Independent and Decentralized Learning in Markov Potential Games*

Chinmay Maheshwari†, Manxi Wu†, Druv Pai, and Shankar Sastry ‡

Abstract

We propose a multi-agent reinforcement learning dynamics, and analyze its convergence properties in infinite-horizon discounted Markov potential games. We focus on the independent and decentralized setting, where players can only observe the realized state and their own reward in every stage. Players do not have knowledge of the game model, and cannot coordinate with each other. In each stage of our learning dynamics, players update their estimate of a perturbed Q-function that evaluates their total contingent payoff based on the realized one-stage reward in an asynchronous manner. Then, players independently update their policies by incorporating a smoothed optimal one-stage deviation strategy based on the estimated Q-function. A key feature of the learning dynamics is that the Q-function estimates are updated at a faster timescale than the policies. We prove that the policies induced by our learning dynamics converge to a stationary Nash equilibrium in Markov potential games with probability 1. Our results build on the theory of two timescale asynchronous stochastic approximation, and new analysis on the monotonicity of potential function along the trajectory of policy updates in Markov potential games.

1 Introduction

Multi-agent Reinforcement Learning (MARL) focuses on analyzing the strategic interactions among multiple players in a dynamic environment, where the utilities and state transition are

*This version: May, 2022.

†Equal contribution and alphabetical order.

‡C. Maheshwari (chinmay_maheshwari@berkeley.edu), D. Pai (druvpai@berkeley.edu), and S. Sastry (shankar_sastry@berkeley.edu) are with the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. M. Wu (manxiwu@cornell.edu) is with the School of Operations Research and Information Engineering at Cornell University, and the Simons Institute for Theory of Computing at the University of California, Berkeley.

jointly determined by players' actions. Such problems are of practical importance in many areas including autonomous driving (Shalev-Shwartz et al. (2016)), adaptive traffic control (Prabuchandran et al. (2014); Bazzan (2009)), e-commerce (Kutschinski et al. (2003)), and AI training in real-time strategy games (Vinyals et al. (2019); Brown and Sandholm (2018)).

Markov games and the associated solution concept – stationary Nash equilibrium – have been proposed to analyze the outcomes of multi-agent interactions in dynamic environment (Shapley (1953); Littman (1994); Hu and Wellman (2003); Fudenberg and Tirole (1991)). Researchers have extensively studied various equilibrium computation algorithms and equilibrium learning dynamics in Markov games. The proposed approaches can be roughly classified into two categories: The first category includes the *coordinated* approaches that require a centralized coordinator to enable communication among players, and to jointly determine the policy updates (Littman et al. (2001); Perolat et al. (2015); Wei et al. (2017); Bai and Jin (2020); Bai et al. (2020); Leonardos et al. (2021); Sidford et al. (2020)). The second category includes the *independent or decentralized* approaches, where no coordination is allowed among players. Studies that fall into this category, although relatively sparse, have attracted more and more attention in recent years (Perolat et al. (2018); Daskalakis et al. (2020); Sayin et al. (2020, 2021)).

Contributions. In this article, we propose an independent and decentralized multi-agent reinforcement learning dynamics, and prove its convergence to stationary Nash equilibrium in infinite-horizon discounted Markov potential games. In our learning dynamics, each player maintains a q-estimate for every state-action pair in every stage, which evaluates the player's total reward contingent on playing that action in the first stage with that state while following the current policy for the rest of the game (represented as the Q-function defined in Sec. 2). In each stage, players update their q-estimate based on the state at that stage, their own action and reward. Then, players update their policies based on the updated q-estimate. In particular, our learning dynamics has the following features:

1. The dynamics is *decentralized*, and only requires the *minimum information*: Players only observe the state and their realized payoff in every stage. They do not need to know the existence of other players, and do not need any knowledge of the utility functions or the state transition matrix.

2. Players *independently* adjust their policies without communication or coordination with others.

3. Players are *self-interested* in that their updated policy incorporates a *smoothed optimal one-stage deviation* that maximize the expected contingent payoff derived from the current

q-estimate. Additionally, the update of the q-estimates adds a *small perturbation* in every stage reward. This reward perturbation turns out to be crucial for the convergence of our dynamics.

4. The dynamics is *asynchronous*. In every stage, only the q-estimate of the realized state-action pair is updated, and only the policy corresponding to the realized state is updated. The remaining elements in q-estimate and policy remain unchanged.[1] Furthermore, each player counts the number of times a state and an action is realized. The stepsize of updating the element corresponds to each state and action is asynchronously adjusted according to the counter.

5. The dynamics have *two timescale*: the q-estimate is updated at a faster timescale compared to the policy update. That is, the speed of players learning their estimated contingent payoff is asymptotically faster than the speed of policy updates.

We prove that our learning dynamics leads to an approximate Nash equilibrium with probability 1, and this approximate Nash equilibrium becomes a Nash equilibrium when the payoff perturbation goes to zero. This convergence result indicates that players' self-interested policy adjustment can naturally leads to an equilibrium in Markov potential game under the minimum information environment. No coordination is needed so long as players' policies are updated slower than their q-estimates.

In the proof of convergence theorem, we apply the theory of asynchronous stochastic approximation to exploit the time separation between the updates of q-estimate and policy (Borkar (2009); Tsitsiklis (1994); Perkins and Leslie (2013)). In particular, we show that our setting satisfies a set of verifiable conditions provided by Perkins and Leslie (2013), which guarantee that the analysis of the discrete time dynamics can be analyzed by the corresponding continuous time dynamical system. In this dynamical system, the fast dynamics – update of the q-estimates – can be analyzed while viewing the policy updates (the slow dynamics) as static. Therefore, the update of q-estimates convergence reduces to the convergence of standard Q-learning in a single-agent reinforcement problem, and we prove that the q-estimate of each player converges to their perturbed Q-function.

On the other hand, the analysis of the slow dynamics – the policy updates – can take the q-estimate as convergent. We construct a perturbed potential function to serve as the Lyapunov function of the continuous time dynamical system associated with the policy

[1] As a result, if a state or an action is never visited, the corresponding q-estimate and policy remain unchanged. We ensure that every state-action pair is repeatedly realized for each agent by imposing a mild irreducibility assumption on the state transition matrix, and adopting the smoothed optimal one-stage deviation in policy update that as a full support on the action set.

updates. We prove that the value of the perturbed potential function increases along the trajectory of policy updates. As a result, the policy updates lead to a global optimum of the perturbed potential function, which is an approximate Nash equilibrium. The approximate Nash equilibrium becomes a Nash equilibrium as the payoff perturbation goes to zero.

Literature review (brief).[2] Our proposed learning dynamics and convergence result contribute to the growing literature of multi-agent reinforcement learning in Markov potential games (Leonardos et al. (2021); Zhang et al. (2021); Song et al. (2021)). In a Markov potential game, there exists a potential function that describes the change of each player' expected total payoff under unilateral policy deviation. As a result, a stationary Nash equilibrium policy can be solved as the global optimum of the potential function, and gradient-based algorithms have been proposed to compute the stationary Nash equilibrium (Zhang et al. (2021); Leonardos et al. (2021)) for discounted infinite horizon setting and for finite horizon episodic setting (Mao et al. (2021); Song et al. (2021)). Our focus is different in that we analyze how stationary Nash equilibrium can naturally arise through repeated policy adjustment by self-interested players in a decentralized manner.

The paper Perolat et al. (2018) considered a similar actor-critic based decentralized learning dynamics, and showed its converges to Nash equilibrium in cooperative multistage games – a special case of Markov games such that each stage is visited only once, and the state transition has a tree structure with finite depth. There are two key differences between our learning dynamics and theirs: (a) we consider updates with asynchronous stepsizes that are adjusted based on counters of each state and each state-action pair; (b) the q-estimate update in our dynamics introduces a reward perturbation in each stage. It turns out that these two differences are crucial for us to achieve equilibrium convergence in a much more general setting – Markov potential games with infinite stages and no restrictions on state transition other than irreducibility. Moreover, since the proof techniques developed in Perolat et al. (2018) exploits the special structure of multistage games, they cannot be applied in our setting. Instead, we develop a new approach that involves recent development of asynchronous stochastic approximation theory, a novel construction of perturbed potential function, and new policy convergence proof in infinite-horizon Markov games.

Additionally, the MARL dynamics proposed by Sayin et al. (2021) for zero-sum discounted Markov games shares similar features with ours – their MARL dynamics is also decentralized, independent, asynchronous, and has two timescale. Our learning dynamics and convergence analysis differ from Sayin et al. (2021) in two aspects:

- The dynamics in Sayin et al. (2021) updates q-estimates at a slower timescale, and update

[2] A complete literature review is provided in the appendix.

policies faster, while our policy update is slower compared to the q-estimate update. The analysis in their paper exploits the properties of zero-sum Markov games to demonstrate the convergence of policy updates while holding q-estimate as constant.

- The dynamics in Sayin et al. (2021) only counts the number of times a state is visited, but does not count the number of times a state-action pair is visited. The step-size in their q-estimate update for each state-action pair is adjusted based on the counter of states, and is normalized by the probability of taking that action given the policy in that stage. On the other hand, our step-size in q-estimate update is asynchronously adjusted for every state-action pair based on the associated counter. As a result, their analysis of asynchronous updates and time separation is different from ours, and requires a different set of assumptions on stepsizes.

Finally, our results also advances the rich literature of learning in stateless potential game that includes continuous and discrete time best response dynamics (Monderer and Shapley (1996b); Swenson et al. (2018)), fictitious play (Monderer and Shapley (1996a); Hofbauer and Sandholm (2002); Marden et al. (2009)), replicator dynamics(Panageas and Piliouras (2016); Hofbauer and Sigmund (2003)), no-regret learning (Heliou et al. (2017); Krichene et al. (2014)), and payoff-based learning (Cominetti et al. (2010)). In particular, our learning dynamics share similar spirit with the payoff-based learning dynamics in stateless potential game (Cominetti et al. (2010)). In a payoff-based learning dynamics, players update the estimate of payoffs associated with each action in a decentralized manner only based on their own payoff in each stage, and update their mixed strategy by incorporating a smoothed best response given the payoff estimate. In a Markov potential game, due to the joint evolvement between state and actions, we need to introduce the time separation, and keep track of the payoff (q-estimate) for every state-action pair in an asynchronous manner. Additionally, the smoothed best response mixed strategy becomes the smoothed optimal one-stage deviation policy. We emphasize that our analysis approach is completely different from that in stateless potential games. Furthermore, our construction of perturbed potential function is new, and our policy convergence proof uses tools in reinforcement learning. Therefore, our results in Markov potential games are not direct extension of the ones in stateless potential games.

Organization. In Sec. 2, we define the Markov potential game and stationary Nash equilibrium. We present our learning dynamics and convergence result in Sec. 3. We present a numerical example in Sec. 4. Proofs of the results are included in the appendix.

2 Game Model

Markov games. Consider a game $\mathcal{G}$ with finite set of players I who interact with one another in a discrete-time dynamic environment over an infinite time horizon. We characterize the Markov game by a tuple $\mathcal{G} = \langle S, (A_i)_{i\in I}, (u_i)_{i\in I}, P, \delta\rangle$, where

- S is a finite set of states s.
- A_i is a finite set of actions 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.
- $u_i : S \times A \to \mathbb{R}$ is the one-stage payoff (i.e. utility) function of player i, where $u_i(s, a)$ is the one-stage payoff of player i when the state is $s \in S$, and the action profile is $a \in A$.[3] Players *may not* know their own utility function or the utility functions of the opponents.
- $P = (P(s'|s, a))_{s,s'\in S,a\in A}$ is the state transition matrix, where $P(s'|s, a)$ is the probability that state changes from s to s' with action profile a. Players *may not* know the transition matrix.
- $\delta \in (0, 1)$ 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 with 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, we 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 $\pi_{-i} = (\pi_j)_{j\in I\setminus\{i\}} \in \Pi_{-i} = \times_{j\in I\setminus\{i\}}\Pi_j$.

The game proceeds in discrete-time stages indexed by $k = \{0, 1, ...\}$. At $k = 0$ the initial state s^0 is sampled from a distribution μ. At every time k, given the state s^k, each player'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 stage s^{k+1} is realized according to the 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 of each player $i \in I$ is given by:

$$V_i(\mu, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k u_i(s^k, a^k)\right], \tag{1}$$

[3]All our results hold when the reward is randomly realized according to $u_i(s, a) + \xi_i(s, a)$, where $\xi_(s, a)$ is a noise term with zero mean and bounded support.

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, we use $V_i(s, \pi)$ to denote the expected total utility of player i when the initial state is a fixed state $s \in S$. Then, we have

$$V_i(\mu, \pi) = \sum_{s \in S} \mu(s) V_i(s, \pi). \tag{2}$$

We assume that the initial state distribution μ has full support on S, and there exists at least one action profile such that the induced Markov process of state transition is irreducible.

Assumption 2.1. *The initial state distribution $\mu(s) > 0$ for all $s \in S$. Additionally, there exists a joint action $a \in A$ such that the induced Markov chain with probability transition $(P(s'|s, a))_{s,s' \in S}$ is irreducible and aperiodic.*

Q-function, and one-stage optimal deviation. Given any policy $\pi \in \Pi$, and any initial state $s \in S$, we define the following *Q-function* for each player $i \in I$ and action $a_i \in A_i$:

$$Q_i(s, a_i; \pi) = \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \left(u_i(s, a_i, a_{-i}) + \delta \sum_{s' \in S} P(s'|s, a_i, a_{-i}) V_i(s', \pi) \right). \tag{3}$$

In (3), player i's expected utility in the first stage with state s is derived from playing action a_i and her opponents' play π_{-i}. The expected total utility starting from stage 2 is derived from all players following policy π. Therefore, the Q-function $Q_i(s, a_i; \pi)$ represents player i's expected total utility when the game starts in state s, and player i deviates from her policy to play a_i in the first stage. Then, player i's *optimal one-stage deviation* from policy π in state s is $\mathrm{br}_i(s; \pi)$ given by:

$$\mathrm{br}_i(s; \pi) = \arg\max_{\hat{\pi}_i(s)} \sum_{a_i \in A_i} \hat{\pi}_i(s, a_i) Q_i(s, a_i, \pi), \quad \forall s \in S. \tag{4}$$

We denote player i's optimal one-stage deviation for all states as $\mathrm{br}_i(\pi) = (\mathrm{br}_i(s; \pi))_{s \in S}$.

Nash equilibrium and Markov potential games.

Definition 2.2 (Stationary Nash equilibrium policy)**.** *A policy profile π^* is a stationary ϵ-Nash equilibrium policy of $\mathcal{G}$ with $\epsilon \geq 0$ if for any $i \in I$, any $\pi_i \in \Pi_i$, and any $\mu \in \Delta(S)$,*

$$V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon.$$

As $\epsilon \to 0$, an ϵ-Nash equilibrium becomes a Nash equilibrium.

That is, the equilibrium policy π_i^* of each player i maximizes their expected total utility given the opponents' policy π_{-i}^*. A stationary Nash equilibrium policy always exists in a Markov game with finite states and actions (Fudenberg and Tirole (1991)).

Definition 2.3 (Markov potential games — Leonardos et al. (2021)). *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 $s \in S$,*

$$\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i}), \tag{5}$$

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

Definition 2.3 shows that in a Markov potential game, there exists a potential function such that if a single agent changes their policy, then the change of the value of the potential function equals to the change of the deviator's expected total utility.[4] Moreover, given the initial state distribution μ, we define $\Phi(\mu, \pi) := \sum_{s \in S} \mu(s)\Phi(s, \pi)$. Then, from (5), we have:

$$\Phi(\mu, \pi_i', \pi_{-i}) - \Phi(\mu, \pi_i, \pi_{-i}) = V_i(\mu, \pi_i', \pi_{-i}) - V_i(\mu, \pi_i, \pi_{-i}),\ \forall i \in I,\ \forall \pi_i, \pi_i^{'} \in \Pi_i,\ \forall \pi_{-i} \in \Pi_{-i}.$$

As a result, the global maximizer $\pi^* = \arg\max_{\pi \in \Pi} \Phi(\mu, \pi)$ must be a stationary Nash equilibrium policy of the Markov potential game. However, directly computing the Nash equilibrium as the maximizer of $\Phi(\mu, \pi)$ is challenging due to the fact that $\Phi(\mu, \pi)$ can be non-linear and non-concave. Also, optimization tools that are used for computing the global maximizer typically require centralized coordination or communication among agents. Instead, we show that a stationary Nash equilibrium policy naturally arises from our proposed two timescale learning dynamics when players are learning in an independent and decentralized manner.

3 Independent and decentralized learning dynamics

In this section, we present the independent and decentralized learning dynamics, and analyze its convergence property.

Information environment of the learning dynamics. Each player $i \in I$ knows the state set S, and their own action set A_i. Players may not know the state transition probability matrix P, their own or others utility functions $\{u_i\}_{i \in I}$. Players may not even know the

[4]We can analogously define the weighted Markov potential game: A Markov game $\mathcal{G}$ is a weighted Markov potential game if there exists a state-dependent weighted potential function $\Phi(s, \pi)$ and a weight vector $w = (w_i)_{i \in I} \in \mathbb{R}_+^I$ such that $\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = w_i(V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i}))$ for any $i \in I$, any $\pi_i, \pi_i' \in \Pi_i$, and any $\pi_{-i} \in \Pi_{-i}$. All results presented in our paper can be extended to weighted Markov potential games.

existence of other players. In each stage $k = 0, 1, 2, \dots$, players observe the realized state s^k, and their own realized reward $r_i^k = u_i(s^k, a^k)$. Players do not observe the actions or the rewards of their opponents.

Learning dynamics (Algorithm 1). In our learning dynamics, each player $i \in I$ updates four components $\{n^k, \tilde{n}_i^k, \tilde{q}_i^k, \pi_i^k\}_{k=0}^{\infty}$ in every stage k. In particular, $n^k = (n^k(s))_{s \in S}$ is the vector of state counters, where $n^k(s)$ is the number of times state s is realized before stage k. For each player $i \in I$, $\tilde{n}_i^k = (\tilde{n}_i^k(s, a_i))_{s \in S, a_i \in A_i}$ is the counter of state-action tuple, where $\tilde{n}_i^k(s, a_i)$ is the number of times that player i has played action a_i in state s before stage k. Additionally, $\tilde{q}_i^k = (\tilde{q}_i^k(s, a_i))_{s \in S, a_i \in A_i}$ is player i's q-estimate of her perturbed Q-function (to be defined next), and $\pi_i^k = (\pi_i^k(s, a_i))_{s \in S, a_i \in A_i}$ is player i's policy in stage k. Before discussing the updates of $\{\tilde{q}_i^k, \pi_i^k\}_{k=0}^{\infty}$, we first describe the perturbed Markov game as preparation.

Perturbed Q-function, and smoothed one-stage optimal deviation. We define a perturbed Markov game $\tilde{\mathcal{G}}$, where the one-stage expected utility of player $i \in I$ with state s and policy $\pi(s)$ is

$$\tilde{u}_i(s, \pi) = \mathbb{E}_{a \sim \pi(s)}[u_i(s, a)] - \tau \nu_i(s, \pi_i),$$

where $\tau > 0$, and $\nu_i(s, \pi_i)$ is a regularizer given by

$$\nu_i(s, \pi_i) = \sum_{a_i \in A_i} \pi_i(s, a_i) \log(\pi_i(s, a_i)). \tag{10}$$

Analogous to (1) and (3), given any initial state s and any policy profile π, the expected total discounted utility function of player i is given by:

$$\tilde{V}_i(s, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i)\right)\right],$$

where $a^k \sim \pi(s^k)$, and $s^k \sim P(\cdot | s^{k-1}, a^{k-1})$.

Additionally, the perturbed Q-function $\tilde{Q}_i(s, a_i; \pi)$ for each $i \in I$, $s \in S$ and $a_i \in A_i$ is given by:

$$\tilde{Q}_i(s, a_i; \pi) = \sum_{a_{-i} \in A_{-i}} \pi_{-i}(s, a_{-i}) \left(u_i(s, a_i, a_{-i}) - \tau \nu_i(s, \pi_i) + \delta \sum_{s' \in S} P(s' | s, a_i, a_{-i}) \tilde{V}_i(s', \pi) \right). \tag{11}$$

We note that as $\tau \to 0$, $\tilde{V}_i(s, \pi) \to V_i(s, \pi)$ and $\tilde{Q}_i(s, a_i; \pi) \to Q_i(s, a_i; \pi)$. The *smoothed*

Algorithm 1 Independent and decentralized learning dynamics

Initialization: $n^0(s) = 0, \forall s \in S$; $\tilde{n}_i^0(s, a_i) = 0, \tilde{q}_i^0(s, a_i) = 0$ and $\pi_i^0(s, a_i) = 1/|A_i|,\ \forall (i, a_i, s)$.

In stage 0, each player observes s^0, choose their action $a_i^0 \sim \pi_i^0(s^0)$, and observe $r_i^0 = u_i(s^0, a^0)$.

In every step $k = 1, 2, ...$, each player observes s^k, and independently updates $\{n_i^k, \tilde{n}_i^k, \tilde{q}_i^k, \pi_i^k\}$.

Update $n^k, \tilde{n}_i^k$:

$$n^k(s^{k-1}) = n^{k-1}(s^{k-1}) + 1, \quad n^k(s) = n^{k-1}(s),\ \forall s \neq s^{k-1}, \tag{6}$$

$$\tilde{n}_i^k(s^{k-1}, a_i^{k-1}) = \tilde{n}_i^{k-1}(s^{k-1}, a_i^{k-1}) + 1, \quad \tilde{n}_i^k(s, a) = \tilde{n}_i^{k-1}(s, a),\ \forall\ (s, a) \neq (s^{k-1}, a_i^{k-1}). \tag{7}$$

Update $\tilde{q}_i^k$:

$$\begin{aligned} \tilde{q}_i^k(s^{k-1}, a_i^{k-1}) = \tilde{q}_i^{k-1}(s^{k-1}, a_i^{k-1}) + \alpha(\tilde{n}_i^k(s^{k-1}, a_i^{k-1})) \cdot \Bigg(r_i^{k-1} \\ - \tau\nu_i(s^{k-1}, \pi_i^{k-1}) + \delta \sum_{a_i \in A_i} \pi_i^{k-1}(s^k, a_i)\tilde{q}_i^{k-1}(s^k, a_i) - \tilde{q}_i^{k-1}(s^{k-1}, a_i^{k-1}) \Bigg), \end{aligned} \tag{8a}$$

$$\tilde{q}_i^k(s, a_i) = \tilde{q}_i^{k-1}(s, a_i), \quad \forall\ (s, a_i) \neq (s^{k-1}, a_i^{k-1}), \tag{8b}$$

where $r_i^{k-1} = u_i(s^{k-1}, a^{k-1})$, and $\nu_i(s^{k-1}, \pi_i^{k-1})$ is given by (10).

Update π_i^k:

$$\begin{aligned} \pi_i^k(s^{k-1}, a_i) = \pi_i^{k-1}(s^{k-1}, a_i) + \beta(n^k(s^{k-1}))\cdot \\ \left(\frac{\exp\left(\tilde{q}_i^{k-1}(s^{k-1}, a_i)/\tau\right)}{\sum_{a_i' \in A_i} \exp\left(\tilde{q}_i^{k-1}(s^{k-1}, a_i')/\tau\right)} - \pi_i^{k-1}(s^{k-1}, a_i) \right),\ \forall a_i \in A_i, \end{aligned} \tag{9a}$$

$$\pi_i^k(s, a_i) = \pi_i^{k-1}(s, a_i), \quad \forall\ s \neq s^{k-1}. \tag{9b}$$

At the end of stage k, each player chooses their action $a_i^k \sim \pi_i^k(s^k)$, observes their own realized reward $r_i^k = u_i(s^k, a^k)$.

optimal one-stage deviation from policy π in state s, denoted as $\widetilde{\mathrm{br}}_i(s;\pi)$, is:

$$\widetilde{\mathrm{br}}_i(s;\pi) = \arg\max_{\hat{\pi}_i(s)} \left(\sum_{a_i \in A_i} \hat{\pi}_i(s,a_i)\tilde{Q}_i(s,a_i,\pi) - \tau \sum_{a_i \in A_i} \hat{\pi}_i(s,a_i)\log(\hat{\pi}_i(s,a_i)) \right), \ \forall s \in S. \tag{12}$$

Additionally, we can prove the following properties of the perturbed game $\tilde{\mathcal{G}}$:

Lemma 3.1.

(a) *The smoothed optimal one-stage deviation for player i given by policy π in state s is:*

$$\widetilde{\mathrm{br}}_i(s,a_i;\pi) = \frac{\exp\left(\tilde{Q}_i(s,a_i;\pi)/\tau\right)}{\sum_{a_i' \in A_i}\exp\left(\tilde{Q}_i(s,a_i';\pi)/\tau\right)}, \quad \forall a_i \in A_i. \tag{13}$$

For any $\tau > 0$ and any $\pi \in \Pi$, $\widetilde{\mathrm{br}}_i(s,a_i;\pi) > 0$ for all $a_i \in A_i$ and all $s \in S$.

(b) *The game $\tilde{\mathcal{G}}$ is a Markov potential game with the following potential function:*

$$\tilde{\Phi}(s,\pi) = \Phi(s,\pi) - \tau\mathbb{E}\left[\sum_{i\in I}\sum_{k=0}^{\infty}\delta^k \nu_i(s^k,\pi_i)\right]. \tag{14}$$

(c) *For any $\epsilon > 0$, there exists a threshold $\tau^\dagger$ such that for any $\tau < \tau^\dagger$, the stationary Nash equilibrium $\tilde{\pi}^*$ policy in $\tilde{\mathcal{G}}$ with τ is an ϵ-Nash equilibrium of $\mathcal{G}$. Furthermore, as $\tau \to 0$, $\tilde{\pi}^* \to \pi^*$.*

Algorithm analysis. In Algorithm 1, $\tilde{q}_i^k$ can be viewed as player i's estimate of her perturbed Q-function in stage k. Given (s^{k-1}, a_i^{k-1}), the estimate $\tilde{q}_i^k(s^{k-1}, a_i^{k-1})$ is updated in (8) as a linear combination of the estimate $\tilde{q}_i^{k-1}(s^{k-1}, a_i^{k-1})$ in the previous stage, and a new estimate that is comprised of the realized perturbed one-stage payoff $r_i^{k-1} - \tau\nu_i(s^{k-1}, \pi_i^{k-1})$ and the estimated total discounted payoff from the next stage. On the other hand, the policy $\pi_i^k(s^{k-1})$ updated in (9) is a linear combination of the policy $\pi_i^{k-1}(s^{k-1})$ in the previous stage, and player i's smoothed optimal one-stage deviation as in (13). Particularly, the smoothed optimal one-stage deviation is computed using the updated q-estimate $\tilde{q}_i^k$ instead of the actual $\tilde{Q}_i$, which is unknown to the player.

We note that updates of $\tilde{q}_i^k$ (resp. π_i^k) in each stage only change the element that corresponds to the realized state and action (s^{k-1}, a^{k-1}) (resp. state s^{k-1}), and the remaining elements remain unchanged. Since players do not have information of the utility functions or the transition matrix, they must update their q-estimate for every $(s, a_i) \in S \times A_i$ infinitely often. Indeed, we know from *(a)* in Lemma 3.1 that the smoothed optimal one-stage deviation policy has full support on A_i for every i. Therefore, each player i will take every action $a_i \in A_i$ with a positive probability in every stage. Furthermore, Assumption 2.1 also guarantees that every state $s \in S$ is realized infinitely often.[5]

Additionally, the update speed of $\tilde{q}_i^k(s^{k-1}, a_i^{k-1})$ (resp. $\pi_i^k(s^{k-1})$) is governed by the step size sequence $\alpha(n)$ (resp. $\beta(n)$) corresponds to the state-action counter $n = \tilde{n}_i^k(s^{k-1}, a_i^{k-1})$ (resp. state counter $n = n^k(s^{k-1})$). Therefore, the update is *asynchronous* in that the stepsizes are different across the elements associated with different states and actions.

Furthermore, we make the following assumption on the stepsizes $\{\alpha(n)\}_{n=1}^{\infty}$ and $\{\beta(n)\}_{n=1}^{\infty}$:

Assumption 3.2. *The step sizes* $\{\alpha(n) \in (0,1)\}_{n=0}^{\infty}$ *and* $\{\beta(n) \in (0,1)\}_{n=0}^{\infty}$ *satisfies*

(i) $\sum_n \alpha(n) = \infty, \sum_n \beta(n) = \infty$*,* $\lim_{n\to\infty} \alpha(n) = \lim_{n\to\infty} \beta(n) = 0$ *and* $\{\alpha(n)\}, \{\beta(n)\}$ *are non-increasing sequence*

(ii) For some $q, q' \geq 2$*,* $\sum_n \alpha(n)^{1+q/2} < \infty$ *and* $\sum_n \beta(n)^{1+q'/2} < \infty$

(iii) For any $x \in (0,1)$*,* $\sup_n \alpha([xn])/\alpha(n) < \infty$*,* $\sup_n \beta([xn])/\beta(n) < \infty$

(iv) $\lim_{n\to\infty} \beta(n)/\alpha(n) = 0$

Assumption 3.2(i)-(ii) are standard in stochastic approximation theory (Borkar (2009); Benaïm (1999); Tsitsiklis (1994)), and Assumption 3.2(iii) is a technical condition required in asynchronous updates (Perkins and Leslie (2013)). Assumption 3.2(iv) implies that our learning dynamics have two timescales: the update of $\{\tilde{q}_i^k\}_{k=0}^{\infty}$ is asymptotically faster than the update of $\{\pi_i^k\}_{k=0}^{\infty}$. One example of stepsizes that satisfies Assumption 3.2 is $\alpha(n) = n^{-c_1}$ and $\beta(n) = n^{-c_2}$ where $0 < c_1 \leq c_2 \leq 1$.

Theorem 3.3. *Suppose that Assumptions 2.1 and 3.2 are satisfied. Given any* $\epsilon > 0$*, there exists* $\tau^\dagger > 0$ *such that for any* $\tau \in (0, \tau^\dagger)$*, the sequence of policy profiles* $(\pi^k)_{k=0}^{\infty}$ *induced by Algorithm 1 converges to an* ϵ*-Nash equilibrium of the Markov potential game* $\mathcal{G}$ *with probability 1. Furthermore, as* $\tau \to 0$*, the convergent policy becomes a Nash equilibrium.*

To prove Theorem 3.3, we first apply the two-timescale asynchronous stochastic approximation theory to decouple the q-estimate updates and policy updates. We show that under Assumptions 2.1 and 3.2, our learning dynamics satisfies the set of conditions introduced in Perkins and Leslie (2013), so that $\{\tilde{q}_i^k, \pi_i^k\}_{k=0}^{\infty}$ can be analyzed by the associated continuous time dynamical system. In the appendix, we provide a brief review of the two-timescale

[5]Detailed discussion is provided in the appendix

asynchronous stochastic approximation theory, and provide detailed discussions on our setup under Assumptions 2.1 and 3.2

Secondly, we show in Lemma 3.4 that the fast dynamics of $\{\tilde{q}_i^k\}_{k=0}^\infty$ converges to each player i's actual perturbed Q-function. The proof of this lemma relies on the continuous time dynamical system at the fast timescale, where the the policy drifts are treated as asymptotically negligible errors. Then, we prove the global convergence of the continuous time dynamical system using the contraction property of Bellman operator.

Lemma 3.4. *For any $s \in S$, $a_i \in A_i$ and $i \in I$, $\lim_{k\to\infty} |\tilde{q}_i^k(s, a_i) - \tilde{Q}_i(s, a_i, \pi^k)| = 0$ with probability 1.*

Next, we analyze the policy updates with respect to the convergent values of the q-estimates provided by the fast dynamics as in Lemma 3.4. Particularly, the policy $\pi_i^k(s^{k-1})$ in (9) becomes a linear combination of $\pi_i^{k-1}(s^{k-1})$, and the smoothed optimal one-stage deviation $\widetilde{\mathrm{br}}_i(s^{k-1}; \pi^{k-1})$ based on the actual perturbed Q-function as in (13). Under Assumption 2.1 and Assumption 3.2, the convergence properties of $\{\pi^k\}_{k=0}^\infty$ can be analyzed using the following continuous time differential equation, where $t \in [0, \infty)$ is a continuous time index, and $\varpi^t : [0, \infty) \to \Pi$ is a continuous time policy function:

$$\dot{\varpi}_i^t(s, a_i) = \gamma_i(s, a_i) \left(\frac{\exp(\tilde{Q}_i(s, a_i; \varpi^t)/\tau)}{\sum_{a_i'} \exp(\tilde{Q}_i(s, a_i'; \varpi^t)/\tau)} - \varpi_i^t(s, a_i) \right), \quad \forall\ (s, a_i), \quad \forall i \in I, \tag{15}$$

and $\gamma_i(s, a_i)$ is lower bounded by a positive number $\eta > 0$ for all $(s, a_i) \in S \times A_i$, and all $i \in I$.

To establish the convergence of (15), we define a Lyapunov function as follows :

$$\phi(t) = \max_{\varpi} \sum_{s\in S} \mu(s)\tilde{\Phi}(s, \varpi) - \sum_{s\in S} \mu(s)\tilde{\Phi}(s, \varpi^t), \quad \forall t \in [0, \infty), \tag{16}$$

which is the difference of the perturbed potential function at its maximizer with that of its value at ϖ^t. Note that $\phi(t) \geq 0$, and $\phi(t) = 0$ if and only if ϖ^t is the global maximizer of the perturbed potential function. In addition, we show that $\phi(t)$ strictly decreases as long as ϖ^t is not a Nash equilibrium policy. Since the value of the perturbed potential function is finite, $\phi(t)$ converges to zero along the trajectories of ϖ^t, and thus ϖ^t converges to a Nash equilibrium of the perturbed game.

Lemma 3.5. $\frac{d}{dt}(\phi(t)) < 0$ *if and only if ϖ^t is not a Nash equilibrium. Moreover, $\lim_{t\to\infty} \varpi^t = \tilde{\pi}^*$, where $\tilde{\pi}^*$ is a Nash equilibrium of $\tilde{\mathcal{G}}$.*

We note that the potential function given by (14) incorporates the stage reward perturbation across all stages. We show in the proof of Lemma 3.5 that this way of constructing potential function is crucial to guarantee that $\phi(t)$ is a Lyapunov function of the continuous time dynamical system. Moreover, since the perturbed potential function is non-cancave in each player's policy, we develop a new approach to demonstrate that $\phi(t)$ decreases along the trajectory of ϖ^t.

Finally, based on the stochastic approximation theory, $\{\pi^k\}$ follow the trajectories of (15) in the limit. From Lemma 3.5, we know that the discrete time policy updates converge to a perturbed Nash equilibrium $\tilde{\pi}^*$.

Lemma 3.6. $\lim_{k\to\infty} \pi^k = \tilde{\pi}^*,\ w.p.1.$.

We conclude from Lemmas 3.4 – 3.5 that policy sequence converges to a Nash equilibrium $\tilde{\pi}^*$ of the perturbed game $\tilde{\mathcal{G}}$. Additionally, from *(iii)* in Lemma 3.1, we know that $\tilde{\pi}^*$ is an ϵ-Nash equilibrium when $\tau < \tau^*$, and converges to a Nash equilibrium π^* as $\tau \to 0$.

4 Numerical experiments

In this section, we demonstrate the performance of the proposed learning dynamics (Algorithm 1) in a Markov routing game (inspired by the example presented in Leonardos et al. (2021)). Consider a parallel link network comprising of L links which is repeatedly used by a set of travelers (i.e. players) denoted by I. At every stage k, each player i picks a link $a_i^k \in [L]$ to commute.

The state of the network is $s = (s_\ell)_{\ell\in[L]}$, where $s_\ell = 1$ represents that link ℓ is unsafe and $s_\ell = 0$ represents that ℓ is safe. The probability that the state of a link transitions to unsafe for next time step $k+1$ is λ_1 if the number of players using that link in step k is larger than or equal to the threshold T. Moreover, the probability that the state of a link transitions to safe for next time step $k+1$ is λ_2 if the number of players using that link in step k is larger than or equal to the threshold T.

Given network state s and joint action a, the utility of player i is

$$u_i(s,a) = \sum_{j=1}^{L} \mathbb{I}(a_i = \ell)\Big(b_\ell - m_\ell(1+s_\ell)\sum_{p\in I}\mathbb{I}(a_p = \ell)\Big),$$

where b_ℓ, m_ℓ are positive scalars for every $\ell \in [L]$. Here, b_ℓ is the fixed utility of using link $\ell \in [L]$, and $m_\ell(1+s_\ell)\sum_{p\in I}\mathbb{I}(a_p = j)$ represents the cost of using ℓ given the total number of players using link ℓ and the state of the link ℓ. The goal of every player $i \in I$ is to choose a policy $\pi_i : S \to \Delta([L])$ to minimize the long run expected discounted payoff $\mathbb{E}[\sum_{k=0}^{\infty} \delta^k u_i(s^k, a^k)]$. For the purpose of numerical experiments we run the algorithm for K iterations.

We present the numerical results for $L = 2$, $T = 2$, $m_1 = 2$, $b_1 = 9$, $m_2 = 4$, $b_2 = 16$, $\delta = 0.5$, $\lambda_1 = 0.8$, $\lambda_2 = 0.2$. The step size sequence we use are $\alpha(n) = (n+1)^{-0.6}, \beta(n) = (n+1)^{-1}$. We study the performance of Algorithm 1 in the following three settings

(S1) we set $|I| = 4$, $\tau = 10^{-3}$, $K = 5 \times 10^3$;

(S2) we set $|I| = 4$, $\tau = 10^{-6}$, $K = 5 \times 10^3$;

(S3) we set $|I| = 8, \tau = 10^{-6}$, $K = 10^3$.

We observe that the sequences of q-estimate and policy of every player converge in all of **(S1)** (Figure 1(a)-1(b)), **(S2)** (Figure 2(a)-2(b)) and **(S3)** (Figure 3(a)-3(b)). To show that the converged policies are close to a Nash equilibrium, we study the function:

$$\|V_i(\cdot, \pi_i^k, \pi_{-i}^K) - \max_{\pi_i' \in \Pi_i} V_i(\cdot, \pi_i', \pi_{-i}^K)\|_1,$$

where π_i^K is the policy update after K iterations and π_i^k is k^{th} policy update. We see that this measure goes to zero[6] in all of **(S1)-(S3)** in Figure 1(c), 2(c), 3(c). Moreover, we see that decreasing τ increases the convergence speed (refer Fig 1(c) and Fig 2(c)). Furthermore, as expected we observe that increasing the number of players decreases the convergence speed (refer Fig 3(c)).

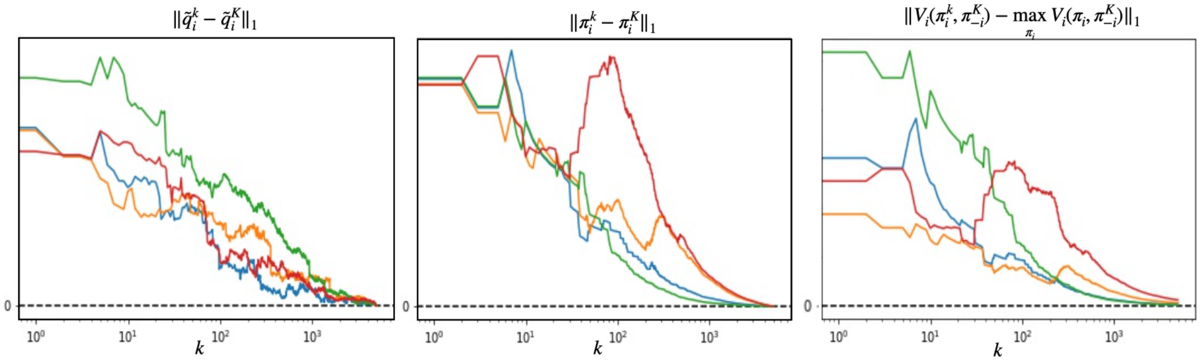
5 Conclusions

We presents a decentralized learning dynamics with two timescales, where self-interested players independently update their estimate of contingent payoffs and their policies based on received reward in each stage. We prove that a Nash equilibrium naturally arises as the long run outcome for our learning dynamics in Markov potential games with infinite horizons. Our results demonstrate that Nash equilibrium can be achieved in Markov potential games under the minimum information environment without coordination among players.

References

Arslan, G. and Yüksel, S. (2016). Decentralized q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4):1545–1558.

[6]Note that in Figure 3(c) it is not exactly at zero because plotting this graph requires too much of computational overhead as we need to compute $V_i(\pi_i^k, \pi_{-i}^K)$ for all k. However, the iterations of $\tilde{q}_i^k, \pi_i^k$ comparably runs a lot faster. The goal of numerical study is to show a proof of concept.

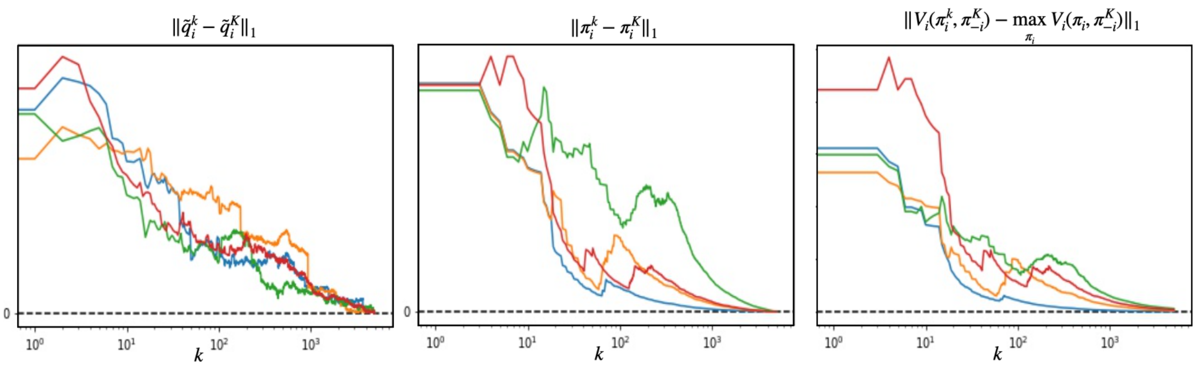


(a) Perturbed q-estimate ($\tau = 10^{-3}$)

(b) Policy update ($\tau = 10^{-3}$)

(c) Distance to Nash equilibrium ($\tau = 10^{-3}$)

Figure 1: Convergence of q-estimate (8) and policies (9) at $\tau = 10^{-3}$ after 5×10^3 steps of Algorithm 1. First two figures corresponds to convergence of iterates in Algorithm 1. While the last figure corresponds to how quickly the strategies approach a Nash equilibrium strategy. In each of the figures the four curves correspond to four players. Blue curves corresponds to player 1, orange curves corresponds to Player 2, green curves corresponds to player 3 and red curves corresponds to player 4.

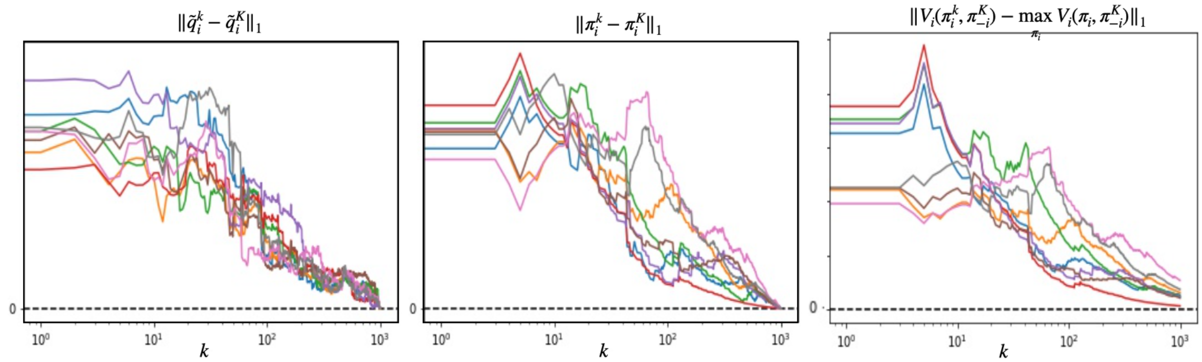


(a) Perturbed q-estimate ($\tau = 10^{-6}$)

(b) Policy update ($\tau = 10^{-6}$)

(c) Distance to Nash equilibrium ($\tau = 10^{-6}$)

Figure 2: Convergence of q-estimate (8) and policies (9) at $\tau = 10^{-6}$ after 5×10^3 steps of Algorithm 1. First two figures corresponds to convergence of iterates in Algorithm 1. While the last figure corresponds to how quickly the strategies approach a Nash equilibrium strategy. In each of the figures the four curves correspond to four players. Blue curves corresponds to player 1, orange curves corresponds to Player 2, green curves corresponds to player 3 and red curves corresponds to player 4.



(a) Perturbed q-estimate ($\tau = 10^{-6}$)

(b) Policy update ($\tau = 10^{-6}$)

(c) Distance to Nash equilibrium ($\tau = 10^{-6}$)

Figure 3: Convergence of q-estimate (8) and policies (9) at $\tau = 10^{-6}$ after 5×10^3 steps of Algorithm 1. First two figures corresponds to convergence of iterates in Algorithm 1. While the last figure corresponds to how quickly the strategies approach a Nash equilibrium strategy. In each of the figures the eight curves correspond to eight players.

Bai, Y. and Jin, C. (2020). Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR.

Bai, Y., Jin, C., and Yu, T. (2020). Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170.

Bazzan, A. L. (2009). Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems*, 18(3):342–375.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer.

Borkar, V. S. (2002). Reinforcement learning in markovian evolutionary games. *Advances in Complex Systems*, 5(01):55–72.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

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

Cominetti, R., Melo, E., and Sorin, S. (2010). A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70(1):71–83.

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

Filar, J. and Vrieze, K. (2012). *Competitive Markov decision processes.* Springer Science & Business Media.

Fudenberg, D. and Tirole, J. (1991). *Game theory.* MIT press.

Guo, H., Fu, Z., Yang, Z., and Wang, Z. (2021). Decentralized single-timescale actor-critic on zero-sum two-player stochastic games. In *International Conference on Machine Learning*, pages 3899–3909. PMLR.

Heliou, A., Cohen, J., and Mertikopoulos, P. (2017). Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30.

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

Hofbauer, J. and Sigmund, K. (2003). Evolutionary game dynamics. *Bulletin of the American mathematical society*, 40(4):479–519.

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

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

Konda, V. and Tsitsiklis, J. (1999). Actor-critic algorithms. *Advances in neural information processing systems*, 12.

Krichene, W., Drighès, B., and Bayen, A. (2014). On the convergence of no-regret learning in selfish routing. In *International Conference on Machine Learning*, pages 163–171. PMLR.

Kutschinski, E., Uthmann, T., and Polani, D. (2003). Learning competitive pricing strategies by multi-agent reinforcement learning. *Journal of Economic Dynamics and Control*, 27(11-12):2207–2218.

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.*

Leslie, D. S., Perkins, S., and Xu, Z. (2020). Best-response dynamics in zero-sum stochastic games. *Journal of Economic Theory*, 189:105095.

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

Littman, M. L. et al. (2001). Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328.

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

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

Marden, J. R., Arslan, G., and Shamma, J. S. (2009). Joint strategy fictitious play with inertia for potential games. *IEEE Transactions on Automatic Control*, 54(2):208–220.

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

Mazumdar, E., Ratliff, L. J., Jordan, M. I., and Sastry, S. S. (2019). Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. *arXiv preprint arXiv:1907.03712*.

Monderer, D. and Shapley, L. S. (1996a). Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265.

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

Panageas, I. and Piliouras, G. (2016). Average case performance of replicator dynamics in potential games via computing regions of attraction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 703–720.

Perkins, S. and Leslie, D. S. (2013). Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems*, 2(2):409–446.

Perolat, J., Piot, B., and Pietquin, O. (2018). Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, pages 919–928. PMLR.

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

Prabuchandran, K., AN, H. K., and Bhatnagar, S. (2014). Multi-agent reinforcement learning for traffic signal control. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2529–2534. IEEE.

Sastry, S. (2013). *Nonlinear systems: analysis, stability, and control*, volume 10. Springer Science & Business Media.

Sayin, M., Zhang, K., Leslie, D., Basar, T., and Ozdaglar, A. (2021). Decentralized q-learning in zero-sum markov games. *Advances in Neural Information Processing Systems*, 34.

Sayin, M. O., Parise, F., and Ozdaglar, A. (2020). Fictitious play in zero-sum stochastic games. *arXiv preprint arXiv:2010.04223*.

Schoenmakers, G., Flesch, J., and Thuijsman, F. (2007). Fictitious play in stochastic games. *Mathematical Methods of Operations Research*, 66(2):315–325.

Shah, D., Somani, V., Xie, Q., and Xu, Z. (2020). On reinforcement learning for turn-based zero-sum markov games. *arXiv preprint arXiv:2002.10620*.

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

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

Sidford, A., Wang, M., Yang, L., and Ye, Y. (2020). Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR.

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*.

Swenson, B., Murray, R., and Kar, S. (2018). On best-response dynamics in potential games. *SIAM Journal on Control and Optimization*, 56(4):2734–2767.

Szepesvári, C. and Littman, M. L. (1999). A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 11(8):2017–2060.

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202.

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., Hong, Y.-T., and Lu, C.-J. (2017). Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. (2021). Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on Learning Theory*, pages 4259–4299. PMLR.

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

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. *Advances in Neural Information Processing Systems*, 33:1166–1178.

Zhang, K., Yang, Z., and Basar, T. (2019). Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32.

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

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

Appendix

The appendix is organized as follows. In Sec A we present a detailed description of existing related works. In Sec B we review the theory of two-timescale asynchronous stochastic approximation from Perkins and Leslie (2013). In Sec C we verify that our dynamics (8)-(9) satisfy the set of conditions that enable us to use the theory stated in Sec B. In Sec D we present the proofs of our results in Sec 3.

A Related works

Markov games and the associated solution concept – stationary Nash equilibrium – have been proposed to analyze the outcomes of multi-agent interactions in dynamic environment (Shapley (1953); Littman (1994); Filar and Vrieze (2012); Fudenberg and Tirole (1991)). Markov games have been treated in a variety of settings which can be broadly categorized as infinite horizon discounted reward setting Littman (1994); Szepesvári and Littman (1999); Littman et al. (2001); Hu and Wellman (2003); Arslan and Yüksel (2016); Borkar (2002); Leslie et al. (2020); Perolat et al. (2018); Sidford et al. (2020); Sayin et al. (2020, 2021); Daskalakis et al. (2020); Wei et al. (2021); Perolat et al. (2015); Shah et al. (2020); Zhao et al. (2021); Guo et al. (2021); Zhang et al. (2020); Leonardos et al. (2021), infinite horizon time-averaged reward setting Schoenmakers et al. (2007); Filar and Vrieze (2012); Wei et al. (2017), finite horizon episodic setting Filar and Vrieze (2012); Jin et al. (2021); Bai et al. (2020); Liu et al. (2021); Bai and Jin (2020); Xie et al. (2020); Mao et al. (2021); Song et al. (2021) and others e.g. continuous state and action settings Mazumdar et al. (2019); Zhang et al. (2019). Out of these, in this work we focus on infinite horizon discounted reward setting.

Shapley (1953) presented minimax-value iteration in zero-sum Markov games, which they showed converges to the equilibrium. Littman (1994) proposed a model free algorithm – namely minmax Q-learning algorithm – whose asymptotic convergence guarantee was established in Szepesvári and Littman (1999). Hu and Wellman (2003); Littman et al. (2001) extended the Q-learning algorithm from zero-sum games Littman (1994) to general-sum games under some conditions. Moreover, majority of literature in multi-agent RL deal with zero-sum Markov games Sidford et al. (2020); Shah et al. (2020); Zhang et al. (2020); Zhao et al. (2021); Leslie et al. (2020); Sayin et al. (2020, 2021); Daskalakis et al. (2020); Guo et al. (2021) where they exploit the min-max dynamic programming framework Shapley (1953).

Our proposed learning dynamics and convergence result contribute to the growing literature of multi-agent reinforcement learning in Markov potential games (Leonardos et al. (2021); Zhang et al. (2021); Song et al. (2021); Mao et al. (2021)). In a Markov potential game, there exists a potential function that describes the change of each player' expected total payoff under unilateral policy deviation. As a result, a stationary Nash equilibrium policy can be solved as the global optimum of the potential function, and gradient-based algorithms have been proposed to compute the stationary Nash equilibrium (Zhang et al. (2021); Leonardos et al. (2021)) for discounted infinite horizon setting. Our focus is different in that we analyze how stationary Nash equilibrium can naturally arise through independent policy adjustments by self-interested players in a decentralized manner.

One of the main challenges associated with developing provably converging independent and decentralized algorithm is that of non-stationarity induced in the environment as the players learn. Indeed, naively deploying single-agent RL algorithms to obtain independent and decentralized multi-agent algorithm may lead to non-convergence to Nash equilibrium Tan (1993); Matignon et al. (2012); Mazumdar et al. (2019). In Arslan and Yüksel (2016) the authors presented a decentralized MARL algorithm albeit with coordination in *acyclic Markov games*, which subsume Markov team games. In this setup the players take a fixed policy for extended period of times (referred as exploration phase). This induces stationarity in the environment of every player which ensures that it forms accurate estimates of the Q-functions. In the context of two-player zero-sum games some recent studies have also developed decentralized and independent learning algorithms Daskalakis et al. (2020); Guo et al. (2021) but these algorithms also require coordination between players either in form of asymmetric step sizes Daskalakis et al. (2020) or fixing policies of one of the players to learn the best response of other player Guo et al. (2021).

Two-timescale based algorithms present a nice framework to overcome the non-stationarity in the environment Borkar (2002); Perolat et al. (2018); Sayin et al. (2020, 2021). Borkar (2002) proposed an actor-critic based algorithm Konda and Tsitsiklis (1999), where players update policies at slower timescale than the value function estimate. The author showed that certain weighted empirical distribution of actions played converges to a *generalized Nash equilibrium*, which need not imply convergence to Nash equilibrium. In Perolat et al. (2018) the authors considered a similar actor-critic based decentralized learning dynamics, and showed its converges to Nash equilibrium in cooperative multistage games – a special case of Markov games such that each stage is visited only once, and the state transition has a tree structure with finite depth. There are two key differences between our learning dynamics and theirs:

(a) we consider updates with asynchronous stepsizes that are adjusted based on counters of each state and each state-action pair;

(b) the q-estimate update in our dynamics introduces a reward perturbation in each stage.

It turns out that these two differences are crucial for us to achieve equilibrium convergence in a much more general setting – Markov potential games with infinite stages and no restrictions on state transition other than irreducibility. Moreover, since the proof techniques developed in Perolat et al. (2018) exploits the special structure of multistage games, they cannot be applied in our setting. Instead, we develop a new approach that involves recent development of asynchronous stochastic approximation theory, a novel construction of perturbed potential function, and new policy convergence proof in infinite-horizon discounted Markov games.

In Sayin et al. (2021) for zero-sum discounted Markov games shares similar features with ours – their MARL dynamics is also decentralized, independent, asynchronous, and has two timescale. Our learning dynamics and convergence analysis differ from Sayin et al. (2021) in two aspects:

(a) The dynamics in Sayin et al. (2021) updates q-estimates at a slower timescale, and update policies faster, while our policy update is slower compared to the q-estimate update. The analysis in their paper exploits the properties of zero-sum Markov games to demonstrate the convergence of policy updates while holding q-estimate as constant;

(b) The dynamics in Sayin et al. (2021) only counts the number of times a state is visited, but does not count the number of times a state-action pair is visited. The step-size in their q-estimate update for each state-action pair is adjusted based on the counter of states, and is normalized by the probability of taking that action given the policy in that stage. On the other hand, our step-size in q-estimate update is asynchronously adjusted for every state-action pair based on the associated counter. As a result, their analysis of asynchronous updates and time separation is different from ours, and requires a different set of assumptions on stepsizes.

Finally, our results also advances the rich literature of learning in stateless potential game that includes continuous and discrete time best response dynamics (Monderer and Shapley (1996b); Swenson et al. (2018)), fictitious play (Monderer and Shapley (1996a); Hofbauer and Sandholm (2002); Marden et al. (2009)), replicator dynamics(Panageas and Piliouras (2016); Hofbauer and Sigmund (2003)), no-regret learning (Heliou et al. (2017); Krichene et al. (2014)), and payoff-based learning (Cominetti et al. (2010)). In particular, our learning dynamics share similar spirit with the payoff-based learning dynamics in stateless potential game (Cominetti et al. (2010)). In a payoff-based learning dynamics, players update the estimate of payoffs associated with each action in a decentralized manner only based on their own payoff in each stage, and update their mixed strategy by incorporating a smoothed best response given the payoff estimate. In a Markov potential game, due to the joint evolvement between state and actions, we need to introduce the time separation, and keep track of the payoff (q-estimate) for every state-action pair in an asynchronous manner. Additionally, the smoothed best response mixed strategy becomes the smoothed optimal one-stage deviation policy. We emphasize that our analysis approach is completely different from that in stateless potential games. Furthermore, our construction of perturbed potential function is new, and our policy convergence proof uses tools in reinforcement learning. Therefore, our results in Markov potential games are not direct extension of the ones in stateless potential games.

B Review of Two-timescale asynchronous stochastic approximation

In this section we review the results from Perkins and Leslie (2013) on two-timescale asynchronous stochastic approximation. Note that we do not state their results in full generality but only to the extend necessary for this paper.

Let $\{x^k\}_{k=1}^{\infty}, \{y^k\}_{k=1}^{\infty}$ stochastic approximation update. Let $x^k \in \mathbb{R}^X$, $y^k \in \mathbb{R}^Y$ for all $k \in \{1, 2, ..\}$. Let $\bar{X} \subset [X]$ (resp. $\bar{Y} \subset [Y]$) be the elements of x update (resp. y update) that have positive probability of being updated in the asynchronous update process. At stage k, let $\bar{X}^k \subset \bar{X}$ and $\bar{Y}^k \subset \bar{Y}$ be the elements that are updated.

Let

$$\tilde{n}^k(\mathfrak{i}) = \sum_{p=1}^{k} \mathbb{1}(\mathfrak{i} \in \bar{X}^p), \quad n^k(\mathfrak{j}) = \sum_{p=1}^{k} \mathbb{1}(\mathfrak{j} \in \bar{Y}^p)$$

Consider the following asynchronous stochastic approximation updates indexed by $k \in \{1, 2, ..\}$

$$\begin{aligned} x^k(\mathfrak{i}) &= x^{k-1}(\mathfrak{i}) + \alpha(\tilde{n}^k(\mathfrak{i}))\mathbb{1}(\mathfrak{i} \in \bar{X}^k)[F(\mathfrak{i}; x^{k-1}, y^{k-1}) + \tilde{M}^k(\mathfrak{i}) + d^k(\mathfrak{i})], \quad \forall\, \mathfrak{i} \in [X] \\ y^k(\mathfrak{j}) &= y^{k-1}(\mathfrak{j}) + \beta(n^k(\mathfrak{j}))\mathbb{1}(\mathfrak{j} \in \bar{Y}^k)[G(\mathfrak{j}; x^{k-1}, y^{k-1}) + M^k(\mathfrak{j}) + e^k(\mathfrak{j})], \quad \forall\, \mathfrak{j} \in [Y] \end{aligned} \tag{17}$$

where

(i) for any $x \in \mathbb{R}^X, y \in \mathbb{R}^Y$, $F(x, y) = (F(\mathfrak{i}; x, y))_{\mathfrak{i} \in [X]} \in \mathbb{R}^X$ and $G(x, y) = (G(\mathfrak{j}; x, y))_{\mathfrak{j} \in [Y]} \in \mathbb{R}^Y$

(ii) $\{\tilde{M}^k = (\tilde{M}^k_{\mathfrak{i}})_{\mathfrak{i} \in [X]}\}, \{M^k = (M^k_{\mathfrak{j}})_{\mathfrak{j} \in [Y]}\}$ be martingale difference processes defined on

$\mathbb{R}^X, \mathbb{R}^Y$ respectively

(iii) $\{d^k = (d^k(\mathfrak{i}))_{\mathfrak{i}\in\mathbb{R}^X}, e^k = (e^k(\mathfrak{j}))_{\mathfrak{j}\in\mathbb{R}^Y}\}$ are asymptotically negligible error terms

(iv) $\{\alpha(n)\}_{n=0}^{\infty}, \{\beta(n)\}_{n=0}^{\infty}$ are the step sizes

(v) $x^0 \in \mathbb{R}^X, y^0 \in \mathbb{R}^Y$ are initialized at some values.

Define

$$
\begin{aligned}
\bar{\alpha}^k &= \max_{\mathfrak{i}\in\bar{X}^k} \alpha(\tilde{n}^k(\mathfrak{i})), & \mu^k(\mathfrak{i}) &= \frac{\alpha(\tilde{n}^k(\mathfrak{i}))}{\bar{\alpha}^k}\mathbb{1}(\mathfrak{i}\in\bar{X}^k) \\
\bar{\beta}^k &= \max_{\mathfrak{j}\in\bar{Y}^k} \beta(\tilde{n}^k(\mathfrak{j})), & \sigma^k(\mathfrak{j}) &= \frac{\beta(\tilde{n}^k(\mathfrak{j}))}{\bar{\beta}^k}\mathbb{1}(\mathfrak{j}\in\bar{Y}^k) \\
\tilde{\mathcal{D}}^k &= \mathsf{diag}([\mu^k(1), \mu^k(2), ..., \mu^k(X)]), & \mathcal{D}(k) &= \mathsf{diag}([\sigma^k(1), \sigma^k(2), ..., \sigma^k(Y)])
\end{aligned}
$$

Using these notations we can concisely write (17) as

$$
\begin{aligned}
x^k &= x^{k-1} + \bar{\alpha}^k\tilde{\mathcal{D}}^k\left(F(x^{k-1}, y^{k-1}) + \tilde{M}^k + d^k\right) \\
y^k &= y^{k-1} + \bar{\beta}^k\mathcal{D}^k\left(G(x^{k-1}, y^{k-1}) + M^k + e^k\right)
\end{aligned} \tag{18}
$$

We now state some assumption that are crucial to study the asymptotic property of the stochastic approximation (18). First, we introduce some important notations. Define $\bar{H} = \bar{X} \times \bar{Y}$ such that if $i \in \bar{X}, j \in \bar{Y}$ then $(i, j) \in \bar{H}$ if and only if i, j have positive probability of occuring simultaneously. At iteration k, $\bar{H}^k \in \bar{H}$ be the updated component in $[X] \times [Y]$. Furthermore $z^k = (x^k, y^k)$ be the joint update. Let $\mathcal{F}^k = \sigma(\{\bar{H}^m\}_m, \{z^m\}_m, \{\tilde{n}^m(\mathfrak{i})\}, \{n^m(\mathfrak{j})\}\} \ \forall\ m \le k, \mathfrak{i} \in [X], \mathfrak{j} \in [Y])$ be sigma algebra containing all information upto stage k. Define $\Omega_K^{eta} = \{\mathsf{diag}(\omega(1), ..., \omega(K)) : \omega(K) \in [\eta, 1] \ \forall i = 1, 2, .., K\}$.

Next, we present the assumptions required in Perkins and Leslie (2013) to study the asymptotic behavior of two-timescale asynchronous stochastic approximation update (17).

Assumption B.1. *Let the following assumptions hold*

(A1) For compact sets $\tilde{\mathcal{S}} \subset \mathbb{R}^X, \mathcal{S} \subset \mathbb{R}^Y$, $x^k \in \tilde{\mathcal{S}}, y^k \in \mathcal{S}$ *for all* $k \in \{0, 1, ...\}$

(A2) $\{d^k\}, \{e^k\}$ *are bounded sequence such that* $\lim_{k\to\infty} d(k) = \lim_{k\to\infty} e(k) = 0$

(A3) Following must be true for the stepsizes:

(i) $\sum_n \alpha(n) = \infty, \sum_n \beta(n) = \infty$, $\lim_{n\to\infty} \alpha(n) = \lim_{n\to\infty} \beta(n) = 0$ *and* $\{\alpha(n)\}, \{\beta(n)\}$ *are non-increasing sequence*

(ii) For any $\theta \in (0,1)$, $\sup_n \alpha([\theta n])/\alpha(n) < \infty$, $\sup_n \beta([\theta n])/\beta(n) < \infty$

(iii) $\lim_{n\to\infty} \beta(n)/\alpha(n) = 0$

(A4) The maps $F(\cdot,\cdot), G(\cdot,\cdot)$, *referred as* mean fields, *are such that*

(i) $F : \tilde{\mathcal{S}} \times \mathcal{S} \to \mathcal{S}$ *is upper semi-continuous and* $\|F(z)\| \leq c(1 + \|z\|)$ *where* c *is a constant.*

(ii) $G : \tilde{\mathcal{S}} \times \mathcal{S} \to \tilde{\mathcal{S}}$ *is upper semi-continuous and for all* $x \in \tilde{\mathcal{S}}$ *we have*[7] $\|G(x,y)\| \leq c(1 + \|y\|)$ *where* c *is a constant.*

(A5) (i) for all $z \in \tilde{\mathcal{S}} \times \mathcal{S}$ *and* $h^{k-1}, h^k \in \bar{H}$,

$$\Pr\left(\bar{H}^k = h^k | \mathcal{F}^{k-1}\right) = \Pr(\bar{H}^k = h^k | \bar{H}^{k-1} = h^{k-1}, z^{k-1} = z)$$

(ii) For all $z \in \tilde{\mathcal{S}} \times \mathcal{S}$ *the transition probability*

$$\mathcal{Q}(z; h^k, h^{k-1}) := \Pr(\bar{H}^k = h^k | \bar{H}^{k-1} = h^{k-1}, z^{k-1} = z) \tag{19}$$

form aperiodic, irreducible Markov chain over $\bar{H}$

(iii) the map $z \mapsto \mathcal{P}(z; h^k, h^{k-1})$ *is Lipschitz.*

(A6) The processes $\tilde{M}^k, \bar{X}^k$ *are uncorrelated given* $\mathcal{F}^{k-1}$*. For some* $q \geq 2$, $\sum_n \alpha(n)^{1+q/2} < \infty$ *and* $\sup_k \mathbb{E}\left[\|\tilde{M}^k\|^q\right] < \infty$*. Similarly the processes* $M^k, \bar{Y}^k$ *are uncorrelated given* $\mathcal{F}^{k-1}$*. For some* $q' \geq 2$, $\sum_n \beta(n)^{1+q'/2} < \infty$ *and* $\sup_k \mathbb{E}\left[\|M^k\|^q\right] < \infty$.

(A7) For all $y \in \mathcal{S}$ *and every* $\eta > 0$ *the differential equation*

$$\dot{x}^t = \Omega_X^\eta \cdot F(x^t, y),$$

has unique global attractor $\Lambda(y)$ *where* $\Lambda : \mathbb{R}^Y \to \mathbb{R}^X$ *is bounded, continuous and single-valued for all* $y \in \mathcal{S}$.

Under the preceding assumptions, we first state a convergence result for the fast dynamics, $\{x(k)\}$, in the following theorem:

[7] The actual condition imposed in Perkins and Leslie (2013) is that both $F(\cdot,\cdot)$, $G(\cdot,y)$ (for every y) be a *Marchaud map* which is required to impose suitable regularity conditions when F, G are set-valued. However in the context of single valued functions, as in our case, those requirements are trivially satisfied.

Theorem B.2 (Fast-timescale convergence)**.** *Under assumption (A1)-(A7) in Assumption B.1,*

$$(x^k, y^k) \to \{(\Lambda(y), y) : y \in D\} \quad as \quad k \to \infty,$$

with probability 1.

Next, we present the corresponding convergence results for the slow dynamics, $\{y(k)\}$. Define $G^{\Lambda} : \mathbb{R}^J \to \mathbb{R}^J$ as $G^{\Lambda,\eta}(y) = G(\Lambda(y), y)$. Furthermore, let $\bar{G}^{\Lambda} = \Omega_Y^{\eta} G^{\Lambda}(y)$. Consider the following differential equation

$$\dot{y}^t = \bar{G}^{\Lambda,\eta}(y^t) \tag{20}$$

Theorem B.3 (Slow-timescale convergence)**.** *If for all $\eta > 0$, there is a global attractor $\mathcal{A} \subset \mathcal{S}$ for the differential equation* (20) *and assumptions (A1)-(A7) in Assumption B.1 are satisfied then $\{y^k\}_{k=0}^{\infty}$ will almost surely converge to $\mathcal{A}$.*

C Verifying conditions for two-timescale asynchronous stochastic approximation from Section B.

Before verifying the conditions for two-timescale asynchronous stochastic approximation stated in Section B we introduce some notations that helps in clear presentation.

First, we define the following notation that will be used in the proofs

$$u_i(s, a_i, \pi_{-i}) \triangleq \sum_{a_{-i}} \pi_{-i}(s, a_{-i}) u_i(s, a_i, a_{-i}), \quad P(s'|s, a_i, \pi_{-i}) \triangleq \sum_{a_{-i}} \pi_{-i}(s, a_{-i}) P(s'|s, a_i, a_{-i}),$$
$$P(s'|s, \pi) \triangleq \sum_{a_i} \sum_{a_{-i}} \pi_i(s, a_i) \pi_{-i}(s, a_{-i}) P(s'|s, a_i, a_{-i}), \quad \tilde{Q}_i(s, \pi_i'; \pi) \triangleq \sum_{a_i} \pi_i'(s, a_i) \tilde{Q}_i(s, a_i; \pi)$$

For any $\tilde{q}$, $\pi \in \Pi$, $i \in I, a_i \in A_i, s \in S$, we define

$$\mathcal{T}_i^{\pi} \tilde{q}_i(s, a_i) := u_i(s, a_i, \pi_{-i}) - \tau \nu_i(s, \pi_i) + \delta \sum_{s'} P(s'|s, a_i, \pi_{-i}) \sum_{a_i'} \pi_i(s', a_i') \tilde{q}(s', a_i'), \tag{21}$$

which is an analogous to Bellman operator in the setup of this paper. Furthermore, let's define

$$\hat{\mathcal{T}}_i^{\pi} \tilde{q}_i(s, a_i) := u_i(s, a_i, a_{-i}) - \tau \nu_i(s, \pi_i) + \delta \sum_{a_i'} \pi_i(s', a_i') \tilde{q}_i(s', a_i'), \tag{22}$$

where $a_{-i} \sim \pi_{-i}(s)$ and $s' \sim \sum_{a_{-i}} \pi_{-i}(s, a_{-i}) P(\cdot|s, a_i, a_{-i})$. Moreover, for any $s \in S, i \in I, a_i \in A_i$ we define

$$\widetilde{\mathrm{br}}_i^{k-1}(s, a_i) := \frac{\exp\left(\tilde{q}_i^{k-1}(s, a_i)/\tau\right)}{\sum_{a_i' \in A_i} \exp\left(\tilde{q}_i^{k-1}(s, a_i')/\tau\right)}. \tag{23}$$

Under the above introduced notations we re-write (8)-(9) as

$$\begin{aligned} \tilde{q}_i^k(s, a_i) = \tilde{q}_i^{k-1}(s, a_i) + \alpha(\tilde{n}^k(s, a_i)) \mathbb{1}((s, a_i) = (s^{k-1}, a_i^{k-1})) \cdot \\ \left(\mathcal{T}_i^{\pi^{k-1}} \tilde{q}_i^{k-1}(s, a_i) - \tilde{q}_i^{k-1}(s, a_i) + \tilde{M}^k(s, a_i)\right) \end{aligned} \tag{24a}$$

$$\pi_i^k(s, a_i) = \pi_i^{k-1}(s, a_i) + \beta(n^k(s)) \mathbb{1}(s = s^{k-1}) \cdot \left(\widetilde{\mathrm{br}}_i^{k-1}(s, a_i) - \pi_i^{k-1}(s, a_i)\right) \tag{24b}$$

for all $(s, a_i) \in S \times A_i$, $i \in I$, where

$$\tilde{M}^k(s, a_i) = \hat{\mathcal{T}}_i^{\pi^{k-1}} \tilde{q}_i^{k-1}(s, a_i) - \mathcal{T}_i^{\pi^{k-1}} \tilde{q}_i^{k-1}(s, a_i) \tag{25a}$$

Note that $\mathbb{E}[\tilde{M}^k(s, a_i)|\mathcal{F}^{k-1}] = 0$ where $\mathcal{F}^{k-1} = \sigma(\{(s^m, a^m)\}_m, \{\tilde{q}_i^m\}_m, \{\pi_i^m\}_m : m \leq k-1, i \in I)$ is the sigma-algebra comprising of history till stage $k-1$. Consequently, $\{\tilde{M}^k\}$ is a martingale difference sequence.

Note that the asynchronous q-estimate updates (24a) and the policy updates (24b) both have $\Pi_{i=1}^I S \times A_i$ components. One can observe that corresponding to all of components of joint q-estimate and policy update which are updated at any instant of time there exists a unique element in the set $\bar{H} := \{((s, a_1), (s, a_2), .., (s, a_I)) : s \in S, a_i \in A_i, i \in I\}$.

We now verify Assumption B.1 (A1)-(A7) one by one

(i) First we show that (A1) in Assumption B.1 is satisfied with $(\tilde{q}^k, \pi^k)$ update (24a)-(24b). Let $\bar{u} = \max_{i,s,a,\pi} u_i(s, a) - \tau\nu_i(s, \pi_i)$. This value is finite as $\nu_i(s, \pi_i) \in [-|A_i|/e, 0]$ for all $s \in S, i \in I, \pi_i \in \Pi_i$. Moreover let $\bar{R} = \max\{\bar{u}/(1-\delta), \max_i \|\tilde{q}_i^0\|_\infty\}$. Then we claim that $\|\tilde{q}_i^k\|_\infty \leq \bar{R}$ for all $k = \{0, 1, 2, ..\}$. We show this by induction. It holds for $k = 0$ by construction. Suppose it holds for $k = m-1$ for some m then we show that it also holds for $k = m$. Indeed, we note from (24a) that $\tilde{q}_i^k$ is a convex combination[8] of $\tilde{q}_i^{k-1}$ and $\mathcal{T}_i^{\pi^{k-1}} \tilde{q}_i^{k-1}(s, a_i) + \tilde{M}^k(s, a_i)$. Using (22) and (25a) we see that

$$\begin{aligned} \|\mathcal{T}_i^{\pi^{k-1}} \tilde{q}_i^{k-1}(s, a_i) + \tilde{M}^k(s, a_i)\| = \|\hat{\mathcal{T}}_i^{\pi^{k-1}} \tilde{q}^{k-1}\|_\infty &\leq \bar{u} + \delta\bar{R} \\ &\leq (1-\delta)\bar{R} + \delta\bar{R} = \bar{R} \end{aligned}$$

[8]This is because we assume that $\alpha(n) \in (0, 1)$ in Assumption 3.2.

This shows that $\|\tilde{q}_i^k\|_\infty \leq \bar{R}$. Moreover note that $\pi^k \in \Pi$ which is product simplex and is always compact.

(ii) Since we do not have error terms in the asynchronous updates, AssumptionB.1-(A2) is immediately satisfied

(iii) Next we note that (A2) in Assumption B.1 is satisfied due to Assumption 3.2.

(iv) Now we show Assumption B.1-(A4) is satisfied. First, we concisely write the mean fields of (24a)-(24b) as follows

$$F((s,a_i);\tilde{q},\pi) := \mathcal{T}_i^\pi \tilde{q}_i(s,a_i) - \tilde{q}_i(s,a_i), \quad \forall\ s \in S, i \in I, a_i \in A_i$$
$$G((s,a_i);\tilde{q},\pi) := \frac{\exp\left(\tilde{q}_i(s,a_i)/\tau\right)}{\sum_{a_i' \in A_i} \exp\left(\tilde{q}_i(s,a_i')/\tau\right)} - \pi_i(s,a_i) \quad \forall\ s \in S, i \in I, a_i \in A_i$$

respectively. Define $F(\tilde{q},\pi) = (F((s,a_i);\tilde{q},\pi))_{s\in S, i\in I, a_i\in A_i}$, $G(\tilde{q},\pi) = (G((s,a_i);\tilde{q},\pi))_{s\in S, i\in I, a_i\in A_i}$. We note that both F, G are continuous as demanded in Assumption B.1-(A4). Furthermore, observe that

$$\|F(\tilde{q},\pi)\|_\infty \leq \|\mathcal{T}_i^\pi \tilde{q}_i\|_\infty + \|\tilde{q}\|_\infty \leq \bar{u} + \delta\|\tilde{q}\|_\infty + \|\tilde{q}\|_\infty \leq \tilde{c}(1 + \|\tilde{q}\|_\infty),$$

where $\tilde{c} = \max\{\bar{u}, 1+\delta\}$. Also note that

$$\|G(\tilde{q},\pi)\|_\infty \leq 1 + \|\pi\|_\infty.$$

Thus we conclude that (A4) in Assumption B.1 is satisfied.

(v) We now verity Assumption B.1-(A5). Consider $h, h' \in \bar{H}$ such that $h = ((s,a_1),(s,a_2),..(s,a_I))$ and $h' = ((s',a_1'),(s,a_2'),...(s',a_I'))$. Moreover, let $z = (\tilde{q},\pi)$ then

$$\mathcal{P}(z;h,\tilde{h}) = P(s'|s,a)\prod_{i\in I} \pi_i(s',a_i') \tag{26}$$

where $a = (a_i)_{i\in I}$ and the function $\mathcal{P}(z;h,\tilde{h})$ is defined in (19). Note that from Lemma 3.1(i) we have $\pi_i(s,a_i) > 0$ for all $s \in S, i \in I, a_i \in A_i$. Moreover, we impose Assumption 2.1 on transition matrix. Thus (A5)-(i) and (A5)-(ii) in Assumption B.1 are satisfied. Finally (A5)-(iii) is satisfied by noting that (26) is Lipschitz in π and therefore in z.

(vi) Assumption -B.1:(A6) is satisfied by noting that (a) $\tilde{M}$ is a bounded martingale difference sequence and (b) the step size condition in Assumption 3.2-(ii) holds.

(vii) For any $\eta > 0$, $\pi \in \Pi$ consider the differential equation

$$\dot{\tilde{q}}_i^t = \Omega_{A_i}^\eta \left(\mathcal{T}_i^\pi \tilde{q}_i^t - \tilde{q}_i^t \right), \quad \forall\, i \in I \tag{27}$$

where $\Omega_{A_i}^\eta = \{\mathsf{diag}(\omega(1), ..., \omega(A_i)) : \omega(A_i) \in [\eta, 1]\ \forall i = 1, 2, .., A_i\}$. In order to verify Assumption B.1-(A7), we show that (27) has unique global attractor for every $\pi \in \Pi$.

We first note that $\mathcal{T}_i^\pi$ is a contraction for every $\pi \in \Pi$. Indeed

$$\begin{aligned} &\mathcal{T}_i^\pi q_i(s, a_i) - \mathcal{T}_i^\pi \bar{q}_i(s, a_i) \\ =&\delta \sum_{a_{ii}} \pi_{-i}(s, a_{-i}) \sum_{s'} P(s'|s, a_i, a_{-i}) \sum_{a_i' \in A_i} \pi_i(s', a_i') \left(q_i(s', a_i') - \bar{q}_i(s', a_i') \right). \end{aligned}$$

Thus, for every $s \in S, i \in I, a_i \in A_i$, we have

$$|\mathcal{T}_i^\pi q_i(s, a_i) - \mathcal{T}_i^\pi \bar{q}_i(s, a_i)| \leq \delta \|q_i - \bar{q}_i\|_\infty$$

Consequently this $\mathcal{T}_i^\pi$ is a contraction. That is,

$$\|\mathcal{T}_i^\pi q_i - \mathcal{T}_i^\pi \bar{q}_i\|_\infty \leq \delta \|q_i - \bar{q}_i\|_\infty$$

Since $\mathcal{T}_i^\pi$ is a contraction, this means that (27) has unique global attractor which is the fixed point of $\mathcal{T}_i^\pi$. In fact it follows from definition that $\tilde{Q}_i(\cdot, \cdot; \pi)$ (defined in (11)) is the unique global attractor. That is,

$$\mathcal{T}_i^\pi \tilde{Q}_i(s, a_i; \pi) = \tilde{Q}_i(s, a_i; \pi), \quad \forall s \in S, i \in I, a_i \in A_i.$$

Define $\bar{u} := \max_{i,s,a,\pi} u_i(s, a) - \tau \nu_i(s, \pi_i)$. Then $\|\tilde{Q}_i(\cdot, \cdot; \pi)\|_\infty \leq \frac{\bar{u}}{1-\delta}$. Additionally, $\tilde{Q}_i(\cdot, \cdot; \pi)$ is also continuous in π. Finally, the convergence of (27) follows by (Borkar, 2009, Chapter 7). Thus Assumption B.1-(A7) is satisfied.

D Proofs of Results in Section 3

Proof of Lemma 3.1. We prove *(a)* – *(c)* in sequence:

(a) From (12), for each $s \in S$ and each $\pi \in \Pi$, player i's optimal one-stage deviation policy

$\widetilde{\mathrm{br}}_i(s;\pi)$ for state s can be written as the optimal solution of the following problem:

$$\begin{aligned}
&\max_{\hat{\pi}_i(s)} f(\hat{\pi}_i(s)) \triangleq \left(\sum_{a_i \in A_i} \hat{\pi}_i(s,a_i)\tilde{Q}_i(s,a_i,\pi) - \tau \sum_{a_i\in A_i} \hat{\pi}_i(s,a_i)\log(\hat{\pi}_i(s,a_i)) \right) \\
&s.t. \quad \sum_{a_i\in A_i} \hat{\pi}_i(s,a_i) = 1, \quad \hat{\pi}_i(s,a_i) \geq 0, \quad \forall\, a_i \in A_i.
\end{aligned} \tag{28}$$

We can check that the objective function $f(\hat{\pi}_i(s))$ of (28) is strictly concave:

$$\frac{\partial^2 f(\hat{\pi}_i(s))}{\partial \hat{\pi}_i(s,a_i)\partial\hat{\pi}_i(s,a_i')} = -\mathbb{1}(a_i = a_i')\frac{\tau}{\hat{\pi}_i(s,a_i)}, \quad \forall a_i, a_i' \in A_i. \tag{29}$$

That is, the Hessian matrix of the objective function f is a diagonal matrix with negative entries. Thus, $\widetilde{\mathrm{br}}_i(s;\pi)$ is the unique solution to the optimization problem (28). Let $\lambda \in \mathbb{R}$ be a Lagrange multiplier corresponding to the equality constraint and $\mu = (\mu(a_i))_{a_i\in A_i} \in \mathbb{R}_+^{|A_i|}$ be the Lagrange multiplier corresponding to the inequality constraints. The Lagrangian of (28) is given as follows:

$$\mathcal{L}(\hat{\pi}_i(s),\lambda,\mu;\pi) = f(\hat{\pi}_i(s)) + \lambda(1 - \sum_{a_i\in A_i}\hat{\pi}_i(s,a_i)) - \sum_{a_i\in A_i}\mu(a_i)\hat{\pi}_i(s,a_i)$$

Then, we know that there must exist $\lambda^\star, \mu^\star$ such that $(\widetilde{\mathrm{br}}_i(s;\pi), \lambda^\star, \mu^\star)$ satisfies the following conditions:

$$\frac{\partial \mathcal{L}(\widetilde{\mathrm{br}}_i(s;\pi),\lambda^\star,\mu^\star;\pi)}{\partial \widetilde{\mathrm{br}}_i(s;\pi)} = 0, \quad \forall\, a_i \in A_i, \tag{30a}$$

$$\frac{\partial \mathcal{L}(\widetilde{\mathrm{br}}_i(s;\pi),\lambda^\star,\mu^\star;\pi)}{\partial \lambda} = 1 - \sum_{a_i\in A_i}\widetilde{\mathrm{br}}_i(s,a_i;\pi) = 0, \tag{30b}$$

$$\mu^\star(a_i)\widetilde{\mathrm{br}}_i(s,a_i) = 0, \quad \mu^\star(a_i) \geq 0, \quad \forall a_i \in A_i. \tag{30c}$$

From (30a), we can derive that

$$\widetilde{\mathrm{br}}_i(s;\pi) = \exp\left(\tilde{Q}_i(s,a_i;\pi)/\tau\right)\exp\left(-1 - \lambda^\star/\tau - \mu^\star(a_i)/\tau\right), \ \forall\, a_i \in A_i, \tag{31}$$

Therefore, $\widetilde{\mathrm{br}}_i(s,a_i) > 0$ for all $a_i \in A_i$. From (30c), we know that $\mu^\star(a_i) = 0$ for all $a_i \in A_i$. Then, from (30b) and (31), we have that

$$\widetilde{\mathrm{br}}_i(s,a_i;\pi) = \frac{\exp\left(\tilde{Q}_i(s,a_i;\pi)/\tau\right)}{\sum_{a_i'\in A_i}\exp\left(\tilde{Q}_i(s,a_i';\pi)/\tau\right)}, \quad \forall a_i \in A_i,$$

as in (13).

(b) We note that

$$\tilde{V}_i(s,\pi_i,\pi_{-i}) = \mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(u_i(s^k,a^k) - \tau\nu_i(s^k,\pi_i)\right)\right]$$
$$= V_i(s,\pi_i,\pi_{-i}) - \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i)\right)\right], \quad \forall s\in S. \tag{32}$$

Then, from (14), for any $s \in S$, any $\pi, \pi^{'} \in \Pi$, any $\pi_{-i} \in \Pi_{-i}$, and any $i \in I$, we have:

$$\begin{aligned}
&\tilde{\Phi}(s,\pi_i',\pi_{-i}) - \tilde{\Phi}(s,\pi_i,\pi_{-i})\\
=&\Phi(s,\pi_i',\pi_{-i}) - \Phi(s,\pi_i,\pi_{-i}) - \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i^{'})\right] + \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i)\right]\\
=&V_i(s,\pi_i',\pi_{-i}) - V_i(s,\pi_i,\pi_{-i}) - \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i^{'})\right] + \tau\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\nu_i(s^k,\pi_i)\right]\\
\overset{(32)}{=}&\tilde{V}_i(s,\pi_i',\pi_{-i}) - \tilde{V}_i(s,\pi_i,\pi_{-i})
\end{aligned}$$

That is, $\tilde{\Phi}(s,\pi)$ as in (14) is a potential function of $\widetilde{G}$.

(c) Let $\tilde{\pi}_i^*$ be a Nash equilibrium for $\tilde{\mathcal{G}}$. Using (32) and (2), for any $\pi_i \in \Pi_i$ and any $i \in I$, we have

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

Since $\tilde{\pi}_i^*$ is a Nash equilibrium of $\tilde{\mathcal{G}}$, we have $\tilde{V}_i(\mu,\pi_i,\tilde{\pi}_{-i}^*) - \tilde{V}_i(\mu,\pi_i^*,\pi_{-i}^*) \leq 0$. Thus,

$$\begin{aligned}
&\tilde{V}_i(\mu,\pi_i,\tilde{\pi}_{-i}^*) - \tilde{V}_i(\mu,\pi_i^*,\pi_{-i}^*) + \tau\left(\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i)\right)\right] - \mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\tilde{\pi}_i^*)\right)\right]\right)\\
\leq&\tau\left(\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i)\right)\right] - \mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\tilde{\pi}_i^*)\right)\right]\right)\\
\leq&\tau\left|\mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\pi_i)\right)\right] - \mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\left(\nu_i(s^k,\tilde{\pi}_i^*)\right)\right]\right|\\
\leq&2\tau\max_{s,\pi_i}|\nu_i(s,\pi_i)|\sum_{k=0}^{\infty}\delta^k = \frac{2\tau}{e}\frac{|A_i|}{1-\delta},
\end{aligned} \tag{33}$$

where the last equality is due to the fact that $\max_{s,\pi_i} |\nu_i(s,\pi_i)| = \frac{|A_i|}{e}$ and $\sum_{k=0}^{\infty} \delta^k = \frac{1}{1-\delta}$.[9] For any $\epsilon > 0$, we set $\tau^\dagger = \epsilon e(1-\delta)/(2(\max_i |A_i|))$. Then, for any $\tau \in (0, \tau^\dagger)$

$$V_i(\mu, \pi_i, \tilde{\pi}^*_{-i}) - V_i(\mu, \tilde{\pi}^*_i, \tilde{\pi}^*_{-i}) = \frac{2\tau}{e}\frac{|A_i|}{1-\delta} \leq \epsilon, \quad \forall\, i \in I, \pi_i \in \Pi_i.$$

That is, $\tilde{\pi}^*_i$ is an ϵ-Nash equilibrium of game $\mathcal{G}$. Furthermore if $\tau \to 0$, then from (33), we can see that $\tilde{\pi}^*_i$ of $\tilde{\mathcal{G}}$ is also a Nash equilibrium of game $\mathcal{G}$.

Proof of Lemma 3.4. We have shown in Section C that Assumption B.1 (A1)-(A7) are satisfied then from Theorem B.2 implies that there exists $\pi^\dagger$ such that $\tilde{q}^k_i(s, a_i) \to \tilde{Q}_i(s, a_i; \pi^\dagger)$ and $\pi^k \to \pi^\dagger$ as $k \to \infty$. We claim that $\lim_{k\to\infty} |\tilde{q}^k_i(s,a_i) - \tilde{Q}_i(s,a;\pi^k)| = 0$. Indeed, it follows from

$$|\tilde{q}^k_i(s,a_i) - \tilde{Q}_i(s,a;\pi^k)| \leq |\tilde{q}^k_i(s,a_i) - \tilde{Q}_i(s,a;\pi^\dagger)| + |\tilde{Q}_i(s,a_i;\pi^k) - \tilde{Q}_i(s,a;\pi^\dagger)|$$

by noting that first term converges to zero as $k \to \infty$. Furthermore, the second term goes to zero as $\tilde{Q}_i(\cdot,\cdot;\pi)$ is continuous in π and $\pi^k \to \pi^\dagger$ as $k \to \infty$.

Before proving Lemma 3.5, we need to present the following three results (Lemmas D.1, D.2, and D.3) as preparation:

Lemma D.1 (Policy gradient theorem for perturbed game)**.** *For any $\mu \in \Delta(S), s \in S, \pi \in \Pi, i \in I, a_i \in A_i$ the gradient of value function $\tilde{V}_i(\mu,\pi)$ with respect to $\pi_i(s,a_i)$ is given by*

$$\frac{\partial \tilde{V}_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \frac{1}{1-\delta} d^\pi_\mu(s) \left(\tilde{Q}_i(s,a_i;\pi) + \tau\nu_i(s,\pi_i) - \tau(1 + \log(\pi_i(s,a_i)))\right)$$

where d^π_μ is the discounted state visitation frequency given by

$$d^\pi_\mu(s) := (1-\delta) \sum_{s^0\in S} \mu(s^0) \sum_{k=0}^{\infty} \delta^k \Pr(s^k = s|s^0), \tag{34}$$

and the sequence of states is governed by the Markov chain induced by the transition matrix P and the policy π. We drop the dependency of conditional probability on π for notational simplicity.

Proof. For any $s \in S, i \in I, a_i \in A_i$ define $t(s,a_i) = \tilde{Q}_i(s,a_i;\pi) + \tau\nu_i(s,\pi_i) - \tau(1 +$

[9] $\max_{s,\pi_i} |\nu_i(s,\pi_i)| = \max_{x_i\in[0,1]^{|A_i|}} |\sum_i x_i \log(x_i)| = \frac{|A_i|}{e}$.

$\log(\pi_i(s,a_i)))$. We claim that for any integer $K \geq 0$, $\mu \in \Delta(S), s \in S, i \in I, a_i \in A_i$,

$$\frac{\partial \tilde{V}_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \mathbb{E}\left[\sum_{k=0}^{K} \delta^k \mathbb{1}(s^k = s)\right] t(s,a_i) + \delta^{K+1}\mathbb{E}\left[\frac{\partial \tilde{V}_i(s^{K+1},\pi)}{\partial \pi_i(s,a_i)}\right], \tag{35}$$

where $s_0 \sim \mu, a^{k-1} \sim \pi(s^{k-1}), s^k \sim P(\cdot|s^{k-1}, a^{k-1})$. We prove this claim by induction. Indeed, this holds for $K = 0$ by noting that

$$\begin{aligned}
&\frac{\partial \tilde{V}_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \frac{\partial}{\partial \pi_i(s,a_i)} \sum_{\bar{s}\in S} \mu(\bar{s}) \left(\sum_{a_i \in A_i} \pi_i(\bar{s},a_i)\tilde{Q}_i(\bar{s},a_i;\pi)\right) \\
=&\frac{\partial}{\partial \pi_i(s,a_i)}\bigg(\sum_{\bar{s}} \mu(\bar{s}) \sum_{\bar{a}_i} \pi_i(\bar{s},\bar{a}_i)\Big(u_i(\bar{s},\bar{a}_i,\pi_{-i}(\bar{s})) - \tau\nu_i(\bar{s},\pi_i) + \delta\sum_{s'} P(s'|\bar{s},\bar{a}_i,\pi_{-i}(\bar{s}))\tilde{V}_i(s',\pi)\Big)\bigg) \\
=&\mu(s)\left(u_i(s,a_i,\pi_{-i}(s)) - \tau(1+\log(\pi_i(s,a_i))) + \delta\sum_{s'} P(s'|s,a_i,\pi_{-i}(s))\tilde{V}_i(s',\pi)\right) \\
&\quad + \delta\sum_{\bar{s}} \mu(\bar{s})\sum_{s'} P(s'|\bar{s},\pi)\frac{\partial \tilde{V}_i(s',\pi)}{\partial \pi_i(s,a_i)} \\
&= \mu(s)\left(\tilde{Q}_i(s,a_i;\pi) + \tau\nu_i(s,\pi_i) - \tau(1+\log(\pi_i(s,a_i)))\right) + \delta\sum_{\bar{s}}\mu(\bar{s})\sum_{s'} P(s'|\bar{s},\pi)\frac{\partial \tilde{V}_i(s',\pi)}{\partial \pi_i(s,a_i)} \\
&= \mathbb{E}\left[\mathbb{1}(s^0 = s)\right] t(s,a_i) + \delta\mathbb{E}\left[\frac{\partial \tilde{V}_i(s^1,\pi)}{\partial \pi_i(s,a_i)}\right]
\end{aligned}$$

We now suppose that the claim holds for some integer K and then show that it holds for $K+1$, that is we have

$$\begin{aligned}
&\frac{\partial \tilde{V}_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \mathbb{E}\left[\sum_{k=0}^{K}\delta^k\mathbb{1}(s^k = s)\right] t(s,a_i) + \delta^{K+1}\mathbb{E}\left[\frac{\partial \tilde{V}_i(s^{K+1},\pi)}{\partial \pi_i(s,a_i)}\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{K}\delta^k\mathbb{1}(s^k = s)\right] t(s,a_i) + \delta^{K+1}\mathbb{E}\left[\frac{\partial}{\partial \pi_i(s,a_i)}\left(\sum_{a_i}\pi_i(s^{K+1},a_i)\tilde{Q}_i(s^{K+1},a_i;\pi)\right)\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{K}\delta^k\mathbb{1}(s^k = s)\right] t(s,a_i) + \delta^{K+1}\left(\mathbb{1}(s^{K+1} = s)t(s,a_i) + \delta\left(\sum_{s'} P(s'|s^{K+1},\pi)\frac{\partial \tilde{V}_i(s',\pi)}{\partial \pi_i(s,a_i)}\right)\right) \\
&= \mathbb{E}\left[\sum_{k=0}^{K+1}\delta^k\mathbb{1}(s^k = s)\right] t(s,a_i) + \delta^{K+2}\mathbb{E}\left[\frac{\partial \tilde{V}_i(s^{K+2},\pi)}{\partial \pi_i(s,a_i)}\right].
\end{aligned}$$

This completes the proof of (35). Now if we let $K \to \infty$ in (35) then we obtain

$$\frac{\partial \tilde{V}_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \mathbb{E}\left[\sum_{k=0}^{\infty}\delta^k\mathbb{1}(s^k = s)\right] t(s,a_i) = \sum_{s^0\in S}\mu(s^0)\sum_{k=0}^{\infty}\Pr(s^k = s|s^0)t(s,a_i) = \frac{1}{1-\delta}d_\mu^\pi(s)t(s,a_i)$$

$\square$

Lemma D.2 (Multiagent Performance Difference Lemma)**.** *For any policy* $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi_{-i}) \in \Pi$ *and any* $\mu \in \Delta(S)$ *the following holds:*

$$\tilde{V}_i(\mu, \pi) - \tilde{V}_i(\mu, \pi') = \frac{1}{1-\delta} \sum_{s'} d^{\pi}_{\mu}(s') \left(\Gamma_i(s, \pi_i; \pi') + \tau \nu_i(s, \pi'_i) - \tau \nu_i(s, \pi_i) \right), \tag{36}$$

where $\Gamma_i(s, a_i; \pi)$ *is the advantage function defined as the advantage gained from one stage deviation from the current policy* π *to play* a_i *in state* s*:*

$$\Gamma_i(s, a_i; \pi) \coloneqq \tilde{Q}_i(s, a_i; \pi) - \tilde{V}_i(s, \pi), \quad \forall i \in I, \ \forall s \in S, \ \forall a_i \in A_i, \ \forall \pi \in \Pi. \tag{37}$$

Proof. For any initial state distribution μ and joint policy $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi_{-i}) \in \Pi$, it holds that

$$\tilde{V}_i(\mu, \pi) - \tilde{V}_i(\mu, \pi') = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i) \right)\right] - \tilde{V}_i(\mu, \pi')$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i) - \tilde{V}_i(s^k, \pi') + \tilde{V}_i(s^k, \pi') \right)\right] - \tilde{V}_i(\mu, \pi')$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i) - \tilde{V}_i(s^k, \pi') \right)\right] + \mathbb{E}\left[\tilde{V}_i(s^0, \pi')\right] + \mathbb{E}\left[\sum_{k=1}^{\infty} \delta^k \tilde{V}_i(s^k, \pi')\right] - \tilde{V}_i(\mu, \pi'),$$

where $s^0 \sim \mu, a^{k-1} \sim \pi(s^{k-1}), s^k \sim P(\cdot | s^{k-1}, a^{k-1})$. We note that

$$\mathbb{E}\left[\tilde{V}_i(s^0, \pi')\right] = \tilde{V}_i(\mu, \pi'), \quad \mathbb{E}\left[\sum_{k=1}^{\infty} \delta^k \tilde{V}_i(s^k, \pi')\right] = \delta \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \tilde{V}_i(s^{k+1}, \pi')\right].$$

Therefore,

$$\tilde{V}_i(\mu, \pi) - \tilde{V}_i(\mu, \pi') = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i) - \tilde{V}_i(s^k, \pi') \right)\right] + \delta \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \tilde{V}_i(s^{k+1}, \pi')\right]$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i) - \tilde{V}_i(s^k, \pi') + \delta \tilde{V}_i(s^{k+1}, \pi') \right)\right]$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi'_i) + \delta \tilde{V}_i(s^{k+1}, \pi') + \tau \nu_i(s^k, \pi'_i) - \tau \nu_i(s^k, \pi_i) - \tilde{V}_i(s^k, \pi') \right)\right]$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi'_i) + \delta \sum_{s'} P(s' | s^k, a^k) \tilde{V}_i(s', \pi') + \tau \nu_i(s^k, \pi'_i) - \tau \nu_i(s^k, \pi_i) - \tilde{V}_i(s^k, \pi') \right)\right]$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(\tilde{Q}_i(s^k, a^k_i; \pi') - \tilde{V}_i(s^k, \pi') + \tau \nu_i(s^k, \pi'_i) - \tau \nu_i(s^k, \pi_i) \right)\right]$$
$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(\Gamma_i(s^k, a^k_i; \pi') + \tau \nu_i(s^k, \pi'_i) - \tau \nu_i(s^k, \pi_i) \right)\right]$$

$$=\frac{1}{1-\delta}\sum_{s'} d_\mu^\pi(s')\left(\Gamma_i(s,\pi_i;\pi')+\tau\nu_i(s,\pi_i')-\tau\nu_i(s,\pi_i)\right).$$

□

Lemma D.3. *Any policy $\tilde{\pi}^*$ that satisfies $\tilde{\pi}_i^*(s)=\widetilde{\mathrm{br}}_i(s;\tilde{\pi}^*)$ for all $i\in I$ and all $s\in S$ is a stationary Nash equilibrium of $\tilde{G}$.*

Proof. For any $s\in S,\pi_i\in\Pi_i,\pi_{-i}\in\Pi_{-i}$, we define $\tilde{u}_i(s,\pi_i,\pi_{-i}):=u(s,\pi_i,\pi_{-i})-\tau\nu_i(s,\pi_i)$. We note that

$$\begin{aligned}
&\tilde{\pi}_i^*(s)=\widetilde{\mathrm{br}}_i(s;\tilde{\pi}^*)=\arg\max_{\hat{\pi}}\sum_{a_i}\hat{\pi}_i(s,a_i)\tilde{Q}_i(s,a_i;\tilde{\pi}^*)-\tau\nu_i(s,\hat{\pi}_i)\\
=&\arg\max_{\hat{\pi}}\sum_{a_i}\hat{\pi}_i(s,a_i)\left(u_i(s,a_i,\tilde{\pi}_{-i}^*)-\tau\nu_i(s,\tilde{\pi}_i^*)+\delta\sum_{s'}P(s'|s,a_i,\tilde{\pi}_{-i}^*)\tilde{V}(s',\tilde{\pi}^*)\right)-\tau\nu_i(s,\hat{\pi}_i)\\
=&\arg\max_{\hat{\pi}}\sum_{a_i}\hat{\pi}_i(s,a_i)\left(u_i(s,a_i,\tilde{\pi}_{-i}^*)-\tau\nu_i(s,\tilde{\pi}_i^*)+\delta\sum_{s'}P(s'|s,a_i,\tilde{\pi}_{-i}^*)\tilde{V}(s',\tilde{\pi}^*)-\tau\nu_i(s,\hat{\pi}_i)\right)\\
=&\arg\max_{\hat{\pi}}\sum_{a_i}\hat{\pi}_i(s,a_i)\left(u_i(s,a_i,\tilde{\pi}_{-i}^*)-\tau\nu_i(s,\hat{\pi}_i)+\delta\sum_{s'}P(s'|s,a_i,\tilde{\pi}_{-i}^*)\tilde{V}(s',\tilde{\pi}^*)\right)\\
=&\arg\max_{\hat{\pi}}\ \tilde{u}_i(s,\hat{\pi}_i,\tilde{\pi}_{-i}^*)+\delta\sum_{s'}P(s'|s,\hat{\pi}_i,\tilde{\pi}_{-i}^*)\tilde{V}(s',\tilde{\pi}^*).
\end{aligned}\tag{38}$$

To prove that $\tilde{\pi}^*$ is a Nash equilibrium of $\tilde{\mathcal{G}}$, we need to show that for every $\pi_i'\in\Pi_i$,

$$\tilde{V}_i(s,\tilde{\pi}_i^*,\tilde{\pi}_{-i}^*)\geq\tilde{V}_i(s,\pi_i',\pi_{-i}^*),\quad\forall\ i\in I,\ \forall s\in S\tag{39}$$

Before proving (39), we first show that for any integer $K\geq1$, any $s\in S$, any $i\in I$, and any $\pi_i'\in\Pi_i$

$$\tilde{V}_i(s,\pi_i^*,\pi_{-i}^*)\geq\mathbb{E}\left[\sum_{k=0}^{K-1}\delta^k\tilde{u}_i(s^k,\pi_i',\pi_{-i}^*)+\delta^K\tilde{V}_i(s^K,\tilde{\pi}^*)\right],\tag{40}$$

where $s^0=s,a_i^k\sim\pi_i'(s^k),a_{-i}^k\sim\pi_{-i}^*(s^k),s^k\sim P(\cdot|s^{k-1},a^{k-1})$. Consider $K=1$, for any $s\in S$, any $i\in I$, any $\pi_i'\in\Pi_i$ we have

$$\begin{aligned}
\tilde{V}_i(s,\tilde{\pi}_i^*,\tilde{\pi}_{-i}^*)&=\sum_{a_i}\pi_i^*(s,a_i)\left(u_i(s,a_i,\tilde{\pi}_{-i}^*)-\tau\nu_i(s,\tilde{\pi}_i^*)+\delta\sum_{s'}P(s'|s,a_i,\tilde{\pi}_{-i}^*)\tilde{V}(s',\tilde{\pi}^*)\right)\\
&=\tilde{u}_i(s,\pi_i^*,\tilde{\pi}_{-i}^*)+\delta\sum_{s'}P(s'|s,\tilde{\pi}_i^*,\tilde{\pi}_{-i}^*)\tilde{V}(s',\tilde{\pi}^*)\qquad(41)\\
&\geq\tilde{u}_i(s,\pi_i',\tilde{\pi}_{-i}^*)+\delta\sum_{s'}P(s'|s,\pi_i',\tilde{\pi}_{-i}^*)\tilde{V}(s',\tilde{\pi}^*)
\end{aligned}$$

$$= \mathbb{E}\left[\tilde{u}_i(s^0, \pi_i', \pi_{-i}^*) + \delta \tilde{V}_i(s^1, \tilde{\pi}^*)\right],$$

where, again, $s^0 = s$, $a_i^0 \sim \pi_i^{'}(s^0)$, $a_{-i}^0 \sim \pi_{-i}^*(s^0)$, $s^1 \sim P(\cdot|s^0, a^0)$.

Next, suppose that (40) holds for some integer K, we consider $K+1$:

$$\begin{aligned}
\tilde{V}_i(s, \pi_i^*, \pi_{-i}^*) &\underset{(a)}{\geq} \mathbb{E}\left[\sum_{k=0}^{K-1} \delta^k \tilde{u}_i(s^k, \pi_i', \pi_{-i}^*) + \delta^K \tilde{V}_i(s^K, \tilde{\pi}^*)\right] \\
&\underset{(b)}{=} \mathbb{E}\left[\sum_{k=0}^{K-1} \delta^k \tilde{u}_i(s^k, \pi_i', \pi_{-i}^*) + \delta^K \left(\tilde{u}_i(s^K, \pi_i^*, \pi_{-i}^*) + \delta \sum_{s'} P(s'|s^K, \pi_i^*, \pi_{-i}^*) \tilde{V}_i(s', \tilde{\pi}^*)\right)\right] \\
&\underset{(c)}{\geq} \mathbb{E}\left[\sum_{k=0}^{K-1} \delta^k \tilde{u}_i(s^k, \pi_i', \pi_{-i}^*) + \delta^K \left(\tilde{u}_i(s^K, \pi_i', \pi_{-i}^*) + \delta \sum_{s'} P(s'|s^K, \pi_i', \pi_{-i}^*) \tilde{V}_i(s', \tilde{\pi}^*)\right)\right] \\
&\underset{(d)}{=} \mathbb{E}\left[\sum_{k=0}^{K} \delta^k \tilde{u}_i(s^k, \pi_i', \pi_{-i}^*) + \delta^{K+1} \tilde{V}_i(s^{K+1}, \tilde{\pi}^*)\right],
\end{aligned}$$

where (a) is by induction hypothesis, (b) is due to (41), (c) is due to (38) and (d) is by rearrangement of terms. Thus, by mathematical induction, we have established that (40) holds for all K. Let $K \to \infty$ in (40), we have

$$\tilde{V}_i(s, \pi_i^*, \pi_{-i}^*) \geq \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \tilde{u}_i(s^k, \pi_i', \pi_{-i}^*)\right] = \tilde{V}_i(s, \pi_i', \pi_{-i}^*), \quad \forall s \in S,\ \forall i \in I,\ \forall \pi_i' \in \Pi.$$

Thus, we have proved (39), i.e. $\tilde{\pi}^*$ is a Nash equilibrium of game $\tilde{\mathcal{G}}$. □

We are now ready to proof Lemma 3.5.

Proof of Lemma 3.5. Recall from Lemma 3.1-(a), we can re-write (15) as follows:

$$\dot{\varpi}_i^t(s, a_i) = \gamma_i(s, a_i)\left(\widetilde{\mathrm{br}}_i(s, a_i; \varpi^t) - \varpi_i^t(s, a_i)\right), \quad \forall\ (s, a_i), \quad \forall i \in I. \tag{42}$$

Furthermore, for every $i \in I, s \in S$ all of the actions $a_i \in A_i$ are updated simultaneously, this implies $\gamma_i(s, a_i) = \gamma_i(s, a_i') = \gamma_i(s)$ for all $a_i, a_i' \in A_i$ Perkins and Leslie (2013). Lemma D.3 ensures that the stationary point of (42) is a Nash equilibrium policy for game $\tilde{\mathcal{G}}$.

We compute the derivative of $\phi(t)$ with respect to t, where $\phi(t)$ is given by (16).

$$\begin{aligned}
\frac{d}{dt}\phi(t) = -\sum_i \sum_s \sum_{a_i} \frac{\partial \tilde{\Phi}(\mu, \varpi^t)}{\partial \varpi_i(s, a_i)} \frac{d\varpi_i^t(s, a_i)}{dt} &\underset{(a)}{=} -\sum_{i,s,a_i} \frac{\partial \tilde{V}_i(\mu, \varpi^t)}{\partial \varpi_i(s, a_i)} \frac{d\varpi_i^t(s, a_i)}{dt} \\
\underset{(b)}{=} -\frac{1}{1-\delta} \sum_{i,s,a_i} d_\mu^{\varpi^t}(s)\left(\tilde{Q}_i(s, a_i; \varpi^t) + \tau\nu_i(s, \varpi_i^t) - \tau(1 + \log(\varpi_i^t(s, a_i)))\right) \cdot& \\
\gamma_i(s)\left(\widetilde{\mathrm{br}}_i(s, a_i; \varpi^t) - \varpi_i^t(s, a_i)\right),&
\end{aligned} \tag{43}$$

where (a) is due to the fact that $\tilde{\Phi}$ is a potential function of the perturbed game (recall Lemma 3.1-(b)), and (b) is due to the policy gradient theorem for the perturbed game (Lemma D.1) and equation 42.

Recall the proof of Lemma 3.1-(a). From (28), we know that $\widetilde{\mathrm{br}}_i(s;\pi) = \underset{\hat{\pi}_i(s)\in\Delta(A_i)}{\arg\max} f(\hat{\pi}_i(s))$, where $f(\hat{\pi}_i(s)) = \left(\sum_{a_i\in A_i} \hat{\pi}_i(s,a_i)\tilde{Q}_i(s,a_i,\pi) - \tau\nu_i(s,\hat{\pi})\right)$ is strictly concave in $\hat{\pi}_i(s)$. From the first order conditions of constrained optimality, we know that

$$\left\langle \nabla_{\hat{\pi}_i} f(\widetilde{\mathrm{br}}_i(s;\pi)), \hat{\pi}_i(s) - \widetilde{\mathrm{br}}_i(s;\pi)\right\rangle \leq 0. \tag{44}$$

Moreover, from the strict concavity, we have

$$\left\langle \nabla_{\hat{\pi}_i} f(\hat{\pi}_i(s)) - \nabla_{\hat{\pi}_i} f(\widetilde{\mathrm{br}}_i(s;\pi)), \hat{\pi}_i(s) - \widetilde{\mathrm{br}}_i(s;\pi)\right\rangle < 0. \tag{45}$$

Combining (44) and (45) we obtain

$$\left\langle \nabla_{\hat{\pi}_i} f(\hat{\pi}_i(s)), \widetilde{\mathrm{br}}_i(s;\pi) - \hat{\pi}_i(s)\right\rangle > 0.$$

That is,

$$\sum_{a_i\in A_i} \left(\tilde{Q}_i(s,a_i,\pi) - \tau(1+\log(\hat{\pi}_i(s,a_i)))\right)\left(\widetilde{\mathrm{br}}_i(s,a_i;\pi) - \hat{\pi}_i(s,a_i)\right) > 0, \quad \forall i\in I, \quad \forall s\in S.$$

Additionally,

$$\sum_{a_i\in A_i} \tau\nu_i(s,\pi_i)\left(\widetilde{\mathrm{br}}_i(s,a_i;\pi) - \hat{\pi}_i(s,a_i)\right) = \tau\nu_i(s,\pi_i)\left(\sum_{a_i\in A_i}\left(\widetilde{\mathrm{br}}_i(s,a_i;\pi) - \hat{\pi}_i(s,a_i)\right)\right) = 0 \tag{46}$$

Therefore,

$$\sum_{a_i\in A_i} \left(\tilde{Q}_i(s,a_i,\pi) + \tau\nu_i(s,\pi_i) - \tau\left(1+\log(\hat{\pi}_i(s,a_i))\right)\right)\left(\widetilde{\mathrm{br}}_i(s,a_i;\pi) - \hat{\pi}_i(s,a_i)\right) > 0 \tag{47}$$

From (43) – (47) and the fact that $\gamma_i(s) > \eta$ for all $i\in I, s\in S$, we have

$$\frac{d}{dt}\phi(t) \leq -\frac{\eta}{1-\delta}\sum_{i,s,a_i} d_\mu^{\varpi^t}(s)\left(\tilde{Q}_i(s,a_i;\varpi^t) + \tau\nu_i(s,\varpi_i^t) - \tau(1+\log(\varpi_i^t(s,a_i)))\right)\cdot \left(\widetilde{\mathrm{br}}_i(s,a_i;\varpi^t) - \varpi_i^t(s,a_i)\right)$$

$$\overset{(46)}{=} -\frac{\eta}{1-\delta}\sum_{i,s,a_i} d_\mu^{\varpi^t}(s)\left(\tilde{Q}_i(s,a_i;\varpi^t) - \tau(1+\log(\varpi_i^t(s,a_i)))\right)\cdot\left(\widetilde{\mathrm{br}}_i(s,a_i;\varpi^t) - \varpi_i^t(s,a_i)\right)$$

$$= -\frac{\eta}{1-\delta}\sum_{i,s} d_\mu^{\varpi^t}(s)\left(\nabla_{\hat{\pi}_i} f(\varpi_i^t)\right)^\top\left(\widetilde{\mathrm{br}}_i(s;\varpi^t) - \varpi_i^t(s)\right)$$

$$\leq -\frac{\eta}{1-\delta}\sum_{i,s} d_\mu^{\varpi^t}(s)\left(f(\widetilde{\mathrm{br}}_i(s;\varpi^t)) - f(\varpi_i^t)\right) \tag{48}$$

where the last inequality is due to concavity of f.

Additionally, let $\pi_i^\dagger \in \arg\max_{\pi_i} \tilde{V}_i(\mu,\pi_i,\varpi_{-i}^t)$ be a best response of player i if the joint strategy of other players is ϖ_{-i}^t. Note that $\pi_i^\dagger$ maximizes the total payoff instead of just maximizing the payoff of one-stage deviation. Therefore, $\pi_i^\dagger$ is different from the optimal one-stage deviation policy. We drop the dependence of $\pi_i^\dagger$ on ϖ_{-i} for notational simplicity.

We define $D = \frac{1}{1-\delta}\max_i \left\|\frac{d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}}{\mu}\right\|_\infty$. We note that D is finite under the assumption that μ has full support (Assumption 2.1). Additionally, from (34), we have $d_\mu^{\varpi^t}(s) \geq (1-\delta)\mu(s)$.

$$\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}(s)\max_{\hat{\varpi}_i(s)}\left(f(\hat{\varpi}_i) - f(\varpi_i^t)\right) = \sum_{i,s} d_\mu^{\varpi^t}(s)\frac{d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}(s)}{d_\mu^{\varpi^t}(s)}\max_{\hat{\varpi}_i(s)}\left(f(\hat{\varpi}_i) - f(\varpi_i^t)\right)$$

$$\leq \sum_{i,s} d_\mu^{\varpi^t}(s)\left\|\frac{d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}}{d_\mu^{\varpi^t}}\right\|_\infty \max_{\hat{\varpi}_i(s)}\left(f(\hat{\varpi}_i) - f(\varpi_i^t)\right)$$

$$\leq \frac{1}{1-\delta}\sum_{i,s} d_\mu^{\varpi^t}(s)\left\|\frac{d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}}{\mu}\right\|_\infty \max_{\hat{\varpi}_i(s)}\left(f(\hat{\varpi}_i) - f(\varpi_i^t)\right)$$

$$= D\sum_{i,s} d_\mu^{\varpi^t}(s)\max_{\hat{\varpi}_i(s)}\left(f(\hat{\varpi}_i) - f(\varpi_i^t)\right) = D\sum_{i,s} d_\mu^{\varpi^t}(s)\left(f(\widetilde{\mathrm{br}}_i(s;\varpi^t)) - f(\varpi_i^t)\right), \tag{49}$$

Then, from (48) and (49), we have

$$\frac{d}{dt}\phi(t) \leq -\frac{\eta}{1-\delta}\sum_{i,s} d_\mu^{\varpi^t}(s)\left(f(\widetilde{\mathrm{br}}_i(s;\varpi^t)) - f(\varpi_i^t)\right)$$

$$\leq -\frac{\eta}{D(1-\delta)}\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}(s)\max_{\hat{\varpi}_i(s)}\left(f(\hat{\varpi}_i) - f(\varpi_i^t)\right)$$

$$\leq -\frac{\eta}{D(1-\delta)}\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}(s)\left(f(\pi_i^\dagger) - f(\varpi_i^t)\right)$$

$$= -\frac{\eta}{D(1-\delta)}\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}(s)\left(\tilde{Q}_i(s,\pi_i^\dagger;\varpi^t) - \tilde{Q}_i(s,\varpi_i^t;\varpi^t) - \tau\nu_i(s,\pi_i^\dagger) + \tau\nu_i(s,\varpi_i^t)\right), \tag{50}$$

where the last equation follows from the definition of function f.

We now analyze the right-hand-side of (50)

$$\begin{aligned}
&\tilde{Q}_i(s,\pi_i^\dagger;\varpi^t) - \tilde{Q}_i(s,\varpi_i^t;\varpi^t) - \tau\nu_i(s,\pi_i^\dagger) + \tau\nu_i(s,\varpi_i^t) \\
=&\sum_{a_i} \tilde{Q}_i(s,a_i;\varpi^t)\left(\pi_i^\dagger(s,a_i) - \varpi_i^t(s,a_i)\right) - \tau\nu_i(s,\pi_i^\dagger) + \tau\nu_i(s,\varpi_i^t) \\
=&\sum_{a_i} \left(\tilde{Q}_i(s,a_i;\varpi^t) - \tilde{V}_i(s,\varpi^t)\right)\left(\pi_i^\dagger(s,a_i) - \varpi_i^t(s,a_i)\right) - \tau\nu_i(s,\pi_i^\dagger) + \tau\nu_i(s,\varpi_i^t) \\
\overset{(37)}{=}&\sum_{a_i} \Gamma_i(s,a_i;\varpi^t)\pi_i^\dagger(s,a_i) - \tau\nu_i(s,\pi_i^\dagger) + \tau\nu_i(s,\varpi_i^t) \\
=&\Gamma_i(s,\pi_i^\dagger;\varpi^t) - \tau\nu_i(s,\pi_i^\dagger) + \tau\nu_i(s,\varpi_i^t) \qquad (51)
\end{aligned}$$

Combining (50) and (51), we have

$$\begin{aligned}
\frac{d\phi(t)}{dt} &\leq -\frac{\eta}{D(1-\delta)}\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^t}(s)\left(\Gamma_i(s,\pi_i^\dagger;\varpi^t) - \tau\nu_i(s,\pi_i^\dagger) + \tau\nu_i(s,\varpi_i^t)\right) \\
&\overset{(36)}{=} -\frac{\eta}{D(1-\delta)}\sum_i \left(\tilde{V}_i(\mu,\pi_i^\dagger,\varpi_{-i}^t) - \tilde{V}_i(\mu,\varpi^t)\right).
\end{aligned}$$

Note that since $\pi_i^\dagger$ is a best response corresponding to ϖ^t this means $\tilde{V}_i(\mu,\pi_i^\dagger,\varpi_{-i}^t) - \tilde{V}_i(\mu,\varpi^t) \geq 0$ for all i. Thus $d\phi(t)/dt \leq 0$. However, if ϖ^t is not a Nash equilibrium then there exists i such that

$$\tilde{V}_i(\mu,\pi_i^\dagger,\varpi_{-i}^t) - \tilde{V}_i(\mu,\varpi^t) > 0$$

which implies $d\phi(t)/dt < 0$. Equivalently, if $d\phi(t)/dt = 0$ then ϖ^t is a Nash equilibrium.

Moreover, if ϖ^t is a Nash equilibrium then from Lemma D.3, we know that

$$\widetilde{\mathrm{br}}_i(s,a_i;\varpi^t) = \frac{\exp(\tilde{Q}_i(s,a_i;\varpi^t)/\tau)}{\sum_{a_i'}\exp(\tilde{Q}_i(s,a_i';\varpi^t)/\tau)} = \varpi_i^t(s,a_i), \quad \forall s \in S, \quad \forall a_i \in A_i, \quad \forall i \in I.$$

Therefore, the right-hand-side of (15) is zero. Thus, $d\phi/dt = 0$. Finally, the convergence of policy updates ϖ^t to a Nash equilibrium policy $\tilde{\pi}^*$ follows from LaSalle's invariance principle Sastry (2013) with ϕ as the Lyapunov function. □

Proof of Lemma 3.6. Recall from Sec C, we have shown that all required assumption in two-timescale asynchronous stochastic approximation are satisfied. Thus, we can apply Theorem B.3 to show that the convergence of the discrete time dynamics $(\pi^k)_{k=0}^\infty$ induced by Algorithm 1 is the same as that of the continuous time dynamical system in (15). From Lemma 3.5, we know that any solution of the continuous time dynamical system (15) must converge to $\tilde{\pi}^*$. Thus, $\lim_{k\to\infty}\pi^k = \tilde{\pi}^*$ with probability 1. □

Proof of Theorem 3.3. For every $\epsilon > 0$, we know from Lemma 3.1(c) that there exists $\tau^\dagger$ such for all $\tau \in (0, \tau^\dagger)$ the stationary Nash equilibrium $\tilde{\pi}^*$ policy in $\tilde{\mathcal{G}}$ is an $\epsilon-$Nash equilibrium policy of $\mathcal{G}$. For any given $\tau \in (0, \tau^\dagger)$, Lemma 3.6 guarantees that $(\pi^k)_{k=0}^\infty$ induced by Algorithm 1 converges to $\tilde{\pi}^*$.

Furthermore, from (33), we obtain for every $\mu \in \Delta(S), i \in I, \pi_i \in \Pi_i$

$$V_i(\mu, \pi_i, \tilde{\pi}^*_{-i}) - V_i(\mu, \tilde{\pi}^*_i, \tilde{\pi}^*_{-i}) \leq \frac{2\tau}{e} \frac{|A_i|}{1-\delta}$$

Thus as $\tau \to 0$, the right-hand-side goes to zero and therefore $\tilde{\pi}^*$ becomes a Nash equilibrium of game $\mathcal{G}$. $\square$