|}}{\sqrt{T}}, \\
&\leq D_1 D_2\sqrt{\log(\bar{A})} + D_1\sqrt{D_2 D_3 \log(|\bar{A}|)}\left(\frac{|I|}{T}\right)^{1/4} + \sqrt{D_1 D_2} D_3 + \frac{D_1 D_3\sqrt{|I|}}{\sqrt{T}}
\end{aligned}$$

Note that $D_3 \geq 1$ and additionally, if we assume that $D_1 \geq 1$ (assuming C is large enough) then it follows that

$$\begin{aligned}
&\text{Nash-Regret}(T) \leq 2D_1 D_3\sqrt{\log(\bar{A})}\left(D_2 + \sqrt{D_2}\left(1 + \left(\frac{|I|}{T}\right)^{1/4}\right) + \sqrt{\frac{|I|}{T}}\right) \\
&\leq \frac{16C\sqrt{\bar{A}}}{\sqrt{\bar{\mu}}(1-\delta)^2}\sqrt{\frac{2\log(\bar{A})}{(1-\delta)}}\sqrt{\log(\bar{A})}\left(\left(\alpha + \frac{C_{\tilde{\Phi}}}{T}\right)^{1/2} + \left(\alpha + \frac{C_{\tilde{\Phi}}}{T}\right)^{1/4}\left(1 + \left(\frac{|I|}{T}\right)^{1/4}\right) + \sqrt{\frac{|I|}{T}}\right) \\
&\leq \frac{16C\sqrt{\bar{A}}}{\sqrt{\bar{\mu}}(1-\delta)^2}\sqrt{\frac{2\log(\bar{A})}{(1-\delta)}}\sqrt{\log(\bar{A})}\left(\left(\alpha + \frac{C_{\tilde{\Phi}}}{T}\right)^{1/2} + \left(\alpha + \frac{C_{\tilde{\Phi}}}{T}\right)^{1/4}\left(1 + \left(\frac{|I|}{T}\right)^{1/4}\right) + \sqrt{\frac{|I|}{T}}\right) \\
&\leq \mathcal{O}\left(\frac{\sqrt{|I|\bar{A}}\log(\bar{A})}{(1-\delta)^{5/2}\sqrt{\bar{\mu}}}\left(\left(\alpha + \frac{C_{\tilde{\Phi}}}{T}\right)^{1/2} + \left(\alpha + \frac{C_{\tilde{\Phi}}}{T}\right)^{1/4}\right)\right) \\
&\leq \mathcal{O}\left(\frac{\sqrt{|I|\bar{A}}\log(\bar{A})}{(1-\delta)^{5/2}\sqrt{\bar{\mu}}}\left(\sqrt{\alpha} + \sqrt{\frac{C_{\tilde{\Phi}}}{T}} + (\alpha)^{1/4} + \left(\frac{C_{\tilde{\Phi}}}{T}\right)^{1/4}\right)\right) \\
&\leq \mathcal{O}\left(\frac{\sqrt{|I|\bar{A}}\log(\bar{A})}{(1-\delta)^{5/2}\sqrt{\bar{\mu}}}\left(\max\{\sqrt{\alpha}, (\alpha)^{1/4}\} + \left(\frac{C_{\tilde{\Phi}}}{T}\right)^{1/4}\right)\right).
\end{aligned}$$

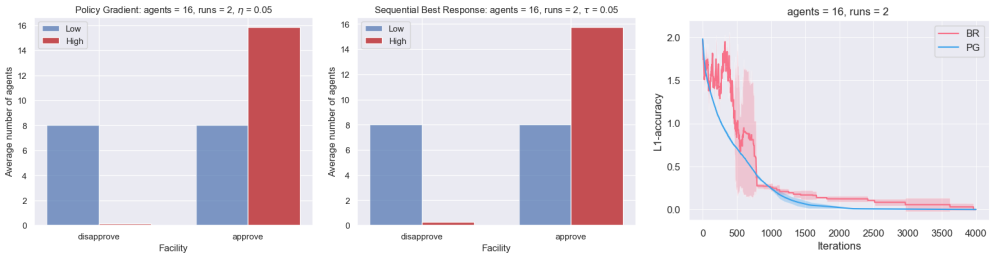
This completes the proof. □

9 Additional Experiments

In this section, we expand on the numerical experiments discussed in Section 5 and investigate the impact of random transitions on the performance of Algorithm 1 and 2. To this end, we simulate a PMTG with the same parameters as before, i.e., $|I| = 16$ agents, and $|A_i| = 2$ possible actions for each agent $i \in I$. However, we introduce a stochastic transition rule based on a logistic function:

$$\begin{aligned}
P\left(\text{High}|\text{Low}\right) &= \frac{1}{1 + \exp\left(-\kappa\left(n(a) - \frac{|I|}{2}\right)\right)} \\
P\left(\text{High}|\text{High}\right) &= \frac{1}{1 + \exp\left(-\kappa\left(n(a) - \frac{|I|}{4}\right)\right)}
\end{aligned}$$

where $n(a)$ denotes the number of agents who approve the project, and κ modulates the steepness of the transition function as it passes through its midpoint.



(a) (b) (c)

Figure 3: **Perturbed Markov team game ($\kappa = 50$):** (a) and (b) are distributions of players taking actions in all states: (a) using policy gradient with stepsize $\eta = 0.05$; (b) using sequential best response with regularizer $\tau_t = 0.9975^t \cdot 0.05$. (c) is mean L1-accuracy with shaded region of one standard deviation over all runs.

We set $\kappa = 50$ as the parameter in the logistic transition function. We apply a regularizer of the form $\tau_t = 0.9975^t \cdot 0.05$ in Algorithm 2 and a fixed step size $\eta = 0.05$ in Algorithm 1. As shown in Figures 3a and 3b, both algorithms successfully converged to deterministic Nash policies, despite the randomness in transitions. Figure 3c further illustrates that the rate of convergence for the two algorithms is similar in this particular problem setting.