Fictitious Play in Markov Games with Single Controller

MUHAMMED O. SAYIN, Electrical and Electronics Engineering, Bilkent University, Turkey
KAIQING ZHANG, Laboratory for Information & Decision Systems, Massachusetts Institute of Technology
ASUMAN OZDAGLAR, Laboratory for Information & Decision Systems, Massachusetts Institute of Technology

Certain but important classes of strategic-form games, including zero-sum and identical-interest games, have the *fictitious-play-property* (FPP), i.e., beliefs formed in fictitious play dynamics always converge to a Nash equilibrium (NE) in the repeated play of these games. Such convergence results are seen as a (behavioral) justification for the game-theoretical equilibrium analysis. Markov games (MGs), also known as stochastic games, generalize the repeated play of strategic-form games to dynamic multi-state settings with Markovian state transitions. In particular, MGs are standard models for multi-agent reinforcement learning – a reviving research area in learning and games, and their game-theoretical equilibrium analyses have also been conducted extensively. However, whether certain classes of MGs have the FPP or not (i.e., whether there is a behavioral justification for equilibrium analysis or not) remains largely elusive. In this paper, we study a new variant of fictitious play dynamics for MGs and show its convergence to an NE in n-player identical-interest MGs in which a single player controls the state transitions. Such games are of interest in communications, control, and economics applications. Our result together with the recent results in [42] establishes the FPP of two-player zero-sum MGs and n-player identical-interest MGs with a single controller (standing at two different ends of the MG spectrum from fully competitive to fully cooperative).

CCS Concepts: • **Theory of computation** → **Convergence and learning in games**.

Additional Key Words and Phrases: Fictitious play, Markov games, identical-interest games, zero-sum games

ACM Reference Format:
Muhammed O. Sayin, Kaiqing Zhang, and Asuman Ozdaglar. 2022. Fictitious Play in Markov Games with Single Controller. In *Proceedings of the 23rd ACM Conference on Economics and Computation (EC '22), July 11–15, 2022, Boulder, CO, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3490486.3538289

1 INTRODUCTION

Markov games (MGs), also known as stochastic games, since their introduction in [45], have been broadly used to model strategic interactions of multiple agents in dynamic environments with multiple states. The players' actions affect not only their immediate *stage-payoffs*, but also the state transitions, and therefore, their future stage-payoffs.[1] This powerful framework to model the sequential decision-making of multiple agents finds broad applications in both Engineering and Economics [3, 37]. Moreover, MGs also serve as the fundamental framework for multi-agent reinforcement learning [10, 29, 48].

Nash equilibrium (NE) [36], on the other hand, has been broadly used as a solution concept in game theory. One important justification of NE is that it is the natural outcome of the myopic learning dynamics of players that take greedy best response actions. Such a perspective has been extensively studied in strategic-form games (also known as normal-form or one-shot games), for

[1] Hereafter, we use *players* and *agents* interchangeably.

EC '22, July 11–15, 2022, Boulder, CO, USA

ACM ISBN 978-1-4503-9150-4/22/07.
https://doi.org/10.1145/3490486.3538289

best-response and fictitious-play types of learning dynamics [23, 26, 31, 47]. In particular, these non-equilibrium adaptation dynamics are referred to as being *uncoupled/independent*, and these games where fictitious-play dynamics converge are referred to as having *fictitious-play-property* (FPP) [34, 35, 41, 44]. For strategic-form games, it is well-known that several important classes of games enjoy the FPP, ranging from fully competitive to fully cooperative ones, with no modification of the fictitious play dynamics being used. This is especially a desired property for independent learning with uncoupled dynamics, where the players are oblivious to the structure of the underlying game while learning.

In stark contrast, the FPP of MGs remains largely elusive. Limited results have been established on uncoupled learning dynamics of non-equilibrium adaptation for MGs, as well as using it as the justifications for the equilibrium therein. Recently, [4, 28, 42] are the first set of results along this line, with focuses on either zero-sum or identical-interest MGs. Moreover, some learning dynamics [4, 28] are not fully independent in that all the players track a common set of parameters. This naturally leads to the following open question we are interested in:

Can we design independent learning dynamics with uncoupled update rules, which enjoy the fictitious-play-property for more than one class of Markov games?

To shed light on this open problem, we study the same (synchronous- and model-based version) learning dynamic in [42], an uncoupled fictitious-play dynamic that provably converges for zero-sum MGs, and investigate its convergence property in an important class of games: identical-interest MGs with single-controller. We summarize our contributions as follows.

Contributions. We study *two-timescale fictitious-play dynamics* for MGs, with independent and uncoupled update rules that combine the classical fictitious-play in the repeated play of strategic-form games with the Q-learning in solving Markov decision processes. We show that this natural learning dynamic converges to an NE in both n-player identical-interest MGs (with single-controller) and two-player zero-sum MGs. In other words, these MGs, standing at two different ends of the MG spectrum, have the FPP. To the best of our knowledge, this appears to be the first fictitious-play type learning dynamics for MGs that enjoys this property. To establish the results, we develop new techniques to handle the challenges due to: 1) non-uniqueness of the NE value and the non-contracting property of the NE operator in identical-interest games; 2) non-monotonicity of the value function estimates when studying the discrete-time updates directly; 3) the deviation from the identical-interest structure of stage-games during learning, caused by the independent and local updates by each player.

1.1 Related work

We summarize the most related literature as follows.

Fictitious-play dynamics/property. Fictitious-play, a simple and independent learning dynamic that has been extensively studied for the repeated play of strategic-form games, was first introduced by [9]. The dynamic has then been shown to converge to an equilibrium in multiple classes of strategic-form games, including zero-sum [41], identical-interest [35], and certain general-sum games [6, 7, 34, 44]. Recall that these games are referred to as having the FPP [35].

For MGs, the FPP has not been understood until recently in [4, 28, 42], which are the most related works to the present one. [28] presents a continuous-time best-response dynamic for zero-sum MGs and embeds the discrete-time update into a continuous-time one. A single continuation payoff (common among the players) is maintained by all players, which makes the update rule not fully decoupled. [42] proposes fictitious play dynamics with uncoupled update rules, also for the zero-sum setting, where the continuation payoffs are updated locally using each player's own belief, yielding a more natural dynamic. Our learning dynamic is thus also based on that in [42]. Very recently, [4] studies fictitious play for identical-interest MGs. The learning dynamic also uses a common continuation payoff, and the discrete-time dynamic with convergence guarantees follows a *single-timescale* update rule. It is unclear if the same learning dynamic converges in other types of MGs. In fact, studying the convergence of *two-timescale* learning dynamics with local updates has been posted as an open question in [4], which is one of the main focuses of the present work.

Independent learning in MGs. Besides the fictitious-play dynamics in [4, 28, 42], other independent learning dynamics have also been proposed for MGs. [1] studies decentralized Q-learning for MGs by focusing only on stationary pure strategies (saying which pure action to play at which state). This restriction allows them to transform the underlying MG into a strategic-form game in which actions correspond to stationary pure strategies (which are finitely many contrary to stationary mixed strategies). Players can learn the payoffs of the associated strategic-form game (without observing others' actions) with coordinated exploration phases in which they do not change their strategies to create a stationary environment. The dynamic presented can converge to a (stationary pure-strategy) equilibrium if the associated normal-form game is weakly acyclic with respect to best (or better) response dynamics. The finite-sample complexity of the algorithm is also established recently in [20]. In contrast, our learning dynamic can converge to a stationary mixed-strategy equilibrium, which is essential for a global convergence result across the MG spectrum, as a pure-strategy equilibrium does not exist in general, e.g., in zero-sum games. [39] develops actor-critic learning dynamics that are decentralized, for a special class of MGs with a "multistage" structure, where each state is assumed to be visited at most once. In [13], independent policy gradient methods with a two-timescale (asymmetric) stepsizes between players have been studied for the zero-sum setting, with non-asymptotic convergence guarantees. Later, [43] developed decentralized Q-learning dynamic that is symmetric, but with only asymptotic convergence guarantees in the zero-sum setting. More recently, for Markov potential games, which also includes identical-interest MGs as an example, such independent policy gradient algorithms are also shown to converge [15, 19, 25, 49]. For episodic MGs, [24, 30, 46] establish the regret guarantees of decentralized learning algorithms in the online exploration setting.

MGs with single controller. An important subclass of MGs is the ones with single controller [17, 38], where one of the players dominates and controls the transitions of the system dynamics (though the reward functions are still affected jointly by all players). Such a model finds applications in communications, control, and economics [2, 16]. It also has natural connection with sequential (or online) learning [12, 21]. Learning in single-controller MGs are mostly focused on the zero-sum case [8, 21, 40]. [8] studies a model-based approach with polynomial time complexity in achieving near-optimal return. [21] investigates the relationship between regret minimization and solving single-controller MGs, by reducing this model to an online linear optimization problem. [40] develops a policy optimization algorithm based on the idea of fictitious play, with regret guarantees in the episodic setting. It is unclear yet if these algorithms also converge to an NE in other classes of MGs.

1.2 Organization

The rest of the paper is organized as follows. We provide a formulation of MGs (with single controller) in §2 and describe the fictitious play dynamic in MGs in §3. We present the main convergence results and the proof of convergence in §4. We conclude the paper in §5 with some remarks.

2 MARKOV GAMES WITH SINGLE CONTROLLER

Consider an n-player MG described by a tuple $\langle S, A, \{r^i\}_{i\in[n]}, p, \gamma\rangle$.[2] The game has *finitely* many states and S denotes the set of states. At each state $s \in S$, each player i can take an action a^i from a *finite* action set A^i, and $A = \bigtimes_i A^i$ denotes the set of action profiles $a = (a^i)_{i\in[n]}$.[3] Over discrete-time $k = 0, 1, 2, \ldots$, the state of the game, s, transitions to a state s' according to the transition probability $p(s'|s, a)$ depending only on the current state s and action profile a. At each stage k, each player i receives a *stage-payoff* $r^i(s, a)$ depending only on the current state s and action profile a while the players take actions simultaneously. Their objective is to maximize the discounted sum of their expected stage-payoffs over infinite horizon with the discount factor $\gamma \in [0, 1)$.

MGs can be viewed as an extension of Markov decision processes to multi-agent settings. [45] (and later [18]) showed that there always exists a *Markov stationary* equilibrium in two-player zero-sum (and n-player general-sum) MGs such that players take actions according to stationary (possibly mixed) strategies depending only on the current state.[4] We denote the stationary mixed-strategy of player i by $\pi^i : S \to \Delta(A^i)$.[5] Correspondingly, $\pi = (\pi^i)_{i\in[n]}$ denotes the strategy profile and Π denotes the space of strategy profiles, i.e., $\pi \in \Pi$. We define

$$u^i(s; \pi) := \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k) \,\middle|\, s_0 = s\right\}, \quad \forall s \in S \text{ and } \pi \in \Pi, \tag{1}$$

where (s_k, a_k) denotes the state and action profile at stage k, and the expectation is taken with respect to the randomness induced from the stochastic state transitions and mixed strategies of players. With slight abuse of notation, we let $u^i(\pi) := \mathbb{E}\{u^i(s_0, \pi)\}$, where the expectation is taken with respect to the initial state distribution. Therefore, $u^i(\pi)$ corresponds to the discounted sum of expected stage-payoffs of player i under strategy profile π.

Definition 1 (Markov Stationary Nash Equilibrium). We say that strategy profile $\pi_* \in \Pi$ is a Markov stationary Nash equilibrium of the n-player MG provided that

$$u^i(\pi_*) \geq u^i(\pi^i, \pi_*^{-i}), \quad \forall \pi^i \text{ and } i = 1, \ldots, n. \tag{2}$$

Hereafter, NE refers to Markov stationary Nash equilibrium. We say that an MG has *zero-sum* or *identical-interest* structure if $\sum_i r^i(s, a) = 0$ or $r^i(s, a) = r(s, a)$ for all (s, a) for some $r : S \times A \to \mathbb{R}$, respectively. In this paper, we focus on *single-controller* MGs where state transitions probabilities depend on the actions of a single player, e.g.,

$$p(s'|s, a) = p(s'|s, a^i), \quad \forall (s, a, s'). \tag{3}$$

Note that since the reward functions, $r^i(s, a)$'s, are affected by the joint action of all players, the accumulated expected payoff of player i still depends on the joint strategy of all players. Hence, when the strategy of other players changes over time, the environment faced by one player is still *non-stationary.* This is the key challenge in establishing the convergence of learning in MGs.

Indeed, single-controller MGs are common models in the literature [17, 38], and find broad applications in communications [16] and traveling inspector problems [17, Chapter 6]. They also have natural connections with regret minimization for sequential (or online) learning [12, 21].

[2] For easy referral, we set player i as the typical player while $-i := \{j \in [n] \mid j \neq i\}$ corresponds to the set of players other than player i.

[3] The formulation can be extended to state-dependent action sets straightforwardly.

[4] Such equilibrium is also referred to as *Markov perfect equilibrium* [32, 33].

[5] We denote the probability simplex over the set A by $\Delta(A)$.

3 FICTITIOUS PLAY IN MARKOV GAMES

Within an MG, stage-wise interactions among players can be viewed as they are playing *auxiliary stage-games* specific to each state whenever the associated state gets visited. In each stage-game, players simultaneously take actions while they can mix their actions independently. Players observe the joint action of all players and receive the associated immediate stage-payoff. However, the payoffs of these stage-games consist of immediate stage-payoffs and continuation payoffs (due to the objectives (1) defined over infinite horizon). The players can compute the continuation payoff based on the observations they make. We focus on the question that whether *non-equilibrium adaptation* of learning agents can converge to a stationary (mixed-strategy) equilibrium of the underlying MG or not if they adopt learning dynamics similar to the ones studied for strategic-form games with repeated play, such as fictitious play and its variants.

Formally, if player i knew that players $-i$ would play according to π^{-i} starting from the next stage, player i's payoff in the auxiliary stage-game associated with state s, denoted by $Q^i(s,a;\pi^{-i})$ and called *Q-function*, would satisfy the following fixed-point equation

$$Q^i(s,a;\pi^{-i}) = r^i(s,a) + \gamma \cdot \sum_{\tilde{s}} p(\tilde{s}|s,a) \max_{\tilde{a}^i \in A^i} \mathbb{E}_{\tilde{a}^{-i}\sim\pi^{-i}(s)} \left\{ Q^i(\tilde{s},\tilde{a};\pi^{-i}) \right\} \quad \forall (s,a). \tag{4}$$

This follows from the backward induction principle that player i would always take the actions maximizing her expected utility in (1). Correspondingly, the value of state s, denoted by $v^i(s;\pi^{-i})$ and called *value function*, would be given by

$$v^i(s;\pi^{-i}) = \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)} \left\{ Q^i(s,a;\pi^{-i}) \right\}, \quad \forall s. \tag{5}$$

Furthermore, if player i also knew that players $-i$ would play according to π^{-i} in the current auxiliary game, she would take the best response action, denoted by $a_*^i : S \to A^i$, satisfying

$$a_*^i(s) \in \operatorname*{argmax}_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)} \left\{ Q^i(s,a;\pi^{-i}) \right\}, \quad \forall s. \tag{6}$$

Neither the opponent's strategy nor the Q-function are directly available to player i in these auxiliary games. Therefore, each player i can form beliefs on every other player's stationary (mixed) strategy and her (local) Q-function based on an erroneous assumption that they are stationary as in the classical fictitious play. Then, they can update these beliefs independently based on the observations they make within the underlying MG. For the ease of exposition, we consider that every player follow the same learning dynamic with the same learning rates (or step sizes) and initializations. Hence, all players $-i$ form the same belief on the stationary strategy of player i. We denote this belief by $\pi_k^i : S \to \Delta(A^i)$ at stage k. Similarly, we denote the belief of player i on her Q-function at stage k by $Q_k^i : S \times A \to \mathbb{R}$. For notational convenience, we also introduce the value function estimates given by

$$v_k^i(s) := \max_{a^i\in A^i} \; \mathbb{E}_{a^{-i}\sim\pi_k^{-i}(s)} \{Q_k^i(s,a^i,a^{-i})\} \tag{7}$$

and the best response action given by

$$a_k^i(s) \in \operatorname*{argmax}_{a^i\in A^i} \; \mathbb{E}_{a^{-i}\sim\pi_k^{-i}(s)} \{Q_k^i(s,a^i,a^{-i})\}. \tag{8}$$

Correspondingly, we have $v_k^i(s) = \mathbb{E}_{a^{-i}\sim\pi_k^{-i}(s)}\{Q_k^i(s,a_k^i(s),a^{-i})\}$.

The players always take the best response (8) according to the beliefs they form and they update their beliefs according to an update rule combining the classical fictitious play and Q-learning together. From player i's viewpoint, the update rule is given by

$$\pi^j_{k+1}(s) = \pi^j_k(s) + \alpha_k\Big(a^j_k(s) - \pi^j_k(s)\Big), \quad \forall j \neq i \text{ and } s \in S, \tag{9a}$$

$$Q^i_{k+1}(s,a) = Q^i_k(s,a) + \beta_k\left(r^i(s,a) + \gamma\sum_{\tilde{s}} p(\tilde{s}|s,a)v^i_k(\tilde{s}) - Q^i_k(s,a)\right), \quad \forall (s,a), \tag{9b}$$

where $\{\alpha_k, \beta_k \in (0,1)\}_{k\geq 0}$ are step sizes and the beliefs are initialized as, e.g., $\pi^j(s) = \frac{1}{|A^j|}\mathbf{1}$ and $Q^i_0(s,a) = 0$ for all (s,a).[6] In (9a), π^j_k gets updated to a convex combination of the current action and the previous belief. On the other hand, in (9b), Q^i_k gets updated to a convex combination of the Q-function realized (according to one step iteration of the fixed-point equation (4) based on the value function estimate (7)) and the previous belief. The weights of the new observations in these convex combinations are determined according to the step sizes $\{\alpha_k, \beta_k\}_{k\geq 0}$.

Note that if there was a single state, then the underlying MG would reduce to the repeated play of a strategic-form game, and correspondingly, (9) would reduce to (9a) for which $Q^i_k \equiv r^i$, which is indeed the classical fictitious play dynamic. On the other hand, if there was a single player, then the MG would reduce to a Markov decision process, and correspondingly, (9) would reduce to (9b), which is indeed the Q-value iteration with smoothing updates (whose model-free version is known as Q-learning). The learning dynamic in (9) combines them together with different step sizes $\{\alpha_k, \beta_k\}_{k\geq 0}$ for learning in MGs.

Remark 2 (Comparison to existing related learning dynamics). *The learning dynamic in* (9) *is a synchronous and model-based version of the fictitious play dynamics in [42] focusing on learning in zero-sum MGs.*[7] *Two important features of the learning dynamic are: 1) the belief update and the Q-function update are performed in a two-timescale fashion; 2) each player maintains her own local estimates of the Q-functions, which are generally not common among players. In contrast, in the closely related recent works [4, 28], a single continuation payoff (common among players) is assumed to be maintained during learning. This way, the stage-games encountered during learning are always zero-sum [28] or identical-interest [4], and some implicit coordination among the players is required. Our learning dynamic is coordination-free and completely uncoupled, and are thus believed to be more natural. In fact, studying this two-timescale learning dynamic with independent Q-function updates has been posted as an interesting open question in [4], with non-trivial technical challenges to address. Finally, another motivation of studying* (9) *is to find a unified learning dynamic that converges for both zero-sum and identical-interest MGs, i.e., being agnostic to the types of games, a desired property of uncoupled dynamics.*

4 CONVERGENCE RESULTS AND PROOFS

Recall that the classical fictitious play is known to converge to an NE in certain but important classes of strategic-form games played repeatedly, such as zero-sum and identical-interest ones. The following theorem shows that the two-timescale fictitious-play dynamic in Eq. (9) possesses similar universality by converging to an NE in both two-player zero-sum and multi-player identical-interest MGs with single controller. The proof is provided later in §4.1.

[6] Consider actions as pure strategies, i.e., $a^i \in A^i \subset \Delta(A^i)$.

[7] The update (9) is more like a computational method similar to the ones in [4, 28] contrary to [42] since players play the auxiliary stage-game associated with each state at every stage. This yields a relaxation on the convergence guarantees by not requiring the underlying Markov chain to ensure infinitely often visit at every state.

Theorem 3. *The update* (9) *converges to an NE, described in Definition 1, in single-controller MGs with two-player zero-sum or multi-player identical-interest structure provided that the step sizes satisfy the usual two-timescale learning conditions that*

(i) *Vanishing rates:* $\alpha_k \longrightarrow 0$ *and* $\beta_k \longrightarrow 0$*, as* $k \longrightarrow \infty$*,*
(ii) *Sufficiently slow decay:* $\sum_{k\geq 0} \alpha_k = \infty$ *and* $\sum_{k\geq 0} \beta_k = \infty$*,*
(iii) *Sufficiently fast decay:* $\sum_{k\geq 0} \alpha_k^2 < \infty$*,*
(iv) *Two-timescale rates:* $\alpha_k \geq \beta_k$ *for all* $k \geq 0$ *and* $\beta_k/\alpha_k \longrightarrow 0$*, as* $k \longrightarrow \infty$*.*

Particularly, there exists $Q_* : S \times A \longrightarrow \mathbb{R}$ *such that*

$$\lim_{k\to\infty} Q_k^i(s,a) = Q_*(s,a) \quad \forall (i,s,a)$$

and Q_* *corresponds to the Q-function associated with some stationary equilibrium* $\pi_* = (\pi_*^i)_{i\in[n]}$ *of the underlying game and*

$$\lim_{k\to\infty} \pi_k^i(s) = \pi_*^i(s), \quad \forall (i,s).$$

We emphasize that the conditions listed are sufficient to ensure convergence of the update (9) in both classes of games. For example, the dynamic in (9) can converge to an equilibrium in two-player zero-sum MGs without Assumption (iii) on sufficiently fast decay of α_k. On the other hand, (9) can converge to an equilibrium in identical-interest MGs with single controller also in the single-timescale scheme where $\alpha_k = \beta_k$.

The following corollary to Theorem 3 shows that the convergence result can be generalized to the case where the stage-payoffs satisfy the following potential-game-like condition similar to the case in one-shot games. The proof is deferred to Appendix E.

Corollary 4. *Suppose that the step sizes satisfy the conditions listed in Theorem 3 and the stage-payoff functions satisfy*

$$r^j(s,\tilde{a}^j,a^{-j}) - r^j(s,a) = r^i(s,\tilde{a}^j,a^{-j}) - r^i(s,a), \quad \forall (s,a), \tilde{a}^j, \text{ and } j \neq i \tag{10}$$

given that player i is the single controller. Then, the update (9) *converges to an equilibrium in single-controller MGs. Particularly, there exists* $Q_*^i : S \times A \longrightarrow \mathbb{R}$ *for each i, which is not necessarily common now, such that*

$$\lim_{k\to\infty} Q_k^i(s,a) = Q_*^i(s,a) \quad \forall (i,s,a).$$

and $\{Q_*^i\}_{i\in[n]}$ *correspond to the Q-functions associated with some stationary equilibrium* $\pi_* = (\pi_*^i)_{i\in[n]}$ *of the underlying game and*

$$\lim_{k\to\infty} \pi_k^i(s) = \pi_*^i(s), \quad \forall (i,s).$$

Zero-sum games and identical-interest games stand at the two extreme ends of the game spectrum from fully competitive to fully cooperative. They possess distinct features. For example, the equilibrium value of a zero-sum (strategic-form) game is *unique* even though there may exist multiple equilibria. Furthermore, minimax value of a game is a non-expansive function like the maximum value. Therefore, [45] could introduce a contraction operator for two-player zero-sum MGs as a counterpart of the Bellman operator in Markov decision processes. Later [28, 42] showed that the contraction property in the evolution of the value function estimates could be approximated with asymptotically negligible error also in *non-equilibrium* learning dynamics through a two-timescale learning scheme.

4.1 Proof of Theorem 3

The proof of Theorem 3 for two-player zero-sum MGs (with single controller) follows from the identical steps in [42, Theorem 4.3] where the convergence properties of the asynchronous version of (9) is characterized. On the other hand, equilibrium values of an identical-interest game are not necessarily unique. In the absence of powerful non-expansiveness and correspondingly contraction property, we need a different technical tool to characterize its convergence properties.

The main premise behind the proof for n-player identical-interest case is that the limiting differential inclusion of the (9a) is the continuous-time best response dynamic in an identical-interest game due to the two-timescale framework. Therefore, the maximum expected values of auxiliary stage games are monotonically non-decreasing in this continuous-time approximation. Correspondingly, if the value function estimates were monotonically non-decreasing in the original discrete-time updates, then the Q-function estimates would also be monotonically non-decreasing. Hence, we could have concluded their convergence since they are bounded by the update rule (9b). However, the discrete-time dynamic does not necessarily lead to an increase in the value function estimates across subsequent stages in general. To address this challenge, we consider the deviation from the monotonicity across multiple stages rather than just subsequent ones, as in [4].

Remark 5 (Challenge due to *Independent* Q-update). *Similar to the zero-sum case in [42], another challenge arises due to the deviation from the identical-interest structure in the auxiliary stage-games since players update their beliefs on the Q-function according to* (9b) *via the maximum expected continuation payoff they believe they would get, as described in* (7). *This challenge would not be observed if players had a common Q-function estimates, i.e., $Q_k^i \equiv Q_k$ for all i for some Q_k. For example, [4] uses an update similar to*

$$Q_{k+1}^i(s,a) = Q_k^i(s,a) + \beta_k \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) \mathbb{E}_{a' \sim \pi_k(s')}\{Q_k^i(s',a')\} - Q_k^i(s,a) \right), \tag{11}$$

for all (i,s,a), rather than (9b). *Such an update guarantees that each auxiliary stage game has identical-interest structure provided that $Q_0^i = Q_0$ for all i. However,* (11) *is still prone to deviation from the identical-interest structure without a common initialization. Furthermore, computation of $\mathbb{E}_{a' \sim \pi_k(s')}\{Q_k^i(s',a')\}$ by player i implies that player i forms belief π^i on her own strategy as if she is playing according to a stationary mixed-strategy even though she always takes (greedy) best response actions against her opponents. Therefore, the independent Q-update* (9b) *contrary to the coupled one* (11) *is a relatively more natural dynamic for practical applications. The characterization of its convergence properties can provide a stronger justification for equilibrium analysis in MGs. To address this challenge, we focus on single-controller MGs, where the auxiliary stage games are strategically equivalent to identical-interest games if the beliefs on Q-functions are initialized the same and become strategically equivalent to identical-interest games at a sufficiently fast rate if they do not have common initialization.*

Given the update (9b), we define

$$\Upsilon_k^i(s,a) := r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) v_k^i(s') - Q_k^i(s,a), \quad \forall (i,s,a) \tag{12}$$

such that

$$Q_{k+1}^i(s,a) = Q_k^i(s,a) + \beta_k \Upsilon_k^i(s,a) \quad \forall (i,s,a). \tag{13}$$

Note that $Q_k^i(s,a)$, for each (i,s,a), is bounded from above by $\frac{1}{1-\gamma} \max_{(s,a)} r(s,a)$ by the definition of the update (9b) and (7) since the step size $\{\beta_k \in (0,1)\}_{k \geq 0}$.

The following proposition provides a monotonicity-like condition on the changes of the estimates accumulated across multiple stages (not just the subsequent ones) to prove the convergence of $\{Q_k^i(s,a)\}_{k\geq 0}$. The proof is deferred to Appendix A.

Proposition 6. *Consider a (real-valued) bounded sequence $\{Q_k^i(s,a)\}_{k\geq 0}$ for each $(i,s,a) \in [n] \times S \times A$ (with $n < \infty$ and $|S \times A| < \infty$) evolving according to*

$$Q_{k+1}^i(s,a) = Q_k^i(s,a) + \beta_k \Upsilon_k^i(s,a), \quad \forall (i,s,a) \tag{14}$$

for some $\{\Upsilon_k^i(s,a)\}_{k\geq 0}$ for each (i,s,a) and step size $\beta_k \geq 0$. If we have

$$\liminf_{k_1\to\infty} \inf_{k_2\geq k_1} \sum_{k=k_1}^{k_2} \beta_k \Upsilon_k^i(s,a) \geq 0, \quad \forall (i,s,a), \tag{15}$$

then there exists $Q_*^i : S \times A \to \mathbb{R}$ such that

$$\lim_{k\to\infty} Q_k^i(s,a) = Q_*^i(s,a), \quad \forall (i,s,a). \tag{16}$$

Henceforth, we focus on proving (15) (through a more tractable lower bound) to conclude the convergence of (9b). To this end, we define an auxiliary parameter bounding $\Upsilon_k^i(s,a)$ for each (s,a) from below as

$$\underline{u}_k^i := \min_{(s,a)} \left\{ r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) \mathbb{E}_{a'\sim\pi_k(s)}\{Q_k^i(s',a')\} - Q_k^i(s,a) \right\}, \tag{17}$$

i.e., we have $\underline{u}_k^i \leq \Upsilon_k^i(s,a)$ for all (s,a), since $v_k^i(s') - \mathbb{E}_{a'\sim\pi_k(s)}\{Q_k^i(s',a')\} \geq 0$ for each s' by the definition of v_k^i, as described in (7). Instead of (15), we can, now, focus on proving the asymptotic non-negativity of the more tractable lower bound:

$$\boxed{\liminf_{k_1\to\infty} \inf_{k_2\geq k_1} \sum_{k=k_1}^{k_2} \beta_k \underline{u}_k^i \geq 0}. \tag{18}$$

If we could have shown that there exists some κ such that $\underline{u}_k^i \geq 0$ for all $k \geq \kappa$, then we would have concluded (18). We do not necessarily have it. On the other hand, showing the asymptotic non-negativity of $\underline{u}_k^i$ would not be sufficient to conclude (18). Hence, we look for some stronger conditions. The following lemma provides a characterization of the evolution of $\{\underline{u}_k^i\}_{i\geq 0}$ (from below) in terms of some *absolutely summable* sequence. The proof is deferred to Appendix B.

Lemma 7. *Given that players follow the dynamic described in* (9) *in an identical-interest MG with single controller i, the evolution of $\underline{u}_k^i$, described in* (17)*, satisfies the following inequality:*

$$\boxed{\underline{u}_{k+1}^i \geq \underline{u}_k^i(1-(1-\gamma)\beta_k) + \underline{e}_k} \tag{19}$$

for all $k \geq 0$ and for some absolutely summable sequence $\{\underline{e}_k\}_{k\geq 0}$.

Remark 8. *We emphasize that the single-controller identical-interest structure plays an important role in ensuring that there exists such an absolutely summable sequence.*

Given that $\{\underline{e}_k\}_{k\geq 0}$ are absolutely summable, we can next invoke the following lemma showing that $\{\underline{u}_k^i\}_{k\geq 0}$ satisfying (19) also satisfies (18). Hence, it can also be of interest on its own. The proof is deferred to Appendix C.

Lemma 9. *Given step sizes $\{\beta_k \in (0,1)\}_{k\geq 0}$ vanishing, i.e., $\beta_k \to 0$ as $k \to \infty$, sufficiently slowly such that $\sum_{k\geq 0} \beta_k = \infty$, consider a sequence $\{\underline{u}_k^i\}_{k\geq 0}$ satisfying* (19) *for some discount factor $\gamma \in [0,1)$ and some absolutely summable error term $\underline{e}_k$, i.e., $\sum_{k\geq 0} |\underline{e}_k| < \infty$. Then, we have* (18).

Lemmas 7 and 9 imply (18), and therefore, (15) by the definition of $\underline{u}_k^i$, as described in (17). Hence, Proposition 6 yields that $\{Q_k^i\}_{i\in[n]}$, i.e., (9b), is convergent. In other words, there exists some $Q_*^i : S \times A \to \mathbb{R}$ such that

$$\lim_{k\to\infty} Q_k^i(s,a) = Q_*^i(s,a). \tag{20}$$

On the other hand, for players other than the controller, say player $j \neq i$, the payoff in the auxiliary stage game $Q_k^j(s,a)$ is always strategically equivalent to $Q_k^i(s,a)$, as shown in the proof of Lemma 7. Note that we have not characterized the limit of Q_k^i yet.

Lastly, we can conclude that the update (9) indeed converges to an equilibrium based on the following lemma (which can be viewed as a corollary to [27, Theorem 4]). This lemma characterizes the limit set of the fictitious-play in terms of its limiting continuous-time best response dynamic for the cases where the underlying game becomes stationary asymptotically. Hence, it can also be of interest on its own. The proof is deferred to Appendix D.

Lemma 10. *Given step sizes $\{\alpha_k \in (0,1)\}_{k\geq 0}$ vanishing, i.e., $\alpha_k \to 0$ as $k \to \infty$, sufficiently slowly such that $\sum_{k\geq 0} \alpha_k = \infty$, consider the update of $\pi^i \in \Delta(A^i)$ for each $i \in [n]$ and finite set A^i, given by*

$$\pi_{k+1}^i = \pi_k^i + \alpha_k \left(a_k^i - \pi_k^i\right), \quad \forall i, \tag{21}$$

where $a_k^i \in \Delta(A^i)$ satisfies

$$a_k^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\{Q_k^i(a^i, a^{-i})\}. \tag{22}$$

Suppose that $Q_k^i(a) \to Q_*^i(a)$ for all (i,a) as $k \to \infty$ for some $Q_*^i : A \to \mathbb{R}$. Then, the limit set of (21) *is a connected internally chain-recurrent set of the following best response differential inclusion*

$$\dot{\pi}^i + \pi^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}}\{Q_*^i(a^i, a^{-i})\}. \tag{23}$$

Based on (20) and the strategic equivalence of Q_k^j for each $j \neq i$ to Q_k^i, Lemma 10 yields that the limit set of (9a) is contained in the connected internally chain-recurrent set of the differential inclusion

$$\dot{\pi}^j(s) + \pi^j(s) \in \operatorname*{argmax}_{a^j \in A^j} \mathbb{E}_{a^{-j}\sim\pi^{-j}(s)}\{Q_*^i(s, a^j, a^{-j})\}, \quad \forall j, \tag{24}$$

which is the continuous-time best response dynamic in an identical-interest game with the payoff $Q_*^i(s,\cdot)$. [5, Theorem 5.5 and Remark 5.6] yield that the limit set of every solution of (24), and therefore (9a), is a connected set of equilibria along which $\mathbb{E}_{a\sim\pi(s)}\{Q_*^i(s,a)\}$ is constant. In other words, the beliefs $\{\pi_k^i(s)\}_{i\in[n]}$ converge to an equilibrium, for each s. Given the convergence of these beliefs, the update (9b) yields that the Q-function estimates of every player also converge to the Q-function associated with the equilibrium strategies and

$$\lim_{k\to\infty} Q_k^j(s,a) = Q_*^i(s,a), \quad \forall (s,a) \text{ and } j \neq i. \tag{25}$$

This completes the proof. □

5 DISCUSSIONS AND CONCLUSIONS

In this paper, we investigated the convergence properties of a new variant of fictitious play dynamics for n-player identical-interest MGs with single controller. Together with the fact that the same learning dynamic also converges to an equilibrium in two-player zero-sum MGs, we established, to the best of our knowledge, the first universal-type fictitious-play-property for more than one class of MGs. The results have thus further justified (Markov stationary) NE in MGs as an outcome of myopic non-equilibrium adaptation. We believe our results have opened up fruitful research directions for future work.

- **Fictitious-play-property for other classes of MGs.** Our results illustrate the promise of our two-timescale fictitious-play dynamic in achieving universal-type convergence in more than one class of MGs. It is interesting to further expand the types of MGs that enjoys the fictitious-play-property, mirroring the results for strategic-form games (cf. [6, 7, 34, 44]).
- **Model-free learning with asynchronous updates.** With a focus on the uncoupled learning dynamics with independent Q-updates, we studied the synchronous-update rule with the knowledge of the transition dynamics. As a standard model for multi-agent reinforcement learning, it is imperative to investigate the convergence of our dynamics in the model-free asynchronous setting. Note that with common Q-updates and single-timescale update-rule, [4] has studied the asynchronous update case with small enough discount factor γ. The model-free learning for fictitious-play in MGs beyond the zero-sum case remains largely open.
- **Convergence rate characterization & faster rates.** It is known that in the worse-case, fictitious play can have exponentially-slow rate when learning in strategic-form games [14]. It would be interesting to understand and compare the convergence rates of the two-timescale fictitious-play in our work and [42], with that of the single-timescale one in [4]. It is also worth exploring the effectiveness of regularization to accelerate convergence of fictitious play dynamics, as in strategic-form games [12].

ACKNOWLEDGMENTS

M. O. Sayin was supported by TUBITAK BIDEB 2232-B International Fellowship for Early Stage Researchers grant 121C124. K. Zhang and A. Ozdaglar were supported by DSTA grant 031017-00016.

REFERENCES

[1] G. Arslan and S. Yuksel. 2017. Decentralized Q-learning for stochastic teams and games. *IEEE Trans. Automat. Control* 62, 4 (2017), 1545–1558.

[2] T. Başar. 1986. *Dynamic Games and Applications in Economics.* Vol. 265. Springer Science & Business Media.

[3] T. Başar and G. J. Olsder. 1998. *Dynamic Noncooperative Game Theory.* SIAM.

[4] L. Baudin. 2021. Best-Response Dynamics and Fictitious Play in Identical Interest Stochastic Games. *arXiv preprint arXiv:2111.04317* (2021).

[5] M. Benaim, J. Hofbauer, and S. Sorin. 2005. Stochastic Approximations and Differential Inclusions. *SIAM J. Control Optim.* 44, 1 (2005), 328–348.

[6] U. Berger. 2005. Fictitious play in 2xn games. *Journal of Economic Theory* 120, 2 (2005), 139–154.

[7] U. Berger. 2008. Learning in games with strategic complementarities revisited. *Journal of Economic Theory* 143, 1 (2008), 292–301.

[8] R. I. Brafman and M. Tennenholtz. 2000. A near-optimal polynomial time algorithm for learning in certain classes of stochastic games. *Artificial Intelligence* 121, 1-2 (2000), 31–47.

[9] G. W. Brown. 1951. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation* 13, 1 (1951), 374–376.

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

[11] O. Candogan, A. Ozdaglar, and P. A. Parrilo. 2013. Dynamics in near-potential games. *Games and Economic Behavior* 82 (2013), 66–90.

[12] N. Cesa-Bianchi and G. Lugosi. 2006. *Prediction, Learning, and Games.* Cambridge University Press.

[13] C. Daskalakis, D. J. Foster, and N. Golowich. 2020. Independent Policy Gradient Methods for Competitive Reinforcement Learning. In *Advances in Neural Information Processing Systems.*

[14] C. Daskalakis and Q. Pan. 2014. A counter-example to Karlin's strong conjecture for fictitious play. In *IEEE Annual Symposium on Foundations of Computer Science.* IEEE, 11–20.

[15] D. Ding, C. Wei, K. Zhang, and M. Jovanovic. 2022. Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence. *arXiv preprint arXiv:2202.04129* (2022).

[16] A. Eldosouky, W. Saad, and D. Niyato. 2016. Single controller stochastic games for optimized moving target defense. In *IEEE International Conference on Communications*. IEEE, 1–6.

[17] J. Filar and K. Vrieze. 2012. *Competitive Markov Decision Processes*. Springer Science & Business Media.

[18] A. M. Fink. 1964. Equilibrium in stochastic n-person game. *Journal of Science Hiroshima University Series A-I* 28 (1964), 89–93.

[19] R. Fox, S. McAleer, W. Overman, and I. Panageas. 2021. Independent Natural Policy Gradient Always Converges in Markov Potential Games. *arXiv preprint arXiv:2110.10614* (2021).

[20] Z. Gao, Q. Ma, T. Başar, and J. R. Birge. 2021. Finite-Sample Analysis of Decentralized Q-Learning for Stochastic Games. *arXiv preprint arXiv:2112.07859* (2021).

[21] P. Guan, M. Raginsky, R. Willett, and D. Zois. 2016. Regret minimization algorithms for single-controller zero-sum stochastic games. In *IEEE Conference on Decision and Control*. IEEE, 7075–7080.

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

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

[24] C. Jin, Q. Liu, Y. Wang, and T. Yu. 2021. V-Learning–A Simple, Efficient, Decentralized Algorithm for Multiagent RL. *arXiv preprint arXiv:2110.14555* (2021).

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

[26] D. S. Leslie and E. J. Collins. 2005. Individual Q-learning in normal form games. *SIAM J. Control Optim.* 44, 2 (2005), 495–514.

[27] D. S. Leslie and E. J. Collins. 2006. Generalized weakened fictitious play. *Games and Economic Behavior* 56, 2 (2006), 285–298.

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

[29] M. L. Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning (ICML)*.

[30] W. Mao and T. Başar. 2022. Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games and Applications* (2022), 1–22.

[31] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma. 2009. Payoff-based dynamics for multiplayer weakly acyclic games. *SIAM Journal on Control and Optimization* 48, 1 (2009), 373–396.

[32] E. Maskin and J. Tirole. 1988. A theory of dynamic oligopoly, I: Overview and quantity competition with large fixed costs. *Econometrica: Journal of the Econometric Society* (1988), 549–569.

[33] E. Maskin and J. Tirole. 1988. A theory of dynamic oligopoly, II: Price competition, kinked demand curves, and Edgeworth cycles. *Econometrica: Journal of the Econometric Society* (1988), 571–599.

[34] K. Miyasawa. 1961. On the convergence of the learning process in a 2x2 non-zero-sum game. *Economic Research Program, Princeton University, Research Memorandum* 33 (1961).

[35] D. Monderer and L. S. Shapley. 1996. Fictitious play property for games with identical interests. *Journal of Economic Theory* 68, 1 (1996), 258–265.

[36] J. Nash. 1951. Non-cooperative games. *Annals of Mathematics* (1951), 286–295.

[37] A. Neyman and S. Sorin. 2003. *Stochastic Games and Applications*. Vol. 570. Springer Science & Business Media.

[38] T. Parthasarathy and T. Raghavan. 1981. An orderfield property for stochastic games when one player controls transition probabilities. *Journal of Optimization Theory and Applications* 33, 3 (1981), 375–392.

[39] J. Pérolat, B. Piot, and O. Pietquin. 2018. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*. 919–928.

[40] S. Qiu, X. Wei, J. Ye, Z. Wang, and Z. Yang. 2021. Provably Efficient Fictitious Play Policy Optimization for Zero-Sum Markov Games with Structured Transitions. In *International Conference on Machine Learning*. PMLR, 8715–8725.

[41] J. Robinson. 1951. An iterative method of solving a game. *Annals of Mathematics* 24 (1951), 296–301.

[42] M. O. Sayin, F. Parise, and A. Ozdaglar. in print. Fictitious play in zero-sum stochastic games. *SIAM J. Control Optim.* (in print).

[43] M. O. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar. 2021. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems* 34 (2021).

[44] A. Sela. 1999. Fictitious play in 'one-against-all' multi-player games. *Economic Theory* 14 (1999), 635–651.

[45] L. S. Shapley. 1953. Stochastic games. *Proceedings of National Academy of Science USA* 39, 10 (1953), 1095–1100.

[46] Z. Song, S. Mei, and Y. Bai. 2021. When Can We Learn General-Sum Markov Games with a Large Number of Players Sample-Efficiently? *arXiv preprint arXiv:2110.04184* (2021).

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

[48] K. Zhang, Z. Yang, and T. Başar. 2020. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *Studies in Systems, Decision and Control, Handbook on RL and Control* (2020).

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

A PROOF OF PROPOSITION 6

Note that if $\Upsilon_k^i(s,a) \geq 0$, then $\{Q_k^i(s,a)\}_{k\geq 0}$ would form a non-decreasing bounded sequence, which implies the existence of its limit. However, we do not necessarily have $\Upsilon_k^i(s,a) \geq 0$. As in [4], we can check monotonicity across multiple stages (from k_1 to k_2+1). For example, we have

$$Q_{k_2+1}^i(s,a) - Q_{k_1}^i(s,a) = \sum_{k=k_1}^{k_2} \beta_k \Upsilon_k^i(s,a), \quad \forall (i,s,a), \tag{26}$$

where the right-hand side is still not necessarily non-negative. Showing the difference goes to zero as $k_1 \to \infty$ would imply that $\{Q_k^i(s,a)\}_{k\geq 0}$ forms a Cauchy sequence, and therefore, it is convergent in the underlying Banach space. A relatively mild alternative (aligned with the intuition on monotonicity) is to show that the right-hand side becomes non-negative asymptotically as $k_1 \to \infty$ for any $k_2 \geq k_1$, i.e.,

$$\liminf_{k_1\to\infty} \inf_{k_2\geq k_1} \sum_{k=k_1}^{k_2} \beta_k \Upsilon_k^i(s,a) \geq 0 \quad \Rightarrow \quad \liminf_{k_1\to\infty} \left(\inf_{k_2\geq k_1} Q_{k_2+1}^i(s,a) - Q_{k_1}^i(s,a) \right) \geq 0, \tag{27}$$

for all (i,s,a). Then, (27) yields that for any $\epsilon > 0$, there exists κ such that

$$Q_k^i(s,a) \geq Q_\kappa^i(s,a) - \epsilon, \quad \forall k \geq \kappa, \tag{28}$$

for each (i,s,a).[8] This completes the proof due to the boundedness of the estimates.

B PROOF OF LEMMA 7

For the ease of notation, we define

$$Y_k^i(s,a) := r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) u_k^i(s') - Q_k^i(s,a), \tag{29}$$

where $u_k^i(s') := \mathbb{E}_{a'\sim\pi_k(s')}\{Q_k^i(s',a')\}$. Then, we have $\underline{u}_k^i = \min_{(s,a)}\{Y_k^i(s,a)\}$ for all i. Due to this dependence, it is instructive to examine the evolution of $Y_k^i(s,a)$, given by

$$\begin{aligned} Y_{k+1}^i(s,a) - Y_k^i(s,a) &= \gamma \sum_{s'\in S} p(s'|s,a)(u_{k+1}^i(s') - u_k^i(s')) - (Q_{k+1}^i(s,a) - Q_k^i(s,a)) \\ &= \gamma \sum_{s'\in S} p(s'|s,a)(u_{k+1}^i(s') - u_k^i(s')) - \beta_k \Upsilon_k^i(s,a) \\ &= \gamma \sum_{s'\in S} p(s'|s,a)(u_{k+1}^i(s') - u_k^i(s')) - \beta_k \left(Y_k^i(s,a) + \gamma \sum_{s'\in S} p(s'|s,a)\Delta_k^i(s') \right) \\ &= \gamma \sum_{s'\in S} p(s'|s,a)(u_{k+1}^i(s') - u_k^i(s') - \beta_k \Delta_k^i(s')) - \beta_k Y_k^i(s,a), \end{aligned} \tag{30}$$

[8] We can have a uniform $\epsilon > 0$ since there are only finitely many (i,s,a) triples.

where $\Delta^i_k(s') := v^i_k(s') - u^i_k(s') \geq 0$ for each s'. By the definition of u^i_k, the difference term in the parenthesis can be written as

$$\begin{aligned} u^i_{k+1}(s') - u^i_k(s') &= \mathbb{E}_{a\sim\pi_{k+1}(s')}\{Q^i_{k+1}(s',a)\} - \mathbb{E}_{a\sim\pi_k(s')}\{Q^i_k(s',a)\} \\ &\overset{(a)}{=} \mathbb{E}_{a\sim\pi_{k+1}(s')}\{Q^i_k(s',a)\} - \mathbb{E}_{a\sim\pi_k(s')}\{Q^i_k(s',a)\} + \beta_k\mathbb{E}_{a\sim\pi_{k+1}(s')}\{\Upsilon^i_k(s',a)\} \\ &\overset{(b)}{=} \alpha_k\left(\Delta^i_k(s') + \sum_{i\neq j}\Gamma^{ij}_k(s')\right) + O(\alpha_k^2) + \beta_k\mathbb{E}_{a\sim\pi_{k+1}(s')}\{\Upsilon^i_k(s',a)\}, \end{aligned} \tag{31}$$

where (a) follows from the update of Q^i_k, as described in (9b), (b) follows from the update of π^j_k for each $j \in [n]$, as described in (9a), and we define

$$\Gamma^{ij}_k(s') := \mathbb{E}_{a\sim\pi_k(s')}\{Q^i_k(s', a^j_k(s'), a^{-j}) - Q^i_k(s',a)\}, \quad \forall j \neq i. \tag{32}$$

Note that $\underline{u}^i_k \leq Y^i_k(s,a) \leq \Upsilon^i_k(s,a)\}$ for each (s,a). Therefore, combining (30) and (31), we obtain

$$Y^i_{k+1}(s,a) \geq \underline{u}^i_k(1-(1-\gamma)\beta_k) + \gamma\sum_{s'\in S}p(s'|s,a)e^i_k(s'), \tag{33}$$

where we define

$$\boxed{e^i_k(s') := \alpha_k\left(\left(1-\frac{\beta_k}{\alpha_k}\right)\Delta^i_k(s') + \sum_{i\neq j}\Gamma^{ij}_k(s')\right) + O(\bar{\alpha}_k(s')^2)}. \tag{34}$$

If we can find an absolutely summable lower bound on $e^i_k(\cdot)$, then the inequality (33) yields (19).

Next, we formulate an absolutely summable lower bound on $e^i_k(\cdot)$ based on the single-controller property of the underlying MG. Since a^i_k, as described in (8), is a best response action, we can write $\Delta^i_k(s') = v^i_k(s') - u^i_k(s')$ also as

$$\Delta^i_k(s') = \mathbb{E}_{a\sim\pi_k(s')}\{Q^i_k(s', a^i_k(s'), a^{-i}) - Q^i_k(s',a)\}. \tag{35}$$

We highlight the differences between $\Delta^i_k(s) \geq 0$, as described in (35), and $\Gamma^{ij}_k(s')$, as described in (32). In particular, a^j_k is a best response of player j according to her payoff function Q^j_k in the associated auxiliary stage-game and her belief π^{-j}_k about her opponents' strategies. Therefore, Γ^{ij}_k is not necessarily non-negative quite contrary to $\Delta^i_k \geq 0$ if we do not have $Q^i_k \equiv Q^j_k$ (e.g., see Remark 5).

On the other hand, by (32) and (35), the term Γ^{ij}_k can be written as

$$\Gamma^{ij}_k(s) = \Delta^j_k(s) + \mathbb{E}_{a\sim\pi_k(s)}\{\delta Q^{ij}_k(s, a^j_k(s), a^{-j}) - \delta Q^{ij}_k(s,a)\}, \tag{36}$$

where the first term on the right hand side is non-negative and we define $\delta Q^{ij}_k(s,a) := Q^i_k(s,a) - Q^j_k(s,a)$ for all (s,a). We can show that

$$\mathbb{E}_{a\sim\pi_k(s)}\{\delta Q^{ij}_k(s,a^j_k(s),a^{-j}) - \delta Q^{ij}_k(s,a)\} = 0 \quad \Rightarrow \quad \Gamma^{ij}_k \equiv \Delta^j_k \geq 0, \quad \forall j \neq i. \tag{37}$$

in MGs with single controllers. Particularly, the update (9b) yields that[9]

$$Q^i_{k+1}(s,a) = r(s,a)\sum_{l=0}^{k}\beta_l\left(\prod_{m=l+1}^{k}(1-\beta_m)\right) + \gamma\sum_{s'\in S}p(s'|s,a)\sum_{l=0}^{k}v^i_l(s')\beta_l\left(\prod_{m=l+1}^{k}(1-\beta_m)\right), \tag{38}$$

[9] We use the convention that $\prod_{m=l}^{k} c_m = 1$ if $k < l$.

for all $k \geq 0$, since the beliefs are initialized by $Q_0^i(s,a)$ for all (i,s,a). This implies that

$$\delta Q_k^{ij}(s,a) = \gamma \sum_{s'\in S} p(s'|s,a) \sum_{l=0}^{k-1} (v_l^i(s') - v_l^j(s'))\beta_l \left(\prod_{m=l+1}^{k-1} (1-\beta_m) \right). \tag{39}$$

Recall that if player i is the single controller, then we have

$$p(s'|s,\tilde{a}^j,a^{-j}) - p(s'|s,a) = p(s'|s,a^i) - p(s'|s,a^i) = 0, \quad \forall a, \tilde{a}^j \text{ and } j \neq i. \tag{40}$$

Hence, (39) and (40) yield (37). Correspondingly, we have

$$\gamma \sum_{s'} p(s'|s,a) e_k^i(s') \geq O(\alpha_k^2) =: \underline{e}_k. \tag{41}$$

Note that $\{\alpha_k^2\}_{k\geq 0}$ is absolutely summable by Assumption (iii) listed in Theorem 3. Hence, (33) and (41) lead to (19). This completes the proof. □

C PROOF OF LEMMA 9

Given (19), we can formulate a lower bound on $\underline{u}_k^i$ in terms of $\underline{u}_0^i$ and $\{\underline{e}_k\}$:

$$\begin{aligned}
\underline{u}_k^i &\geq \underline{u}_{k-1}^i(1-\tilde{\beta}_{k-1}) + \underline{e}_{k-1} \\
&\geq \underline{u}_{k-2}^i(1-\tilde{\beta}_{k-2})(1-\tilde{\beta}_{k-1}) + \underline{e}_{k-2}(1-\tilde{\beta}_{k-1}) + \underline{e}_{k-1} \\
&\cdots \\
&\geq \underline{u}_0^i \prod_{l=0}^{k-1}(1-\tilde{\beta}_l) + \sum_{l=0}^{k-1} \underline{e}_l \prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m),
\end{aligned} \tag{42}$$

where $\tilde{\beta}_k := (1-\gamma)\beta_k$ for notational convenience. We can incorporate the lower bound (42) on $\underline{u}_k^i$ into the summation in (18) as

$$\begin{aligned}
\sum_{k=k_1}^{k_2} \beta_k \underline{u}_k^i &\geq \sum_{k=k_1}^{k_2} \beta_k \left(\underline{u}_0^i \prod_{l=0}^{k-1}(1-\tilde{\beta}_l) + \sum_{l=0}^{k-1} \underline{e}_l \prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m) \right) \\
&\geq -\sum_{k=k_1}^{k_2} \beta_k \left(|\underline{u}_0^i| \prod_{l=0}^{k-1}(1-\tilde{\beta}_l) + \sum_{l=0}^{k-1} |\underline{e}_l| \prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m) \right) \\
&= -\frac{|\underline{u}_0^i|}{1-\gamma} \sum_{k=k_1}^{k_2} \tilde{\beta}_k \prod_{l=0}^{k-1}(1-\tilde{\beta}_l) - \frac{1}{1-\gamma} \sum_{k=k_1}^{k_2} \sum_{l=0}^{k-1} |\underline{e}_l| \tilde{\beta}_k \prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m).
\end{aligned} \tag{43}$$

By changing the order of summation at the second term, we have:

$$\begin{aligned}
-\frac{1}{1-\gamma} \sum_{k=k_1}^{k_2} \sum_{l=0}^{k-1} |\underline{e}_l| \tilde{\beta}_k \prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m) = &-\frac{1}{1-\gamma} \sum_{l=0}^{k_1-2} |\underline{e}_l| \sum_{k=k_1}^{k_2} \tilde{\beta}_k \prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m) \\
&-\frac{1}{1-\gamma} \sum_{l=k_1-1}^{k_2-1} |\underline{e}_l| \sum_{k=l+1}^{k_2} \tilde{\beta}_k \prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m).
\end{aligned} \tag{44}$$

We are interested in proving (18) and

$$\begin{aligned}\liminf_{k_1\to\infty}\inf_{k_2\geq k_1}\sum_{k=k_1}^{k_2}\beta_k\underline{u}_k^i \geq\ & \liminf_{k_1\to\infty}\inf_{k_2\geq k_1}\left(-\frac{|\underline{u}_0^i|}{1-\gamma}\sum_{k=k_1}^{k_2}\tilde{\beta}_k\prod_{l=0}^{k-1}(1-\tilde{\beta}_l)\right)\\ &+\liminf_{k_1\to\infty}\inf_{k_2\geq k_1}\left(-\frac{1}{1-\gamma}\sum_{l=0}^{k_1-2}|\underline{e}_l|\sum_{k=k_1}^{k_2}\tilde{\beta}_k\prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m)\right)\\ &+\liminf_{k_1\to\infty}\inf_{k_2\geq k_1}\left(-\frac{1}{1-\gamma}\sum_{l=k_1-1}^{k_2-1}|\underline{e}_l|\sum_{k=l+1}^{k_2}\tilde{\beta}_k\prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m)\right). \qquad (45)\end{aligned}$$

Therefore, showing the non-negativity of each term at the right-hand side implies (18). To this end, the following lemma enables us to rewrite the inner summations in (43) and (44) as a difference of two partial products.

Lemma 11. *We have*

$$\sum_{k=k_1}^{k_2}\beta_k\prod_{l=k_0}^{k-1}(1-\beta_l)=\prod_{l=k_0}^{k_1-1}(1-\beta_l)-\prod_{l=k_0}^{k_2}(1-\beta_l) \qquad (46)$$

and

$$\sum_{k=k_1}^{k_2}\beta_k\prod_{l=k+1}^{k_0}(1-\beta_l)=\prod_{l=k_1+1}^{k_0}(1-\beta_l)-\prod_{l=k_2}^{k_0}(1-\beta_l). \qquad (47)$$

Proof. By adding and subtracting one to the term β_k, we obtain

$$\begin{aligned}\sum_{k=k_1}^{k_2}\beta_k\prod_{l=k_0}^{k-1}(1-\beta_l) &= \sum_{k=k_1}^{k_2}(1-(1-\beta_k))\prod_{l=k_0}^{k-1}(1-\beta_l)\\ &=\sum_{k=k_1}^{k_2}\left(\prod_{l=k_0}^{k-1}(1-\beta_l)-\prod_{l=k_0}^{k}(1-\beta_l)\right)\\ &=\prod_{l=k_0}^{k_1-1}(1-\beta_l)-\prod_{l=k_0}^{k_2}(1-\beta_l), \qquad (48)\end{aligned}$$

and

$$\begin{aligned}\sum_{k=k_1}^{k_2}\beta_k\prod_{l=k+1}^{k_0}(1-\beta_l) &= \sum_{k=k_1}^{k_2}(1-(1-\beta_k))\prod_{l=k+1}^{k_0}(1-\beta_l)\\ &=\sum_{k=k_1}^{k_2}\left(\prod_{l=k+1}^{k_0}(1-\beta_l)-\prod_{l=k}^{k_0}(1-\beta_l)\right)\\ &=\prod_{l=k_1+1}^{k_0}(1-\beta_l)-\prod_{l=k_2}^{k_0}(1-\beta_l), \qquad (49)\end{aligned}$$

where (48) and (49) follow from telescoping the series. □

Based on Lemma 11, the first term on the right-hand side of (45) is non-negative because the summation is bounded from below by

$$\begin{aligned}-\frac{|\underline{u}_0^i|}{1-\gamma}\sum_{k=k_1}^{k_2}\tilde{\beta}_k\prod_{l=0}^{k-1}(1-\tilde{\beta}_l) &= -\frac{|\underline{u}_0^i|}{1-\gamma}\left(\prod_{l=0}^{k_1-1}(1-\tilde{\beta}_l)-\prod_{l=0}^{k_2}(1-\tilde{\beta}_l)\right)\\ &\geq -\frac{|\underline{u}_0^i|}{1-\gamma}\prod_{l=0}^{k_1-1}(1-\tilde{\beta}_l), \end{aligned}\tag{50}$$

which does not depend on k_2 and goes to zero as $k_1 \to \infty$ due to Assumption (ii) listed in Theorem 3. Similarly, the second term is also non-negative because the summation is bounded from below by

$$\begin{aligned}-\frac{1}{1-\gamma}\sum_{l=0}^{k_1-2}|\underline{e}_l|\sum_{k=k_1}^{k_2}\tilde{\beta}_k\prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m) &= -\frac{1}{1-\gamma}\sum_{l=0}^{k_1-2}|\underline{e}_l|\left(\prod_{m=l+1}^{k_1-1}(1-\tilde{\beta}_m)-\prod_{m=l+1}^{k_2}(1-\tilde{\beta}_m)\right)\\ &\geq -\frac{1}{1-\gamma}\sum_{l=0}^{k_1-2}|\underline{e}_l|\prod_{m=l+1}^{k_1-1}(1-\tilde{\beta}_m), \end{aligned}\tag{51}$$

which does not depend on k_2 and goes to zero as $k_1 \to \infty$. Particularly, the absolute summability of $\{\underline{e}_k\}$ and Assumption (ii) yields that $\{\underline{e}_k\}$ decays faster than $\{\beta_k\}$ and there exists k_0 such that $|\underline{e}_k| \leq \beta_k$ for all $k \geq k_0$. Therefore, for $k_1 \geq k_0$, we have

$$\begin{aligned}-\frac{1}{1-\gamma}\sum_{l=0}^{k_1-2}|\underline{e}_l|\prod_{m=l+1}^{k_1-1}(1-\tilde{\beta}_m) = &-\frac{1}{1-\gamma}\left(\prod_{m=k_0}^{k_1-1}(1-\tilde{\beta}_m)\right)\sum_{l=0}^{k_0-1}|\underline{e}_l|\prod_{m=l+1}^{k_0-1}(1-\tilde{\beta}_m)\\ &-\frac{1}{1-\gamma}\sum_{l=k_0}^{k_1-2}|\underline{e}_l|\prod_{m=l+1}^{k_1-1}(1-\tilde{\beta}_m), \end{aligned}\tag{52}$$

where the first-term goes to zero as $k_1 \to \infty$ due to Assumption (ii), and based on Lemma 11, the second term is bounded from below by

$$\begin{aligned}-\frac{1}{1-\gamma}\sum_{l=k_0}^{k_1-2}|\underline{e}_l|\prod_{m=l+1}^{k_1-1}(1-\tilde{\beta}_m) &\geq -\frac{1}{1-\gamma}\sum_{l=k_0}^{k_1-2}\tilde{\beta}_l\prod_{m=l+1}^{k_1-1}(1-\tilde{\beta}_m)\\ &= \prod_{l=k_0+1}^{k_1-1}(1-\tilde{\beta}_l)-\prod_{l=k_1-2}^{k_1-1}(1-\tilde{\beta}_l)\\ &\geq \prod_{l=k_0+1}^{k_1-1}(1-\tilde{\beta}_l), \end{aligned}\tag{53}$$

which goes to zero as $k_1 \to \infty$ by Assumption (ii). Finally, the third term is also non-negative because the summation is bounded from below by

$$
\begin{aligned}
-\frac{1}{1-\gamma}\sum_{l=k_1-1}^{k_2-1}|\underline{e}_l|\sum_{k=l+1}^{k_2}\tilde{\beta}_k\prod_{m=l+1}^{k-1}(1-\tilde{\beta}_m) &= -\frac{1}{1-\gamma}\sum_{l=k_1-1}^{k_2-1}|\underline{e}_l|\left(\prod_{m=l+1}^{l}(1-\beta_l)-\prod_{m=l+1}^{k_2}(1-\beta_l)\right)\\
&\geq -\frac{1}{1-\gamma}\sum_{l=k_1-1}^{k_2-1}|\underline{e}_l|\\
&\geq -\frac{1}{1-\gamma}\sum_{l=k_1-1}^{\infty}|\underline{e}_l|, \qquad (54)
\end{aligned}
$$

which goes to zero as $k_1 \to \infty$ since $\{\underline{e}_k\}$ is absolutely summable. This completes the proof. □

D PROOF OF LEMMA 10

The proof follows from [27, Theorem 4]. Particularly, we can view (21) as a weakened fictitious play dynamic in a game with payoffs $Q_*^i(\cdot)$ for each i since the action a_k^i satisfies

$$
\mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\{Q_k^i(a_k^i,\pi_k^{-i})\} = \max_{a^i\in A^i}\mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\{Q_k^i(a^i,a^{-i})\} \geq \max_{a^i\in A^i}\mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\{Q_*^i(a^i,a^{-i})\} - \epsilon_k \qquad (55)
$$

for some $\epsilon_k \to 0$ as $k \to \infty$ since $Q_k^i(a) \to Q_*^i(a)$ for all (i,a) as $k \to \infty$. The asymptotic negligibility of the error term follows since

$$
\begin{aligned}
&\left|\max_{a^i\in A^i}\mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\{Q_k^i(a^i,a^{-i})\} - \max_{a^i\in A^i}\mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\{Q_*^i(a^i,a^{-i})\}\right|\\
&\qquad\qquad \leq \max_{a^i\in A^i}\left|\mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\{Q_k^i(a^i,a^{-i}) - Q_*^i(a^i,a^{-i})\}\right| \qquad (56)
\end{aligned}
$$

and the right-hand side goes to zero due to the convergence of Q_k^i to Q_*^i. This completes the proof. □

E PROOF OF COROLLARY 4

The proof follows from the observation that based on (38) and (40), Γ_k^{ij}, as described in (32), can be written as

$$
\begin{aligned}
\Gamma_k^{ij}(s) = \mathbb{E}_{a\sim\pi_k(s)}\{r^i(s,a_k^j(s),a^{-j}) - r^i(s,a)\}\sum_{l=0}^{k}\beta_l\left(\prod_{m=l+1}^{k}(1-\beta_m)\right)&\\
+\gamma\sum_{s'\in S}\mathbb{E}_{a\sim\pi_k(s)}\{p(s'|s,a^i) - p(s'|s,a^i)\}\sum_{l=0}^{k}v_l^i(s')\beta_l\left(\prod_{m=l+1}^{k}(1-\beta_m)\right)& \qquad (57)\\
= \mathbb{E}_{a\sim\pi_k(s)}\{r^j(s,a_k^j(s),a^{-j}) - r^j(s,a)\}\sum_{l=0}^{k}\beta_l\left(\prod_{m=l+1}^{k}(1-\beta_m)\right) \geq 0.& \qquad (58)
\end{aligned}
$$

This completes the proof. □