Proceedings of Machine Learning Research vol 144:1–15, 2022

On the Heterogeneity of Independent Learning Dynamics in Zero-sum Stochastic Games

Muhammed O. Sayin SAYIN@EE.BILKENT.EDU.TR
Bilkent University, Electrical and Electronics Engineering Department, Ankara, Turkey 06800

K. Alperen Cetiner ALPEREN.CETINER@BILKENT.EDU.TR
Bilkent University, Electrical and Electronics Engineering Department, Ankara, Turkey 06800
Aselsan A.Ş., Ankara, Turkey 06800

Abstract

We analyze the convergence properties of the two-timescale fictitious play combining the classical fictitious play with the Q-learning for two-player zero-sum stochastic games with player-dependent learning rates. We show its almost sure convergence under the standard assumptions in two-timescale stochastic approximation methods when the discount factor is less than the product of the ratios of player-dependent step sizes. To this end, we formulate a novel Lyapunov function formulation and present a one-sided asynchronous convergence result.

Keywords: Q-learning, stochastic games, multi-agent learning, heterogenous systems

1. Introduction

Multi-agent reinforcement learning has become the frontier of many advancements in artificial intelligence systems, where autonomous agents make decisions in dynamic environments (e.g., see Zhang et al. (2021) and the references therein). Heterogeneity and independence of the learning dynamics adopted by these autonomous agents are inevitable in practical applications of multi-agent systems. However, there has been very limited progress addressing it in the context of stochastic games (introduced by Shapley (1953)) - a canonical model for dynamic multi-agent interactions.

Independent learning in strategic-form games played repeatedly has been studied extensively with many well-established results, e.g., see (Fudenberg and Levine, 1998; Young, 2004; Fudenberg and Levine, 2009). On the other hand, for stochastic games, Arslan and Yuksel (2017); Wei et al. (2017) presented learning dynamics with double-loop-like update rules necessitating coordination among players that may not be inline with their best interests. Recently, Leslie et al. (2020) has drawn a two-timescale learning framework in which continuous-time best-response dynamics could also convergence to an equilibrium of a two-player zero-sum stochastic game though the players' learning dynamics are not completely independent since they track a common parameter together. Within the two-timescale learning framework, Sayin et al. (2020) presented independent learning dynamics for stochastic games and analyzed its almost-sure convergence also in two-player zero-sum stochastic games. Note that Ozdaglar et al. (2021) provides an overview of studies on independent learning dynamics in stochastic games.

Heterogenous learning in games has also been studied, however, with a specific focus on multi-timescale learning, e.g., (Leslie and Collins, 2003, 2005) for strategic-form games with repeated play and recently (Daskalakis et al., 2020) for two-player zero-sum stochastic games. Particularly, dynamics of different players evolve at different timescales, e.g., one player's dynamics evolve slower than the others, contrary to (Leslie et al., 2020; Sayin et al., 2020) where the two-timescale framework is at player-level. On heterogenous rates that may not lead to different timescales, Zhu et al. (2011) studies heterogenous learning in a special case of zero-sum stochastic games while Chasnov et al. (2020) studies heterogenous gradient-based learning dynamics in continuous games.

In this paper, we address the heterogeneity and independence of learning in stochastic games by characterizing the convergence properties of the independent learning dynamics presented in (Sayin et al., 2020) with player-dependent learning rates. This dynamics is a new variant of fictitious play combining the classical fictitious play (Fudenberg and Levine, 1998) with the Q-learning (Watkins and Dayan, 1992) while they evolve at two different timescales. The key idea is that the underlying stochastic game can be viewed as a collection of auxiliary stage-games specific to each state whose payoff functions are the Q-functions. Though these auxiliary stage-games are not necessarily stationary, the slow evolution of Q-function estimates make them relatively stationary. The key challenge is the deviation of these auxiliary stage-games from the zero-sum structure due to the independent update of the Q-function estimates, and heterogenous learning rates boost this deviation further.

We show the almost sure convergence of the dynamics under the usual two-timescale stochastic approximation assumptions when the discount factor is less than the product of the ratios of player-dependent step sizes. We elaborate on the implications and high-level interpretation of this result later in Section 3. To show this result, we formulate a novel Lyapunov function formulation which reduces to the one presented in (Sayin et al., 2020) in the homogenous case, and present a one-sided asynchronous convergence result, which has a similar flavor with (Tsitsiklis, 1994, Theorem 1).

The rest of the paper is organized as follows. In Section 2, we introduce stochastic games and describe the learning dynamics. We present the assumptions and main convergence result in Section 3 and the proof of the main convergence result in Section 4. In Section 5, we provide an illustrative example. We conclude the paper with some remarks in Section 6. Appendices A-E include the proofs of the technical lemmas used in Section 4.

2. Independent Learning in (Zero-sum) Stochastic Games

Formally, a two-player stochastic game is characterized by a tuple $\langle S, A, r^1, r^2, p, \gamma \rangle$.[1] The *finite* set of states is denoted by S while $A = A^1 \times A^2$ with A^i denoting the *finite* set of actions that player i can take at any state.[2] The *stage payoff function* of player i is denoted by $r^i : S \times A \to \mathbb{R}$. In zero-sum case, we have $r^1(s,a) + r^2(s,a) = 0$ for all $(s,a) \in S \times A$. At any stage $k = 0, 1, \ldots$, if players play the action profile $a = (a^1, a^2) \in A$, then the state of the game, s, transits to another state, s', according to the transition probability $p(s'|s,a)$. Player i's objective is to maximize her expected sum of discounted stage-payoffs with the discount factor $\gamma \in [0,1)$.

Shapley (1953) (and Fink (1964)) showed that in two-player zero-sum (and n-player general-sum) stochastic games, there always exists a *stationary* equilibrium where players play *stationary* strategies depending only on the current state. Let $\pi^i : S \to \Delta(A^i)$ denote the stationary strategy of player i such that $\pi^i(s) \in \Delta(A^i)$ corresponds to her mixed strategy at state s, and $\pi = (\pi^1, \pi^2)$

1. For easy referral, we index players as player 1 and player 2. Furthermore, player i is the typical player and player $-i$ is her opponent.
2. The formulation can be extended to state-variant action sets straightforwardly.

denote the (stationary) strategy profile of players.[3] Then, the expected discounted sum of stage payoff of player i under the strategy profile π is given by

$$U^i(\pi) := \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k)\right\}, \tag{1}$$

where $a_k \sim \pi(s_k)$ denotes the action profile at stage k while $\{s_k\}_{k\geq 0}$ is a stochastic process such that s_k represents the state at stage k. The expectation is taken with respect to all randomness.

Definition 1 (Stationary Nash Equilibrium) *We say that a stationary strategy profile π is a stationary mixed-strategy equilibrium of the two-player stochastic game provided that*

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

We consider the same (independent) learning dynamics presented in (Sayin et al., 2020) but with player-dependent learning rates to examine the robustness of the convergence result to such heterogeneity. Particularly, we can view the stage-wise interaction among players as they are playing *auxiliary stage-games* specific to current state. For example, if player i knew that the opponent will play according to the stationary strategy π^{-i} in future stages, then her payoff function in the auxiliary stage-game specific to state s, denoted by $Q^i(s,\cdot) : A \to \mathbb{R}$, would satisfy the following fixed-point condition

$$Q^i(s,a) = r^i(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s')}\{Q^i(s',a)\}, \quad \forall a \in A. \tag{3}$$

This follows from backward induction that player i would always look for maximizing her utility, as described in (1). Note that the dependence on π^{-i} is implicit for notational convenience. The function $Q^i(\cdot)$ is known as *Q-function* in the MDP or reinforcement learning literature (Filar and Vrieze, 1997; Sutton and Barto, 2018), and it is well-defined due to the contraction property of the Bellman operator. Therefore, the auxiliary stage game is the tuple $\langle A, Q^1(s,\cdot), Q^2(s,\cdot)\rangle$.

Players form a belief about their Q-function and adopt fictitious play in auxiliary stage-games by forming also a belief on the opponent strategy. We denote the beliefs of player i at stage k about the opponent strategy by π_k^{-i} and about her Q-function by Q_k^i. At each stage k, player i always take the best response action in the auxiliary stage-game based on her beliefs $\pi_k^{-i}(s_k)$ and $Q_k^i(s_k) := Q_k^i(s_k,\cdot)$. Therefore, her action $a_k^i \in A^i$ always satisfies

$$a_k^i \in \underset{a^i\in A^i}{\operatorname{argmax}}\, \mathbb{E}_{a^{-i}\sim\pi_k^{-i}(s_k)}\{Q_k^i(s_k,a)\}. \tag{4}$$

Player i can observe the opponent's action $a_k^{-i} \in A^{-i}$ and update her beliefs according to

$$\pi_{k+1}^{-i}(s) = \pi_k^{-i}(s) + \mathbb{I}_{\{s=s_k\}}\alpha^i_{c_k(s)}(a_k^{-i} - \pi_k^{-i}(s)), \tag{5a}$$

$$Q_{k+1}^i(s,a) = Q_k^i(s,a) + \mathbb{I}_{\{s=s_k\}}\beta^i_{c_k(s)}\Big(r^i(s,a) + \gamma\sum_{s'\in S} p(s'|s,a)v_k^i(s') - Q_k^i(s,a)\Big), \tag{5b}$$

3. Given a set A, we denote the probability simplex over A by $\Delta(A)$.

for all (s,a), where $\mathbb{I}_{\{s=s_k\}}$ is the indicator function, we let the pure action a_k^{-i} be a deterministic strategy in the probability simplex $\Delta(A^{-i})$, the value function estimate $v_k^i : S \to \mathbb{R}$ is defined by

$$v_k^i(s) = \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi_k^{-i}(s)} \left\{ Q_k^i(s,a) \right\}, \tag{6}$$

and the player-dependent step sizes $\alpha_c^i \in (0,1)$ and $\beta_c^i \in (0,1)$ vanish with $c_k(s)$, the number of times state s is visited until stage k.

3. Convergence Results

The dynamics in (5) allow player-dependent and belief-dependent step sizes. In this section, we identify the conditions under which such a heterogenous two-timescale learning dynamics is guaranteed to converge to an equilibrium of the underlying two-player zero-sum stochastic game.

Assumption 2 (Markov Chain) *Every state is visited infinitely often with probability* 1.

Note that beliefs associated with a state gets updated only if that state is visited. Therefore, this assumption ensures that beliefs associated with each state gets updated infinitely often. Furthermore, Assumption 2 holds if the underlying stochastic game is *irreducible*, e.g., transition probabilities between any pair of states are positive for any joint action as in Leslie et al. (2020). This can be a restrictive assumption in practical application when players take deterministic best responses. However, it can be relaxed further as discussed in Ozdaglar et al. (2021) if players choose strategies in which every action is taken with some positive probability, e.g., due to smoothed best response or exploration, especially in the model-free cases where players do not know the stage-payoff and state transition kernel. We leave them as future research directions.

Assumption 3 (Step Sizes) *The step sizes satisfy the following conditions:*

(a) *For each $i = 1,2$, the step sizes $\alpha_c^i \to 0$, $\beta_c^i \to 0$, and $\beta_c^i/\alpha_c^i \to 0$ as $c \to \infty$, and the series $\sum_{c=0}^{\infty} \alpha_c^i = \sum_{c=0}^{\infty} \beta_c^i = \infty$.*

(b) *Let $i,j \in \{1,2\}$ denote arbitrary player indices with $i \neq -i$ and $j \neq -j$. Then, for some $d_\alpha, d_\beta \in (0,1]$, we have $\lim_{c\to\infty} \alpha_c^i/\alpha_c^{-i} = d_\alpha$ and $\lim_{c\to\infty} \beta_c^j/\beta_c^{-j} = d_\beta$.*[4]

In Assumption 3, part-(a) is standard in multi-timescale stochastic approximation, e.g., see (Borkar, 2008, Chapter 6). The part-(b) says that the ratios of the step sizes have a non-zero limit at each timescale. The limit assumption can be relaxed into conditions on limit inferior and limit superior, necessitating more involved analysis. We leave it as a future research direction.

Theorem 4 (Convergence Result) *Given a two-player zero-sum stochastic game, suppose that players follow the heterogenous and independent learning dynamics (5). Under Assumptions 2 and 3, we have $\pi_k^i \to \pi_*^i$ and $Q_k^i \to Q_*^i$ for each $i = 1,2$ as $k \to \infty$ with probability 1, for some stationary equilibrium π_* and the associated Q-functions (Q_*^1, Q_*^2) provided that $\gamma \leq d_\alpha d_\beta$.*

4. This is without loss of generality because if $\lim_{c\to\infty} \alpha_c^1/\alpha_c^2 \geq 1$ then we have $\lim_{c\to\infty} \alpha_c^2/\alpha_c^1 \leq 1$.

We attribute the upper bound on the discount factor as the heterogeneity increases how much the auxiliary stage-games can deviate from the zero-sum structure. However, a small discount factor will compensate this by restraining the deviation since the stage-payoffs have zero-sum structure. Furthermore, when players have (asymptotically) common step sizes, i.e., $d_\alpha = d_\beta = 1$, the bound on the discount factor becomes $\gamma < 1$, which is inherent to the discounted stochastic games. Therefore, Theorem 4 reduces to the convergence result (Sayin et al., 2020, Theorem 4.3) for the special case of homogeneous learning rates.

We can interpret Theorem 4 at a high level as players could reach to an equilibrium through heterogenous and independent learning dynamics if they are sufficiently myopic so that they discount the impact of future stages more in their utilities. We can also interpret the discount factor as the continuation probability of the stochastic game with indefinite horizon length (Shapley, 1953). Therefore, the heterogenous learning dynamics is guaranteed to converge to an equilibrium if the game has sufficiently short expected termination time.

Note also that stochastic games turn into strategic-form games with repeated play if there is only one state and $\gamma = 0$. In that case, the assumption on γ is always satisfied and the auxiliary stage-games are always zero-sum. Therefore, we have the following corollary to Theorem 4.

Corollary 5 (Heterogenous Fictitious Play) *Consider a two-player zero-sum strategic-form game played repeatedly. Suppose that players follow the fictitious play dynamics with player-dependent learning rates α_k^1, α_k^2 such that their ratio has non-zero limit. Then, the beliefs formed about the opponent strategy converge to an equilibrium of the game.*

4. Proof of the Convergence Result

In two-player zero-sum stochastic games, Shapley (1953) provided a (minimax) value iteration to compute equilibrium values associated with a stationary equilibrium. The operator used in Shapley's value iteration can be transformed into

$$(\mathcal{F}^i Q^i)(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)\mathrm{val}^i(Q^i(s',\cdot)), \quad \forall (s,a) \in S \times A, \tag{7}$$

as in (Szepesvari and Littman, 1999), where the minimax value function $\mathrm{val}^i(\cdot)$ is defined by

$$\mathrm{val}^i(Q^i(s,\cdot)) := \max_{\mu^i \in \Delta(A^i)} \min_{\mu^{-i} \in \Delta(A^{-i})} \mathbb{E}_{(a^i,a^{-i}) \sim (\mu^i,\mu^{-i})}\{Q^i(s,a)\}. \tag{8}$$

The operator $\mathcal{F}^i$ is a contraction and its unique fixed point $Q_*^i = \mathcal{F}^i Q_*^i$ is the equilibrium Q-function of the underlying stochastic game.

In this proof, we look for identifying the conditions under which

$$e_k^i(s) := v_k^i(s) - \mathrm{val}^i(Q_k^i(s)) \to 0 \quad \text{and} \quad \widetilde{Q}_k(s,a) := Q_k^i(s,a) - Q_*^i(s,a) \to 0, \tag{9}$$

for each (s,a) and $i = 1, 2$, by using stochastic differential inclusion theory while formulating a novel Lyapunov function and one-sided convergence result to address the heterogeneity. To this end, firstly, the following lemma establishes the connection between (5) and its (continuous-time) limiting differential inclusion based on (Benaim et al., 2005).[5] The proof is deferred to Appendix A.

5. Without loss of generality, suppose that $\lim_{c \to \infty} \alpha_c^1/\alpha_c^2 = d_\alpha \in (0, 1]$.

Lemma 6 *For each state s, the limiting differential inclusion of (5) is given by*

$$\dot{\pi}^1 + d_\alpha \pi^1 \in d_\alpha \operatorname*{argmax}_{a^1 \in A^1} \mathbb{E}_{a^2 \sim \pi^2}\{Q^1(a)\} \tag{10a}$$

$$\dot{\pi}^2 + \pi^2 \in \operatorname*{argmax}_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \pi^1}\{Q^2(a)\} \tag{10b}$$

$$\dot{Q}^1(a) = 0 \quad \text{and} \quad \dot{Q}^2(a) = 0, \quad \forall a, \tag{10c}$$

where $\pi^i : [0,\infty) \to \Delta(A^i)$ and $Q^i(a) : [0,\infty) \to \mathbb{R}$, for each $i = 1,2$, are continuous-time functions where we drop the dependence on s for notational convenience.

We note that there always exists an absolutely continuous solution to (10) since the best response satisfies the conditions listed in (Benaim et al., 2005, Hypothesis 1.1). Then, (Benaim et al., 2005, Theorem 3.6 and Proposition 3.27) yield that we can characterize the convergence properties of the discrete-time dynamics (5) in terms of the zero-set of a Lyapunov function to the differential inclusion (10).

Though (10) resembles to continuous-time best response dynamics in the auxiliary stage-game with time-invariant payoff functions Q^1 and Q^2, there are two challenges: (i) the heterogeneity when $d_\alpha \in (0,1)$ and (ii) the deviation from the zero-sum structure since $Q_k^1(s,a) + Q_k^2(s,a)$ is not necessarily zero for all (s,a) and k when they are updated independently according to (5b). Our candidate Lyapunov function is defined by

$$V(\pi, Q) := \left(d_\alpha \Delta^1(\pi^2, Q^1) + \Delta^2(\pi^1, Q^2) - \Xi(Q^1, Q^2)\right)_+, \tag{11}$$

where we define

$$\Delta^i(\pi^{-i}, Q^i) := \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}}\{Q^i(a)\} - \mathrm{val}^i(Q^i) \geq 0 \tag{12}$$

$$\Xi(Q^1, Q^2) := \lambda \|Q^1 + Q^2\| - (\mathrm{val}^1(Q^1) + \mathrm{val}^2(Q^2)), \tag{13}$$

where $\lambda \in (1, d_\alpha d_\beta/\gamma)$, $\|\cdot\|$ is the maximum norm, i.e., $\|Q\| = \max_a |Q(a)|$ and the positive function $(x)_+ = \max\{0, x\}$. The candidate function is non-negative by its definition. Its zero-set $\{(\pi, Q) : V(\pi, Q) = 0\}$ is given by

$$\left\{(\pi, Q) : \Delta^1(\pi^2, Q^1) + \Delta^2(\pi^1, Q^2) \leq \Xi(Q^1, Q^2) + (1 - d_\alpha)\Delta^1(\pi^2, Q^1)\right\} \tag{14}$$

and Δ^i is the continuous-time counterpart of the tracking error (9). Therefore, the convergence to the zero-set (14) would provide an (asymptotic) upper bound on the sum of tracking errors. Note also that when $d_\alpha = 1$, minimax values disappear and $V(\cdot)$ reduces to the one presented in (Sayin et al., 2020).

Since Q^i's are time-invariant, $\mathrm{val}^i(Q^i)$'s and Ξ are also time-invariant. As shown in the following lemma, these time-invariant terms together with the positive function play an important role for the validity of the candidate $V(\cdot)$ as a Lyapunov function to (10) for the zero-set (14) and later in characterizing the convergence properties of the tracking error. The proof is deferred to Appendix B.

Lemma 7 (Lyapunov Function) *The candidate function $V(\cdot)$ is a Lyapunov function of the differential inclusion (10) for the zero-set $\{(\pi, Q) : V(\pi, Q) = 0\}$. In other words, for any absolutely continuous solution to (10), we have*

- *$V(\pi(t'), Q(t')) < V(\pi(t), Q(t))$ for all $t' > t$ if $V(\pi(t), Q(t)) > 0$,*
- *$V(\pi(t'), Q(t')) = 0$ for all $t' > t$ if $V(\pi(t), Q(t)) = 0$.*

Based on the stochastic differential inclusion theory Benaim et al. (2005), Lemma 7, the definition of v_k^i, as described in (6), and the zero-set (14), we can conclude that

$$(\bar{v}_k(s) - \lambda \|\bar{Q}_k(s)\| - (1 - d_\alpha) e_k^1(s))_+ \to 0, \tag{15}$$

where we define $\bar{v}_k(s) := v_k^1(s) + v_k^2(s)$ and $\bar{Q}_k(s) := Q_k^1(s) + Q_k^2(s)$ with the tracking error e_k^1, as described in (9). Since we let $i = 1$ arbitrarily for notational convenience, (15) can be written as

$$\bar{v}_k(s) \leq \lambda \|\bar{Q}_k(s)\| + (1 - d_\alpha) e_k^i(s) + \epsilon_k(s), \tag{16}$$

where $\epsilon_k(s) \to 0$ as $k \to \infty$ almost surely. The upper bound on $\bar{v}_k(s)$ is in terms of $\|\bar{Q}_k(s)\|$ and the tracking error e_k^i. The following lemma provides an upper bound on the tracking error so that we can obtain an upper bound in terms of $\bar{Q}_k(s)$ only. The proof is deferred to Appendix C.

Lemma 8 *We have $0 \leq e_k^i(s) = v_k^i(s) - \mathrm{val}^i(Q_k^i(s)) \leq \bar{v}_k(s) - \min_{a \in A} \bar{Q}_k(s, a)$ for every s.*

Based on (16) and Lemma 8, we obtain

$$\min_{a \in A} \bar{Q}_k(s, a) \leq \bar{v}_k(s) \leq \frac{\lambda}{d_\alpha} \|\bar{Q}_k(s)\| - \frac{1 - d_\alpha}{d_\alpha} \min_{a \in A} \bar{Q}_k(s, a) + \frac{1}{d_\alpha} \epsilon_k(s), \tag{17}$$

where the lower bound follows since $v_k^i(s) \geq \mathbb{E}_{a \sim \pi_k(s)}\{Q_k^i(s, a)\}$ for $i = 1, 2$. We emphasize that in the homogenous case, the evolution of $\bar{Q}_k(s, a)$ is given by

$$\bar{Q}_{k+1}(s, a) = \bar{Q}_k(s, a) + \mathbb{I}_{\{s = s_k\}} \beta_{c_k(s)} \left(\gamma \sum_{s' \in S} p(s'|s, a) \bar{v}_k(s') - \bar{Q}_k(s, a) \right) \tag{18}$$

due to the symmetry that $\beta_c^i = \beta_c$. However, the characterization of the convergence properties of $\bar{Q}_k(s, a)$ in the heterogenous case requires more involved analysis where we will address the limit inferior and limit superior of $\bar{Q}_k(s, a)$ separately.

Recall the definition of $\widetilde{Q}_k^i$ in (9) and note that the fixed points Q_*^1 and Q_*^2 satisfy $Q_*^1(s, a) + Q_*^2(s, a) = 0$ for every (s, a). Therefore, we also have $\bar{Q}_k = \widetilde{Q}_k^1 + \widetilde{Q}_k^2$. Based on (5b) and the definition of the fixed point Q_*^i, the evolution of $\widetilde{Q}_k^i$ can be written as

$$\widetilde{Q}_{k+1}^i(s, a) = \widetilde{Q}_k^i(s, a) + \bar{\beta}_k^i(s) \left(\gamma \sum_{s' \in S} p(s'|s, a)(v_k^i(s') - \mathrm{val}^i(Q_*^i(s'))) - \widetilde{Q}_k^i(s, a) \right), \tag{19}$$

where $\bar{\beta}_k^i(s) := \mathbb{I}_{\{s = s_k\}} \beta_{c_k(s)}^i$ and $r^i(s, a)$ disappears since $Q_*^i(s, a) = (\mathcal{F}^i Q_*^i)(s, a)$. We are interested in the limit inferior of $\widetilde{Q}_k^i$, for each i, so that we can formulate the limit inferior of $\bar{Q}_k$, which will play an important role in (17).

The following lemma characterizes the limit inferior of an iterate whose evolution satisfies one-sided contraction-like condition. The proof is deferred to Appendix D.

Lemma 9 (One-sided Asynchronous Discrete-time Convergence) *Consider a sequence of vectors* $\{y_k\}_{k=0}^{\infty}$ *such that the* n*th entry, denoted by* $y_k(n)$*, satisfies the following lower bound:*

$$y_{k+1}(n) \geq y_k(n) + \beta_k(n) \left(\gamma \min_m y_k(m) - y_k(n) + \epsilon_k(n)\right), \tag{20}$$

where $\gamma \in (0,1)$*, the vanishing (possibly random) step size* $\beta_k(n) \in [0,1]$ *satisfies* $\beta_k(n) \to 0$ *and* $\sum_{k=0}^{\infty} \beta_k(n) = \infty$*, and* $\liminf_k \epsilon_k(n) \geq 0$ *for each* n*, with probability* 1*. Suppose that* $\min_n y_k(n) \geq M$ *for all* k*. Then, we have* $\liminf_k y_k(n) \geq 0$ *for all* n*, with probability* 1*.*

By the definition of the tracking error (9), we have

$$\begin{aligned}\sum_{s' \in S} p(s'|s,a)(v_k^i(s') - \mathrm{val}^i(Q_*^i(s'))) &= \sum_{s' \in S} p(s'|s,a)(e_k^i(s') + \mathrm{val}^i(Q_k^i(s')) - \mathrm{val}^i(Q_*^i(s'))) \\ &\geq \min_{(s,a)} \widetilde{Q}_k^i(s,a) \end{aligned} \tag{21}$$

where the inequality follows since $e_k^i(s') \geq 0$ and $\mathrm{val}^i(Q_k^i(s')) - \mathrm{val}^i(Q_*^i(s')) \geq \min_{a \in A} \widetilde{Q}_k^i(s',a)$ for all s'. Under Assumptions 2 and 3, the step size $\bar{\beta}_k(s) \in [0,1]$ vanishes and $\sum_{k=0}^{\infty} \bar{\beta}_k(n) = \infty$ with probability. Therefore, we can invoke Lemma 9, for every $i = 1,2$, and obtain

$$\liminf_{k \to \infty} \widetilde{Q}_k^i(s,a) \geq 0 \quad \Rightarrow \quad \liminf_{k \to \infty} \bar{Q}_k(s,a) \geq 0, \tag{22}$$

almost surely for every (s,a). Combined with (17), the bound (22) yields that

$$\underline{\epsilon}_k(s) \leq \bar{v}_k(s) \leq \frac{\lambda}{d_\alpha} \|\bar{Q}_k(s)\| + \bar{\epsilon}_k(s), \quad \forall s, \tag{23}$$

for some error terms $\underline{\epsilon}_k(s) \to 0$ and $\bar{\epsilon}_k(s) \to 0$ as $k \to \infty$ for each s almost surely.

The heterogeneity of β_c^i's has not played a role up to this point. Suppose that $\lim_{c \to \infty} \beta_c^1/\beta_c^2 = d_\beta \in (0,1]$ without loss of generality. Then, the iteration (19) can be written as

$$\begin{bmatrix} \widetilde{Q}_{k+1}^1 \\ \widetilde{Q}_{k+1}^2 \end{bmatrix} = \begin{bmatrix} \widetilde{Q}_k^1 \\ \widetilde{Q}_k^2 \end{bmatrix} + \bar{\beta}_k(s) \begin{bmatrix} d_\beta \left(\gamma \sum_{s'} p(s'|\cdot)(v_k^1(s') - \mathrm{val}^1(Q_*^1(s'))) - \widetilde{Q}_k^1 + \zeta_k\right) \\ \gamma \sum_{s'} p(s'|\cdot)(v_k^2(s') - \mathrm{val}^2(Q_*^2(s'))) - \widetilde{Q}_k^2 \end{bmatrix}, \tag{24}$$

where we dropped the argument (s,a) for notational convenience and the error term is given by

$$\zeta_k(s,a) := \frac{1}{d_\beta} \left(\frac{\beta_{c_k(s)}^1}{\beta_{c_k(s)}^2} - d_\beta\right) \left(\gamma \sum_{s'} p(s'|s,a)(v_k^1(s') - \mathrm{val}^1(Q_*^1(s'))) - \widetilde{Q}_k^1(s,a)\right), \tag{25}$$

which is asymptotically negligible by Assumptions 2 and 3, and by the boundedness of the iterates. For notational convenience, we also define

$$\Gamma_k(s,a) = \widetilde{Q}_k^1(s,a) + d_\beta \widetilde{Q}_k^2(s,a). \tag{26}$$

Then, the weighted combination of the entries in (24) yield that

$$\Gamma_{k+1}(s,a) = \Gamma_k(s,a) + \bar{\beta}_k(s) d_\beta \left(\gamma \sum_{s' \in S} p(s'|s,a) \bar{v}_k(s') - \bar{Q}_k(s,a) + \zeta_k\right) \tag{27}$$

since $\mathrm{val}^1(Q_*^1(s)) + \mathrm{val}^2(Q_*^2(s)) = 0$ for each s. By (22) and (27), we also have

$$\liminf_{k\to\infty} \Gamma_k(s,a) \geq 0, \quad \forall\,(s,a) \tag{28}$$

with probability 1. In order to characterize the convergence properties of Γ_k within the framework of Lemma 9, we introduce the following lemma formulating a bound on $\bar{Q}_k$ in terms of Γ_k. The proof is deferred to Appendix E.

Lemma 10 *For each (s,a), we have*

$$\overline{\eta}_k(s,a) + \frac{1}{d_\beta}\Gamma_k(s,a) \geq |\bar{Q}_k(s,a)| \geq \bar{Q}_k(s,a) \geq \Gamma_k(s,a) + \underline{\eta}_k(s,a) \tag{29}$$

for some $\underline{\eta}_k(s,a) \to 0$ and $\overline{\eta}_k(s,a) \to 0$ with probability 1.

Based on (23), (27), and Lemma 10, we obtain

$$\Gamma_{k+1}(s,a) \leq \Gamma_k(s,a) + \bar{\beta}_k(s) d_\beta \left(\frac{\gamma\lambda}{d_\alpha d_\beta} \max_{(s',a')} \Gamma_k(s',a') - \Gamma_k(s,a) + \varepsilon_k(s,a) \right), \tag{30}$$

where $\varepsilon_k(s,a)$ is an asymptotically negligible error almost surely. We can invoke Lemma 9 for $\{-\Gamma_k\}_{k\geq 0}$ and obtain that

$$\liminf_{k\to\infty} -\Gamma_k(s,a) = -\limsup_{k\to\infty} \Gamma_k(s,a) \geq 0 \quad \Rightarrow \quad \limsup_{k\to\infty} \Gamma_k(s,a) \leq 0. \tag{31}$$

Combined with (28), the bound (31) yields that

$$\lim_{k\to\infty} \Gamma_k(s,a) = 0 \quad \Rightarrow \quad \lim_{k\to\infty} \widetilde{Q}_k^i(s,a) = 0 \tag{32}$$

for all (s,a) and $i = 1,2$ with probability 1. Since $\bar{Q}_k = \widetilde{Q}_k^1 + \widetilde{Q}_k^2$, we also have $\bar{Q}_k(s,a) \to 0$ for each (s,a). Then, we obtain $\bar{v}_k(s) \to 0$ for each s by (23). Therefore, Lemma 8 yields that the tracking error $e_k^i(s) \to 0$ for each s, with probability 1.

5. Numerical Example

In this section, we examine the convergence properties of the heterogenous learning dynamics numerically in an (irreducible) zero-sum stochastic game whose configuration is selected arbitrarily as in (Sayin et al., 2020). For example, there are three states, four actions per state, and the discount factor is 0.8. The player-dependent step sizes are set as $\alpha_c^1 = (1+c)^{0.5}$ and $\alpha_c^2 = (1+0.81\cdot c)^{0.5}$ while $\beta_c^1 = (1+c)^{-1}$ and $\beta_c^2 = (1+0.95\cdot c)^{-1}$. In Figure 1, we plot the evolution of the value function estimates of both player in addition to $\bar{v}_k$ to illustrate that the auxiliary stage-games become zero-sum also in the heterogenous case. Since the underlying game is irreducible, Assumption 2 holds. For these step sizes, we have $d_\alpha = 0.9$ and $d_\beta = 0.95$. Assumption 3 also holds. Therefore, Theorem 4 says that the dynamics should converge to an equilibrium of the game and we have observed the convergence of the value function estimates to the equilibrium values of the game, as expected from Theorem 4.

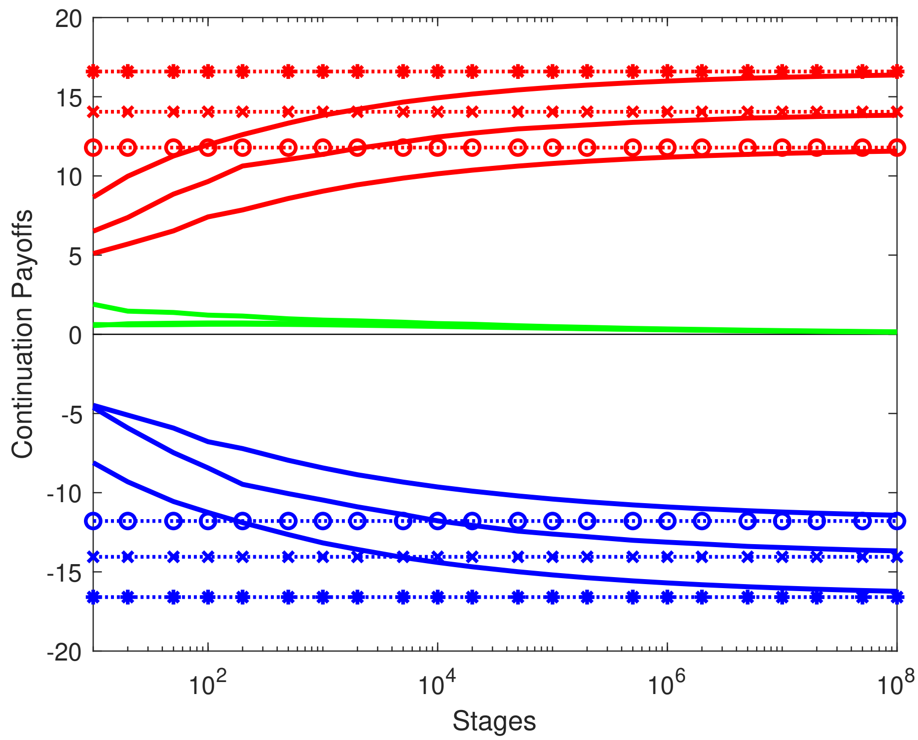

Figure 1: Evolution of v_k^1, v_k^2, and $\bar{v}_k$ for each s, respectively, in **red**, **blue**, and **green**. The flat line represents the zero-level to illustrate the convergence of $\bar{v}_k$ to zero. The dotted line $\ldots/\ldots$ are the actual equilibrium values associated with each state (computed via Shapley's iteration (Shapley, 1953)). We emphasize that the horizontal axis is logarithmic with markers at instances $10, 20, 50, 100, 200, \ldots$ in this order for clear demonstration.

6. Conclusion

We showed the almost sure convergence of two-timescale fictitious play with heterogeneous learning rates in two-player zero-sum stochastic games under the standard assumptions in two-timescale stochastic approximation methods when the discount factor is less than the product of the ratios of the player-dependent step-sizes. Since strategic-form games played repeatedly is a special case of stochastic games with single state and zero discount factor, this result also implied the almost sure convergence of heterogenous fictitious play in zero-sum strategic-form games. We attributed the bound on the discount factor to the deviation of the auxiliary stage-games from the zero-sum structure, which becomes multifold with the player-dependent rates while sufficiently small discount rates can compensate it.

Interesting future research directions include characterizing the convergence properties of heterogenous and independent learning dynamics in stochastic games other than zero-sum, and in model-free and minimal information cases, e.g., as in (Sayin et al., 2021).

Appendix A. Proof of Lemma 6

The dynamics specific to state s, i.e., (5a) and (5b), can be written as

$$\begin{bmatrix}\pi^1_{k+1}(s)\\ \pi^2_{k+1}(s)\end{bmatrix} = \begin{bmatrix}\pi^1_k(s)\\ \pi^2_k(s)\end{bmatrix} + \bar{\alpha}_k(s)\left(\begin{bmatrix}d_\alpha(a^1_k - \pi^1_k(s))\\ a^2_k - \pi^2_k(s)\end{bmatrix} + \begin{bmatrix}\varepsilon_k(s)\\ \mathbf{0}\end{bmatrix}\right), \tag{33}$$

and

$$\begin{bmatrix}Q^1_{k+1}(s,a)\\ Q^2_{k+1}(s,a)\end{bmatrix} = \begin{bmatrix}Q^1_k(s,a)\\ Q^2_k(s,a)\end{bmatrix} + \bar{\alpha}_k(s)\begin{bmatrix}E^1_k(s,a)\\ E^2_k(s,a)\end{bmatrix}, \tag{34}$$

where $\mathbf{0}$ is a zero vector, the step size $\bar{\alpha}_k(s) := \mathbb{I}_{\{s=s_k\}}\alpha^2_{c_k(s)} \in [0,1]$, and the error terms are defined by

$$\varepsilon_k(s) := \left(\frac{\alpha^1_{c_k(s)}}{\alpha^2_{c_k(s)}} - d_\alpha\right)(a^1_k - \pi^1_k(s)) \tag{35a}$$

$$E^1_k(s,a) := \frac{\alpha^1_{c_k(s)}}{\alpha^2_{c_k(s)}}\frac{\beta^1_{c_k(s)}}{\alpha^1_{c_k(s)}}\left(r^1(s,a) + \gamma\sum_{s'\in S} p(s'|s,a)v^1_k(s') - Q^1_k(s,a)\right) \tag{35b}$$

$$E^2_k(s,a) := \frac{\beta^2_{c_k(s)}}{\alpha^2_{c_k(s)}}\left(r^2(s,a) + \gamma\sum_{s'\in S} p(s'|s,a)v^2_k(s') - Q^2_k(s,a)\right). \tag{35c}$$

Note that the iterates are bounded since $\gamma \in (0,1)$, the stage-payoffs have compact support and the step sizes are in $(0,1]$. Therefore, Assumptions 2 and 3 yield that the error terms (35) are all asymptotically negligible for all (s,a). Note also that under Assumption 2 and 3, we have $\sum_{k=0}^\infty \bar{\alpha}_k(s) = \infty$ while $\bar{\alpha}_k(s) \to 0$ as $k \to \infty$ with probability 1. Furthermore, the best response satisfies the conditions for the stochastic differential inclusion theory (Benaim et al., 2005, Hypothesis 1.1). Therefore, its limiting differential inclusion is given by (10).

Appendix B. Proof of Lemma 7

For fixed absolutely continuous solution $(\pi^1(t), \pi^2, Q^1, Q^2)$ to (10), the argument of the positive function in (11) is given by

$$L(t) := d_\alpha\Delta^1(\pi^2(t), Q^1) + \Delta^2(\pi^1(t), Q^2) - \Xi(Q^1, Q^2), \tag{36}$$

which is also an absolutely continuous function since $\max$ and addition satisfy the Lipschitz condition (Bogachev and Smolyanov, 2020, Lemma 4.3.2). Therefore, we can compute its derivative almost everywhere as in Harris (1998) and obtain

$$\begin{aligned}\frac{d}{dt}\left(d_\alpha\Delta^1 + \Delta^2 - \Xi\right) &= d_\alpha(a^1)^T Q^1\frac{d\pi^2}{dt} + \left(\frac{d\pi^1}{dt}\right)^T Q^2 a^2\\ &= d_\alpha\left((a^1)^T Q^1(a^2 - \pi^2) + (a^1 - \pi^1)^T Q^2 a^2\right)\\ &= d_\alpha\left((a^1)^T(Q^1 + Q^2)a^2 - (\Delta^1 + \Delta^2) - (\mathrm{val}^1(Q^1) + \mathrm{val}^2(Q^2))\right),\end{aligned} \tag{37}$$

where

$$a^1 := \frac{1}{d_\alpha}\frac{d\pi^1}{dt} + \pi^1 \quad \text{and} \quad a^2 := \frac{d\pi^2}{dt} + \pi^2. \tag{38}$$

Note that if it is zero-sum, we have $\dot{L} = -d_\alpha(\Delta^1 + \Delta^2) \leq 0$ and equal to zero only if $\Delta^i = 0$, which corresponds to the equilibrium. Suppose that it is not zero-sum. Then, the time derivative of L is not necessarily negative almost everywhere. However, we can write (37) as

$$\frac{d}{dt}\left(d_\alpha \Delta^1 + \Delta^2 - \Xi\right) = -d_\alpha \overbrace{\left(d_\alpha \Delta^1 + \Delta^2 - \Xi\right)}^{=L} + d_\alpha((a^1)^T(Q^1+Q^2)a^2 - \lambda\|Q^1+Q^2\|) \\ - d_\alpha(1-d_\alpha)\Delta^1. \tag{39}$$

The second term at the right-hand side is negative since $\lambda > 1$ and it is not zero-sum. Furthermore, the last term in (39) is non-positive since $d_\alpha \in (0,1]$ and $\Delta^1 \geq 0$. Therefore, we have

$$\dot{L} < -d_\alpha L, \tag{40}$$

almost everywhere. Therefore, the absolutely continuous $L(t)$ is strictly decreasing when $L(t) \geq 0$ and $(-\infty, 0]$ is a positively invariant set for $L(t)$. Correspondingly, $V(\pi(t), Q) = (L(t))_+$ is strictly decreasing when $V(\pi(t), Q) > 0$ and $(-\infty, 0]$ is also a positively invariant set for $V(\pi(t), Q)$.

Note that we can pick λ such that $\lambda \in (1, d_\alpha d_\beta/\gamma)$ since it is arbitrary and $\gamma < d_\alpha d_\beta$, as stated in Theorem 4. This completes the proof.

Appendix C. Proof of Lemma 8

The proof follows from the saddle point inequality:

$$v_k^i(s) = \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi_k^{-i}(s)}\{Q_k^i(s,a)\} \geq \mathrm{val}^i(Q_k^i(s)) \geq \min_{a^{-i} \in A^{-i}} \mathbb{E}_{a^i \sim \pi_k^i(s)}\{Q_k^i(s)\} \tag{41}$$

because the right-most term is bounded from below by

$$\min_{a^{-i}} \mathbb{E}_{a^i \sim \pi_k^i(s)}\{\bar{Q}_k(s) - Q_k^{-i}(s)\} \geq \min_{a^{-i}} \mathbb{E}_{a^i \sim \pi_k^i(s)}\{\bar{Q}_k(s)\} + \min_{a^{-i}} \mathbb{E}_{a^i \sim \pi_k^i(s)}\{-Q_k^{-i}(s)\} \\ = \min_{a^{-i}} \mathbb{E}_{a^i \sim \pi_k^i(s)}\{\bar{Q}_k(s)\} - \max_{a^{-i}} \mathbb{E}_{a^i \sim \pi_k^i(s)}\{Q_k^{-i}(s)\}. \tag{42}$$

Therefore, we obtain

$$v_k^i(s) \geq \mathrm{val}^i(Q_k^i(s)) \geq -v_k^{-i}(s) + \min_{a \in A} \bar{Q}_k(s,a). \tag{43}$$

The difference between the first and the second term is bounded from above by the difference between the first and the third term. This completes the proof.

Appendix D. Proof of Lemma 9

Note that if $M \geq 0$, the lower bound, $y_k(n) \geq M$ for all $k \geq 0$ and n, already implies that $\liminf_{k\to\infty} y_k(n) \geq 0$ for all n.

Suppose that $M < 0$. Then, the proof has a flavor similar to (Tsitsiklis, 1994, Theorem 1) and (Sayin et al., 2020, Theorem 5.1) but only for one-side to characterize the limit inferior of the sequence based on the assumption that all iterates are bounded from below and update has a one-sided contraction-like structure.

Define the negative sequence $\{M^t < 0\}_{t\geq 0}$ over a separate timescale by

$$M^{t+1} = (\gamma + 2\epsilon)M^t, \quad \forall\, t \geq 0, \tag{44}$$

and $M^0 = M$, where $\epsilon \in (0, (1-\gamma)/2)$. Since $\gamma + 2\epsilon \in (0,1)$, we have $M^t \to 0$ monotonically (from below) as $t \to \infty$.

Since $\liminf_k \epsilon_k(n) \geq 0$ for each n and $y_k(n) \geq M^0$ for all $k \geq 0$, there exists k^0 such that

$$y_k(n) \geq M^0, \quad \text{and} \quad \epsilon_k(n) > \epsilon M^0 \quad \forall\, k \geq k^0. \tag{45}$$

Suppose that for some $t \geq 0$, there exists k^t such that

$$y_k(n) \geq M^t, \quad \text{and} \quad \epsilon_k(n) > \epsilon M^t \quad \forall\, k \geq k^t. \tag{46}$$

Then, we can define an auxiliary sequence $\{Y_k^t\}_{k\geq k^t}$ by

$$Y_{k+1}^t(n) = Y_k^t(n)(1-\beta_k(n)) + \beta_k(n)(\gamma+\epsilon)M^t, \quad \forall\, k \geq k^t, \tag{47}$$

and $Y_{k^t}^t(n) = M^t$, for each n such that

$$Y_k^t(n) \leq y_k(n), \quad \forall\, k \geq k^t, \tag{48}$$

for each n, by its definition. Note that $Y_k^t(n) \to (\gamma+\epsilon)M^t$ for each n as $k \to \infty$ almost surely. Since $(\gamma + 2\epsilon)M^t < (\gamma+\epsilon)M^t$, there exists $k^{t+1} \geq k^t$ such that

$$y_k(n) \geq (\gamma+2\epsilon)M^t = M^{t+1} \quad \text{and} \quad \epsilon_k(n) > \epsilon M^{t+1}, \quad \forall k \geq k^{t+1} \tag{49}$$

for each n almost surely.

By induction, we can conclude that for each $t \geq 0$, there exists k^t such that $y_k(n) \geq M^t$ for all $k \geq k^t$ and each n. Since $M^t \to 0$ as $t \to \infty$, we obtain $\liminf_k y_k(n) \geq 0$ for each n.

Appendix E. Proof of Lemma 10

The proof follows from (22), which implies that given (s,a), we have $\widetilde{Q}_k^i(s,a) \geq (1-d_\beta)^{-1}\underline{\eta}_k(s,a)$ for all k for some asymptotically negligible term $\underline{\eta}_k(s,a) \to 0$ almost surely. Particularly, we have

$$\bar{Q}_k = \widetilde{Q}_k^1 + \widetilde{Q}_k^2 \geq \widetilde{Q}_k^1 + d_\beta \widetilde{Q}_k^2 + \underline{\eta}_k = \Gamma_k + \underline{\eta}_k, \tag{50}$$

and

$$\Gamma_k = \widetilde{Q}_k^1 + d_\beta \widetilde{Q}_k^2 \geq d_\beta(\widetilde{Q}_k^1 + \widetilde{Q}_k^2) - d_\beta \overline{\eta}_k = d_\beta \bar{Q}_k - d_\beta \overline{\eta}_k \tag{51}$$

where we drop the arguments (s,a) for notational convenience and $\overline{\eta}_k(s,a) = -\underline{\eta}_k(s,a)/d_\beta$. By (50) and (51), we obtain

$$\overline{\eta}_k + \frac{1}{d_\beta}\Gamma_k \geq \bar{Q}_k \geq \Gamma_k + \underline{\eta}_k. \tag{52}$$

The proof is completed since $\liminf_k \bar{Q}_k(s) \geq 0$ implies that $\lim_k(|\bar{Q}_k(s)| - \bar{Q}_k(s)) = 0$.

References

G. Arslan and S. Yuksel. Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4):1545–1558, 2017.

M. Benaim, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control Optim.*, 44(1):328–348, 2005.

V. Bogachev and O. G. Smolyanov. *Real and Functional Analysis*. Springer Nature, 2020.

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, 2008.

Benjamin Chasnov, Lillian Ratliff, Eric Mazumdar, and Samuel Burden. Convergence analysis of gradient-based learning in continuous games. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pages 935–944, 2020.

C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer Verlag, 1997.

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

D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, MA, 1998.

D. Fudenberg and D. K. Levine. Learning and equilibrium. *The Annual Review of Economics*, 1: 385–419, 2009.

C. Harris. On the rate of convergence of continuous-time fictitious play. *Games and Economic Behavior*, 22:238–259, 1998.

D. S. Leslie and E. J. Collins. Convergent multi-timescales reinforcement learning algorithms in normal form games. *The Annals of Applied Probability*, 13(4):1231–1251, 2003.

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

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

A. Ozdaglar, M. O. Sayin, and K. Zhang. Independent learning in stochastic games. *ArXiv: 2111.11743*, 2021.

M. O. Sayin and K. A. Cetiner. On the heterogeneity of independent learning dynamics in zero-sum stochastic games. *Extended version available at https://gdn.bilkent.edu.tr/l4dc/*, 2021.

M. O. Sayin, F. Parise, and A. Ozdaglar. Fictitious play in zero-sum stochastic games. *ArXiv*, 2020.

M. O. Sayin, K. Zhang, D. Leslie, A. Ozdaglar, and T. Başar. Decentralized Q-learning in zero-sum stochastic games. In *Advances in Neural Information Processing Systems*, 2021.

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

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2018.

C. Szepesvari and M. Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Computing*, 11:2017–2060, 1999.

J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16: 185–202, 1994.

C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

C.-Y. Wei, Y.-T. Hong, and C.-J. Lu. Online reinforcement learning in stochastic games. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS)*, 2017.

H. P. Young. *Strategic Learning and Its Limits*. Oxford University Press, 2004.

K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handb. Rein. Learn. Cont.*, volume 325. Springer, 2021.

Q. Zhu, H. Tembine, and T. Başar. Heterogeneous learning in zero-sum stochastic games with incomplete information. In *IEEE Conf. Decision and Control*, pages 219–224, 2011.