SIAM J. CONTROL OPTIM.
Vol. 60, No. 4, pp. 2095–2114




# FICTITIOUS PLAY IN ZERO-SUM STOCHASTIC GAMES*

MUHAMMED O. SAYIN†, FRANCESCA PARISE‡, AND ASUMAN OZDAGLAR§

**Abstract.** We present a novel variant of fictitious play dynamics combining classical fictitious play with $Q$-learning for stochastic games and analyze its convergence properties in two-player zero-sum stochastic games. Our dynamics involves players forming beliefs on the opponent strategy and their own continuation payoff ($Q$-function), and playing a greedy best response by using the estimated continuation payoffs. Players update their beliefs from observations of opponent actions. A key property of the learning dynamics is that update of the beliefs on $Q$-functions occurs at a slower timescale than update of the beliefs on strategies. We show that in both the model-based and model-free cases (without knowledge of player payoff functions and state transition probabilities), the beliefs on strategies converge to a stationary mixed Nash equilibrium of the zero-sum stochastic game.

**Key words.** stochastic games, fictitious play, $Q$-learning, two-timescale learning

**MSC codes.** 91A15, 91A26, 68T05

**DOI.** 10.1137/21M1426675

## 1. Introduction.

A common justification for Nash equilibrium is that it arises from the learning dynamics of myopic players taking greedy best response actions. This perspective has been investigated for various classes of strategic-form games (also referred to as one-shot or normal-form games) mostly focusing on best response type dynamics (including fictitious play) [12, 16, 25, 26]. Nevertheless, study of learning dynamics in the context of stochastic games has been limited.

Stochastic games were introduced by [24] to model interactions among multiple players in a multistate dynamic environment. Players' actions determine not only the payoffs at the current state, *stage payoffs*, but also the transition probability to the next state, and hence the continuation payoffs. The decision problem of a player thus involves trading off current stage payoff for estimated continuation payoffs while forming predictions on the opponent's strategy. This dynamic trade-off makes the analysis of learning in stochastic games potentially challenging.

### 1.1. Contributions.

In this paper, we present a novel variant of fictitious play combining classical fictitious play with $Q$-learning [31] for stochastic games and analyze its convergence properties in two-player zero-sum stochastic games. Each player forms a belief on the opponent's (stationary) mixed strategy and her own *$Q$-function*, which corresponds to her continuation payoff given the opponent's strategy. Players play a greedy best response strategy in an *auxiliary game* with player payoffs given by the sum of the stage payoffs and estimated continuation payoffs. At each stage of the game over the infinite horizon, the players update their beliefs on the opponent strat-

*Received by the editors June 14, 2021; accepted for publication (in revised form) April 20, 2022; published electronically July 13, 2022.

https://doi.org/10.1137/21M1426675

**Funding:** This research was supported by the U.S. Army Research Office (ARO), grant W911NF-18-1-0407.

†Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey, 06800 (sayin@ee.bilkent.edu.tr).

‡Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (fp264@cornell.edu).

§Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (asuman@mit.edu).

egy from the observation of the other player's action. The update of the $Q$-function is then constructed as the maximum payoff that can be attained in the auxiliary game over all possible actions (given beliefs on opponent play and $Q$-function).

A key property of our learning dynamics is that the beliefs on opponent strategy and $Q$-functions are updated simultaneously though the update of the latter is at a slower timescale than the former. This is consistent with the literature on evolutionary game theory, e.g., see [7, 21] and also the closely related recent paper [14], which we discuss below.

We show that beliefs on the opponent strategy converge to a (stationary) Nash equilibrium of zero-sum stochastic games under the assumption that each state is visited infinitely often in both the model-based case and model-free case. In the model-free case, players do not; know their own payoff functions and the underlying state transition probabilities, however, each player can still observe her realized stage payoff and the current state of the game, as in the reinforcement learning literature. Similarly, beliefs on $Q$-functions converge to the $Q$-functions associated with the equilibrium strategies in both cases.

The fictitious play dynamics presented here reduce to the classical fictitious play when there is only one state and inherit the following features of the classical fictitious play: (i) The dynamics do not require knowledge of the underlying game's type and is not specific to any specific class of games. (ii) Following this scheme, players attain the best performance against an opponent following an asymptotically stationary strategy. (iii) If the dynamics converge, it must converge to an equilibrium of the underlying game.

**1.1.1. Challenges.** As a first challenge, we note that even though the players play best response strategies in the auxiliary games associated with each state, the players do not necessarily play the same auxiliary game repeatedly because their payoff matrices, i.e., $Q$-functions, are time-varying and depend on the play at other states. Since these auxiliary game are not time-invariant, the convergence results for fictitious play or best response dynamics in zero-sum strategic-form games with repeated play, e.g., [20, 9], are not directly applicable to zero-sum stochastic games. We remove this dependency by approximating the discrete-time update via a differential inclusion (specific to each state) at the timescale of the beliefs on strategies as if the beliefs on $Q$-functions are time-invariant. To this end, we interpret an appropriate affine interpolation of the discrete-time update as a perturbed solution to a certain differential inclusion and characterize its limit set via a Lyapunov function argument, as shown by [3].

As a second challenge, we note that when the players do not share a common belief about the $Q$-function, their individual beliefs do not necessarily sum to zero due to the independent updates. Hence, the auxiliary games are non-zero-sum in general. This is problematic since it is well-known that fictitious play (or any uncoupled learning dynamic that does not incorporate the opponent's objective) cannot converge to an equilibrium in every class of non-zero-sum games [10]. To address this challenge, we exploit the structure of the stochastic game and construct a new Lyapunov function for both zero-sum and non-zero-sum games.

For zero-sum games, our Lyapunov function reduces to the Lyapunov function presented by [9] for continuous-time best response dynamics in zero-sum strategic-form games. For non-zero-sum games, this new Lyapunov function enables us to characterize the limit set of the beliefs on strategies in terms of how much the sum of beliefs on $Q$-functions deviates from zero. We exploit this characterization via the asynchronous stochastic approximation methods, provided by [28], to show that the *beliefs* on $Q$-functions sum to zero asymptotically (although they do not necessarily

sum to zero in finite time). We emphasize that the zero-sum structure of stage payoffs of the underlying stochastic game is crucial for this result to hold.

**1.2. Related works.** Our paper is most closely related to the recent paper [14], which presents and studies a continuous-time best response dynamics for zero-sum stochastic games. They consider dynamics where each player selects *a mixed strategy in an auxiliary game* and updates her strategy in the direction of her best response to her opponent's current mixed strategy in the auxiliary game. A single continuation payoff (common among the players) is updated at a slower speed representing the time average of the auxiliary game payoffs up to time $t$. The common update ensures that the auxiliary game is always zero-sum. This allows building on the convergence analysis provided by [9] since two-timescale learning enables the mixed strategies to track an equilibrium associated with the estimates of the continuation payoffs. In [14], the authors generalized the convergence result in [29] (which extends the convergence result in [24] to continuous-time dynamics) to settings with asymptotically negligible (tracking) error, thus establishing convergence of their dynamics in zero-sum stochastic games.

Their dynamics involve updating mixed strategies at every state at every time. Hence, the authors study also an alternative update rule by considering a continuous-time embedding of the actual play of the stochastic game where the game transitions according to a controlled continuous-time Markov chain. Our paper instead considers dynamics where each player follows a best response pure action in the auxiliary game (without any specific tie-breaking rule) while updating her $Q$-function using her belief on the opponent strategy and her current $Q$-function estimate. Therefore, estimates of $Q$-functions do not necessarily sum to zero, leading to an auxiliary game that is not necessarily zero-sum. Furthermore, players update their beliefs on opponent strategy only for the current state within the course of a stochastic game without need for such a continuous-time embedding.

Other related papers include [24], [30], and [23]. In [24], the author presented and studied the minimax-value iteration in zero-sum stochastic games, which can be viewed as a generalization of value iteration in Markov decision problems to zero-sum stochastic game settings by replacing the optimization with the minimax-value of the auxiliary zero-sum game. [24] showed that the minimax-value iteration converges to a unique point, establishing existence of a stationary equilibrium in zero-sum stochastic games. The minimax value iteration necessitates computation of the minimax value of the auxiliary game at each stage. This can be done by solving a collection of linear programs (LPs), one per state at each iteration, which can be computationally demanding.

To mitigate the need to solve LPs at each iteration, in [30], the authors presented and studied a fictitious play like the discrete-time update rule to find an equilibrium in zero-sum stochastic games without solving LPs. This update rule involves each player playing a best-response in an auxiliary game with a *common* continuation payoff as in [14], which preserves the zero-sum nature of the auxiliary game. However, unlike [14], in [30], the players update the continuation payoff using the payoff estimate of one of the players, which simplifies the analysis, but is not a natural update process. The dynamics reduce to fictitious play applied to a convergent sequence of zero-sum strategic-form games for each state.

For *time-averaged* stochastic games, in [23], the authors focused on a special class with two players, two states, and two actions (per state) and provided an example in which the fictitious play presented does not necessarily converge to a stationary equilibrium. This is in contrast with results for two-player two-action strategic-form

games, where fictitious play is known to converge to an equilibrium with a certain tie-breaking rule (see [17]), or when the game has the "diagonal property" for any tie-breaking rule (see [18, 19]).

**1.2.1. Model-free case.** Our paper is also related to a number of papers on multiagent reinforcement learning, e.g., see the survey in [33] and the references therein. Particularly noteworthy is [15], which presented a model-free version of [24]'s minimax-value iteration via a Q-learning-type algorithm, called Minimax-Q. Similar to Shapley's minimax value iteration, Minimax-Q assumes a zero-sum structure and therefore is specific to zero-sum games.

Alternative to Minimax-Q, in [27], the author presented a fictitious-play-like dynamics, called Hyper-Q, which applies beyond zero-sum games. Hyper-Q has dynamics similar to ours, however, evolves over a single timescale without any convergence guarantee in any specific class of stochastic games. On the other hand, in [5], the author presented an actor-critic-type learning algorithm that is also not specific to zero-sum games. Contrary to Hyper-Q, there players do not seek to learn opponent strategies based on actions taken. He showed that a certain (weighted) empirical distribution of the joint actions taken converges to the set of (a modified version of) *generalized Nash equilibria* in stochastic games provided that each state-action pair is visited infinitely often and frequently enough. This is a weaker sense of convergence compared to our result although it is for stochastic games beyond zero-sum.

Other than the papers reviewed above, there are several other multiplayer reinforcement learning algorithms that are shown to have good convergence properties in stochastic games with respect to certain performance measures provided that every player follows rules which at times may not align with their best interests. For example, in [6] and more recently [2], the authors focused on scenarios where players play in a coordinated manner a *finite-horizon version* of a zero-sum stochastic game within repeated episodes, referred to as *episodic reinforcement learning*, even though player payoffs are defined over an infinite horizon. In another line of work, in [32] and [1], the authors presented and studied algorithms that update policies only at certain time instances while keeping them *fixed* in between—even when players may have an incentive to change their actions—in order to create a stationary environment for learning the underlying model or estimating the associated $Q$-functions.

**1.3. Organization.** The rest of the paper is organized as follows. In sections 2 and 3, we model stochastic games and our fictitious play scheme, respectively. We present the assumptions and the convergence results in section 4. The proofs of the main convergence results in the model-based and model-free settings are provided, respectively, in sections 5 and 6. In section 7, we provide an illustrative example. We conclude the paper with some remarks in section 8.

**2. Stochastic games.** Consider two players that interact with each other by taking actions in a dynamic environment over an infinite horizon with discrete time $k = 0, 1, \ldots$. The players collect a *stage payoff* depending on their actions and the current state of the environment, which also determines the next state. A two-player zero-sum stochastic game is a tuple $\langle S, A, r, p, \gamma \rangle$ constructed as follows.

- Let $S$ be a set of *finitely* many states.
- Let $A^i$ be the set of *finitely* many actions that player $i$ can take at any state $s \in S$.[1] Furthermore, $A := A^1 \times A^2$ denotes the set of action profiles $a = (a^1, a^2)$ for $a^i \in A^i$, $i = 1, 2$.

[1] This can be generalized to the case where action spaces depend on state straightforwardly.

- Let $r^i : S \times A \to \mathbb{R}$ denote the *stage payoff function* of player $i$ at state $s$. Since it is a zero-sum game, we have $r^1(s,a) + r^2(s,a) = 0$ for all $(s,a) \in S \times A$.
- For any pair of states $(s, \tilde{s})$ and action profile $a \in A$, we define $p(\tilde{s}|s,a)$ as the *transition probability* from $s$ to $\tilde{s}$ given action profile $a$.
- Let $\gamma \in (0,1)$ denote a *discount factor* that affects the importance of future stage payoffs.

We focus on *stationary (Markov) strategies*, meaning that at each stage each player plays a mixed action that depends only on the current state (and not, for example, on time). This does not cause any loss of generality due to the existence result in [24]. More specifically, for each $i = 1, 2$, we denote by $\pi^i(s, a^i) \in [0,1]$ the probability that player $i$ takes action $a^i$ at state $s$ and the stationary strategy of player $i$ by $\pi^i$. Let us also denote the strategy profile by $\pi := \{\pi^1, \pi^2\}$. Correspondingly, $a_k = (a_k^1, a_k^2)$ is the action profile at stage $k$.

We define the *expected utility* of player $i$ under the strategy profile $\pi$ as the expected discounted sum of stage payoffs

$$U^i(\pi^1, \pi^2) := \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k)\right\}, \tag{2.1}$$

where $\{s_k \sim p(\cdot|s_{k-1}, a_{k-1})\}_{k>0}$ and $\{a_k \sim \pi(s_k, \cdot)\}_{k \geq 0}$ are stochastic processes, respectively, representing the state and the action profile at each stage $k$ and the expectation is taken with respect to all randomness induced by the initial state distribution $s_0 \sim p_o \in \Delta(S)$, the state transition kernel, and strategy profile $\pi$.[2]

A strategy profile $(\tilde{\pi}^1, \tilde{\pi}^2)$ is an $\varepsilon$-Nash equilibrium of the stochastic game with $\varepsilon \geq 0$ provided that

$$U^1(\tilde{\pi}^1, \tilde{\pi}^2) \geq U^1(\pi^1, \tilde{\pi}^2) - \varepsilon \quad \forall \quad \pi^1, \tag{2.2a}$$

$$U^2(\tilde{\pi}^1, \tilde{\pi}^2) \geq U^2(\tilde{\pi}^1, \pi^2) - \varepsilon \quad \forall \quad \pi^2. \tag{2.2b}$$

Correspondingly, $(\tilde{\pi}^1, \tilde{\pi}^2)$ is a Nash equilibrium if (2.2) holds with $\varepsilon = 0$.

**2.1. Auxiliary stage-games in a stochastic game.** At each stage of a stochastic game, the action profile determines the current stage payoff and the stage payoffs that will be received in future stages by determining the next state (since stage payoffs also depend on the state). Correspondingly, if player $i$ knew that the opponent $-i$ is playing according to the stationary strategy $\pi^{-i}$, then the value of the action profile $a \in A$ at current state $s$, denoted by $Q^i(s,a)$ (and known as: *Q-function*), would satisfy the following fixed-point equation:

$$Q^i(s,a) = r^i(s,a) + \gamma \sum_{\tilde{s} \in S} p(\tilde{s}|s,a) \max_{\tilde{a}^i \in A^i} \mathbb{E}_{\tilde{a}^{-i} \sim \pi^{-i}(\tilde{s}, \cdot)}\{Q^i(\tilde{s}, \tilde{a})\}. \tag{2.3}$$

This follows from backward induction based on the principle that player $i$ would look for maximizing her expected utility, as described in (2.1), in future stages. Therefore, a stochastic game can be viewed as a collection of *auxiliary stage-games* specific to each state and represented by $\langle A^1, A^2, Q^1(s,\cdot), Q^2(s,\cdot)\rangle$. For notational convenience, we also define the value function $v^i : S \to \mathbb{R}$ by

$$v^i(s) = \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}(s,\cdot)}\{Q^i(s,a)\}, \tag{2.4}$$

[2]For a set $X$, we denote the probability simplex by $\Delta(X)$.

which corresponds to the maximum value player $i$ would get in the associated auxiliary stage-game. Note that the dependence of $Q^i$ and $v^i$ on $\pi^{-i}$ is implicit in (2.3) and (2.4) for notational convenience.

We also note that in two-player zero-sum stochastic games, there may exist multiple stationary equilibria in two-player zero-sum stochastic games. However, the $Q$-functions and value functions associated with any stationary equilibrium are all the same [24]. We denote them, respectively, by $(Q^1_*, Q^2_*)$ and $(v^1_*, v^2_*)$.

Though a stochastic game can be viewed as a collection of such auxiliary stage-games that are being played repeatedly and asynchronously, the opponent's strategy and the $Q$-function are not readily available to the players. Furthermore, these auxiliary stage-games are not necessarily stationary. However, as in the classical fictitious play, the players can form beliefs on them based on the empirical play as if they are stationary. Given this observation, in the following section, we introduce fictitious-play-type learning dynamics that combines the classical fictitious play with the $Q$-learning for stochastic games.

**3. Fictitious play in stochastic games.** We consider scenarios where players follow fictitious play dynamics in which they form beliefs not only on the opponent's strategy but also on the $Q$-function based on the history of the play. They take the greedy best action in the associated auxiliary stage-game conditioned on their beliefs. We emphasize that the players do not know the opponent's objective. In other words, they do not possess the knowledge that the underlying game is zero-sum.

In the following, we describe the learning dynamics for the typical player 1 in both model-based and model-free settings. The dynamics for the typical opponent player 2 is a mirror of it.

**3.1. Fictitious play for the model-based setting.** At each stage, player 1 has beliefs on the opponent strategy and her $Q$-function, respectively, denoted by $\hat{\pi}^2_k : S \times A \to [0,1]$ and $\hat{Q}^1_k : S \times A \to \mathbb{R}$. For notational convenience, we define $\hat{\pi}^2_k(s) := \hat{\pi}^2_k(s, \cdot)$ and $\hat{Q}^1_k(s) := \hat{Q}^1_k(s, \cdot)$. At stage $k = 0$, she initializes her beliefs *arbitrarily* such that $\hat{\pi}^2_0(s) \in \Delta(A^2)$ and $\hat{Q}^1_0(s) \in \mathbb{R}^{|A^1| \times |A^2|}$ for each $s \in S$.

Let $s \in S$ denote the current state at stage $k \geq 0$. Player 1 and simultaneously player 2 take their greedy best response actions $a^1_k \in A^1$ and $a^2_k \in A^2$. For example, player 1 can take any action satisfying

$$a^1_k \in \underset{a^1 \in A^1}{\operatorname{argmax}} \; \mathbb{E}_{a^2 \sim \hat{\pi}^2_k(s)} \left\{ \hat{Q}^1_k(s, a^1, a^2) \right\}, \tag{3.1}$$

according to arbitrary tie-breaking rules. Without loss of generality, we consider pure actions as degenerate mixed strategies giving probability one to the associated action, i.e., $A^i \subset \Delta(A^i)$. The players can observe the opponent's action. Hence, player 1 updates her belief on player 2's strategy at the current state $s$ according to

$$\hat{\pi}^2_{k+1}(s) = \hat{\pi}^2_k(s) + \alpha_{\#s}(a^2_k - \hat{\pi}^2_k(s)), \tag{3.2}$$

where $\alpha_{\#s} \in (0,1]$ is a step size specific to $\#s$, representing the number of times that $s$ gets visited until (and including) stage $k$.

Furthermore, player 1 updates her belief on her own $Q$-function only for the current state $s$. The update of $\hat{Q}^1_k(s)$ is given by

$$\hat{Q}^1_{k+1}(s,a) = \hat{Q}^1_k(s,a) + \beta_{\#s} \left( r^1(s,a) + \gamma \sum_{\tilde{s} \in S} p(\tilde{s}|s,a) \hat{v}^1_k(\tilde{s}) - \hat{Q}^1_k(s,a) \right) \tag{3.3}$$

TABLE 1
*Fictitious play of the typical player $i$ in stochastic games.*

**Require:** Keep track of $\hat{\pi}_k^{-i}$ and $\hat{Q}_k^i$ for every $(s,a)$.
**for** Each stage $k \geq 0$ **do**
  **Observe** the current state $s_k$.
  **Take action** $a_k^i$ according to (3.1).
  **Observe** the opponent's action $a_k^{-i}$.
  **Update** $\hat{\pi}_k^{-i}(s_k)$ according to (3.2).
  **Update** $\hat{Q}_k^i(s_k,a)$ for all $a \in A$ according to (3.3).
**end for**

for all $a \in A$, where we define the value function estimate $\hat{v}_k^1 : S \to \mathbb{R}$ by

$$\hat{v}_k^1(\tilde{s}) := \max_{\tilde{a}^1 \in A^1} \mathbb{E}_{\tilde{a}^2 \sim \hat{\pi}_k^2(\tilde{s})} \left\{ \hat{Q}_k^1(\tilde{s},\tilde{a}) \right\} \tag{3.4}$$

and again $\beta_{\#s} \in (0,1]$ is a step size specific to $\#s$.

Player 1 does not update her beliefs associated with other states, i.e., $\hat{\pi}_{k+1}^2(s') = \hat{\pi}_k^2(s')$ and $\hat{Q}_{k+1}^1(s') = \hat{Q}_k^1(s')$ if $s' \neq s$, i.e., if $s'$ is not the current state. A description of the dynamics is tabulated in Table 1.

Note that given the definition of $\hat{v}_k^i(s)$ in (3.4), we do not necessarily have $\hat{v}_k^1(s) + \hat{v}_k^2(s) = 0$ for all $s \in S$. Moreover, when $\hat{v}_k^1(s) + \hat{v}_k^2(s) \neq 0$ for some $s \in S$, then by (3.3), we do not necessarily have $\hat{Q}_{k+1}^1(s) + \hat{Q}_{k+1}^2(s)$ equal to the zero matrix for all $s \in S$. Therefore, the auxiliary stage-games need not be zero-sum in this learning dynamics.

Alternatively, consider the scenario where player 1 and player 2 update their beliefs on $Q$-functions as in (3.3) but with

$$\tilde{v}_k^i(\tilde{s}) = \mathbb{E}_{\tilde{a} \sim \hat{\pi}(\tilde{s})} \left\{ \hat{Q}_k^i(\tilde{s},\tilde{a}) \right\}. \tag{3.5}$$

In other words, the players form beliefs on their own strategies (which is not a natural dynamics but we pursue it briefly to illustrate the more tractable mathematical structure this leads to). If the beliefs on $Q$-functions are initialized such that $\hat{Q}_0^1(s) + \hat{Q}_0^2(s)$ is equal to the zero matrix for each $s$, then we have $\tilde{v}_0^1(s) + \tilde{v}_0^2(s) = 0$. Then by induction, it can be shown that $\tilde{v}_k^1(s) + \tilde{v}_k^2(s) = 0$ for all $s$ and $k$, and indeed $\hat{Q}_k^1(s) + \hat{Q}_k^2(s)$ is equal to the zero matrix for all $s$ and $k$. Therefore, the auxiliary stage-games would always be zero-sum.

In the scenarios where the auxiliary stage-games remain always zero-sum, the convergence analysis is a direct application of the two-timescale stochastic approximation theory built on the convergence result for fictitious play in zero-sum strategic-form games (with repeated play) provided by [9] and the convergence result for the minimax value iteration provided by [24]. However, this is not the case when players follow an uncoupled learning dynamics, such as our two-timescale fictitious play, and its convergence analysis necessitates development of new technical tools specific to the structure of stochastic games rather than resorting directly to the two-timescale stochastic approximation theory.

**3.2. Fictitious play for the model-free setting.** Next we consider the scenarios where players do not know their own stage payoff function and the transition probabilities. They can still observe their current stage payoff (realized), current state (visited), and the current action (taken by the opponent). Given the beliefs on $Q$-functions, the players take the actions according to (3.1), while they may also take

TABLE 2
*Model-free fictitious play of the typical player $i$ in stochastic games.*

**Require:** Keep track of $\hat{\pi}_k^{-i}$ and $\hat{Q}_k^i$ for every $(s,a)$.
**for** Each stage $k \geq 0$ **do**
  **Observe** the current state $s_k$.
  **Update** $\hat{Q}_{k-1}^i(s_{k-1},a_{k-1})$ according to (3.7).
  **Take action** $a_k^i$ according to (3.6).
  **Observe** the opponent's action $a_k^{-i}$.
  **Update** $\hat{\pi}_k^{-i}(s_k)$ according to (3.2).
**end for**

some random action with some small probability $\epsilon > 0$ to *experiment* stochastic state transitions, as in [15]. For example, player 1 can take action

$$a_k^1 = \begin{cases} a_*^1 & \text{w.p. } (1-\epsilon), \\ u^1 & \text{w.p. } \epsilon, \end{cases} \tag{3.6}$$

where $a_*^1$ is a greedy best response satisfying (3.1) while $u^1 \sim \mathcal{U}(A^1)$ with $\mathcal{U}(\cdot)$ denoting the uniform distribution over the associated set. We focus on this basic exploration strategy as a proof of concept. The players can also resort to more sophisticated strategies to speed up their exploration, e.g., see [11].

Players still update their beliefs on opponent strategy according to (3.2). However, since player 1 and player 2 cannot update their beliefs on $Q$-functions as in (3.3) without knowing state transition probabilities, they instead follow a $Q$-learning-type of update described as follows.

Player $i$ observes current state $s$, current action profile $a$, and her current stage payoff (denoted by $r_k^i$), and by looking one stage ahead, she also observes the next state $\tilde{s}$. Given the triple $(s,a,\tilde{s})$, she uses an estimate for the continuation payoff for the next state, i.e., $\hat{v}_k^i(\tilde{s})$, as an unbiased estimator of $\sum_{s' \in S} \hat{v}_k^i(s') p(s'|s,a)$. She updates her belief on the $Q$-function only for the current state and action profile $(s,a)$, according to

$$\hat{Q}_{k+1}^i(s,a) = \hat{Q}_k^i(s,a) + \beta_{\#(s,a)} \left( r_k^i + \gamma \hat{v}_k^i(\tilde{s}) - \hat{Q}_k^i(s,a) \right), \tag{3.7}$$

where $\hat{v}_k^i$ is as described in (3.4). Note that here $\beta_{\#(s,a)} \in (0,1]$ is a step size specific to $\#(s,a)$ representing the number of times *state-action* pair $(s,a)$ occurs until (and including) stage $k$. Note also that player 1 does not update $\hat{Q}_k^1$ associated with other state-action pairs, i.e., $\hat{Q}_{k+1}^1(s',a') = \hat{Q}_k^1(s',a')$ if $(s',a') \neq (s,a)$, i.e., if either $s'$ is not the current state or $a'$ is not the current action profile. A description of the model-free dynamics is tabulated in Table 2. Note that $\hat{Q}_{k-1}^i(s_{k-1},a_{k-1})$ gets updated after $s_k$ is observed. Therefore, the updates of beliefs take place at different orders in Tables 1 and 2.

**4. Main result.** In this paper, we focus on whether the beliefs formed on the opponent's strategies and $Q$-functions converge to a stationary equilibrium and the corresponding $Q$-functions in zero-sum stochastic games, or not. The answer is *affirmative* for both model-based and model-free settings under certain assumptions provided below precisely.

*Assumption* 4.1. Each state is visited infinitely often with probability one.

Players update their beliefs associated with a state only when that state is visited. This assumption ensures that players have sufficient time to revise and improve their

beliefs. Furthermore, it holds if the stochastic game is *irreducible*, e.g., transition probabilities between any pair of states are positive for any joint action as in [14].

*Assumption* 4.2. The step sizes $\{\alpha_c \in (0,1]\}_{c=0}^{\infty}$ and $\{\beta_c \in (0,1]\}_{c=0}^{\infty}$ satisfy $\sum_{c=0}^{\infty} \alpha_c = \infty$, $\sum_{c=0}^{\infty} \beta_c = \infty$, and $\lim_{c\to\infty} \alpha_c = \lim_{c\to\infty} \beta_c = 0$. The beliefs on $Q$-functions are updated at a slower timescale compared to the timescale in which the beliefs on strategies are updated, i.e., $\lim_{c\to\infty} \frac{\beta_c}{\alpha_c} = 0$.

Now we are ready to present the convergence results specific to zero-sum stochastic games.

THEOREM 4.3. *Suppose that Assumptions* 4.1 *and* 4.2 *hold. When both players follow the fictitious play dynamics described in Table* 1*, i.e.,* (3.1)–(3.3)*, the beliefs on strategies and Q-functions, respectively, converge to a stationary equilibrium and the corresponding Q-functions in zero-sum stochastic games almost surely. In other words, for some stationary equilibrium* $\pi_* = (\pi_*^1, \pi_*^2)$*, we have* $(\hat{\pi}_k^1, \hat{\pi}_k^2) \to (\pi_*^1, \pi_*^2)$ *and* $(\hat{Q}_k^1, \hat{Q}_k^2) \to (Q_*^1, Q_*^2)$*, as* $k \to \infty$*, with probability* 1*.*

The following corollary to Theorem 4.3 characterizes the convergence properties of the dynamics in two-player *general-sum* stochastic games in terms of how much the stage payoffs deviate from the zero-sum structure. Particularly, it shows convergence of the dynamics to a *near* equilibrium in *near* zero-sum stochastic games.

COROLLARY 4.4. *Suppose that Assumptions* 4.1 *and* 4.2 *hold and both players follow the fictitious play dynamics described in Table* 1*, i.e.,* (3.1)–(3.3)*. Then, in two-player general-sum stochastic games, we have*

$$\limsup_{k\to\infty} |\hat{Q}_k^i(s,a) - Q_d^i(s,a)| \le \frac{d(1+\gamma)}{\gamma(1-\gamma)^2} \quad \forall (s,a), \tag{4.1}$$

*with probability* 1*, where* $Q_d^i : S \times A \to \mathbb{R}$ *satisfies the fixed point equation*

$$Q_d^i(s,a) = r^i(s,a) + \gamma \sum_{\tilde{s}\in S} p(\tilde{s}|s,a) \max_{\tilde{\pi}^i \in \Delta(A^i)} \min_{\tilde{\pi}^{-i} \sim \Delta(A^{-i})} \mathbb{E}_{(\tilde{a}^i, \tilde{a}^{-i}) \sim (\tilde{\pi}^i, \tilde{\pi}^{-i})} \{Q_d^i(\tilde{s}, \tilde{a})\}$$

*for all* $(s,a)$ *and* $d := \max_{(s,a)} |r^1(s,a) + r^2(s,a)|$*.*

In the model-free setting, players can update only a single entry of the belief on $Q$-function corresponding to the current action profile. This strengthens the coupling across states and makes the dynamics difficult to track. Furthermore, the players can only observe the realization of the state transitions, which introduces stochastic approximation errors in the learning dynamics. Hence, we make the following assumption to limit the impact of the coupling and the stochastic approximation errors.

*Assumption* 4.5. The step sizes satisfy $\sum_{c=0}^{\infty} \alpha_c^2 < \infty$ and $\sum_{c=0}^{\infty} \beta_c^2 < \infty$. The sequence $\{\beta_c\}_{c\ge 0}$ is monotonically decreasing. We have $\lim_{c\to\infty} \frac{\beta_{\lfloor mc \rfloor}}{\alpha_c} = 0$ for any $m \in (0,1]$.

The first part of Assumption 4.5 ensures that the stochastic approximation terms are square integrable Martingale difference sequences conditioned on the history. On the other hand, the second part of Assumption 4.5 ensures that $\beta_{\#(s,a)}/\alpha_{\#s}$ gets arbitrarily small asymptotically even though $\#(s,a)$ increases more slowly than $\#s$. We also emphasize that Assumptions 4.2 and 4.5 are properties of step sizes and do not impose further conditions on zero-sum stochastic games for which the results hold. For example, the step sizes given by $\alpha_c = (c+1)^{-\rho_\alpha}$ and $\beta_c = (c+1)^{-\rho_\beta}$, where $1/2 < \rho_\alpha < \rho_\beta \le 1$, satisfy Assumptions 4.2 and 4.5.

The following theorem characterizes the convergence properties of fictitious play for the model-free setting.

THEOREM 4.6. *Suppose that Assumptions* 4.1, 4.2, *and* 4.5 *hold. When both players follow the fictitious play dynamics described in Table* 2, *i.e.,* (3.6), (3.2), *and* (3.7), *the beliefs on strategies and Q-functions converge to a near equilibrium and the equilibrium Q-functions with an approximation level linear in the exploration probability* $\epsilon > 0$, *almost surely. More explicitly, we have*

$$\limsup_{k\to\infty} \left(U^i(\pi^i, \hat{\pi}_k^{-i}) - U^i(\hat{\pi}_k^i, \hat{\pi}_k^{-i})\right) \leq 2\epsilon D \frac{(1+\gamma)^2}{\gamma(1-\gamma)^3} \quad \forall \pi^i, \tag{4.2}$$

$$\limsup_{k\to\infty} |\hat{Q}_k^i(s,a) - Q_*^i(s,a)| \leq \epsilon D \frac{1+\gamma}{(1-\gamma)^2} \quad \forall (s,a), \tag{4.3}$$

*with probability* 1, *where* $D = \frac{1}{1-\gamma}\sum_i \max_{(s,a)} |r^i(s,a)|$.

Note that if the players decrease their exploration probability at a suitable rate, the learning dynamics for the model-free case can also converge to an exact equilibrium of the stochastic game, e.g., see [13, section 5]. Note also that the analysis can be generalized to the case with independent random perturbations of stage-payoffs (with compact support) straightforwardly, as shown in the extended version [22, section 6.3].

We provide the proofs of Theorems 4.3 and 4.6 in sections 5 and 6, respectively.

**5. Proof of Theorem 4.3: Convergence for the model-based setting.** We divide the proof into three main steps: (i) We first decouple the dynamics across states over the fast timescale. (ii) We next zoom into the local dynamics specific to each state and characterize their limit set via a novel Lyapunov function argument addressing the deviation of auxiliary stage games from zero-sum structure in two-player zero-sum stochastic games. (iii) We then zoom out to global dynamics across every state over the slow timescale and show that the beliefs on $Q$-functions sum to zero asymptotically, and then show that estimates of continuation payoffs track the minimax values associated with the beliefs on $Q$-functions. Finally, we show that the beliefs on $Q$-functions converge to the unique $Q$-functions of an equilibrium of the underlying zero-sum stochastic game.

We can now delve into the technical details.[3]

**5.1. Step (i): Decoupling the dynamics at the fast timescale.** Let $s_k$ denote the current state at stage $k$. Based on (3.2), we can write the updates of the beliefs on strategies for state $s \in S$ as

$$\begin{bmatrix} \hat{\pi}_{k+1}^1(s) \\ \hat{\pi}_{k+1}^2(s) \end{bmatrix} = \begin{bmatrix} \hat{\pi}_k^1(s) \\ \hat{\pi}_k^2(s) \end{bmatrix} + \bar{\alpha}_k(s) \begin{bmatrix} a_k^1 - \hat{\pi}_k^1(s) \\ a_k^2 - \hat{\pi}_k^2(s) \end{bmatrix}, \tag{5.1}$$

where $\bar{\alpha}_k(s) := \mathbb{I}_{\{s=s_k\}}\alpha_{\#s} \in [0,1]$.[4] At the same timescale, we can write the updates of the beliefs on $Q$-functions of both players for state $s \in S$ as

$$\begin{bmatrix} \hat{Q}_{k+1}^1(s,a) \\ \hat{Q}_{k+1}^2(s,a) \end{bmatrix} = \begin{bmatrix} \hat{Q}_k^1(s,a) \\ \hat{Q}_k^2(s,a) \end{bmatrix} + \bar{\alpha}_k(s) \begin{bmatrix} \mathcal{E}_k^1(s,a) \\ \mathcal{E}_k^2(s,a) \end{bmatrix} \quad \forall a \in A, \tag{5.2}$$

[3] We omit the qualification "with probability 1" since we have already discarded the suitable set of measure zero, where Assumption 4.1 does not hold.

[4] We represent the indicator function (which is 1 if $s = s_k$ and 0 otherwise) by $\mathbb{I}_{\{s=s_k\}}$.

combining (3.3) for each $i = 1, 2$ together. The error terms $\mathcal{E}_k^i : S \times A \to \mathbb{R}$ are defined by

$$\begin{bmatrix} \mathcal{E}_k^1(s,a) \\ \mathcal{E}_k^2(s,a) \end{bmatrix} := \frac{\beta_{\#s}}{\alpha_{\#s}} \begin{bmatrix} r^1(s,a) + \gamma \sum_{\tilde{s} \in S} p(\tilde{s}|s,a) \hat{v}_k^1(\tilde{s}) - \hat{Q}_k^1(s,a) \\ r^2(s,a) + \gamma \sum_{\tilde{s} \in S} p(\tilde{s}|s,a) \hat{v}_k^2(\tilde{s}) - \hat{Q}_k^2(s,a) \end{bmatrix}, \tag{5.3}$$

where $\hat{v}_k^i(\tilde{s})$ is as described in (3.4). Note that the iterates are bounded since the stage payoffs have compact support, the discount factor $\gamma \in (0,1)$, and step sizes take values only in $[0,1]$. Therefore, Assumption 4.2, i.e., $\lim_{c \to \infty} \frac{\beta_c}{\alpha_c} = 0$, yields that the error matrices in (5.2) are asymptotically negligible.

Note that (5.1) and (5.2) for state $s$ are coupled with the dynamics at other states $s' \neq s$ through the asymptotically negligible error terms. We consider the evolution of the iterates specific to a state, $\mathrm{x}_k(s) := (\hat{\pi}_k^1(s), \hat{\pi}_k^2(s), \hat{Q}_k^1(s), \hat{Q}_k^2(s))$, *separately* by exploiting this weak coupling. Note that we have $\sum_{k=0}^{\infty} \bar{\alpha}_k(s) = \infty$ and $\bar{\alpha}_k(s) \to 0$ as $k \to \infty$ for each $s$ based on Assumptions 4.1 and 4.2. Therefore, the stochastic differential inclusion theory, e.g., see [3, Theorem 3.6 and Proposition 3.27], yields that we can characterize the limit set of $\mathrm{x}_k(s)$ by formulating a Lyapunov function for the following flow:[5]

$$\frac{d\boldsymbol{\pi}^1(t)}{dt} + \boldsymbol{\pi}^1(t) \in \operatorname*{argmax}_{a \in A^1} \left\{ a^T \mathrm{Q}^1(t) \boldsymbol{\pi}^2(t) \right\}, \tag{5.4a}$$

$$\frac{d\boldsymbol{\pi}^2(t)}{dt} + \boldsymbol{\pi}^2(t) \in \operatorname*{argmax}_{a \in A^2} \left\{ \boldsymbol{\pi}^1(t)^T \mathrm{Q}^2(t) a \right\}, \tag{5.4b}$$

$$\frac{d\mathrm{Q}^i(t)}{dt} = O, \quad i = 1, 2, \tag{5.4c}$$

where we drop the dependence on $s$ for notational brevity, use a different notation for the functions $\boldsymbol{\pi}^i : [0, \infty) \to \Delta(A^i)$ and $\mathrm{Q}^i : [0, \infty) \to \mathbb{R}^{|A^1| \times |A^2|}$ for $i = 1, 2$, and represent zero matrices by $O$. Note that (5.4c) yields that $\mathrm{Q}^1(\cdot)$ and $\mathrm{Q}^2(\cdot)$ are time-invariant, e.g., $\tilde{Q}^1 := \mathrm{Q}^1(t)$ and $\tilde{Q}^2 := \mathrm{Q}^2(t)$ for all $t \in [0, \infty)$. Therefore, (5.4a) and (5.4b) correspond to the continuous-time best-response dynamics in a strategic-form game with payoff matrices $\tilde{Q}^1$ and $\tilde{Q}^2$. However, $\tilde{Q}^1 + \tilde{Q}^2$ is not necessarily equal to the zero matrix. In other words, it is not necessarily a zero-sum game. In the next step, we formulate a novel Lyapunov function addressing this challenge.

**5.2. Step (ii): Zooming into the local dynamics at the fast timescale.** It is instructive to examine the continuous-time best response dynamics in zero-sum strategic-form games. For a *fixed* (absolutely continuous) solution to (5.4), e.g., $(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t))$, [9, section 4] showed that

$$V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t)) = \mathrm{h}^1(t) + \mathrm{h}^2(t), \tag{5.5}$$

where

$$\mathrm{h}^1(t) := \max_{a \in A^1} \left\{ a^T \tilde{Q}^1 \boldsymbol{\pi}^2(t) \right\} \quad \text{and} \quad \mathrm{h}^2(t) := \max_{a \in A^2} \left\{ \boldsymbol{\pi}^1(t)^T \tilde{Q}^2 a \right\} \tag{5.6}$$

is a Lyapunov function for (5.4) when $\tilde{Q}^1 + \tilde{Q}^2$ is equal to the zero matrix. Particularly, define functions $\mathrm{a}^1(t)$ and $\mathrm{a}^2(t)$ such that

$$\mathrm{a}^1(t) = \frac{d\boldsymbol{\pi}^1(t)}{dt} + \boldsymbol{\pi}^1(t) \quad \text{and} \quad \mathrm{a}^2(t) = \frac{d\boldsymbol{\pi}^2(t)}{dt} + \boldsymbol{\pi}^2(t), \tag{5.7}$$

[5] Note that $\mathbb{E}_{a^2 \sim \hat{\pi}_k^2(s)} \{ \hat{Q}_k^1(s,a) \} = (a^1)^T \hat{Q}_k^1(s) \hat{\pi}_k^2(s)$ and $\mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)} \{ \hat{Q}_k^2(s,a) \} = \hat{\pi}_k^1(s)^T \hat{Q}_k^2(s) a^2$.

then we have $\mathrm{h}^1(t) = \mathrm{a}^1(t)^T \tilde{Q}^1 \boldsymbol{\pi}^2(t)$ and $\mathrm{h}^2(t) = \boldsymbol{\pi}^1(t)^T \tilde{Q}^2 \mathrm{a}^2(t)$ since

$$\mathrm{a}^1(t) \in \operatorname*{argmax}_{a \in A^1} \left\{ a^T \tilde{Q}^1 \boldsymbol{\pi}^2(t) \right\} \quad \text{and} \quad \mathrm{a}^2(t) \in \operatorname*{argmax}_{a \in A^2} \left\{ \boldsymbol{\pi}^1(t)^T \tilde{Q}^2 a \right\}$$

by (5.4). [9, section 5] showed that for almost every $t \in [0, \infty)$, $\mathrm{h}^1(t)$ and $\mathrm{h}^2(t)$ are differentiable functions of time and we have[6]

$$\frac{d\mathrm{h}^1(t)}{dt} = \mathrm{a}^1(t)^T \tilde{Q}^1 (\mathrm{a}^2(t) - \boldsymbol{\pi}^2(t)) \quad \text{and} \quad \frac{d\mathrm{h}^2(t)}{dt} = (\mathrm{a}^1(t) - \boldsymbol{\pi}^1(t))^T \tilde{Q}^2 \mathrm{a}^2(t). \tag{5.8}$$

Note that (5.8) holds irrespective of whether $\tilde{Q}^1 + \tilde{Q}^2 = O$ or not. Then by the definition of $V_H$, as described in (5.5), we obtain

$$\begin{aligned} \frac{dV_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t))}{dt} &= \mathrm{a}^1(t)^T \tilde{Q}^1 (\mathrm{a}^2(t) - \boldsymbol{\pi}^2(t)) + (\mathrm{a}^1(t) - \boldsymbol{\pi}^1(t))^T \tilde{Q}^2 \mathrm{a}^2(t) \\ &= -V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t)) + \mathrm{a}^1(t)^T (\tilde{Q}^1 + \tilde{Q}^2) \mathrm{a}^2(t) \end{aligned} \tag{5.9}$$

for almost every $t \in [0, \infty)$. Therefore, as can be seen in (5.9), the time derivative of $V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t))$ is *not* necessarily nonpositive for an arbitrary fixed solution $(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t))$. However, when $\tilde{Q}^1 + \tilde{Q}^2 = O$, the time derivative (5.9) reduces to

$$\frac{dV_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t))}{dt} = -V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t)), \tag{5.10}$$

which is nonpositive since $V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t)) \geq 0$ when $\tilde{Q}^1 + \tilde{Q}^2 = O$.[7] Hence $V_H$ is a Lyapunov function when $\tilde{Q}^1 + \tilde{Q}^2 = O$.

Our goal here is to modify $V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t))$ to general-sum settings for which the *possibility* that

$$\mathrm{h}^1(t) + \mathrm{h}^2(t) < \mathrm{a}^1(t)^T (\tilde{Q}^1 + \tilde{Q}^2) \mathrm{a}^2(t) \tag{5.11}$$

poses a challenge as can be seen in (5.9). Hence as a candidate Lyapunov function, instead we propose

$$V(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t), \mathrm{Q}^1(t), \mathrm{Q}^2(t)) := \left[ \mathrm{h}^1(t) + \mathrm{h}^2(t) - \lambda \|\mathrm{Q}^1(t) + \mathrm{Q}^2(t)\|_{\max} \right]_+, \tag{5.12}$$

where $\lambda > 1$ is *arbitrary* and is set such that $\gamma\lambda < 1$ given the discount factor $\gamma \in (0, 1)$. By (5.4c), $V(\cdot)$ reduces to $V_H(\cdot)$ if $\tilde{Q}^1 + \tilde{Q}^2 = O$. When $\|\tilde{Q}^1 + \tilde{Q}^2\|_{\max} > 0$, we always have

$$\mathrm{a}^1(t)^T (\tilde{Q}^1 + \tilde{Q}^2) \mathrm{a}^2(t) < \lambda \|\tilde{Q}^1 + \tilde{Q}^2\|_{\max} \tag{5.13}$$

since $\lambda > 1$. Furthermore, $V(\cdot)$ is a nonnegative continuous function by definition irrespective of whether $\tilde{Q}^1 + \tilde{Q}^2 = O$ or not.

[6] As pointed out by Harris [9], the envelope theorem would have led to (5.8) if certain smoothness conditions held. Inspired by this intuition, he showed (5.8) by using the Taylor series expansion of strategies and by incorporating the closed form solution of the differential equation (5.7).

[7] In particular, $\tilde{Q}^1 + \tilde{Q}^2 = O$ implies that $\mathrm{h}^2(t) = \max_{a \in A^2} \{\boldsymbol{\pi}^1(t)^T (-\tilde{Q}^1) a\}$ and therefore $V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t))$ is bounded from below by

$$V_H(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t)) \geq \min_{\pi^2 \in \Delta(A^2)} \max_{\pi^1 \in \Delta(A^1)} \left\{ (\pi^1)^T \tilde{Q}^1 \pi^2 \right\} - \max_{\pi^1 \in \Delta(A^1)} \min_{\pi_2 \in \Delta(A^2)} \left\{ (\pi^1)^T \tilde{Q}^1 \pi^2 \right\} = 0,$$

where the equality follows from the minimax theorem.

It is also instructive to note that the set $\{x : V(x) = 0\}$ is not necessarily the set of equilibria for the continuous-time best response dynamics, (5.4). However, validity of this candidate function implies that for any solution to the best-response dynamics (5.4), the sum

$$\mathrm{h}^1(t) + \mathrm{h}^2(t) = \max_{a \in A^1} \left\{ a^T \tilde{Q}^1 \boldsymbol{\pi}^2(t) \right\} + \max_{a \in A^2} \left\{ \boldsymbol{\pi}^1(t)^T \tilde{Q}^2 a \right\} \tag{5.14}$$

is asymptotically bounded from above by a scaled version of the maximum norm $\|\tilde{Q}^1 + \tilde{Q}^2\|_{\max}$. Later when we focus on discrete-time updates of beliefs on $Q$-functions (at the slow timescale), we will exploit this asymptotic bound to show that the beliefs on $Q$-functions sum to zero asymptotically.

The following lemma shows that $V(\cdot)$ is indeed a Lyapunov function.

LEMMA 5.1. *The candidate function $V(\cdot)$ is a Lyapunov function of the differential inclusion* (5.4) *for the set $\{x : V(x) = 0\}$. In other words, for any trajectory $\mathrm{x}(t) = (\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t), \mathrm{Q}^1(t), \mathrm{Q}^2(t))$ of* (5.4)*, we have*

- *$V(\mathrm{x}(t')) < V(\mathrm{x}(t))$ for all $t' > t$ if $V(\mathrm{x}(t)) > 0$,*
- *$V(\mathrm{x}(t')) = 0$ for all $t' > t$ if $V(\mathrm{x}(t)) = 0$.*

*Proof.* Fix an arbitrary solution to (5.4), $\mathrm{x}(t)$, and define the function

$$L(t) := \mathrm{h}^1(t) + \mathrm{h}^2(t) - \lambda \|\tilde{Q}^1 + \tilde{Q}^2\|_{\max} \tag{5.15}$$

so that $V(\mathrm{x}(t)) = [L(t)]_+$. Note that $L(\cdot)$ is absolutely continuous because the solution $(\pi^1(\cdot), \pi^2(\cdot))$ is absolutely continuous, $\max\{\cdot\}$ and addition satisfy Lipschitz condition [4, Lemma 4.3.2]. Furthermore, the term $\lambda \|\tilde{Q}^1 + \tilde{Q}^2\|_{\max}$ does not depend on time. Hence we can compute the time derivative almost everywhere by using (5.8), which leads to

$$\begin{aligned} \frac{dL(t)}{dt} &= -(\mathrm{h}^1(t) + \mathrm{h}^2(t)) + \mathrm{a}^1(t)^T (\tilde{Q}^1 + \tilde{Q}^2) \mathrm{a}^2(t) \\ &< -(\mathrm{h}^1(t) + \mathrm{h}^2(t)) + \lambda \|\tilde{Q}^1 + \tilde{Q}^2\|_{\max} = -L(t), \end{aligned} \tag{5.16}$$

where the strict inequality follows since $\lambda > 1$. Therefore, the absolutely continuous $L(t)$ is *strictly* decreasing whenever $L(t) \geq 0$.

On the other hand, by definition, $V(\mathrm{x}(t)) = L(t)$ if $L(t) \geq 0$. Further $V(\mathrm{x}(t)) = 0$ if $L(t) < 0$. Therefore $V(\mathrm{x}(t))$ is strictly decreasing if and only if $V(\mathrm{x}(t)) > 0$. Further $V(\mathrm{x}(t))$ remains constant if and only if $V(\mathrm{x}(t)) = 0$. This is irrespective of whether $L(t) < 0$ increases or decreases since $L(t)$ is strictly decreasing when $L(t) = 0$. More explicitly, the only way $L(t) < 0$ becomes $L(t') > 0$ for some $t' > t$ is by crossing zero, that is, if there exists a time $t'' \in (t, t')$ such that $L(t'') = 0$ since $L(\cdot)$ is a continuous function. However, at that time $t''$, by (5.16) the time derivative is *strictly* negative leading it back inside the set $(-\infty, 0]$ and preventing an escape to positive values. ∎

Lemma 5.1 and the stochastic differential inclusion yield that

$$\lim_{k \to \infty} V(\hat{\pi}_k^1(s), \hat{\pi}_k^2(s), \hat{Q}_k^1(s), \hat{Q}_k^2(s)) = 0 \quad \forall\, s \in S \tag{5.17}$$

and the limit set of the Lyapunov function $\{x : V(x) = 0\}$ is given by

$$\left\{ (\pi^1, \pi^2, Q^1, Q^2) : \max_{a \in A^1} \{a^T Q^1 \pi^2\} + \max_{a \in A^2} \{(\pi^1)^T Q^2 a\} - \lambda \|Q^1 + Q^2\|_{\max} \leq 0 \right\}.$$

In the following step, we characterize the convergence properties of the beliefs on $Q$-functions by using (5.17).

**5.3. Step (iii): Zooming out to the global dynamics at the slow timescale.** We will first show that the sum of the payoff matrices in the auxiliary games, i.e.,

$$\bar{Q}_k(s) := \hat{Q}_k^1(s) + \hat{Q}_k^2(s), \tag{5.18}$$

converges to zero based on the limit set characterization (5.17), and then we will use this result to show that $\hat{v}_k^i(s)$ tracks the saddle point associated with $\hat{Q}_k^i(s)$, denoted by

$$\mathrm{val}^i(\hat{Q}_k^i(s)) := \max_{\pi^i \in \Delta(A^i)} \min_{\pi^{-i} \in \Delta(A^{-i})} \left\{ (\pi^1)^T \hat{Q}_k(s)^i \pi^2 \right\}, \quad i = 1, 2. \tag{5.19}$$

By definition of the Lyapunov function (5.12), we can write (5.17) as

$$\lim_{k\to\infty} \left[ \bar{v}_k(s) - \lambda \|\bar{Q}_k(s)\|_{\max} \right]_+ = 0, \tag{5.20}$$

where we define the sum $\bar{v}_k(s) := \hat{v}_k^1(s) + \hat{v}_k^2(s)$. This implies that the sum $\bar{v}_k(s)$ is less than or equal to $\lambda\|\bar{Q}_k(s)\|_{\max}$ in the limit. On the other hand, we can bound $\bar{v}_k(s)$ from below by $-\lambda\|\bar{Q}_k(s)\|_{\max}$ as

$$\bar{v}_k(s) \geq \hat{\pi}_k^1(s)^T \hat{Q}_k^1(s) \hat{\pi}_k^2(s) + \hat{\pi}_k^1(s)^T \hat{Q}_k^2(s) \hat{\pi}_k^2(s) \geq -\lambda \|\bar{Q}_k(s)\|_{\max} \tag{5.21}$$

since $\lambda > 1$. Combining (5.20) and (5.21), we obtain

$$-\lambda\|\bar{Q}_k(s)\|_{\max} \leq \bar{v}_k(s) \leq \lambda\|\bar{Q}_k(s)\|_{\max} + \bar{\epsilon}_k(s) \ \forall s \in S,\ k \geq 0, \tag{5.22}$$

where $\bar{\epsilon}_k(s)$ is an asymptotically negligible error for each $s \in S$. Based on (5.22) and [22, Theorem 5.1], the following lemma exploits the fact that $r_s^1(a) + r_s^2(a) = 0$ for all $(s, a)$ and shows that the auxiliary games become zero-sum in the *limit* even though they are not necessarily zero-sum in finite time.

LEMMA 5.2. $\|\bar{Q}_k(s)\|_{\max} \to 0$ *as* $k \to \infty$ *for each* $s$.

*Proof.* The proof follows from (i) writing the sum $\bar{Q}_k$ in a recursive form based on the updates of $(\hat{Q}_k^1, \hat{Q}_k^2)$ and then (ii) showing that this recursion satisfies the conditions listed in [22, Theorem 5.1] .

*Step* (a). Based on the fact that $r^1(s, a) + r^2(s, a) = 0$ for all $(s, a)$, the update of the beliefs on $Q$-functions, as described in (3.3), yields that

$$\bar{Q}_{k+1}(s, a) = (1 - \bar{\beta}_k(s))\bar{Q}_k(s, a) + \bar{\beta}_k(s)\gamma \sum_{\tilde{s}\in S} p(\tilde{s}|s, a)\bar{v}_k(\tilde{s}) \quad \forall a, \tag{5.23}$$

where $\bar{\beta}_k(s) := \mathbb{I}_{\{s=s_k\}}\beta_{\#s} \in [0, 1]$. Note that without $r^1(s, a) + r^2(s, a) = 0$ for all $(s, a)$, there would be an additional constant term in the evolution of the sequence $\{\bar{Q}_k(s, a)\}_{k\geq 0}$ preventing its convergence to zero.

*Step* (b). Based on (5.22) and (5.23), we can bound $\bar{Q}_{k+1}(s, a)$ as follows:

$$\bar{Q}_{k+1}(s, a) \leq (1 - \bar{\beta}_k(s))\bar{Q}_k(s, a) + \bar{\beta}_k(s) \left( \bar{\gamma} \max_{s'\in S} \|\bar{Q}_k(s')\|_{\max} + \bar{\varepsilon}_k \right), \tag{5.24a}$$

$$\bar{Q}_{k+1}(s, a) \geq (1 - \bar{\beta}_k(s))\bar{Q}_k(s, a) - \bar{\beta}_k(s)\bar{\gamma} \max_{s'\in S} \|\bar{Q}_k(s')\|_{\max}, \tag{5.24b}$$

where $\bar{\gamma} := \gamma\lambda \in (0, 1)$ and $\bar{\varepsilon}_k = \gamma \max_{s'\in S} \bar{\epsilon}_{s',k}$ is an asymptotically negligible error. We emphasize that $\gamma\lambda \in (0, 1)$ plays an important role in [22, Theorem 5.1]. Based on Assumptions 4.1 and 4.2, the step size $\bar{\beta}_k(s) \in [0, 1]$ satisfies the conditions listed in [22, Theorem 5.1]. Furthermore, the iterates are bounded since the stage payoffs have compact support, $\gamma \in (0, 1)$ and the step sizes take values only in $[0, 1]$. Therefore, we can invoke [22, Theorem 5.1] and complete the proof. ∎

Since the maximum norm of $\bar{Q}_k(s)$ converges to zero by Lemma 5.2, the upper bound on $\bar{v}_k$, as described in (5.22), implies

$$\lim_{k\to\infty} |\bar{v}_k(s)| = 0 \ \forall s \in S. \tag{5.25}$$

Based on (5.25) and Lemma 5.2, the following lemma shows that estimates of continuation payoffs track the minimax values associated with the beliefs on $Q$-functions.

LEMMA 5.3. $|\hat{v}_k^i(s) - \mathrm{val}^i(\hat{Q}_k^i(s))| \to 0$ *as* $k \to \infty$ *for each* $(i, s)$.

*Proof.* Set $i = 1$. Then, the stationary-point inequality says that

$$\hat{v}_k^1(s) = \max_{a\in A^1}\{a^T \hat{Q}_k^1(s)\hat{\pi}_k^2(s)\} \geq \mathrm{val}^1(\hat{Q}_k^1(s)) \geq \min_{a\in A^2}\{\hat{\pi}^1(s)^T \hat{Q}_k^1(s)a\}. \tag{5.26}$$

Since $\bar{Q}_k(s) = \hat{Q}_k^1(s) + \hat{Q}_k^2(s)$, the right-hand side is bounded from below by

$$\begin{aligned}
\min_{a\in A^2}\{\hat{\pi}^1(s)^T \hat{Q}_k^1(s)a\} &\geq \min_{a\in A^2}\{\hat{\pi}^1(s)^T(-\hat{Q}_k^2(s))a\} + \min_{a\in A^2}\{\hat{\pi}^1(s)^T \bar{Q}_k(s)a\} \\
&= -\max_{a\in A^2}\{\hat{\pi}^1(s)^T \hat{Q}_k^2(s)a\} + \min_{a\in A^2}\{\hat{\pi}^1(s)^T \bar{Q}_k(s)a\} \\
&\geq -\max_{a\in A^2}\{\hat{\pi}^1(s)^T \hat{Q}_k^2(s)a\} - \|\bar{Q}_k(s)\|_{\max}. \qquad (5.27)
\end{aligned}$$

Then, the definition of $\hat{v}_k^2(s)$ yields that $\hat{v}_k^1(s) \geq \mathrm{val}^1(\hat{Q}_k^1(s)) \geq -\hat{v}_k^2(s) - \|\bar{Q}_k(s)\|_{\max}$. Since the difference between the first and second terms is bounded from above by the difference between the first and third terms, we have

$$\hat{v}_k^1(s) + \hat{v}_k^2(s) + \|\bar{Q}_k(s)\|_{\max} \geq \hat{v}_k^1(s) - \mathrm{val}^1(\hat{Q}_k^1(s)) \geq 0. \tag{5.28}$$

Since $\bar{v}_k(s) = \hat{v}_k^1(s) + \hat{v}_k^2(s)$, the left-hand side goes to zero by Lemma 5.2 and (5.25). By symmetry, the result can be generalized to $i = 2$, which completes the proof. ∎

Next we introduce the Shapley operator $\mathcal{T}^i$ for each $i = 1, 2$, where

$$(\mathcal{T}^i Q^i)(s, a) := r^i(s, a) + \gamma \sum_{\tilde{s}\in S} p(\tilde{s}|s, a)\mathrm{val}^i(Q^i(\tilde{s})). \tag{5.29}$$

[24] showed that the Shapley operators have a contraction property, i.e.,

$$\max_{(s,a)\in S\times A} |(\mathcal{T}^i Q^i)(s, a) - (\mathcal{T}^i \tilde{Q}^i)(s, a)| \leq \gamma \max_{s\in S} \|Q^i(s) - \tilde{Q}^i(s)\|_{\max}, \tag{5.30}$$

because $\gamma \in (0, 1)$ and

$$|\mathrm{val}^i(Q^i(s)) - \mathrm{val}^i(\tilde{Q}^i(s))| \leq \|Q^i(s) - \tilde{Q}^i(s)\|_{\max} \tag{5.31}$$

for any $Q^i(s)$ and $\tilde{Q}^i(s)$. Therefore, the Shapley operator $\mathcal{T}^i$ has a *unique* fixed point, denoted by $Q_*^i$, corresponding to the $Q$-functions associated with any stationary equilibrium in the underlying zero-sum stochastic game. Furthermore, $Q_*^1(s, a) + Q_*^2(s, a) = 0$ for all $(s, a) \in S \times A$.

At stage $k$ and state $s \in S$, the update of beliefs on $Q$-functions, e.g., (3.3), can be written as

$$\hat{Q}_{k+1}^i(s, a) = (1 - \bar{\beta}_k(s))\hat{Q}_k^i(s, a) + \bar{\beta}_k(s)\left((\mathcal{T}^i \hat{Q}_k^i)(s, a) + \bar{\mathcal{E}}_k^i(s, a)\right), \tag{5.32}$$

where $\bar{\beta}_k(s) = \mathbb{I}_{\{s=s_k\}}\beta_{\#s} \in [0,1]$ and $\bar{\mathcal{E}}_k^i(s,a) := \gamma \sum_{\tilde{s}\in S} p(\tilde{s}|s,a)[\hat{v}_k^i(\tilde{s}) - \mathrm{val}^i(\hat{Q}_k^i(\tilde{s}))]$ is an asymptotically negligible error matrix due to the tracking result in Lemma 5.3. Based on the contraction property of the Shapley operators, the following lemma characterizes the convergence properties of the iteration (5.32) by invoking [22, Theorem 5.1] again.

LEMMA 5.4. $|\hat{Q}_k^i(s,a) - Q_*^i(s,a)| \to 0$ *as* $k \to \infty$ *for each* $(i,s,a)$.

*Proof.* Denote $\widetilde{Q}_k^i := \hat{Q}_k^i - Q_*^i$. If we subtract the fixed point $Q_*^i(s,a)$ from both sides of (5.32), we obtain

$$\widetilde{Q}_{k+1}^i(s,a) = (1-\bar{\beta}_k(s))\widetilde{Q}_k^i(s,a) + \bar{\beta}_k(s)\left((\mathcal{T}^i\hat{Q}_k^i)(s,a) - (\mathcal{T}^i Q_*^i)(s,a) + \bar{\mathcal{E}}_k^i(s,a)\right)$$

since $\mathcal{T}^i Q_*^i = Q_*^i$. Correspondingly, (5.30) yields that

$$\widetilde{Q}_{k+1}^i(s,a) \le (1-\bar{\beta}_k(s))\widetilde{Q}_k^i(s,a) + \bar{\beta}_k(s)\left(\gamma \max_{s'\in S}\|\widetilde{Q}_k^i(s')\|_{\max} + \epsilon_k^i\right), \tag{5.33a}$$

$$\widetilde{Q}_{k+1}^i(s,a) \ge (1-\bar{\beta}_k(s))\widetilde{Q}_k^i(s,a) + \bar{\beta}_k(s)\left(-\gamma \max_{s'\in S}\|\widetilde{Q}_k^i(s')\|_{\max} - \epsilon_k^i\right) \tag{5.33b}$$

for all $a$, where $\epsilon_k := \max_{(s,a)}|\mathcal{E}_k^i(s,a)|$ is an asymptotically negligible error. The proof is completed by invoking [22, Theorem 5.1]. ∎

Based on (5.31), Lemmas 5.3 and 5.4 yield that $|\hat{v}_k^i(s) - \mathrm{val}^i(Q_*^i(s))| \to 0$ as $k \to \infty$ for all $s \in S$ and $i = 1,2$. Therefore the beliefs on the opponent's strategies also converge to an equilibrium of the underlying zero-sum stochastic game. This completes the proof of Theorem 4.3. The proof of Corollary 4.4 is deferred to the extended version [22].

**6. Proof of Theorem 4.6: Convergence for the model-free setting.** The proof of Theorem 4.6 follows similar lines with the proof of Theorem 4.3. However, we face additional challenges such as the players update the beliefs on the $Q$-functions only for the current state and joint action pair, and they explore by taking any action randomly since they do not know their payoff functions and state transition probabilities. In the following, we focus on how to address these challenges.

*Step* (i). *Decoupling the dynamics at the fast timescale.* Similar to (5.2), the update of the belief on $Q$-function at stage $k$ could be written as

$$\hat{Q}_{k+1}^i(s,a) = \hat{Q}_k^i(s,a) + \bar{\alpha}_k(s)\tilde{\mathcal{E}}_k^i(s,a) \tag{6.1}$$

for each $(s,a)$, and the error term $\tilde{\mathcal{E}}_k^i(s,a)$ is now defined by

$$\tilde{\mathcal{E}}_k^i(s,a) = \mathbb{I}_{\{(s,a)=(s_k,a_k)\}}\frac{\beta_{\#(s,a)}}{\alpha_{\#s}}\left(r_k^i + \gamma\hat{v}_k^i(s_{k+1}) - \hat{Q}_k^i(s,a)\right), \tag{6.2}$$

where $s_k, s_{k+1}$ denote the current and next states, and $a_k$ denotes the current action profile. The beliefs on the $Q$-functions remain bounded also in the model-free setting. Particularly, for each $s \in S$, we have

$$\limsup_{k\to\infty}\|\hat{Q}_k^i(s)\|_{\max} \le \frac{1}{1-\gamma}\max_{(s',a')\in S\times A}|r^i(s',a')| =: D^i \tag{6.3}$$

independent of the initialization by $\gamma \in (0,1)$ and Assumptions 4.1 and 4.2. Although we have the ratio $\beta_{\#(s,a)}/\alpha_{\#s}$ in (6.2) instead of $\beta_{\#s}/\alpha_{\#s}$ different from (5.3), the following lemma shows that $\tilde{\mathcal{E}}_k^i(s,a)$ is still asymptotically negligible almost surely.

LEMMA 6.1. *Under Assumptions* 4.1, 4.2, *and* 4.5, *we have* $\tilde{\mathcal{E}}_k^i(s,a) \to 0$ *almost surely for each* $(i,s,a)$ *as* $k \to \infty$.

*Proof.* The proof is provided in the extended version [22]. ∎

*Step* (ii). *Zooming into the local dynamics at the fast timescale.* Lemma 6.1 enables us to decouple dynamics across states as in section 5. Given that player $i$ takes $a_k^i$, (3.6) yields that the update of $\hat{\pi}_k^i$ can be written as

$$\hat{\pi}_{k+1}^i(s) = \hat{\pi}_k^i(s) + \bar{\alpha}_k(s)(\mathbb{E}\{a_k^i\} - \hat{\pi}_k^i + \nu_k^i(s)), \tag{6.4}$$

where the stochastic approximation error induced by the fixed-probability exploration is given by

$$\nu_k^i(s) := a_k^i - \mathbb{E}\{a_k^i\}.$$

Note that the expectation is taken with respect to the random exploration and $\mathbb{E}\{a_k^i\} = a_*^i(1-\epsilon) + \frac{1}{|A^i|}\mathbf{1}\epsilon$, where $\mathbf{1}$ denotes a vector whose entries are all ones with the associated dimension, since $\mathbb{E}\{u^i\} = \frac{1}{|A^i|}\mathbf{1}$ if $u^i \sim \mathcal{U}(A^i)$. By its definition, $\{\nu_k^i(s)\}_{k\geq 0}$ is a square integrable Martingale difference sequence. Since $\sum_{c=0}^{\infty} \alpha_c^2 < \infty$ by Assumption 4.5, the limiting differential inclusion, i.e., the counterpart of (5.4), is now given by

$$\frac{d\boldsymbol{\pi}^i(t)}{dt} = (1-\epsilon)\mathrm{a}^i(t) + \epsilon\bar{u}^i - \boldsymbol{\pi}^i(t), \tag{6.5}$$

due to the exploration and $\bar{u}^i := \frac{1}{|A^i|}\mathbf{1}$ for $i = 1, 2$. To address this, we modify the Lyapunov function (5.12) as follows:

$$\bar{V}(\boldsymbol{\pi}^1(t), \boldsymbol{\pi}^2(t), \tilde{Q}^1, \tilde{Q}^2) := \left[\mathrm{h}^1(t) + \mathrm{h}^2(t) - \lambda\zeta(\tilde{Q}^1, \tilde{Q}^2)\right]_+, \tag{6.6}$$

where $\zeta(\cdot)$ is defined by

$$\zeta(\tilde{Q}^1, \tilde{Q}^2) := (1-\epsilon)\|\tilde{Q}^1 + \tilde{Q}^2\|_{\max} + \epsilon\|\tilde{Q}^1\|_{\max} + \epsilon\|\tilde{Q}^2\|_{\max}, \tag{6.7}$$

and $\mathrm{h}^i(t)$ is as described in (5.6). Note that (6.6) reduces to (5.12) when $\epsilon = 0$, i.e., no experimentation. Its validity can be shown as in Lemma 5.1 if we follow the lines in [9, section 5] but for the dynamics (6.5) instead. For example, we now have

$$\frac{d\mathrm{h}^1(t)}{dt} = \mathrm{a}^1(t)^T\tilde{Q}^1\left((1-\epsilon)\mathrm{a}^2(t) + \epsilon\bar{u}^2 - \boldsymbol{\pi}^2(t)\right), \tag{6.8a}$$

$$\frac{d\mathrm{h}^2(t)}{dt} = \left((1-\epsilon)\mathrm{a}^1(t) + \epsilon\bar{u}^1 - \boldsymbol{\pi}^1(t)\right)^T\tilde{Q}^2\mathrm{a}^2(t) \tag{6.8b}$$

for almost every $t \in [0, \infty)$ and

$$\begin{aligned}\frac{d\mathrm{h}^1}{dt} + \frac{d\mathrm{h}^2}{dt} =& (1-\epsilon)(\mathrm{a}^1)^T(\tilde{Q}^1 + \tilde{Q}^2)\mathrm{a}^2 + \epsilon((\mathrm{a}^1)^T\tilde{Q}^1\bar{u}^2 + (\bar{u}^1)^T\tilde{Q}^2\mathrm{a}^2) - (\mathrm{h}^1 + \mathrm{h}^2)\\ <& \lambda\zeta(\tilde{Q}^1, \tilde{Q}^2) - (\mathrm{h}^1 + \mathrm{h}^2)\end{aligned} \tag{6.9}$$

if $\tilde{Q}^1 + \tilde{Q}^2$ is not equal to the zero matrix. The validity of $\bar{V}(\cdot)$ yields that

$$\lim_{k\to\infty}\left[\bar{v}_k(s) - \lambda\left((1-\epsilon)\|\bar{Q}_k(s)\|_{\max} + \epsilon\sum_{i=1,2}\|\hat{Q}_k^i(s)\|_{\max}\right)\right]_+ = 0, \quad s \in S. \tag{6.10}$$

*Step* (iii). *Zooming out to the global dynamics at the slow timescale.* Let us select the arbitrary $\lambda > 1$ such that $\lambda(1-\epsilon) < 1$ and in addition that $\lambda\gamma < 1$. By (6.10), as a counterpart of (5.22), we have

$$-\|\bar{Q}_k(s)\|_{\max} \leq \bar{v}_k(s) \leq \|\bar{Q}_k(s)\|_{\max} + \bar{\epsilon}_k(s) \quad \forall s \in S, k \geq 0, \tag{6.11}$$

but with an error term satisfying $\limsup_{k\to\infty} |\bar{\epsilon}_k(s)| \leq \lambda\epsilon(D^1 + D^2)$ (and $D^i$ is as described in (6.3)) for each $s \in S$. Furthermore, we need to consider stochastic approximation error terms induced by sampling the underlying state transition probabilities, which are given by

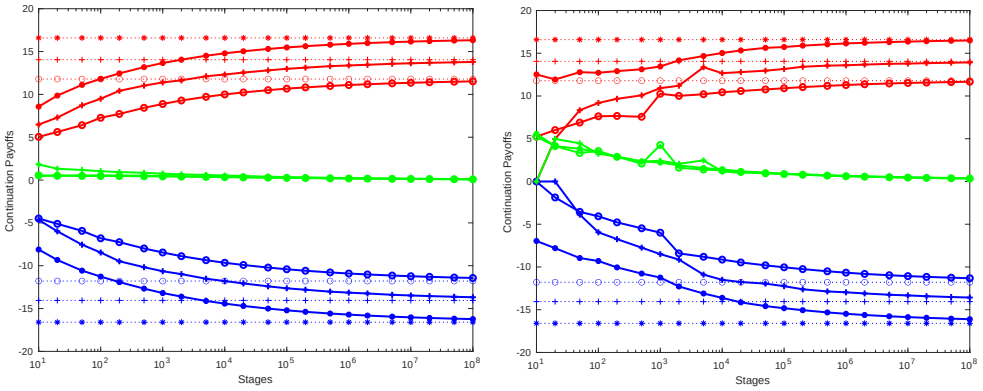
$$\omega_k^i(s,a) := \gamma\hat{v}_k^i(s_{k+1}) - \gamma\sum_{\tilde{s}\in S}\hat{v}_k^i(\tilde{s})p(\tilde{s}|s,a) \tag{6.12}$$

for each $i = 1, 2$, and $\{\omega_k^i(s,a)\}_{k\geq 0}$ is a square integrable Martingale difference sequence by its definition. Note that the sum of approximation errors $\bar{\omega}_k := \omega_k^1 + \omega_k^2$ is also a square integrable Martingale difference sequence. Then, the proof follows after some algebra similar to that in Step (iii) in the proof of Theorem 4.3.

The relevant technical details are deferred to the extended version [22]. This completes the proof of Theorem 4.6.

**7. An illustrative example.** In this section, we examine our fictitious play dynamics numerically in a zero-sum stochastic game whose configuration is selected arbitrarily. For example, there are three states, players have four actions per state, and the discount factor $\gamma = 0.8$. State transition probabilities and stage payoffs are chosen randomly in a way that players can have preferences over the states so that they would face the trade-off between current stage payoff and the continuation payoffs. We set the step sizes as $\alpha_c = 1/(1+c)^{0.51}$ and $\beta_c = 1/(1+c)$ such that they would satisfy both Assumptions 4.2 and 4.5. Furthermore in the model-free setting, players take a random action with probability 0.02 in order to learn the unknown state transition probabilities associated with each action.

In Figures 1(a) and 1(b), we plot the evolution of the continuation payoff estimates of both players for each state in comparison to the equilibrium values, respectively,




(a) Model-based Case (b) Model-free Case

FIG. 1. *Evolution of continuation payoff (or value function) estimates* $\{\hat{v}_{s,k}^1, \hat{v}_{s,k}^2\}_{s\in S}$ *and* $\bar{v}_{s,k} = \hat{v}_{s,k}^1 + \hat{v}_{s,k}^2$, *respectively, converging to positive values, negative values, and zero in* ***red***, ***blue***, *and* ***green***. *Color available in the online version. The red and blue dotted lines denote the actual Nash equilibrium values at each state. The horizontal axis is logarithmic with markers at instances* $10, 20, 50, 100, 200, \ldots$, *in this order.*

in model-based and model-free settings. We also plot the sum of continuation payoffs to observe its expected convergence to zero. As expected, we have observed that the estimates of the continuation payoffs converge to the minimax values of each state in the stochastic game while the convergence is relatively slower and more noisy in the model-free setting.

**8. Concluding remarks.** We presented fictitious play dynamics for stochastic games and analyzed its convergence properties in zero-sum games. In the dynamics presented, players form a belief not only on opponent (stationary) strategy but also on the associated $Q$-functions and update them based on the actions taken by the opponent. The update of beliefs on $Q$-functions evolves at a slower timescale compared to the evolution of beliefs on strategies.

In order to show the convergence of the dynamics, we first approximated the dynamics via a certain differential inclusion at the timescale of the fast update and formulated a novel Lyapunov function for it in order to characterize the limiting behavior of the fast update. Then we used this characterization accompanied with certain contraction arguments at the timescale of the slow update in order to show the almost sure convergence of the dynamics. In particular, we showed that beliefs on strategies and $Q$-functions, respectively, converge to a stationary equilibrium and the corresponding $Q$-functions in the model-based and model-free settings provided that each state is visited infinitely often.

Some of the future research directions include the analyses of this framework (i) in other classes of games, e.g., identical-interest games or zero-sum games with more than two players; (ii) without the conditions on visiting each state infinitely often, e.g., in terms of self-confirming equilibrium, as studied in [8] for learning in extensive-form game; (iii) with function approximation to address computational challenge due to large state and action spaces; and (iv) with nonasymptotic convergence guarantees.

## REFERENCES


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

[2] Y. Bai and C. Jin, *Provable self-play algorithms for competitive reinforcement learning*, in Proceeding of ICML, 2020.

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

[4] V. Bogachev and O. G. Smolyanov, *Real and Functional Analysis*, Springer Nature, New York, 2020.

[5] V. S. Borkar, *Reinforcement learning in Markovian evolutionary games*, Adv. Complex Syst., 5 (2002), pp. 55–72.

[6] R. I. Brafman and M. Tennenholtz, *R-max: A general polynomial time algorithm for near-optimal reinforcement learning*, J. Mach. Learn. Res., 3 (2002), pp. 213–231.

[7] J. C. Ely and O. Yilankaya, *Nash equilibrium and the evolution of preferences*, J. Econom. Theory, 97 (2001), pp. 255–272.

[8] D. Fudenberg and D. Kreps, *Learning in extensive-form games* I. *Self-confirming equilibria*, Games Econom. Behav., 8 (1995), pp. 20–55.

[9] C. Harris, *On the rate of convergence of continuous-time fictitious play*, Games Econom. Behav., 22 (1998), pp. 238–259.

[10] S. Hart and A. Mas-Colell, *Uncoupled dynamics cannot lead to Nash equilibrium*, Amer. Econom. Rev., 93 (2003), pp. 1830–1836.

[11] T. Lattimore and C. Szepesvári, *Bandit Algorithms*, Cambridge University Press, Cambridge, UK, 2020.

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

[13] D. S. Leslie and E. J. Collins, *Generalized weakened fictitious play*, Games Econom. Behav., 56 (2006), pp. 285–298.

[14] D. S. Leslie, S. Perkins, and Z. Xu, *Best-response dynamics in zero-sum stochastic games*, J. Econom. Theory, 189 (2020).
[15] M. L. Littman, *Markov games as a framework for multi-agent reinforcement learning*, in Proceeding ICML, 1994.
[16] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, *Payoff-based dynamics for multiplayer weakly acyclic games*, SIAM J. Control Optim., 48 (2009), pp. 373–396.
[17] K. Miyasawa, *On the Convergence of the Learning Process in a 2x2 Non-Zero-Sum Game*, Research Memorandum, 33, Economic Research Program, Princeton University, 1961.
[18] D. Monderer and A. Sela, *A 2x2 game without the fictitious play property*, Games Econom. Behav., 14 (1996), pp. 144–148.
[19] D. Monderer and L. Shapley, *Fictitious play property for games with identical interests*, J. Econom. Theory, 68 (1996), pp. 258–265.
[20] J. Robinson, *An iterative method of solving a game*, Ann. of Math., 24 (1951), pp. 296–301.
[21] W. H. Sandholm, *Preference evolution, two-speed dynamics, and rapid social change*, Rev. Econom. Dyn., 4 (2001), pp. 637–679.
[22] M. O. Sayin, F. Parise, and A. Ozdaglar, *Fictitious Play in Zero-Sum Stochastic Games*, arXiv:2010.04223, 2020.
[23] G. Schoenmakers, J. Flesch, and F. Thuijsman, *Fictitious play in stochastic games*, Math. Methods Oper. Res., 66 (2007), pp. 315–325.
[24] L. S. Shapley, *Stochastic games*, Proc. Natl. Acad. Sci. USA, 39 (1953), pp. 1095–1100.
[25] B. Swenson, S. Kar, J. Xavier, and D. S. Leslie, *Robustness properties in fictitious-play-type algorithms*, SIAM J. Control Optim., 55 (2017), pp. 3295–3318.
[26] B. Swenson, R. Murray, and S. Kar, *On best-response dynamics in potential games*, SIAM J. Control Optim., 56 (2018), pp. 2734–2767.
[27] G. Tesauro, *Extending Q-learning to general adaptive multi-agent systems*, in Proceedings of NIPS, 2003.
[28] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Mach. Learn., 16 (1994), pp. 185–202.
[29] G. Vigeral, *Evolution equations in discrete and continuous time for nonexpansive operators in Banach spaces*, ESAIM Control Optim. Calc., 16 (2010), pp. 809–832.
[30] O. J. Vrieze and S. H. Tijs, *Fictitious play applied to sequences of games and discounted stochastic games*, Internat. J. Game Theory, 11 (1982), pp. 71–85.
[31] C. J. C. H. Watkins and P. Dayan, *Q-learning*, Mach. Learn., 8 (1992), pp. 279–292.
[32] C. Wei, Y. Hong, and C. Lu, *Online reinforcement learning in stochastic games*, in Proceedings of NIPS, 2017.
[33] K. Zhang, Z. Yang, and T. Başar, *Multi-agent reinforcement learning: A selective overview of theories and algorithms*, Handb. Rein. Learn. Cont., 325 (2021), pp. 321–384.