# INDEPENDENT LEARNING IN STOCHASTIC GAMES

ASUMAN OZDAGLAR, MUHAMMED O. SAYIN, AND KAIQING ZHANG

## ABSTRACT

Reinforcement learning (RL) has recently achieved tremendous successes in many artificial intelligence applications. Many of the forefront applications of RL involve *multiple agents*, e.g., playing chess and Go games, autonomous driving, and robotics. Unfortunately, the framework upon which classical RL builds is inappropriate for multiagent learning, as it assumes an agent's environment is stationary and does not take into account the adaptivity of other agents. In this review paper, we present the model of *stochastic games* [69] for multiagent learning in *dynamic* environments. We focus on the development of *simple* and *independent* learning dynamics for stochastic games: each agent is myopic and chooses best-response type actions to other agents' strategy without any coordination with her opponent. There has been limited progress on developing convergent best-response type independent learning dynamics for stochastic games. We present our recently proposed simple and independent learning dynamics that guarantee convergence in zero-sum stochastic games, together with a review of other contemporaneous algorithms for dynamic multiagent learning in this setting. Along the way, we also reexamine some classical results from both the game theory and RL literature, to situate both the conceptual contributions of our independent learning dynamics, and the mathematical novelties of our analysis. We hope this review paper serves as an impetus for the resurgence of studying independent and natural learning dynamics in game theory, for the more challenging settings with a dynamic environment.

## MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 91A15; Secondary 91A25, 68T05

## KEYWORDS

Stochastic games, learning in games, reinforcement learning

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14


Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5340–5373
DOI 10.4171/ICM2022/152

## 1. INTRODUCTION

Reinforcement learning (RL) in which autonomous agents make decisions in unknown dynamic environments has emerged as the backbone of many artificial intelligence (AI) problems. The frontier of many AI systems emerges in *multiagent* settings, including playing games such as chess and Go [73,74], robotic manipulation with multiple connected arms [30], autonomous vehicle control in dynamic traffic and automated warehouses or production facilities [68,86]. Further advances in these problems critically depend on developing stable and agent incentive-compatible learning dynamics in multiagent environment. Unfortunately, the mathematical framework upon which classical RL depends on is inadequate for multiagent learning, since it assumes an agent's environment is stationary and does not contain any adaptive agents.

The topic of multiagent learning has a long history in game theory, almost as long as the discipline itself. One of the most studied models of learning in games is *fictitious play*, introduced by Brown [14], with first rigorous convergence analysis presented by Robinson [59] for its discrete-time variant and for finite two-player zero-sum games. See also [27,28,33,49,51,70] and others for the analysis of fictitious play. In fictitious play, each agent is myopic (i.e., she does not take into account the fact that her current action will have an impact on the future actions of other players[1]), and therefore chooses a best response to the opponent's strategy, which she estimates to be the empirical distribution of past play. Despite extensive study on learning in repeated play of static complete-information games (also referred to as strategic- or normal-form games) and the importance of the issues, there is limited progress on multiagent learning in dynamic environments (where the environments evolve over time). The key challenge is to estimate the decision rules of other agents that in turn *adapt* their behavior to changing nonstationary environments.

In this review paper, we first present stochastic games, first introduced in [69], as a model for representing dynamic multiagent interactions in Section 3.[2] Stochastic games extend strategic-form games to dynamic settings where the environment changes with players' decisions. They also extend single-agent Markov decision problems (Markov decision processes) to competitive situations with more than one decision-maker. Developing simple and independent learning rules, e.g., the fictitious-play/best-response type dynamics, for stochastic games has been an open question for some time in the literature (see [19,20,78] for some negative nonconvergent results due to nonstationarity).

In the second part of the paper in Section 4, we present recently proposed simple and independent learning rules from [63,64], and show their convergence for zero-sum stochastic games. Crucially, these rules are based on fictitious play-type dynamics and, unlike earlier works, do not require coordination between agents, leading to fully decentralized and independent multiagent learning dynamics. We combine ideas from *g*ame theory and *R*L in developing these learning rules, and consider three different settings: *model-based* setting

1 Hereafter, we use *player* and *agent* interchangeably.

2 The preliminary information on strategic-form games and learning in strategic-form games with repeated play are provided in Section 2.

where players know their payoff functions, transition probabilities of the underlying stochastic games, and observe opponent's actions; *model-free* setting where players do not know payoff functions and transition probabilities but can still observe the opponent's actions; and the *minimal information* setting where players do not even observe opponent's actions. In all three settings, the players do not know the opponent's objective, i.e., they do not possess the knowledge that the underlying game is zero-sum. In the minimal-information setting, the players may not even know the existence of an opponent.

In Section 5, we have also reviewed several other algorithms/learning dynamics, and their convergence results for multiagent learning in stochastic games. We cover both results from the game theory literature that typically assumes knowledge of the model of the players' payoff functions, and the transition probabilities of the underlying stochastic games, and also from the RL literature which posit learning dynamics that perform updates without knowing the transition probabilities. Most of these update rules typically involve coordination and computationally intensive steps for the players. These algorithms can be viewed more as ones for *computing* the Nash equilibrium of the stochastic games, as opposed to natural learning dynamics that would be adopted by self-interested agents interested in maximizing their own payoffs given their inferences (as captured in our learning dynamics). Finally, we conclude the paper with open questions on independent learning in stochastic games in Section 6.

## 2. PRELIMINARIES: STRATEGIC-FORM GAMES

A two-player strategic-form game can be characterized by a tuple $\langle A^1, A^2, r^1, r^2 \rangle$, in which

- the *finite* set of actions that player $i$ can take is denoted by $A^i$,
- the *payoff function* of player $i$ is denoted by $r^i : A \to \mathbb{R}$, where $A := A^1 \times A^2$.[3]

Each player $i$ takes an action from her action set $A^i$ *simultaneously* and receives the payoff $r^i(a^1, a^2)$.

We let players choose a mixed strategy to randomize their actions independently. For example, $\pi^i : A^i \to [0, 1]$ denotes the mixed strategy of player $i$ such that $\pi^i(a^i)$ corresponds to the probability that player $i$ plays $a^i$. Note that we have $\sum_{a^i \in A^i} \pi^i(a^i) = 1$ by its definition.

We represent the strategy profile and action profile of the players by $\pi = (\pi^1, \pi^2)$ and $a = (a^1, a^2)$, respectively. Under the strategy profile $\pi$, the expected payoff of player $i$ is defined by

$$U^i(\pi) := \mathbb{E}_{a \sim \pi}\{r^i(a)\}.$$

3 We can generalize the definition to arbitrary number of players in a rather straightforward way.

Note that the expected payoff of player $i$ is affected by the strategy of the opponent. We next introduce the Nash equilibrium where players do not have any (or large enough) incentive to change their strategies unilaterally.

**Definition 2.1** ($(\varepsilon$-)Nash equilibrium). A strategy profile $\pi_*$ is a mixed-strategy $\varepsilon$-Nash equilibrium with $\varepsilon \geq 0$ if we have

$$U^1\left(\pi_*^1, \pi_*^2\right) \geq U^1\left(\pi^1, \pi_*^2\right) - \varepsilon, \quad \text{for all } \pi^1, \tag{2.1a}$$

$$U^2\left(\pi_*^1, \pi_*^2\right) \geq U^2\left(\pi_*^1, \pi^2\right) - \varepsilon, \quad \text{for all } \pi^2. \tag{2.1b}$$

Furthermore, $\pi_*$ is a mixed-strategy Nash equilibrium if (2.1) holds with $\varepsilon = 0$.

The following is the classical existence result for any strategic-form game (e.g., see [4, THEOREM 3.2]).

**Theorem 2.2** (Existence of an equilibrium in strategic-form games). *In strategic-form games (with finitely many players and finitely many actions), a mixed-strategy equilibrium always exists.*

The key question is whether an equilibrium can be realized or not in the interaction of self-interested decision-makers. In general, finding the best strategy against another decision-maker is not a well-defined optimization problem because the best strategy that reflects the viewpoint of the individual depends on the opponent's strategy. Therefore, players are generally not able to compute their best strategy beforehand. When there exists a unique equilibrium, we can expect the players to identify their equilibrium strategies as a result of an introspective thinking process. For example, what would the opponent choose? What would the opponent have chosen if she knew I am considering what she would pick while choosing my strategy? And so on. However, many empirical analyses suggest that an equilibrium would not typically be realized in one shot even with such reasoning (see, e.g., [29]).

It is instructive to consider the following well-known example: Consider a game played among $n > 1$ students. The teacher asks the students to pick a number between 0 and 100, and submit it within a closed envelope. The winner will be the one who chooses the number closest to the two-thirds of the average of all numbers picked. It can be seen that the unique equilibrium is the strategy profile where every player chooses 0. We would expect the students to pick 0 as a result of an introspective thinking process, however, empirical studies show that they typically pick numbers other than zero such that their average ends up around 30, with its two-thirds around 20 [53]. This results in players who have selected 0 by strategizing their actions introspectively losing the game. However, if the game is played repeatedly with players observing chosen actions, each player will have a tendency to pick numbers closer to the winning number (or its two-thirds if they notice that others can also have such a tendency to pick the number closest to the winning one). This results in convergence to the equilibrium play along repeated play of the game, even when the players have not engaged in any forward-looking strategy.

Many games have multiple equilibria which makes coordination and selection through introspective thinking challenging. On the other hand, empirical studies suggest even in strategic situations equipped with multiple equilibria, individual agents reach an equilibrium as long as they *engage with each other multiple times and receive feedback* to revise their strategies [29].

In the following, we review the canonical models of learning with multiple agents through repeated interactions.

### 2.1. Learning in strategic-form games with repeated play

Suppose that players know the primitives of the game, i.e., $\langle A^1, A^2, r^1, r^2\rangle$. If players knew the opponent's strategy, computation of the best strategy is a simple optimization problem where they pick one of the maxima among linearly ordered finitely many elements. However, players do not know the opponent's strategy. When they play the same game repeatedly and observe the opponent's actions in these games, they have a chance to reason about what the opponent would play in the next repetition of the game. Therefore, they can estimate the opponent's strategy based on the history of the play. However, the opponent is not necessarily playing according to a stationary strategy since she is also a strategic decision-maker who can adapt her strategy according to her best interest.

*Fictitious play* is a simple and stylist learning dynamic where players (erroneously) assume that the opponent plays according to a stationary strategy.[4] This assumption lets players form a belief on the opponent's strategy based on the history of the play, e.g., the empirical distribution of the actions taken. Then, the players can adapt their strategies based on the belief constructed.

Fictitious play, since its first introduction by [14], has become the most appealing best-response type learning dynamics in game theory. Formally, at iteration $k$, player $i$ maintains a *belief* on the opponent's strategy, denoted by $\hat{\pi}_k^{-i} \in \Delta(A^{-i})$.[5] For example, the belief can correspond to the empirical average of the actions taken in the past. Note that we can view an action $a^i$ as a deterministic strategy in which the action is played with probability 1, i.e., $a^i \in \Delta(A^i)$ with slight abuse of notation. Then, the empirical average is given by

$$\hat{\pi}_{k+1}^{-i} = \frac{1}{k+1}\sum_{\kappa=0}^{k} a_\kappa^{-i}. \tag{2.2}$$

The belief $\hat{\pi}_{k+1}^{-i}$ can be computed iteratively using bounded memory according to

$$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k+1}\cdot\left(a_k^{-i} - \hat{\pi}_k^{-i}\right), \tag{2.3}$$

with arbitrary initialization $\hat{\pi}_0^{-i} \in \Delta(A^{-i})$. In other words, players do not have to remember every action taken by the opponent in the past. Moreover, player $i$ selects her action following

$$a_k^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim\hat{\pi}_k^{-i}}\left\{r^i(a^1, a^2)\right\}, \tag{2.4}$$

---

4 It is called *fictitious play* because [14] introduced it as an introspective thinking process that a player can play by herself.

5 We represent the probability simplex over a set $A$ by $\Delta(A)$.

with an arbitrary tie-breaking rule, playing a greedy best-response to the belief she maintains on opponent's strategy.

We say that fictitious play dynamics *converge to an equilibrium* if beliefs formed converge to a Nash equilibrium when all players follow the fictitious play dynamics (2.3)–(2.4). We also say that a class of games has *fictitious play property* if fictitious play converges in every game of that class. The following theorem is about two important classes of games from two extremes of the game spectrum: two-player zero-sum strategic-form games, where $r^1(a) + r^2(a) = 0$ for all $a \in A$, and $n$-player identical-interest strategic-form games, where there exists a common payoff function $r : A \to \mathbb{R}$ such that $r^i(a) = r(a)$ for all $a \in A$ and for each player $i$.

**Theorem 2.3** (Fictitious play property of zero-sum and identical-interest games).

- *The two-player zero-sum strategic-form games have fictitious play property* **[59]**.
- *The n-player identical-interest strategic-form games have fictitious play property* **[51]**.

As an alternative to the insightful proofs in **[59]** and **[51]**, we can establish a connection between fictitious play and continuous-time best response dynamics to characterize its convergence properties. For example, **[31]** provided a proof for the continuous-time best-response dynamics in zero-sum strategic-form games through a Lyapunov function formulation. This convergence result also implies the convergence of fictitious play in repeated play of the same zero-sum strategic-form game. We next briefly describe the approach in **[31]** to convergence analysis for continuous-time best-response dynamics.

In continuous-time best response dynamics, the strategies $(\pi^1, \pi^2)$ evolve according to the following differential inclusion:

$$\frac{d\pi^i}{dt} + \pi^i \in \underset{a^i \in A^i}{\operatorname{argmax}} \mathbb{E}_{a^{-i} \sim \pi^{-i}} \left\{ r^i(a^1, a^2) \right\} \tag{2.5}$$

for $i = 1, 2$. We highlight the resemblance between (2.3) and (2.5) because we can view (2.5) as the limiting flow of (2.3) as $1/(k+1) \to 0$. Note also that there exists an absolutely continuous solution to this differential inclusion **[31]**. To characterize the convergence properties of this flow, **[31]** showed that the function

$$V(\pi) = \sum_{i=1,2} \left( \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}} \left\{ r^i(a^1, a^2) \right\} - \mathbb{E}_{a \sim \pi} \left\{ r^i(a) \right\} \right) \tag{2.6}$$

is a Lyapunov function when $r^1(a) + r^2(a) = 0$ for all $a \in A$.[6] This yields that $V(\pi(t)) \geq V(\pi(t'))$ for all $t' > t$ and $V(\pi(t)) > V(\pi(t'))$ if $V(\pi(t)) > 0$. Correspondingly, we have $V(\pi(t)) \to 0$ as $t \to \infty$. This implies that the continuous-time best response dynamics converge to the equilibrium of the zero-sum game. Since the terms in parentheses

6 Note that $\mathbb{E}_{a \sim \pi}\{r^1(a)\} + \mathbb{E}_{a \sim \pi}\{r^2(a)\} = 0$ when $r^1(a) + r^2(a) = 0$ for all $a \in A$.

in (2.6) are nonnegative, $V(\pi) = 0$ yields that they are equal to zero for each $i = 1, 2$, which is indeed the definition of the Nash equilibrium.

Generally, the convergence of the limiting flow would not lead to the convergence of the discrete-time update. However, based on tools from differential inclusion approximation theory [5], the existence of such a Lyapunov function yields that the fictitious play dynamics converge to an equilibrium since its linear interpolation after certain transformation of the time axis can be viewed as a perturbed solution to the differential inclusion (2.5) with asymptotically negligible perturbation while the existence of Lyapunov function yields that any such perturbed solution also converges to the zero-set of the Lyapunov function, i.e., $\{\pi : V(\pi) = 0\}$.

The fictitious play dynamics enjoy the following desired properties [29]: (i) The dynamics do not require knowledge of the underlying game's class, e.g., the opponent's payoff function, and is not specific to any specific class of games; (ii) Players attain the best-response performance against an opponent following an asymptotically stationary strategy, i.e., the learning dynamics is rational; (iii) If the dynamics converge, it must converge to an equilibrium of the underlying game.

Unfortunately, there exist strategic-form games that do not have fictitious play property as shown by [70] through a counterexample. The classes of strategic-form games with fictitious play property have been studied extensively, e.g., see [6, 7, 48–52, 59, 65]. Variants of fictitious play, including smoothed fictitious play [27] and weakened fictitious play [81] have also been studied extensively. However, all these studies focus on the repeated play of *the same* strategic-form game at every stage. There are very limited results on dynamic games where players interact repeatedly while the game played at a stage (called *stage-game*) evolves with their actions. Note that players need to consider the impact of their actions in their future payoffs as in dynamic programming or optimal control when they have utilities defined over infinite horizon.

In the next section, we introduce stochastic games, a special (and important class) of dynamic games where the stage-games evolve over infinite horizon based on the current actions of players.

## 3. STOCHASTIC GAMES

Stochastic games (also known as *Markov* games), since their first introduction by Shapley [69], have been widely used as a canonical model for dynamic multiagent interactions (e.g., see the surveys [16, 89]). At each time $k = 0, 1, \ldots$, players play a stage game that corresponds to a particular state of a multistate environment. The stage games evolve stochastically according to the transition probabilities of the states controlled jointly by the actions of both players. The players receive a payoff which is some aggregate of the stage payoffs; a typical model is to assume the players receive a discounted sum of stage payoffs over an infinite horizon.

Formally, a two-player stochastic game is characterized by a tuple $\langle S, A^1, A^2, r^1, r^2, p, \gamma \rangle$, in which:

- The *finite* set of states is denoted by $S$.
- The *finite* set of actions that player $i$ can take at any state is denoted by $A^i$.[7]
- The *stage payoff function* of player $i$ is denoted by $r^i : S \times A \to \mathbb{R}$, where $A = A^1 \times A^2$.
- For any pair of states $(s, s')$ and action profile $a \in A$, we define $p(s'|s, a)$ as the *transition probability* from $s$ to $s'$ given action profile $a$.
- The players also discount the impact of future payoff in their utility with the discount factor $\gamma \in [0, 1)$.

The objective of player $i$ is to maximize the expected sum of discounted stage-payoffs collected over infinite horizon, given by

$$\mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k)\right\}, \tag{3.1}$$

where $a_k \in A$ denotes the action profile played at stage $k$, $\{s_0 \sim p_o, s_{k+1} \sim p(\cdot \mid s_k, a_k), k \geq 0\}$ is a stochastic process representing the state at each stage $k$ and $p_o \in \Delta(S)$ is the initial state distribution. The expectation is taken with respect to randomness due to stochastic state transitions and actions mixed independently by the players.

The players can play an *infinite* sequence of (mixed) actions. When they have perfect recall, they can mix their actions independently according to a *behavioral strategy* in which the probability of an action is taken depends on the history of states and action profiles, e.g., $h_k = \{s_0, a_0, s_1, a_1, \ldots, s_{k-1}, a_{k-1}, s_k\}$ at stage $k$. This results in an infinite-dimensional strategy space, and therefore, the universal result for the existence of an equilibrium, Theorem 2.2, does not apply here. On the other hand, stochastic games can also be viewed as a generalization of Markov decision processes (MDPs) to multiagent cases since state transition probabilities depend only on the current state and current action profile of players. Behavioral strategies that depend only on the final state of the history (which corresponds to the current state) are known as *Markov strategies*. Furthermore, we call a Markov strategy by a *stationary strategy* if it does not depend on the stage, e.g., see [71, SECTION 6.2]. In (discounted) MDPs, there always exists an optimal strategy that is stationary, e.g., see [23]. Shapley [69] showed that this can be generalized to two-player zero-sum stochastic games.

We denote the stationary mixed strategy of player $i$ by $\pi^i : S \to \Delta(A^i)$, implying that she takes actions according to the mixed strategy specific to state $s$, i.e., $\pi^i(s) \in \Delta(A^i)$. We represent the strategy profile of players by $\pi := \{\pi^1, \pi^2\}$. Correspondingly, the expected discounted sum of stage payoffs of player $i$ under the strategy profile $\pi$ is defined by

$$U^i(\pi) := \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k)\right\}, \tag{3.2}$$

7 The formulation can be generalized to the case where the action spaces depend on state in a rather straightforward way.

where $a_k \sim \pi(s_k)$, and the expectation is taken with respect to the all randomness. We next introduce the Nash equilibrium (more specifically, Markov perfect equilibrium [44,45]) where players do not gain any utility improvement by unilateral changes in their *stationary* strategies regardless of the initial state, e.g., see [71, SECTION 6.2].

**Definition 3.1** (Stationary ($\varepsilon$-)Nash equilibrium). We say that a stationary strategy profile $\pi$ is a stationary mixed-strategy $\varepsilon$-Nash equilibrium with $\varepsilon \geq 0$ if we have

$$U^1(\pi^1, \pi^2) \geq U^1(\bar{\pi}^1, \pi^2) - \varepsilon \quad \text{for all } \bar{\pi}^1, \tag{3.3a}$$

$$U^2(\pi^1, \pi^2) \geq U^2(\pi^1, \bar{\pi}^2) - \varepsilon \quad \text{for all } \bar{\pi}^2. \tag{3.3b}$$

We say that $\pi$ is a stationary mixed-strategy Nash equilibrium if (3.3) holds with $\varepsilon = 0$.

We next state an important existence result for discounted stochastic games.

**Theorem 3.2** (Existence of a stationary equilibrium in stochastic games [24]). *In stochastic games (with finitely many players, states, and actions, and discount factor $\gamma \in [0, 1)$), a stationary mixed-strategy equilibrium always exists.*

The proof for two-player zero-sum stochastic games is shown by Shapley [69] while its generalization to $n$-player general-sum stochastic games is proven by Fink [24] and Takahashi [77] concurrently. Shapley [69] had also presented an iterative algorithm to compute the unique equilibrium value of a two-player zero-sum stochastic game. To describe the algorithm, let us first note that in a zero-sum strategic-form game, there always exists a unique equilibrium *value* for the players (though there may exist multiple equilibria). For example, given a zero-sum strategic-form game $\langle A^1, A^2, u^1, u^2 \rangle$, we denote the equilibrium values of player 1 and player 2, respectively, by

$$\operatorname{val}^1[u^1] = \max_{\pi^1 \in \Delta(A^1)} \min_{\pi^2 \in \Delta(A^2)} \mathbb{E}_{a \sim (\pi^1, \pi^2)}\{u^1(a)\}, \tag{3.4}$$

$$\operatorname{val}^2[u^2] = \max_{\pi^2 \in \Delta(A^2)} \min_{\pi^1 \in \Delta(A^1)} \mathbb{E}_{a \sim (\pi^1, \pi^2)}\{u^2(a)\}. \tag{3.5}$$

It is instructive to examine the following thought experiment. Imagine that players are at the edge of the infinite horizon. Then the players' continuation payoff would be determined by the stage game at state $s$ since there would not be any future stages to consider. The unique equilibrium values they would get would be $\operatorname{val}^i[r^i(s, \cdot)]$. Then, at the stage just before the last one, they would have played the strategic-form game $\langle A^1, A^2, Q^1(s, \cdot), Q^2(s, \cdot) \rangle$ at state $s$, where

$$Q^i(s, \cdot) = r^i(s, \cdot) + \gamma \sum_{s' \in S} p(s'|s, \cdot) \operatorname{val}^i[r^i(s', \cdot)]. \tag{3.6}$$

Shapley [69] showed that if we follow this *backward induction*, we can always compute the equilibrium values associated with a stationary equilibrium. To this end, he introduced the operator $\mathcal{T}^i$ defined by

$$(\mathcal{T}^i v^i)(s) := \operatorname{val}^i\left[ r^i(s, \cdot) + \gamma \sum_{s' \in S} p(s|s, \cdot) v^i(s') \right], \quad \forall s \in S, \tag{3.7}$$

which is a contraction with respect to the $\ell_\infty$-norm when $\gamma \in (0,1)$ since $\mathrm{val}^i$ is a nonexpansive mapping, i.e.,

$$\left|\mathrm{val}^i(u^i) - \mathrm{val}^i(\tilde{u}^i)\right| \leq \max_{a \in A} \left|u^i(a) - \tilde{u}^i(a)\right|,$$

for any $u^i : A \to \mathbb{R}$ and $\tilde{u}^i : A \to \mathbb{R}$, similar to the maximum function in the Bellman operator. Therefore, the iteration

$$v^i_{(n+1)} = \mathcal{T}^i v^i_{(n)}, \quad \forall n \geq 0, \tag{3.8}$$

starting from arbitrary $v^i_{(0)}$ converges to the unique fixed point of the operator. Further inspection of the fixed point reveals that it is indeed the equilibrium values of states associated with some stationary equilibrium of the underlying two-player zero-sum stochastic game. There does not exist a counterpart of this iteration for the computation of equilibrium values in general-sum stochastic games, since the value of a game is not uniquely defined for general-sum stochastic games, and involves a fixed point operation, which is hard to compute at each stage of an algorithm. However, Shapley's iteration is still a powerful method to compute equilibrium values in a two-player zero-sum stochastic game.

In the following section, we examine whether a stationary equilibrium would be realized as a consequence of nonequilibrium adaptation of learning agents as in Section 2.1 but now for stochastic games instead of repeated play of the same strategic-form game.

## 4. LEARNING IN STOCHASTIC GAMES

Fictitious play dynamics is a best-response type learning dynamics where each player aims to take the best response against the opponent by learning the opponent's strategy based on the history of the play. This stylist learning dynamic can be generalized to stochastic games as players (again erroneously) assume that the opponent plays according to a *stationary* strategy (which depends only on the current state). Hence, they can again form a belief on the opponent's stationary strategy based on the history of the play. Particularly, they can form a belief on the opponent's mixed strategy specific to a state based on the actions taken at that state only due to the stationarity assumption on the opponent's strategy. Given that belief on the opponent's strategy, players can also compute the value of each state-action pair based on *backward induction* since their actions determine both the stage payoff and the continuation payoff by determining the state transitions. Therefore, they essentially play an *auxiliary stage-game* at each stage specific to the current state, which can be represented by $\mathcal{G}_s := \langle A^1, A^2, Q^1(s,\cdot), Q^2(s,\cdot)\rangle$, where the payoff or the *Q-function*, $Q^i(s,\cdot) : A \to \mathbb{R}$ is determined according to the backward induction given $\pi^{-i}$ the belief of player $i$ about player $-i$'s strategy, and therefore, it satisfies the following fixed-point equation:

$$Q^i(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p\big(s'|s,a\big) \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}(s')}\big\{Q^i(s', a^1, a^2)\big\}. \tag{4.1}$$

For notational convenience, we also define the *value function* $v^i : S \to \mathbb{R}$ by

$$v^i(s) := \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}(s)}\big\{Q^i(s, a^1, a^2)\big\}. \tag{4.2}$$

At each stage $k$, player $i$ has a belief on player $-i$'s strategy, which we denote by $\hat{\pi}_k^{-i}$. Player $i$ also forms a belief on the payoff function for the auxiliary game, or the $Q$-function, denoted by $\hat{Q}_k^i$. Let $s$ be the current state of the stochastic game. Then, player $i$ selects her action $a_k^i$ according to

$$a_k^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \hat{\pi}_k^{-i}(s)} \{\hat{Q}_k^i(s, a^1, a^2)\}. \tag{4.3}$$

Observing the opponent's action $a_k^{-i}$, player $i$ forms her belief on player $-i$'s strategy for the current state $s$ as a weighted empirical average, which can be constructed iteratively as

$$\hat{\pi}_{k+1}^{-i}(s) = \hat{\pi}_k^{-i}(s) + \alpha_{c_k(s)}\big(a_k^{-i} - \hat{\pi}_k^{-i}(s)\big). \tag{4.4}$$

Here $\alpha_c \in [0, 1]$ is a step size and it vanishes with $c_k(s)$ indicating the number of visits to state $s$ rather than time. Note that if there was a single state, $c_k(s)$ would correspond to the time, i.e., $c_k(s) = k$, as in the classical fictitious play. The update (4.4) can also be viewed as taking a convex combination of the current belief $\hat{\pi}_k^2(s)$ and the observed action $a_k^2$ while the step size $\alpha_{c_k(s)}$ is the (vanishing) weight of the action observed. Vanishing step size as a function of the number of visits implies that, the players give less weight to their current belief than the observed action by using a large step size if that state has not been visited many times. This means that the players will still give less weight to their current belief even at later stages if the specific state has not been visited many times, and indicating, they have not been able to strengthen their belief enough to rely more on it.

Simultaneously, player $i$ updates her belief on her own $Q$-function for the current state $s$ according to

$$\hat{Q}_{k+1}^i(s,a) = \hat{Q}_k^i(s,a) + \beta_{c_k(s)}\left(r^i(s,a) + \gamma \sum_{s' \in S} p\big(s'|s,a\big)\hat{v}_k^i(s') - \hat{Q}_k^i(s,a)\right), \quad \forall a \in A, \tag{4.5}$$

where we define $\hat{v}_k^i : S \to \mathbb{R}$ as the value function estimate given by

$$\hat{v}_k^i(s) = \max_{a^i} \mathbb{E}_{a_k^{-i} \sim \hat{\pi}_k^{-i}(s)} \{\hat{Q}_k^i(s, a^1, a^2)\}, \tag{4.6}$$

and $\beta_c \in [0, 1]$ is another step size that also vanishes with $c_k(s)$. Similar to (4.4), the update of the belief on the $Q$-function (4.5) can be viewed as a convex combination of the current belief $\hat{Q}_k^i(s,a)$ and the new observation $r^i(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)\hat{v}_k^i(s')$. Such vanishing step size again implies that the players are relying on their beliefs more if they have had many chances to strengthen them.

The *key feature* of this learning dynamic is that the players update their beliefs on their $Q$-functions at a slower timescale than the update of their beliefs on the opponent strategy. This is consistent with the literature on evolutionary game theory [22, 62] (which postulates players' choices to be more dynamic than changes in their preferences) since we can view $Q$-functions in auxiliary games as slowly evolving player preferences. Particularly, the two-timescale learning framework implies that the players take smaller and smaller steps at (4.5) than the steps at (4.4) such that the ratio of the step sizes, $\beta_c/\alpha_c$, goes to zero with the number of visits to the associated state. Note that this implies that $\beta_c$ goes to zero faster

than $\alpha_c$ does, implying slower update of the $Q$-function estimate compared to the opponent's strategy estimate. This weakens the dependence between evolving beliefs on opponent strategy and $Q$-function.

We say that this *two-timescale fictitious play dynamics converge to an equilibrium* if beliefs on opponent strategies converge to a Nash equilibrium which associates with the auxiliary games while the beliefs on $Q$-functions converge to the $Q$-functions for a stationary equilibrium of the underlying stochastic game. Particularly, given an equilibrium $\pi_*$, the associated $Q$-function of player $i$ satisfies

$$Q^i(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p\left(s'|s,a\right) \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi_*^{-i}(s)} \left\{ Q^i(s', a^1, a^2) \right\}, \quad \forall (s,a) \in S \times A.$$

Recall that players are playing a dynamically evolving auxiliary game at each state repeatedly, but update their beliefs on the $Q$-functions and opponent strategies only when that state is visited. Therefore, the players are updating their beliefs on the opponent strategy and $Q$-function specific to that state only during these visits. Hence, we make the following assumption ensuring that players have sufficient time to revise and improve their beliefs specific to a state.

**Assumption 4.1** (Markov chain). Each state is visited infinitely often.

Stochastic games reduce to the repeated play of the same strategic-form game if there exists only one state and the discount factor is zero. Correspondingly, Assumption 4.1 always holds in such a case. However, when there are multiple states, Assumption 4.1 does not necessarily hold, e.g., since some states can be absorbing by preventing transitions to others. In the following, we exemplify four Markov chain configurations with different generality:

- *Case* (i) The probability of transition between any pair of states is positive for *any* action profile. This condition is also known as *irreducible stochastic games* [40].
- *Case* (ii) The probability of transition between any pair of states is positive for *at least one* action profile. Case (ii) includes Case (i) as a special case.[8]
- *Case* (iii) There is positive probability that any state can be reached from any state within a finite number of stages for any sequence of action profiles taken during these stages. Case (iii) includes Case (i) as a special case but not necessarily Case (ii).
- *Case* (iv) There is positive probability that any state can be reached from any state within a finite number of stages for at least one sequence of action pro-

8 Another possibility in between Cases (i) and (ii) is that the probability of transition between any pair of states is positive for at least one action of one player and any action of the opponent. In other words, the opponent cannot prevent the game to transit from any state to any state.

files taken during these stages. Case (iv) includes Cases (ii) and (iii) as special cases.[9]

Note that Assumption 4.1 holds under Case (iii) but not necessarily under Case (ii) or (iv).

Recall that in the classical fictitious play, the beliefs on opponent strategy are formed by the empirical average of the actions taken by the opponent. The players can also form their beliefs as a weighted average of the actions while the weights may give more (or less) importance to recent ones depending on the player's preferences, e.g., as in (4.4). In other words, we let $\alpha_c$ take values other than $1/(c+1)$ for $c = 0, 1, \ldots$ Furthermore, the two-timescale learning scheme imposes that $\beta_c/\alpha_c$ goes to zero as $c$ goes to infinity. In the following, we specify conditions on step sizes that are sufficient to ensure convergence of the two-timescale fictitious play in two-player zero-sum stochastic games under Assumption 4.1.

**Assumption 4.2** (Step sizes). The step sizes $\{\alpha_c\}$ and $\{\beta_c\}$ satisfy the following conditions:

(a) They vanish at a slow enough rate such that

$$\sum_{c\geq 0} \alpha_c = \sum_{c\geq 0} \beta_c = \infty$$

while $\alpha_c \to 0$ and $\beta_c \to 0$ as $c \to \infty$.

(b) They vanish at two separate timescales such that

$$\lim_{c\to\infty} \frac{\beta_c}{\alpha_c} = 0.$$

The following theorem shows that the two-timescale fictitious play converges in two-player zero-sum stochastic games under these assumptions.

**Theorem 4.3** ([63]). *Given a two-player zero-sum stochastic game, suppose that players follow the two-timescale fictitious play dynamics* (4.4) *and* (4.5)*. Under Assumptions* 4.1 *and* 4.2*, we have*

$$(\hat{\pi}_k^1, \hat{\pi}_k^2) \to (\pi_*^1, \pi_*^2) \quad \text{and} \quad (\hat{Q}_k^1, \hat{Q}_k^2) \to (Q_*^1, Q_*^2), \quad \text{with probability } 1, \tag{4.7}$$

*as $k \to \infty$ for some stationary equilibrium $\pi_* = (\pi_*^1, \pi_*^2)$ of the underlying stochastic game and $(Q_*^1, Q_*^2)$ denote the associated $Q$-functions.*

Before delving into the technical details of the proof, it is instructive to compare the two-timescale fictitious play with both the classical fictitious play and the Shapley's iteration. For example, the update of $\hat{\pi}_k^{-i}$, described in (4.4), differs from the classical fictitious play dynamics (2.3) since the auxiliary game depends on the belief $\hat{Q}_k^i$ while the belief (and therefore the payoffs of the auxiliary games) evolves in time with new observations, quite

9 Another possibility in between Cases (iii) and (iv) is that there is positive probability that any state can be reached from any state within a finite number of stages for at least one sequence of actions of one player and for any sequence of actions taken by the opponent during these stages. In other words, the opponent cannot prevent the player to reach any state from any state within a finite number of stages.

contrary to the classical scheme (2.4). In general, this constitutes a challenge in directly adopting the convergence analysis for the classical scheme to stochastic games. However, the two-timescale learning scheme weakens this coupling, enabling us to characterize the asymptotic behavior specific to a state separately from the dynamics in other states as if $(\hat{Q}_k^1(s,\cdot), \hat{Q}_k^2(s,\cdot))$ is stationary.

Moreover, even with the two-timescale learning scheme, we still face a challenge in directly adopting the convergence analysis of fictitious play specific to zero-sum games, e.g., [31, 59]. Particularly, players form beliefs on their $Q$-functions independently based on the backward induction that they will always look for maximizing their utility against the opponent strategy. Due to this independent update, the auxiliary games can deviate from the zero-sum structure even though the underlying game is zero-sum. Hence we do not necessarily have $\hat{Q}_k^1(s,a) + \hat{Q}_k^2(s,a) = 0$ for all $a \in A$ and for each $s \in S$. This poses an important challenge in the analysis since an arbitrary general-sum game does not necessarily have fictitious play property in general.

Next, we compare the two-timescale fictitious play with Shapley's value iteration. We can list the differences between the update of $\hat{Q}_k^i$, described in (4.5), and the Shapley's iteration (3.8) as follows:

- The Shapley's iteration is over the value functions, however, it can be turned into an iteration over the $Q$-functions with the operator
$$(\mathcal{F}^i Q^i)(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p\big(s'|s,a\big)\, \mathrm{val}^i\big(Q^i(s',\cdot)\big), \quad \forall (s,a) \in S \times A, \tag{4.8}$$
as derived in [76]. The transformed iteration is given by $Q^i_{(n+1)} = \mathcal{F}^i Q^i_{(n)}$ starting from arbitrary $Q^i_{(0)}$. Furthermore, the Shapley's iteration does not involve a step size, however, a step size can be included if we view $Q^i_{(n+1)} = \mathcal{F}^i Q^i_{(n)}$ as the one
$$Q^i_{(n+1)} = Q^i_{(n)} + \beta_{(n)}(\mathcal{F}^i Q^i_{(n)} - Q^i_{(n)}) \tag{4.9}$$
with the step size $\beta_{(n)} = 1$ for all $n$.

- The Shapley's iteration updates the value function at every state at each stage while (4.5) takes place only when the state is visited. Therefore, we face the asynchronous update challenge in the convergence analysis of (4.5) together with (4.4), which can take place only when the associated state is visited. To address this, we can resort to the asynchronous stochastic approximation methods, e.g., see [80] (also upcoming Theorem 4.5).

- More importantly, the convergence of the Shapley's iteration benefits from the contraction property of the operator (3.7) (or its transformed version (4.8)) based on the nonexpansive mapping $\mathrm{val}^i(\cdot)$. However, in the update (4.5), we have
$$\hat{v}_k^i(s) = \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \hat{\pi}^{-i}(s)} \big\{ \hat{Q}^i(s, a^1, a^2) \big\}$$
rather than $\mathrm{val}^i(\hat{Q}^i(s,\cdot))$, which need not lead to a contraction.

The proof of Theorem 4.3 follows from exploiting the two-timescale learning scheme to analyze the evolution of the beliefs on opponent strategies specific to a state in isolation as if the beliefs on $Q$-functions are stationary and then showing that $\hat{v}_k^i(s)$ tracks $\mathrm{val}^i(\hat{Q}^i(s,\cdot))$ while addressing the deviation from the zero-sum structure via a novel Lyapunov function construction. The two-timescale learning scheme yields that the limiting flow of the dynamics specific to a state is given by

$$\frac{d\pi^i(s)}{dt} + \pi_s^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a^1,a^2)\}, \tag{4.10}$$

$$\frac{dQ^i(s,a)}{dt} = 0, \tag{4.11}$$

for all $(s,a) \in S \times A$ and $i = 1,2$. The function (2.6) presented in [31] for continuous-time best response dynamics in zero-sum games is no longer a valid Lyapunov function since $\sum_{i=1,2} Q^i(s,a)$ is not necessarily zero for all $s$ and $a$. Therefore, we modify this function to characterize the asymptotic behavior of this flow in terms of the deviation from the zero-sum structure, e.g., $\max_{a\in A} |\sum_{i=1,2} Q^i(s,a)|$. The new function is defined by

$$V_*\big(\pi(s), Q(s,\cdot)\big) := \left(\sum_{i=1,2} \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a^1,a^2)\} - \lambda \max_{a\in A}\left|\sum_{i=1,2} Q^i(s,a)\right|\right)_+,$$

where $\lambda$ is a fixed scalar satisfying $\lambda \in (1, 1/\gamma)$. The lower bound on $\lambda$ plays a role in its validity as a Lyapunov function when $\max_{a\in A} |\sum_{i=1,2} Q^i(s,a)| \neq 0$ while the upper bound will play a role later when we focus on the evolution of $\sum_{i=1,2} Q^i(s,a)$ to show that the sum converges to zero, i.e., the auxiliary stage games become zero-sum, almost surely.

Note that $V_*(\cdot)$ reduces to $V(\cdot)$, described in (2.6), if $\sum_{i=1,2} Q^i(s,a) = 0$ for all $a \in A$. Furthermore, it is a valid Lyapunov function for any $Q^1(s,\cdot)$ and $Q^2(s,\cdot)$ since we have

$$\begin{aligned}\frac{d}{dt}&\left(\sum_{i=1,2} \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a)\}\right)\\ &= \sum_{i=1,2} Q^i(s,a_*) - \sum_{i=1,2} \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a)\},\end{aligned} \tag{4.12}$$

where $a_* = (a_*^1, a_*^2)$ are the maximizing actions in (4.10), and we always have

$$\sum_{i=1,2} Q^i(s,a_*) < \lambda \max_{a\in A}\left|\sum_{i=1,2} Q^i(s,a)\right|$$

if it is not zero-sum, since $\lambda > 1$. In other words, the term inside $(\cdot)_*$ in the new Lyapunov function always decreases along the flow when it is nonnegative and cannot be positive once it becomes nonpositive.

If we let $\bar{v}_k := \hat{v}_k^1 + \hat{v}_k^2$ and $\bar{Q}_k := \hat{Q}_k^1 + \hat{Q}_k^2$, the new Lyapunov function yields that

$$\left(\bar{v}_k(s) - \lambda \max_{a\in A} |\bar{Q}_k(s,a)|\right)_+ \to 0 \tag{4.13}$$

as $k \to \infty$ for each $s \in S$. On the other hand, we always have $\bar{v}_k(s) \geq -\lambda \max_{a \in A} |\bar{Q}_k(s,a)|$ by the definition of $\hat{v}_k^i$. These bounds imply that $\bar{Q}_k(s,a) \to 0$, and therefore $\bar{v}_k(s) \to 0$ for all $(s,a) \in S \times A$, because the evolution of $\bar{Q}_k$ for the current state $s$ is given by

$$\bar{Q}_{k+1}(s,a) = \bar{Q}_k(s,a) + \beta_{c_k(s)}\left(\gamma \sum_{s' \in S} p(s'|s,a)\bar{v}_k(s') - \bar{Q}_k(s,a)\right), \quad \forall a \in A \tag{4.14}$$

by (4.5), while the upper bound on $\lambda$ ensures that $\lambda\gamma \in (0,1)$, and therefore, $\bar{Q}_k(s,a)$ contracts at each stage until it converges to zero for all $s \in S$ and $a \in A$. The asynchronous update and the asymptotic upper bound on $\bar{v}_k$, as described in (4.13), constitute a technical challenge to draw this conclusion, however, they can be addressed via asynchronous stochastic approximation methods, e.g., see [80].

Furthermore, the saddle point equilibrium yields

$$\max_{a^1 \in A^1} \mathbb{E}_{a^2 \sim \hat{\pi}_k^2(s)}\{\hat{Q}_k^1(s,a)\} \geq \mathrm{val}^1\big(\hat{Q}_k^1(s,\cdot)\big) \geq \min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\hat{Q}_k^1(s,a)\}, \tag{4.15}$$

and the right-hand side is bounded from below by

$$\min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\hat{Q}_k^1(s,a)\} \geq \min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{-\hat{Q}_k^2(s,a)\} + \min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\bar{Q}_k(s,a)\} \tag{4.16}$$

$$\geq -\max_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\hat{Q}_k^2(s,a)\} - \max_{a \in A} |\bar{Q}_k(s,a)|. \tag{4.17}$$

These bounds lead to

$$0 \leq \hat{v}_k^i(s) - \mathrm{val}^i\big(\hat{Q}_k^i(s,\cdot)\big) \leq \bar{v}_k(s) + \max_{a \in A} |\bar{Q}_k(s,a)|, \tag{4.18}$$

Since the right-hand side goes to zero as $k \to \infty$, we have that $\hat{v}_k^i(s)$ tracks $\mathrm{val}^i(\hat{Q}_k^i(s,\cdot))$. Based on this tracking result, the update of $\hat{Q}_k^i$ can be viewed as an asynchronous version of the iteration

$$Q_{(n+1)}^i = Q_{(n)}^i + \beta_{(n)}\big(\mathcal{F}^i Q_{(n)}^i + \epsilon_{(n)}^i - Q_{(n)}^i\big), \tag{4.19}$$

where the tracking error $\epsilon_{(n)}^i$ is asymptotically negligible almost surely and the operator $\mathcal{F}$, as described in (4.8), is a contraction similar to the Shapley's operator, described in (3.7). This completes the sketch of the proof for Theorem 4.3.

### 4.1. Model-free learning in stochastic games

We next consider scenarios where players do not know the transition probabilities and their own stage payoff function, however, they can still observe their stage payoffs (associated with the current action profile), the opponent's action, and the current state visited. Therefore, the players can still form beliefs on opponent strategy and their $Q$-functions.

The update of the belief on opponent strategy does not depend on the model knowledge. Therefore, the players can update their beliefs $\hat{\pi}_k^{-i}$ as in (4.4) also in the model-free case. However, the update of $\hat{Q}_k^i$ necessitates the model knowledge by depending on the stage payoff function and transition probabilities. The same challenge arises also in model-free solution of Markov decision processes (MDPs)—a *single* player version of stochastic games.

For example, $Q$-learning algorithm, introduced by [82], can be viewed as a model-free version of the value iteration in MDPs and the update rule is given by

$$\hat{q}_{k+1}(s,a) = \hat{q}_k(s,a) + \beta_k(s,a)\Big(r_k + \gamma \max_{\tilde{a}\in A} \hat{q}_k(\tilde{s},\tilde{a}) - \hat{q}_k(s,a)\Big), \tag{4.20}$$

where the triple $(s,a,\tilde{s})$ denotes the current state $s$, current action $a$, and the next state $\tilde{s}$, respectively, the payoff $r_k$ corresponds to the payoff received, i.e., $r_k = r(s,a)$, and $\beta_k(s,a) \in [0,1]$ is a step size specific to the state-action pair $(s,a)$. The entries corresponding to the pairs $(s',a') \neq (s,a)$ do not get updated, i.e., $\hat{q}_{k+1}(s',a') = \hat{q}_k(s',a')$.

Watkins and Dayan [82] provided an ingenious (direct) proof for the almost sure convergence of $Q$-learning algorithm. Alternatively, it is also instructive to establish a connection between $Q$-learning algorithm and the classical value iteration to characterize its convergence properties. For example, the differences between them can be listed as follows:

- In $Q$-learning, agents use the value function estimate for the next state $\tilde{s}$, i.e., $\hat{v}^i_k(\tilde{s})$, in place of the expected continuation payoff $\sum_{s'\in S} p(s'|s,a)\hat{v}^i_k(s')$. This way, they can sample from the state transition probabilities associated with the current state–action pair by observing the state transitions. Correspondingly, this update takes place only after the environment transitions to the next state.
- The update can take place only for the current state–action pair because the agent can sample only from the transition probabilities associated with the current state-action pair by letting the environment do the experimentation.

Therefore, the $Q$-learning algorithm can be viewed as an asynchronous $Q$-function version of the value iteration

$$\hat{q}_{k+1} = \hat{q}_k + \beta_k(\mathcal{F}_o\hat{q}_k + \omega_{k+1} - \hat{q}_k), \tag{4.21}$$

where the $Q$-function version of the Bellman operator is given by

$$(\mathcal{F}_o\hat{q}_k)(s,a) = r(s,a) + \gamma \sum_{s'\in S} p\big(s'|s,a\big) \max_{a'\in A} \hat{q}_k(s',a') \tag{4.22}$$

and the stochastic approximation error $\omega_{k+1}$ is defined by

$$\omega_{k+1}(s,a) := \gamma\left(\max_{\tilde{a}\in A} \hat{q}_k(\tilde{s},\tilde{a}) - \sum_{s'\in S} p\big(s'|s,a\big) \max_{a'\in A} \hat{q}_k(s',a')\right), \tag{4.23}$$

with $\tilde{s}$ denoting the next state at stage $k$. Note that (4.21) turns into an asynchronous update if $\beta_k(s,a)$ is just zero when $\hat{q}_k(s,a)$ is not updated. Though these error terms $\{\omega_k\}_{k>0}$ do not form an independent sequence, they form a finite-variance martingale difference sequence conditioned on the history of parameters. The following well-known result shows that the weighted sum of such martingale difference sequences vanishes asymptotically almost surely.

**Lemma 4.4** ([58]). *Let $\{\mathcal{F}_k\}_{k\geq 0}$ be an increasing sequence of $\sigma$-fields. Given a sequence $\{\omega_k\}_{k\geq 0}$, suppose that $\omega_{k-1}$ is $\mathcal{F}_k$-measurable random variable satisfying $\mathbb{E}[\omega_k|\mathcal{F}_k] = 0$ and $\mathbb{E}[\omega_k^2|\mathcal{F}_k] \leq K$ for some $K$. Then, the sequence $\{W_k\}_{k\geq 0}$ evolving according to*

$$W_{k+1} = (1-\alpha_k)W_k + \alpha_k\omega_k, \tag{4.24}$$

*vanishes to zero asymptotically almost surely, i.e.,* $\lim_{k\to\infty} W_k = 0$ *with probability* 1*, provided that* $\alpha_k \in [0, 1]$ *is a vanishing step size that is* $\mathcal{F}_k$*-measurable, square-summable* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ *while* $\sum_{k=0}^{\infty} \alpha_k = \infty$ *with probability* 1.

This is a powerful result to characterize the convergence properties of stochastic approximation algorithms having the structure

$$x_{k+1} = x_k + \alpha_k \big(F(x_k) - x_k + \omega_k\big)$$

where $x_k$ is an $n$-dimensional vector, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz function, $\alpha_k \in [0, 1]$ is a step size, and $\omega_k$ is a stochastic approximation error term forming a finite-variance martingale difference sequence conditioned on the history of parameters. Note that every entry of the vector $x_k$ gets updated synchronously. If we also have that the iterate is bounded, we can characterize the convergence properties of this discrete-time update based on its limiting ordinary differential equation via a Lyapunov function formulation [11]. If the entries do not get updated synchronously, the asynchronous update challenge can be addressed based the *averaging techniques* [38]. In the case of $Q$-learning, this corresponds to assuming that different state–action pairs occur at well-defined average frequencies, which can be a restriction in practical applications [80]. Instead, [80] showed that we do not need such an assumption if the mapping $F$ has a contraction-like property based on the asynchronous convergence theory [8, 9].

**Theorem 4.5** ([80]). *Given an MDP, let an agent follow the* $Q$*-learning algorithm, described in* (4.20)*, with vanishing step sizes* $\beta_k(s, a) \in [0, 1]$ *satisfying* $\sum_{k\geq 0} \beta_k(s, a) = \infty$ *and* $\sum_{k\geq 0} \beta_k(s, a)^2 < \infty$ *for each* $(s, a) \in S \times A$*. Suppose that the entries corresponding to each* $(s, a)$ *gets updated infinitely often. Then, we have*

$$\hat{q}_k(s, a) \to q_*(s, a), \quad \text{with probability } 1, \tag{4.25}$$

*for each* $(s, a) \in S \times A$*, as* $k \to \infty$*, where* $q_*$ *is the unique* $Q$*-function solving the MDP.*

Tsitsiklis [80] considered a more general case where agents receive random payoffs. In general, such randomness can result in unbounded parameters. However, this is not the case for $Q$-learning algorithm, i.e., the iterates in the $Q$-learning algorithm remains bounded. Furthermore, the boundedness of the iterates plays a crucial role in the proof of Theorem 4.5. Particularly, consider the deviation between the iterate $\hat{q}_k$ and the unique solution $q_*$, i.e., $\tilde{q}_k = \hat{q}_k - q_*$, which evolves according to

$$\tilde{q}_{k+1} = \tilde{q}_k + \beta_k(\mathcal{F}_0\tilde{q}_k + \omega_{k+1} - \tilde{q}_k) \tag{4.26}$$

by (4.21) and since $\mathcal{F}_0 q_* = q_*$. Boundedness of the iterates $\hat{q}_k$ yields that $\tilde{q}_k$ is also bounded. For example, let $|\tilde{q}_k(s, a)| \leq D$ for all $(s, a)$ and $k$. Furthermore, by the contraction property of $\mathcal{F}_0$ with respect to the maximum norm, we have

$$\max_{(s,a)} \big|(\mathcal{F}_0\tilde{q}_k)(s, a)\big| \leq \gamma \max_{(s,a)} \big|\tilde{q}_k(s, a)\big|.$$

Therefore, we can show that the absolute value of new iterates are bounded from above by

$$\big|\tilde{q}_k(s, a)\big| \leq Y_k(s, a) + W_{k+1}(s, a), \tag{4.27}$$

where $\{Y_k(s,a)\}_{k\geq 0}$ and $\{W_{k+1}(s,a)\}_{k\geq 0}$ are two sequences evolving, respectively, according to

$$Y_{k+1}(s,a) = \big(1-\beta_k(s,a)\big)D + \beta_k(s,a)\gamma D \tag{4.28}$$

starting from $Y_0 = D$, and

$$W_{k+1}(s,a) = \big(1-\beta_k(s,a)\big)W_k(s,a) + \beta_k(s,a)\omega_k(s,a), \tag{4.29}$$

starting from $W_1(s,a) = 0$ for all $(s,a)$. For each $(s,a)$, the sequence $\{Y_k(s,a)\}_{k\geq 0}$ converges to $\gamma D$ while $\{W_{k+1}(s,a)\}_{k\geq 0}$ converges to zero with probability 1 by Lemma 4.4 due to the assumptions on the step size and the infinitely often update of every entry. Letting $k\to\infty$ for both sides of (4.27), we obtain that the shifted iterates are asymptotically bounded from above by $\gamma D$. This yields that there exists a stage where the iterates remain bounded from above by $(\gamma+\epsilon)D$ where $\epsilon > 0$ is sufficiently small such that $\gamma+\epsilon < 1$. By following the same lines, we can find a smaller asymptotic bound on the iterates. Therefore, we can induce that the shifted iterates converge to zero and the iterates converge to the solution of the MDP even with the asynchronous update.

Similar to the generalization of the value iteration to $Q$-learning for model-free solutions, [42] generalized the Shapley's iteration to *Minimax-Q learning* to compute equilibrium values in two-player zero-sum stochastic games in a model-free way. The update rule is given by

$$\hat{Q}^i_{k+1}(s,a) = \hat{Q}^i_k(s,a) + \beta_k(s,a)\big(r^i_k + \gamma\,\mathrm{val}^i\big[\hat{Q}^i_k(\tilde{s})\big] - \hat{Q}^i_k(s,a)\big), \tag{4.30}$$

for the current state $s$, current action profile $a$, and next state $\tilde{s}$ with a step size $\beta_k(s,a)\in[0,1]$ vanishing sufficiently slow such that $\sum_{k\geq 0}\beta_k(s,a) = \infty$ and $\sum_{k\geq 0}\beta_k(s,a)^2 < \infty$ with probability 1. The payoff $r^i_k$ corresponds to the payoff received for the current state and action profile, i.e., $r^i_k = r^i(s,a)$. The Minimax-Q algorithm converges to the equilibrium $Q$-functions of the underlying two-player zero-sum stochastic game almost surely if every state and action profile occur infinitely often.

In model-free methods, the assumption that every state–action pair occur infinitely often can be restrictive for practical applications. A remedy to this challenge is that agents explore at random instances by taking any action with uniform probability. Such random exploration results in that every state-action pair gets realized infinitely often if every state is visited infinitely often. Indeed, random exploration will also yield that each state gets visited infinitely often if there is always positive probability that any state is reachable from any state within a finite number of stages for at least one sequence of actions taken during these stages. This corresponds to Case (iv) described in Section 4.

In the model-free two-timescale fictitious play, players play the best response in the auxiliary game with probability $(1-\epsilon)$ while experimenting with probability $\epsilon$ by playing any action with uniform probability. They still update the belief on the opponent strategy as in (4.4). Furthermore, they update their beliefs on the $Q$-function for the current state $s$,

current action profile $a$ and next state $s'$ triple $(s, a, s')$ according to

$$\hat{Q}^1_{k+1}(s,a) = \hat{Q}^1_k(s,a) + \beta_{c_k(s,a)}\Big(r^1_k + \gamma \max_{a^1 \in A} \mathbb{E}_{a^2 \sim \hat{\pi}^2_k(s')}\{\hat{Q}^1_k(s', a^1, a^2)\} - \hat{Q}^1_k(s,a)\Big), \tag{4.31}$$

where $\beta_{c_k(s,a)} \in [0,1]$ is a step size vanishing with the number of times $(s,a)$ is realized and the payoff $r^1_k$ corresponds to the payoff associated with the current state $s$ and action profile $a$, i.e., $r^1_k = r^1(s,a)$.

Recall that the two-timescale learning scheme plays an important role in the convergence of the dynamics. Particularly, the step size $\alpha_c$ used in the update of the belief $\hat{\pi}^{-i}_k(s)$ goes to zero slower than the step size $\beta_c$ used in the update of the belief $\hat{Q}^i_k(s,\cdot)$. Since both step size depend on the number of visits to the associated state, the assumption that $\beta_c/\alpha_c \to 0$ as $c \to \infty$ is sufficient to ensure this timescale separation. However, in the model-free case, the asynchronous update of $\hat{Q}^i_k(s,a)$ for different action profiles can undermine this timescale separation because the step size $\beta_c$ specific to the update of $\hat{Q}^i_k(s,a)$ depends the number of times the state and action profile $(s,a)$, i.e., $c_k(s,a)$, is realized. Therefore, we make the following assumption ensuring that the step size in the update of $\hat{Q}^i_k(s,a)$ vanishes still faster than the step size in the update of $\hat{\pi}^{-i}_k(s)$ as long as $c_k(s,a)$ is comparable with $c_k(s)$, i.e., $\liminf_{k\to\infty} c_k(s,a)/c_k(s) > 0$ with probability 1.

**Assumption 4.6** (Step sizes). The step sizes $\{\alpha_c\}$ and $\{\beta_c\}$ satisfy the following conditions:

(a) They vanish at a slow enough rate such that

$$\sum_{c\geq 0}\alpha_c = \sum_{c\geq 0}\beta_c = \infty, \quad \text{and} \quad \sum_{c\geq 0}\alpha_c^2 < \infty, \quad \sum_{c\geq 0}\beta_c^2 < \infty$$

while $\alpha_c \to 0$ and $\beta_c \to 0$ as $c \to \infty$.[10]

(b) The sequence $\{\beta_c\}_{c\geq 0}$ is monotonically decreasing. For any $m \in (0,1]$, we have[11]

$$\lim_{c\to\infty} \frac{\beta_{\lfloor mc \rfloor}}{\alpha_c} = 0.$$

When we have $\liminf_{k\to\infty} c_k(s,a)/c_k(s) > 0$ with probability 1 for all $(s,a)$, the second part of Assumption 4.6 ensures that $\lim_{k\to\infty} \frac{\beta_{c_k(s,a)}}{\alpha_{c_k(s)}} = 0$ with probability 1 for all $(s,a)$. Indeed, Assumptions 4.2 and 4.6 are satisfied for the usual (vanishing) step sizes such as

$$\alpha_c = \frac{1}{(c+1)^{\rho_\alpha}} \quad \text{and} \quad \beta_c = \frac{1}{(c+1)^{\rho_\beta}},$$

where $1/2 < \rho_\alpha < \rho_\beta \leq 1$.

---

10 We have the additional assumption that the step size $\beta_c$ is square summable to ensure that the stochastic approximation error terms have finite variance conditioned on the history of the parameters.

11 Perkins and Leslie **[54]** made a similar assumption that $\sup_c \frac{\beta_{\lfloor mc \rfloor}}{\beta_c} < M$ for all $m \in (0,1)$ and $\frac{\beta_c}{\alpha_c} \to 0$ for two-timescale asynchronous stochastic approximation.

When players do random experimentation in the model-free case, they do not take the best response with certain probability. Therefore, we do not have convergence to an exact equilibrium as in the model-based case. However, the players still converge to a near equilibrium of the game with linear dependence on the experimentation probability and the following theorem provides an upper bound on this approximation error.

**Theorem 4.7** ([63]). *Given a two-player zero-sum stochastic game, suppose that players follow the model-free two-timescale fictitious play dynamics with experimentation probability $\epsilon > 0$. Under Assumptions 4.1 and 4.6, we have*

$$\limsup_{k\to\infty} \left| v_k^i(s) - v^i(s) \right| \leq \epsilon D \frac{1+\gamma}{\gamma(1-\gamma)^2}, \tag{4.32}$$

$$\limsup_{k\to\infty} \max_{a\in A} \left| \hat{Q}_k^i(s,a) - Q^i(s,a) \right| \leq \epsilon D \frac{1+\gamma}{(1-\gamma)^2}, \tag{4.33}$$

*with probability* 1, *where* $D = \frac{1}{1-\gamma} \sum_i \max_{(s,a)} |r^i(s,a)|$, *where* $v_*^i$ *and* $Q_*^i$ *denote, respectively, the value function and* $Q$*-function of player* $i$ *for some stationary equilibrium of the stochastic game.*

Even though the random experimentation can prevent convergence to an exact equilibrium, it provides an advantage for the applicability of this near-convergence result because every state gets visited infinitely often, and therefore, Assumption 4.1 holds, if the underlying Markov chain satisfies Case (iv), i.e., there is positive probability that any state can be reached from any state within a finite number of stages for at least one sequence of action profiles taken during these stages.

The dynamics can converge to an exact equilibrium also in the model-free case if players let the experimentation probability vanish at certain rate. However, there are technical details that can limit the applicability of the result for Case (iv).

### 4.2. Radically uncoupled learning in stochastic games

Finally, we consider minimal-information scenarios where players do not even observe the opponent's actions in the model-free case. Each player can still observe its own stage payoff received and the current state visited. The players also do not know the opponent's action set. Indeed, they may even be oblivious to the presence of an opponent. The learning dynamics under such minimal information case is known as *radically uncoupled learning* in the learning in games literature, e.g., see [25].

Without observing the opponent's actions and knowing her action space, players are not able to form beliefs on opponent strategy as in the fictitious play. This challenge is present also in the repeated play of the same strategic-form game. For example, consider the strategic-form game $\langle A^1, A^2, r^1, r^2 \rangle$ and define $q^i : A^i \to \mathbb{R}$ by

$$q^i(a^i) := \mathbb{E}_{a^{-i}\sim\pi^{-i}} \left\{ r^i(a^1, a^2) \right\}, \quad \forall a^i \in A^i \tag{4.34}$$

given the opponent's strategy $\pi^{-i}$. Then, the computation of the best response is a simple optimization problem for player $i$, given by

$$a^i_* \in \operatorname*{argmax}_{a^i \in A^i} q^i(a^i).$$

Player $i$ would be able to compute her best response $a^i_*$ even when she does not know the opponent strategy $\pi^{-i}$ and her payoff function $r^i$ if she knew the function $q^i(\cdot)$. Hence, the question is whether the computation of $q^i(\cdot)$ can be achieved without observing the opponent's action.

Suppose that players are playing the same strategic-form game repeatedly and player $i$ makes the forward induction that the opponent will play as how he has played in the past similar to the fictitious play dynamics. If that were the case, i.e., the opponent were playing according to a stationary strategy $\pi^{-i}$, then at each stage the payoff received by player $i$ would be the realized payoff $r^i(a^1, a^2)$, where $a^{-i} \sim \pi^{-i}$ and $a^i$ is the current action she has taken. Correspondingly, player $i$ can form a belief about $q^i(a^i)$ for all $a^i \in A^i$ and update $q^i(\cdot)$ associated with the current action based on the payoff she received. For example, let $\hat{q}^i_k$, $a^i_k$ and $r^i_k$ denote, respectively, the belief of player $i$ on $q^i$, her current action and the current payoff she received. Similar to the update of the belief on opponent's strategy, the update of $\hat{q}^i_k$ is given by

$$\hat{q}^i_{k+1}(a^i) = \begin{cases} \hat{q}^i_k(a^i) + \alpha_k(a^i)(r^i_k - \hat{q}^i_k(a^i)) & \text{if } a^i = a^i_k, \\ \hat{q}^i_k(a^i) & \text{otherwise,} \end{cases}$$

where $\alpha_k(a^i) \in [0, 1]$ is a vanishing step size specific to the action $a^i$. However, this results in an asynchronous update of $\hat{q}_k$ for different actions quite contrary to the synchronous belief update (2.3) in the fictitious play. There is no guarantee that it would converge to an equilibrium even in the zero-sum case. On the other hand, such an asynchrony issue is not present and the update turns out to be synchronous in expectation if players take *smoothed* best response while normalizing the step size by the probability of the current action taken [39].

Given $\hat{q}^i_k$, the smoothed best response $\overline{\mathrm{BR}}^i_k \in \Delta(A^i)$ is given by

$$\overline{\mathrm{BR}}^i_k := \operatorname*{argmax}_{\mu^i \in \Delta(A^i)} \left( \mathbb{E}_{a^i \sim \mu^i} \{\hat{q}^i_k(a^i)\} + \tau v^i(\mu^i) \right), \tag{4.35}$$

where $v^i : \Delta(A^i) \to \mathbb{R}$ is a smooth and strictly concave function whose gradient is unbounded at the boundary of the simplex $\Delta(A^i)$ [29]. The temperature parameter $\tau > 0$ controls the amount of perturbation on the smoothed best response. Note that the smooth perturbation ensures that there always exists a unique maximizer in (4.35). Since players take smoothed best response rather than best response, we use an equilibrium concept different from the Nash equilibrium. This new definition is known as Nash distribution or quantal response equilibrium [46].

**Definition 4.8** (Nash distribution). We say that a strategy profile $\pi_*$ is a Nash distribution if we have

$$\pi^i_* = \operatorname*{argmax}_{\mu^i \in \Delta(A^i)} \left( \mathbb{E}_{(a^i, a^{-i}) \sim (\mu^i, \pi^{-i}_*)} \{r^i_k(a)\} + \tau v^i(\mu^i) \right) \tag{4.36}$$

for each $i$.

An example to the smooth function is $v^i(\mu^i) := -\mathbb{E}_{a^i\sim\mu^i}\{\log(\mu^i(a^i))\}$, also known as the entropy [34], and the associated smoothed best response has the following analytical form:

$$\overline{\mathrm{BR}}_k^i(a^i) = \frac{\exp(\hat{q}_k^i(a^i)/\tau)}{\sum_{\tilde{a}^i\in A^i}\exp(\hat{q}_k^i(\tilde{a}^i)/\tau)},$$

which is positive for all $a^i \in A^i$.

When player $i$ takes her action according to the smoothed best response $\overline{\mathrm{BR}}_k^i$, any action will be taken with some positive probability $\overline{\mathrm{BR}}_k^i(a^i) > 0$. Hence she can update her belief according to

$$\hat{q}_{k+1}^i(a^i) = \begin{cases} \hat{q}_k^i(a^i) + \overline{\mathrm{BR}}_k^i(a^i)^{-1}\alpha_k(r_k^i - \hat{q}_k^i(a^i)) & \text{if } a^i = a_k^i, \\ \hat{q}_k^i(a^i) & \text{otherwise,} \end{cases} \tag{4.37}$$

where $\alpha_k \in (0,1)$ is a step size vanishing with $k$ and not specific to any action. This asynchronous update rule, also known as *individual $Q$-learning*, turns out to be synchronous in the expectation. Particularly, the new update rule is given by

$$\hat{q}_{k+1}^i(a^i) = \hat{q}_k^i(a^i) + \alpha_k\big(\mathbb{E}_{a^{-i}\sim\overline{\mathrm{BR}}_k^{-i}}\{r^i(a^1,a^2)\} - \hat{q}_k^i(a^i) + \omega_k^i(a^i)\big), \quad \forall a^i \in A^i, \tag{4.38}$$

and $\omega_k^i(a^i)$ is the stochastic approximation error defined by

$$\omega_k^i(a^i) := \mathbf{1}_{\{a^i=a_k^i\}}\overline{\mathrm{BR}}_k^i(a^i)^{-1}\big(r_k^i - \hat{q}_k^i(a^i)\big) - \mathbb{E}_{a\sim\overline{\mathrm{BR}}_k}\big\{\mathbf{1}_{\{a^i=a_k^i\}}\overline{\mathrm{BR}}_k^i(a^i)^{-1}\big(r_k^i - \hat{q}_k^i(a^i)\big)\big\},$$

where $\overline{\mathrm{BR}}_k = (\overline{\mathrm{BR}}_k^1, \overline{\mathrm{BR}}_k^2)$, because we have

$$\mathbb{E}_{a\sim\overline{\mathrm{BR}}_k}\big\{\mathbf{1}_{\{a^i=a_k^i\}}\overline{\mathrm{BR}}_k^i(a^i)^{-1}\big(r_k^i - \hat{q}_k^i(a^i)\big)\big\} = \mathbb{E}_{a^{-i}\sim\overline{\mathrm{BR}}_k^{-i}}\{r^i(a^1,a^2)\} - \hat{q}_k^i(a^i).$$

Furthermore, the stochastic approximation error term forms a martingale difference sequence conditioned on the history of iterates while the *boundedness* of the iterates ensure that it has finite variance. Therefore, we can invoke Lemma 4.4 to characterize the convergence properties of (4.38)—a rewritten version of (4.37) with the stochastic approximation term $\omega_k^i$.

**Theorem 4.9** ([39]). *In two-player zero-sum (or identical-payoff) strategic-form games played repeatedly, if both player follows the individual $Q$-learning algorithm, described in (4.37), then their estimate $\hat{q}_k^i$ converges to $q_*^i$ for all $a^i \in A^i$ satisfying*

$$q_*^i(a^i) = \mathbb{E}_{a^{-i}\sim\pi_*^{-i}}\{r^i(a^1,a^2)\}$$

*for some Nash distribution $\pi_* = (\pi_*^1, \pi_*^2)$ under the assumption that the iterates remain bounded. Correspondingly, their smoothed best response also converges to $\pi^*$.*

Recall that in stochastic games, players are playing an *auxiliary stage-game* specific to the current state $\mathcal{G}_s = \langle A^1, A^2, Q^1(s,\cdot), Q^2(s,\cdot)\rangle$, where $Q^i$ satisfies (4.1). Therefore, in the minimal information case, each player $i$ can form a belief about the associated

$$q^i(s,a^i) := \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a^1,a^2)\},$$

which is now specific to state $s$ contrary to (4.34), and update it based on the stage payoffs received as in the individual $Q$-learning dynamics. We can view $q^i$ as the local $Q$-function

since it is defined over individual actions rather than action profiles. We denote player $i$'s belief on $q^i$ by $\hat{q}^i_k$. Let $s$ be the current state of the stochastic game. Then, player $i$ selects her action $a^i_k$ according to smoothed best response

$$\overline{\mathrm{BR}}^i_k(s,\cdot) = \operatorname*{argmax}_{\mu^i \in \Delta(A^i)} \big(\mathbb{E}_{a^i\sim\mu^i}\{\hat{q}^i_k(s,a^i)\} + \tau v^i(\mu^i)\big),$$

i.e., $a^i_k \sim \overline{\mathrm{BR}}^i_k(s,\cdot)$. The smoothed best response depends only on the belief on the local $Q$-function, i.e., $\hat{q}^i_k(s,\cdot)$. Observing the stage reward $r^i_k$ and the next state $s'$, player $i$ can update her belief according to

$$\hat{q}^i_{k+1}(s,a^i) = \begin{cases} \hat{q}^i_k(s,a^i) + \overline{\mathrm{BR}}^i_k(s,a^i)^{-1}\alpha_{c_k(s)}(r^i_k + \gamma \hat{v}^i_k(s') - \hat{q}^i_k(s,a^i)) & \text{if } a^i = a^i_k, \\ \hat{q}^i_k(s,a^i) & \text{otherwise,} \end{cases} \tag{4.39}$$

where $\alpha_c \in (0,1)$ is a vanishing step size and recall that $c_k(s)$ denotes the number of visits to state $s$ until and including stage $k$. The update (4.39) differs from (4.37) due to the additional term $\gamma\hat{v}^i_k(s')$ corresponding to an unbiased estimate of the continuation payoff in the model-free case. Due to this additional term, the individual $Q$-learning dynamics in auxiliary stage-games specific to each state are coupled with each other. A two-timescale learning framework can weaken this coupling if players estimate $\hat{v}^i_k$ at a slower timescale according to

$$\hat{v}^i_{k+1}(s) = \hat{v}^i_k(s) + \beta_{c_k(s)}\big(\mathbb{E}_{a^i\sim\overline{\mathrm{BR}}^i_k(s,\cdot)}\{\hat{q}^i_k(s,a^i)\} - \hat{v}^i_k(s)\big), \tag{4.40}$$

where $\beta_c \in (0,1)$ is a vanishing step size that goes to zero faster than $\alpha_c$, rather than $\hat{v}^i_k(s) = \mathbb{E}_{a^i\sim\overline{\mathrm{BR}}^i_k(s,\cdot)}\{\hat{q}^i_k(s,a^i)\}$.

This decentralized $Q$-learning dynamics, described in (4.39) and (4.40), have convergence properties similar to the two-timescale fictitious play even in this minimal information case. Furthermore, random exploration is inherent in the smoothed best response. Therefore, Assumption 4.1 holds if the underlying Markov chain satisfies Case (iv). However, due to the smoothed best response, the dynamics does not necessarily converge to an exact Nash equilibrium.

**Theorem 4.10** ([64]). *Given a two-player zero-sum stochastic game, suppose that players follow the decentralized $Q$-learning dynamics. In addition to Assumptions 4.1 and 4.6, we assume that $\sum_{c\geq 0}\alpha_c^2 < \infty$ and the iterates are bounded. Let $Q^i_*$ and $v^i_*$ denote the unique equilibrium $Q$-function and value function of player $i$. Then, we have*

$$\limsup_{k\to\infty} \big|\hat{v}^i_k(s) - v^i_*(s)\big| \leq \tau \log\big(\big|A^1\big|\big|A^2\big|\big) g(\gamma), \tag{4.41}$$

*for all $(i,s) \in \{1,2\}\times S$, with probability 1, where $g(\lambda) := \frac{2+\lambda-\lambda\gamma}{(1-\lambda\gamma)(1-\gamma)}$ with some $\lambda \in (1, 1/\gamma)$.*

*Furthermore, let $\hat{\pi}^i_k(s) \in \Delta(A^i)$ be the weighted time-average of the smoothed best response updated as*

$$\hat{\pi}^i_{k+1}(s) = \hat{\pi}^i_k(s) + \mathbf{1}_{\{s=s_k\}}\alpha_{c_k(s)}\big(\overline{\mathrm{BR}}^i_k(s,\cdot) - \hat{\pi}^i_k(s)\big).$$

*Then, we have*

$$\limsup_{k\to\infty} \max_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim \hat{\pi}_k^{-i}(s)} \{Q_*^i(s,a)\} - v_{\pi_*}^i(s)| \le \tau \log(|A^1||A^2|)h(\gamma), \tag{4.42}$$

*for all* $(i,s) \in \{1,2\} \times S$, *with probability* 1, *where* $h(\gamma) := g(\gamma)(1+\gamma) - 1$. *In other words, these weighted-average strategies converge to near Nash equilibrium strategies of the stochastic game.*

The iterates would be bounded inherently if players update the local $Q$-function (4.37) by thresholding the step $\overline{\mathrm{BR}}_k^i(a^i)^{-1}\alpha_{c_k(s)}$ from above by 1. Furthermore, the dynamics could converge to an exact equilibrium if players let their temperature parameter $\tau > 0$ vanishes over time at a certain rate, e.g., see [64]. With vanishing temperature, Assumption 4.1 holds if the underlying Markov chain satisfies Case (iii).

## 5. OTHER LEARNING ALGORITHMS

Previous sections have focused on a detailed description of best-response/fictitious-play type learning dynamics, together with $Q$-learning dynamics, for stochastic games. In this section, we summarize several other algorithms in the learning in games literature, with a focus on independent/decentralized learning for stochastic games (also belonging to the area of *multiagent reinforcement learning* in the machine learning literature).

### 5.1. Classical algorithms

For stochastic games, other than $Q$-learning-type algorithms presented in Section 4.1, [10] also established the asymptotic convergence of an actor–critic algorithm to a weaker notion of generalized Nash equilibrium. Another early work [13] proposed R-MAX, an optimism-based RL algorithm for average-reward two-player zero-sum stochastic games, with polynomial time convergence guarantees. However, convergence to the actual Nash equilibrium is not guaranteed from the regret definition in the paper.

For strategic-form games, besides fictitious play, several other *decentralized* learning dynamics have also been thoroughly studied. A particular example is the *no-regret* learning algorithms[12] from the online learning literature. It is a folklore theorem that: If both players of a game use some no-regret learning dynamics to adapt their strategies to their opponent's strategies, then the time-average strategies of the players constitute a Nash equilibrium of the zero-sum strategic-form game [18, 61]. Popular no-regret dynamics include multiplicative weights update [26, 41], online gradient descent [91], and their generalizations [47, 67]. These no-regret learning dynamics are *uncoupled* in that a player's dynamics does not explicitly rely on the payoffs of other players [32]. They are also posited to be a rational model of players' rational behavior [60, 75]. In addition, [39] proposed individual $Q$-learning, a fully decentralized learning dynamics where each player's update rule requires no observation of the opponent's actions, with convergence to the Nash equilibrium *distribution* of

12 See [18] for formal definitions and results of no-regret learning.

certain two-player games. Notably, these decentralized learning dynamics are only known to be effective for strategic-form games.

### 5.2. Multiagent reinforcement learning

There has been a flurry of recent works on multiagent RL in stochastic games with focuses on *nonasymptotic* performance guarantees. The authors of [56,57] proposed batch RL algorithms to find an approximate Nash equilibrium using approximate dynamic programming analysis. Wei et al. [83] studied *online* RL, where only one of the player is controlled, and develops the UCSG algorithm with sublinear regret guarantees that improves the results in [13], though still without guarantees of finding the Nash equilibrium. Subsequently, [72] provided near-optimal sample complexity for solving *turn-based* two-player zero-sum finite stochastic games, when a generative model that enables sampling from any state–action pair is available. Under the same setting, the near-optimal sample complexity for general two-player zero-sum finite stochastic games was then established in [87]. Without a generative model, [2, 85] presented optimistic value iteration-based RL algorithms for two-player zero-sum stochastic games, with efficient exploration of the environment, and finite-time regret guarantees. The two players need some coordination to perform the algorithms, and the focus in these two works is the *finite-horizon episodic* setting. Later, [3] and [43] provided tighter regret bounds for the same setting, with model-free and model-based RL methods, respectively. Liu et al. [43] has also studied the general-sum setting, with finite-sample guarantees for finding the Nash equilibrium, assuming some computation oracle for finding the equilibrium of general-sum strategic-form games at each iteration. Contemporaneously, [35, 37] studied multiagent RL with *function approximation* in finite-horizon episodic zero-sum stochastic games, with also the optimism principle and regret guarantees.

In addition, *policy-based* RL algorithms have also been developed for solving stochastic games. The authors of [15, 88] developed double-loop policy gradient methods for solving zero-sum linear quadratic dynamic games, a special case of zero-sum stochastic games with linear transition dynamics and quadratic cost functions, with convergence guarantees to the Nash equilibrium. Later, [90] also studied double-loop policy gradient methods for zero-sum stochastic games with general function approximation. Note that these double-loop algorithms are not symmetric in that they require one of the players to wait for the opponent to update her policy parameter multiple steps while updating her own policy for one step, which necessarily requires some coordination between players. Finally, [66] developed an Explore–Improve–Supervise approach, which combines ideas from Monte Carlo Tree Search and Nearest Neighbors methods, to find the approximate Nash equilibrium value of *continuous-space* turn-based zero-sum stochastic games. The two players are coordinated to learn the minimax value jointly.

Notably, as minimax $Q$-learning, these multiagent RL algorithms are mostly focused on the *computational* aspect of learning in stochastic games: compute the Nash equilibrium without knowing the model, using possibly as few samples as possible. Certain level of coordination among the players is either explicitly or implicitly assumed when implementing these algorithms, even for the zero-sum setting where the players compete against each

other. For human-like self-interested players, these update rules may not be sufficiently rational and natural to execute. Indeed, as per [12], a preferable multiagent RL algorithm should be both *rational* and *convergent*: a rational algorithm ensures that the iterates converge to the opponent's best-response if the opponent converges to a stationary policy; while a convergent algorithm ensures convergence to some equilibrium if all the agents apply the learning dynamics. In general, a rational algorithm, in which each player *adapts* to the (possibly nonstationary) behavior of other players and uses only *local* information she observes without the aid of any central coordinator, does not lead to the equilibrium of the game. In fact, investigating whether a game-theoretical equilibrium can be realized as a result of nonequilibrium adaptation dynamics is the core topic in the literature of *learning in games* [29]. These multiagent RL works have thus motivated our study of independent learning dynamics presented in Section 4.

### 5.3. Decentralized learning in stochastic games

Decentralized learning in stochastic games has attracted increasing research interest lately. In [1], decentralized $Q$-learning has been proposed for *weakly acyclic* stochastic games, which include stochastic teams (identical-interest stochastic games) as a special case. The update rule for each player does not need to observe the opponent players' actions, and is even oblivious to the presence of other players. However, the players are implicitly coordinated to explore every multiple iterations (in the exploration phase) without changing their policies, in order to create a stationary environment for each player. The key feature of the update rule is to restrict player strategies to stationary pure strategies. Since there are only finitely many stationary pure strategy, players can create a huge-game matrix for each stationary pure strategy and a pure-strategy equilibrium always exists when this huge-game is weakly acyclic with respect to best response. However, in the model-free case, players do not know the payoffs of this huge-game and the two-phase update rule addresses this challenge. Perolat et al. [55] developed actor–critic-type learning dynamics that are decentralized and of fictitious-play type, where the value functions are estimated at a faster timescale (in the critic step), and the policy is improved at a slower one (in the actor step). Nonetheless, the learning dynamics only applies to a special class of stochastic games with a "multistage" structure, in which each state can only be visited once. In [21], an independent policy gradient method was investigated for zero-sum stochastic games with convergence rate analysis, where two players use *asymmetric* stepsizes in their updates with one updates faster than the other. This implicitly requires some coordination between players to determine who shall update faster. Contemporaneously, [79] studied *online* RL in unknown stochastic games, where only one player is controlled and the update rule is fully decentralized. The work focused on the efficient *exploration* aspect of multiagent RL, by establishing the regret[13] guarantees of the proposed update rule. The work considered only the finite-horizon episodic setting, and it

13 The regret defined in [79] is weaker than the normal one with the *best-in-hindsight* comparator. See [79, SECT. 2] for a detailed comparison.

is also unclear if the learning dynamics converge to any equilibrium when all players apply it.[14]

With symmetric and decentralized learning dynamics, [17, 40, 84] are, to the best our knowledge, the latest efforts on learning in stochastic games. Leslie et al. [40] studied *continuous-time* best-response dynamics for zero-sum stochastic games, with a *two-timescale* update rule: at the slower timescale, a single continuation payoff (common among the players) is updated, representing time average of auxiliary game payoffs up to time $k$; at the faster timescale, each player updates its strategy in the direction of its best response to opponent's current strategy in the auxiliary game. The common continuation payoff update ensures that the auxiliary game is always zero-sum, allowing the use of the techniques for the strategic-form game setting [31]. The dynamics update the mixed strategies at every state at every time. Alternatively, the work also considered a continuous-time embedding of the *actual play* of the stochastic game where game transitions according to a controlled continuous-time Markov chain. Both [84] and [17] studied the genuine infinite-horizon discounted zero-sum stochastic games, and provided *last-iterate* convergence rate guarantees to approximate Nash equilibrium. To this end, [84] developed an optimistic variant of gradient descent-ascent update rule; while [17] focused on the *entropy-regularized* stochastic games, and advocated the use of policy extragradient methods. Though theoretically strong and appealing, these update rules assume either exact access or sufficiently accurate estimates of the continuation payoffs under instantaneous joint strategies and/or the instantaneous strategy of the opponent. In particular, to obtain finite-time bounds, the players are coordinated to interact multiple steps to estimate the continuation payoffs in the learning setting [84].

By and large, ever since the introduction of fictitious play [14] and stochastic games [69], it remains a long-standing problem whether an equilibrium in a stochastic game can be realized as an outcome of some natural and decentralized nonequilibrium adaptation, e.g., fictitious play (except the contemporaneous work [40] with some continuous-time embeddings). Hence, our solutions in Section 4 serve as an initial attempt towards settling the argument positively.

## 6. CONCLUSIONS AND OPEN PROBLEMS

In this review paper, we introduced multiagent dynamic learning in stochastic games, an increasingly active research area where artificial intelligence, specifically reinforcement learning, meets game theory. We have presented the fundamentals and background of the problem, followed by our recent advances in this direction, with a focus on studying *independent learning* dynamics. We believe our work has opened up fruitful directions for future research, on developing more natural and rational multiagent learning dynamics for

14 The same update rule with different stepsize and bonus choices and a certified policy technique, however, can return a non-Markovian approximate Nash equilibrium policy pair in the zero-sum setting; see [3], and the very recent and more complete treatment [36], for more details.

stochastic games. In particular, several future/ongoing research directions include: (1) establishing convergence guarantees of our independent learning dynamics for other stochastic games, e.g., identical-interest ones; (2) establishing nonasymptotic convergence guarantees of our learning dynamics, or other independent learning dynamics, for stochastic games; (3) developing natural learning dynamics that also account for the large state–action spaces in practical stochastic games, e.g., via function approximation techniques.


## FUNDING

A. Ozdaglar and K. Zhang were supported by DSTA grant 031017-00016 and ARO Project W911NF1810407.


## REFERENCES


[1] G. Arslan and S. Yüksel, Decentralized Q-learning for stochastic teams and games. *IEEE Trans. Automat. Control* **62** (2017), no. 4, 1545–1558.

[2] Y. Bai and C. Jin, Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pp. 551–560, PMLR, 2020.

[3] Y. Bai, C. Jin, and T. Yu, Near-optimal reinforcement learning with self-play. In *Advances in neural information processing systems 33*, pp. 2159–2170, Curran Associates, Inc., 2020.

[4] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory. 2nd edn.* Classics Appl. Math., SIAM, 1999.

[5] M. Benaim, J. Hofbauer, and S. Sorin, Stochastic approximations and differential inclusions. *SIAM J. Control Optim.* **44** (2005), no. 1, 328–348.

[6] U. Berger, Fictitious play in $2 \times n$ games. *J. Econom. Theory* **120** (2005), no. 2, 139–154.

[7] U. Berger, Learning in games with strategic complementarities revisited. *J. Econom. Theory* **143** (2008), no. 1, 292–301.

[8] D. P. Bertsekas, Distributed dynamic programming. *IEEE Trans. Automat. Control* **AC-27** (1982), 610–616.

[9] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice Hall, New Jersey, 1989.

[10] V. S. Borkar, Reinforcement learning in Markovian evolutionary games. *Adv. Complex Syst.* **5** (2002), no. 01, 55–72.

[11] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Hindustan Book Agency, 2008.

[12] M. Bowling and M. Veloso, Rational and convergent learning in stochastic games. In *Proceedings 17th international joint conference on artificial intelligence*, pp. 1021–1026, Morgan Kaufmann Publishers Inc., 2001.

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

[14] G. W. Brown, Iterative solution of games by fictitious play. In *Activity analysis of production and allocation*, pp.374–376, Cowles Commission Monograph 13, Wiley, New York, 1951.

[15] J. Bu, L. J. Ratliff, and M. Mesbahi, Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. 2019, arXiv:1911.04672.

[16] L. Busoniu, R. Babuska, and B. D. Schutter, A comprehensive survey of multiagent reinforcement learning. *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.* **38** (2008), no. 2, 156–172.

[17] S. Cen, Y. Wei, and Y. Chi, Fast policy extragradient methods for competitive games with entropy regularization. 2021, arXiv:2105.15186.

[18] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[19] C. Claus and C. Boutilier, The dynamics of reinforcement learning in cooperative multiagent systems. In *Conference on artificial intelligence*, pp. 746–752, American Association for Artificial Intelligence, 1998.

[20] A. Condon, On algorithms for simple stochastic games. In *Advances in computational complexity theory 13*, pp. 51–72, American Mathematical Society, 1990.

[21] C. Daskalakis, D. J. Foster, and N. Golowich, Independent policy gradient methods for competitive reinforcement learning. In *Advances in neural information processing systems*, Curran Associates, Inc., 2020.

[22] J. C. Ely and O. Yilankaya, Nash equilibrium and the evolution of preferences. *J. Econom. Theory* **97** (2001), no. 2, 255–272.

[23] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer, 2012.

[24] A. M. Fink, Equilibrium in stochastic $n$-person game. *J. Sci. Hiroshima Univ., Ser. A-I* **28** (1964), 89–93.

[25] D. P. Foster and H. P. Young, Regret testing: learning to play Nash equilibrium without knowing you have an opponent. *Theor. Econ.* **1** (2006), 341–367.

[26] Y. Freund and R. E. Schapire, Adaptive game playing using multiplicative weights. *Games Econom. Behav.* **29** (1999), no. 1–2, 79–103.

[27] D. Fudenberg and D. M. Kreps, Learning mixed equilibria. *Games Econom. Behav.* **5** (1993), no. 3, 320–367.

[28] D. Fudenberg and D. K. Levine, Consistency and cautious fictitious play. *J. Econom. Dynam. Control* **19** (1995), no. 5–7, 1065–1089.

[29] D. Fudenberg and D. K. Levine, *The theory of learning in games. 2*. MIT Press, 1998.

[30] S. Gu, E. Holly, T. Lillicrap, and S. Levine, Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *IEEE international conference on robotics and automation*, pp. 3389–3396, IEEE, 2017.

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

[32] S. Hart and A. Mas-Colell, Uncoupled dynamics do not lead to Nash equilibrium. *Am. Econ. Rev.* **93** (2003), no. 5, 1830–1836.

[33] J. Hofbauer and W. H. Sandholm, On the global convergence of stochastic fictitious play. *Econometrica* **70** (2002), no. 6, 2265–2294.

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

[35] B. Huang, J. D. Lee, Z. Wang, and Z. Yang, Towards general function approximation in zero-sum Markov games. 2021, arXiv:2107.14702.

[36] C. Jin, Q. Liu, Y. Wang, and T. Yu, V-learning—a simple, efficient, decentralized algorithm for multiagent RL. 2021, arXiv:2110.14555.

[37] C. Jin, Q. Liu, and T. Yu, The power of exploiter: provable multi-agent RL in large state spaces. 2021, arXiv:2106.03352.

[38] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*. Springer, 1978.

[39] D. S. Leslie and E. J. Collins, Individual Q-learning in normal form games. *SIAM J. Control Optim.* **44** (2005), no. 2, 495–514.

[40] D. S. Leslie, S. Perkins, and Z. Xu, Best-response dynamics in zero-sum stochastic games. *J. Econom. Theory* **189** (2020).

[41] N. Littlestone and M. K. Warmuth, The weighted majority algorithm. *Inform. and Comput.* **108** (1994), no. 2, 212–261.

[42] M. L. Littman, Markov games as a framework for multi-agent reinforcement learning. In *International conference on machine learning*, pp. 157–163, Morgan Kaufmann Publishers Inc., 1994.

[43] Q. Liu, T. Yu, Y. Bai, and C. Jin, A sharp analysis of model-based reinforcement learning with self-play. In *International conference on machine learning*, pp. 7001–7010, PMLR, 2021.

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

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

[46] R. McKelvey and T. Palfrey, Quantal response equilibria for normal form games. *Games Econom. Behav.* **10** (1995), 6–38.

[47] B. McMahan, Follow-the-regularized-leader and mirror descent: equivalence theorems and $l_1$ regularization. In *International conference on artificial intelligence and statistics*, pp. 525–533, PMLR, 2011.

[48] P. Milgrom and J. Roberts, Adaptive and sophisticated learning in normal form games. *Games Econom. Behav.* **3** (1991), 82–100.

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

[50] D. Monderer and A. Sela, A $2 \times 2$ game without the fictitious play property. *Games Econom. Behav.* **14** (1996), 144–148.

[51] D. Monderer and L. Shapley, Fictitious play property for games with identical interests. *Games Econom. Behav.* **68** (1996), 258–265.

[52] D. Monderer and L. Shapley, Potential games. *Games Econom. Behav.* **14** (1996), 124–143.

[53] R. Nagel, Unraveling in guessing games: An experimental study. *Am. Econ. Rev.* **5** (1995), 1313–1326.

[54] S. Perkins and D. S. Leslie, Asynchronous stochastic approximation with differential inclusions. *Stoch. Syst.* **2** (2012), no. 2, 409–446.

[55] J. Pérolat, B. Piot, and O. Pietquin, Actor–critic fictitious play in simultaneous move multistage games. In *International conference on artificial intelligence and statistics*, pp. 919–928, PMLR, 2018.

[56] J. Pérolat, B. Scherrer, B. Piot, and O. Pietquin, Approximate dynamic programming for two-player zero-sum Markov games. In *International conference on machine learning*, pp. 1321–1329, PMLR, 2015.

[57] J. Pérolat, F. Strub, B. Piot, and O. Pietquin, Learning Nash Equilibrium for General-Sum Markov Games from Batch Data. In *International conference on artificial intelligence and statistics*, pp. 232–241, PMLR, 2017.

[58] B. T. Poljak and Y. Z. Tsypkin, Pseudogradient adaptation and training algorithms. *Autom. Remote Control* **12** (1973), 83–94.

[59] J. Robinson, An iterative method of solving a game. *Ann. of Math.* (1951), 296–301.

[60] T. Roughgarden, Intrinsic robustness of the price of anarchy. In *ACM symposium on theory of computing*, pp. 513–522, Association for Computing Machinery, 2009.

[61] T. Roughgarden, Algorithmic game theory. *Commun. ACM* **53** (2010), no. 7, 78–86.

[62] W. H. Sandholm, Preference evolution, two-speed dynamics, and rapid social change. *Rev. Econ. Dyn.* **4** (2001), no. 3, 637–679.

[63] M. O. Sayin, F. Parise, and A. Ozdaglar, Fictitious play in zero-sum stochastic games. 2020, arXiv:2010.04223.

[64] M. O. Sayin, K. Zhang, D. S. Leslie, T. Başar, and A. Ozdaglar, Decentralized Q-learning in zero-sum markov games. In *Thirty-fifth conference on neural information processing systems*, 2021.

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

[66] D. Shah, V. Somani, Q. Xie, and Z. Xu, On reinforcement learning for turn-based zero-sum Markov games. 2020, arXiv:2002.10620.

[67] S. Shalev-Shwartz, Online learning and online convex optimization. *Found. Trends Mach. Learn.* **4** (2011), no. 2, 107–194.

[68] S. Shalev-Shwartz, S. Shammah, and A. Shashua, Safe, multi-agent, reinforcement learning for autonomous driving. 2016, arXiv:1610.03295.

[69] L. S. Shapley, Stochastic games. *Proc. Natl. Acad. Sci. USA* **39** (1953), no. 10, 1095–1100.

[70] L. S. Shapley, Some topics in two-person games. *Adv. Game Theory* **52** (1964), 1–29.

[71] Y. Shoham and K. Leyton-Brown, *Multiagent systems: algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[72] A. Sidford, M. Wang, L. Yang, and Y. Ye, Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International conference on artificial intelligence and statistics*, pp. 2992–3002, PMLR, 2020.

[73] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of Go with deep neural networks and tree search. *Nature* **529** (2016), no. 7587, 484–489.

[74] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of Go without human knowledge. *Nature* **550** (2017), no. 7676, 354–359.

[75] V. Syrgkanis and E. Tardos, Composable and efficient mechanisms. In *ACM symposium on theory of computing*, pp. 211–220, Association for Computing Machinery, 2013.

[76] C. Szepesvári and M. L. Littman, A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Comput.* **11** (1999), no. 8, 2017–2060.

[77] M. Takahashi, Equilibrium points of stochastic non-cooperative $n$-person games. *J. Sci. Hiroshima Univ., Ser. A-I* **28** (1964), 95–99.

[78] M. Tan, Multi-agent reinforcement learning: independent vs. cooperative agents. In *International conference on machine learning*, pp. 330–337, PMLR, 1993.

[79] Y. Tian, Y. Wang, T. Yu, and S. Sra, Online learning in unknown Markov games. In *International conference on machine learning*, pp. 10279–10288, PMLR, 2021.

[80] J. N. Tsitsiklis, Asynchronous stochastic approximation and Q-learning. *Mach. Learn.* **16** (1994), 185–202.

[81] B. Van der Genugten, A weakened form of fictitious play in two-person zero-sum games. *Int. Game Theory Rev.* **2** (2000), no. 4, 307–328.

[82] C. J. C. H. Watkins and P. Dayan, Q-learning. *Mach. Learn.* **8** (1992), no. 3, 279–292.

[83] C.-Y. Wei, Y.-T. Hong, and C.-J. Lu, Online reinforcement learning in stochastic games. In *Advances in neural information processing systems*, pp. 4987–4997, Curran Associates, Inc., 2017.

[84] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo, Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on learning theory 134*, pp. 4259–4299, PMLR, 2021.

[85] Q. Xie, Y. Chen, Z. Wang, and Z. Yang, Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pp. 3674–3682, PMLR, 2020.

[86] Y. Yang, J. Li, and L. Peng, Multi-robot path planning based on a deep reinforcement learning DQN algorithm. *CAAI Trans. Intell. Technol.* **5** (2020), no. 3, 177–183.

[87] K. Zhang, S. Kakade, T. Başar, and L. Yang, Model-based multi-agent RL in zero-sum markov games with near-optimal sample complexity. In *Advances in neural information processing systems 33*, pp. 1166–1178, Curran Associates, Inc., 2020.

[88] K. Zhang, Z. Yang, and T. Başar, Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in neural information processing systems*, pp. 11598–11610, Curran Associates, Inc., 2019.

[89] K. Zhang, Z. Yang, and T. Başar, Multi-agent reinforcement learning: a selective overview of theories and algorithms. In *Handbook of reinforcement learning and control*, pp. 321–384, Stud. Syst. Decis. Control. Springer, 2021.

[90] Y. Zhao, Y. Tian, J. D. Lee, and S. S. Du, Provably efficient policy gradient methods for two-player zero-sum Markov games. 2021, arXiv:2102.08903.

[91] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent. In *International conference on machine learning*, pp. 928–936, PMLR, 2003.

**ASUMAN OZDAGLAR**
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, asuman@mit.edu

**MUHAMMED O. SAYIN**
Electrical and Electronics Engineering Department in Bilkent University, Ankara, Turkey, sayin@ee.bilkent.edu.tr

**KAIQING ZHANG**
LIDS and CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA, kaiqing@mit.edu