Fictitious Play and Best-Response Dynamics in Identical Interest and Zero Sum Stochastic Games

Lucas Baudin [1] **Rida Laraki** [1 2]

Abstract

This paper proposes an extension of a popular decentralized discrete-time learning procedure when repeating a static game called fictitious play (FP) (Brown, 1951; Robinson, 1951) to a dynamic model called discounted stochastic game (Shapley, 1953). Our family of discrete-time FP procedures is proven to converge to the set of stationary Nash equilibria in identical interest discounted stochastic games. This extends similar convergence results for static games (Monderer & Shapley, 1996a). We then analyze the continuous-time counterpart of our FP procedures, which include as a particular case the best-response dynamic introduced and studied by Leslie et al. (2020) in the context of zero-sum stochastic games. We prove the converge of this dynamics to stationary Nash equilibria in identical-interest and zero-sum discounted stochastic games. Thanks to stochastic approximations, we can infer from the continuous-time convergence some discrete time results such as the convergence to stationary equilibria in zero sum and team stochastic games (Holler, 2020).

1. Introduction

Learning Nash equilibria of a static game G after playing repeatedly G is a subject that has been widely studied almost since the beginning of game theory (Fudenberg & Levine, 1998; Hart & Mas-Colell, 2013; Young, 2004; Cesa-Bianchi & Lugosi, 2006; Brown, 1951; Robinson, 1951). In contrast, *learning Nash equilibria of a dynamic model such as a stochastic game* has comparatively been much less developed. Some noticeable exceptions are the two recent papers (Leslie et al., 2020; Sayin et al., 2020) which develop quite similar systems that converge to stationary equilibria in zero-sum discounted stochastic game. Our paper extends the continuous-time dynamics of (Leslie et al., 2020) to a large parametric class and to discrete-time. We show the convergence of the two systems to stationary equilibria in identical interest and also in zero-sum stochastic games.

Discounted stochastic games, introduced by Shapley (1953), model strategic interactions between players with a state variable. Thus, compared to non-stochastic games, actions that players take impact their current payoff but also a state variable that may influence their future payoff. Therefore, this class of games offers a rich framework (Neyman & Sorin, 2003) that is especially well suited for economic applications (see the survey by Amir (2003) and references therein), or engineering applications. In the latter, this belongs to the more general framework of multi-agent reinforcement learning (see Busoniu et al. (2008) for a survey).

Fink (1964) proved that when there are finitely many players and actions (our framework), any discounted stochastic game admits a stationary (mixed) Nash equilibrium (e.g. a decentralized randomized policy that depends only on the state variable). The proof is implied by the fact that a stationary equilibrium is the fixed point of an operator where, given other players stationary strategies, each player strategy is an optimal stationary policy in a Markov Decision Process. A stochastic game admits typically many other Nash equilibria. To see why, if there is only one state, it is a repeated game and a stationary equilibrium consists on playing independently the same mixed strategy. Therefore, the set of stationary Nash equilibrium payoffs of the discounted repeated game coincides with the Nash equilibria of its stage game (which is known to be a semi-algebraic set with several connected components (Laraki et al., 2019)). On the other hand, the famous folk theorem of repeated game (Aumann & Shapley, 1994; Fudenberg & Maskin, 1986; Laraki et al., 2019) shows that when the discount factor is large enough, any feasible and individually rational payoff of the stage game is a Nash equilibrium payoff of the repeated game.

Stationary Nash equilibria are the simplest of all equilibria and it is desirable to construct some natural learning procedures that are provable to converge to them. This is a challenging problem even in identical interest or zero-sum

[1] Université Paris-Dauphine - PSL, France [2] University of Liverpool, United Kingdom. Correspondence to: Lucas Baudin <lucas.baudin@dauphine.eu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

stochastic games (the subject of our paper). The reason is that the payoff function of a player in a discounted stochastic game is not linear, nor concave or quasi-concave when the players are restricted to their stationary strategies and thus, no gradient method is guaranteed to converge, even to a local Nash equilibrium (Daskalakis et al., 2021). Only recently, two extensions of the oldest of learning procedures –fictitious play (FP)– have been proposed and proven to converge to stationary equilibria but only in zero-sum stochastic games (Leslie et al., 2020; Sayin et al., 2020). Our article extends one of the two papers in both discrete and continuous time and shows that our systems converge in both regimes to the set of stationary Nash equilibria in identical-interest and zero-sum discounted stochastic games.

Our family of procedures combines classical fictitious play (a kind of myopic best reply in the local game given some prior constructed from empirical observations) with an elaborate rule in the spirit of Q-learning to update the expected continuation payoff, as in (Leslie et al., 2020; Sayin et al., 2020). Q-learning is a quite famous model-free rule which allows to learn the stationary policy in dynamic programming without even knowing the transition probabilities or the state space). The updating rule we study needs the knowledge of the model, however, as observed in the conclusion, they can be easily modified to become model-free.

The combination of FP and Q-learning is not surprising. A major trend in recent years is the advent of efficient reinforcement learning algorithms (Sutton & Barto, 2018). Q-learning (Watkins & Dayan, 1992) is one of the most successful model-free algorithms (Konda & Borkar, 1999) with numerous extensions (Hasselt, 2010; Kumar et al., 2020). On the game theory side, fictitious play (Brown, 1951; Robinson, 1951) is one of the most studied procedures of learning in games.

The idea of combining concepts from Q-learning and fictitious play emerged only recently with the work of Leslie et al. (2020); Sayin et al. (2020) in the context of zero-sum stochastic games. In these papers, a mechanism inspired by Q-learning is used to learn rewards in future states and fictitious play (or in continuous time, the related best response dynamics) is employed to choose the action in the current state taking into account the interaction with other players. Typically, players learn at a fast rate the actions of the other players in every state but compute the future rewards at a comparatively slower pace. Our work extends Leslie et al. (2020), by introducing other procedures with several time-scales, and prove their convergence to the set of Nash equilibria in identical interest and in zero-sum stochastic games, resulting in a decentralized algorithm for fully cooperative multi-agent reinforcement learning (Busoniu et al., 2008). In contrast with Leslie et al. (2020); Sayin et al. (2021), our algorithm can use the same timescale to learn actions and future rewards, which is of practical interest in implementations. Importantly, we prove convergence to global Nash equilibria and not to local Nash equilibria.

Contributions. Our contributions are as follows:

- We define procedures to play stochastic games in discrete time combining ideas from fictitious play and Q-learning. We prove their convergence to the set of stationary Nash equilibria in identical interest stochastic games. The proof is given directly in discrete time.
- We define the continuous-time counterpart of our procedures. It results in a generalization of the best-response dynamics of Leslie et al. (2020) where the relative timescales of the estimation of continuation and of other players strategies is much less restricted. We prove the convergence of our continuous-time dynamics to the set of stationary equilibria in identical interest and in zero-sum stochastic games.
- The convergence in continuous time allows us to prove some new convergence results in discrete time for zero-sum and team stochastic games (Holler, 2020) (a class which includes identical interest stochastic games).

Outline Section 2 gives initial definitions and assumptions. The next section describes related works. Then, Section 4 introduces two families of fictitious play procedures in discrete time whose convergence is shown in identical interest stochastic games. The continuous time best response dynamics are defined in Section 5 together with their convergence in identical interest and in zero-sum stochastic games. The last section uses the continuous and discrete time results to infer convergence of the discrete time procedures in zero-sum and in team stochastic games (Holler, 2020).

2. Background

Stochastic games We study dynamically interactive multi-agent systems based on the model of discounted stochastic games. Two or more players can take actions over an infinite horizon. Players' actions affect both their current stage payoffs but also the transition probability of the future state, which is the second determining factor of the total discounted average payoff. Therefore, compared to standard repeated games where there is no evolving state variable, stochastic games add a layer of complexity: a player who wants to optimize its payoff should strike a balance between the instantaneous payoff optimization and an advantageous orientation of the state. As in (Leslie et al., 2020; Sayin et al., 2020), we will focus on finite games where the state space, the action sets and the player set are finite.

Definition 2.1. Stochastic games are tuples $G = (S, I, (A^i)_{i\in I}, (r^i_s)_{i\in I, s\in S}, (P_s)_{s\in S})$ where S is the state space (a finite set), I is the finite set of players, A^i is the finite action set of player i, $A := \Pi_{i\in I} A^i$ is the set of action profiles, $r^i_s : A \to \mathbb{R}$ is the stage reward of player i, and $P_s : A \to \Delta(S)$ is the transition probability map (where $\Delta(S)$ is the set of probability distributions on S).

A stochastic game is played in discrete time as follows: it starts in an initial state $s_0 \in S$ and at every time step $n \in \mathbb{N}$, the system state is s_n. Knowing the current state s_n and the past history of states and actions $(s_0, a_0, ..., s_{n-1}, a_{n-1})$, every player $i \in I$ chooses, independently from the other players, an action $a^i_n \in A^i$ (potentially at random) and receives a stage reward of $r^i_{s_n}(a_n)$. The new state s_{n+1} is the realization of a random variable whose distribution is $P_{s_n}(a_n)$. The total payoff of such a sequence of play for player i is $(1-\delta)\sum_{k=0}^{\infty} \delta^k r^i_{s_k}(a_k)$ where $\delta \in (0, 1)$ is the discount factor (the $1-\delta$ factor is the usual normalization).

A behavioral strategy σ^i for player i is a mapping associating with each stage $n \in \mathbb{N}$, history $h_n \in (S \times A)^n$ and current state s, a mixed action $x^i_n = \sigma^i(n, h_n, s)$ in $\Delta(A^i)$. The behavioral strategy is pure if its image is always in A^i. By Kolmogorov's extension theorem, each behavioral strategy profile induces a unique probability distribution on the set of infinite histories, from which one can compute an expected discounted payoff for every player (Neyman & Sorin, 2003). A stationary strategy of player i is the simplest of behavioral strategies. It depends only on the current state s but not on the period n nor on the past history h_n. As such, a stationary strategy can be identified with an element of $\Delta(A^i)^S$ (a mixed action per state interpreted as: whenever the state is s, i plays randomly according to distribution x^i_s). The set of stationary strategy profiles is $\Pi_{i\in I}\Delta(A^i)^S$. For $y^i \in (\Delta(A^i))^S$ and $x \in \Pi_{i\in I}\Delta(A^i)^S$, we denote by (y^i, x^{-i}) the stationary strategy profile where i changes its strategy from x^i to y^i.

Extending a result of (Shapley, 1953) for zero-sum games, (Fink, 1964) proved the existence of stationary Nash equilibria in every finite discounted stochastic game, as a fixed point of the Shapley operator. Fink's characterization implies that a stationary strategy profile forms a Nash equilibrium iff there is no pure stationary profitable deviation.

Proposition 2.2 (Stationary equilibrium characterization)**.** *A stationary profile $x \in \Pi_{i\in I}\Delta(A^i)^S$ is a Nash equilibrium if and only if no pure stationary deviation is profitable: for every player i, its expected total payoff with x is greater or equal than its expected total payoff of strategy (b, x^{-i}) for any $b \in (A^i)^S$.*

Fink's result implies that (1) checking that a stationary profile is a Nash equilibrium is equivalent to solving finitely many polynomial inequalities, implying that the set of stationary strategy profiles is a semi-algebraic set (Neyman & Sorin, 2003); (2) a stationary profile is a Nash equilibrium of our stochastic games if and only if it is a Nash equilibrium of a restricted stochastic game where each player i is restricted to play a stationary strategy. While this restricted game is smooth (strategy spaces are convex/compact and the payoff functions analytic) the payoff functions are not linear, nor concave, nor quasi-concave with respect to a player own-strategy. This makes the computation of an equilibrium very hard: no learning gradient-based method is guaranteed to convergence, even to a local Nash equilibrium, and even in a zero-sum game (Daskalakis et al., 2021). But this is not any non-concave game: its dynamic nature is structured enough to allow learning to occur. Using the dynamic programming principle that stationary Nash equilibria satisfy allowed (Leslie et al., 2020; Sayin et al., 2020) to introduced some learning procedures that converge to stationary equilibria in zero-sum stochastic games. We enlarge the set of procedures in (Leslie et al., 2020) and show convergence to stationary equilibria in team and also zero-sum stochastic games. Formally, *team* stochastic games are such that the payoff functions of the players differ only by a constant (there is $r_s(\cdot) : A \to \mathbb{R}$ such that for every i and s, $r^i_s(\cdot) = r_s(\cdot) + c^i$ for a constant c^i). A special case is *identical-interest* stochastic games where all payoff functions are equal ($r^i_s(\cdot) = r_s(\cdot)$ for all i). A stochastic game is *zero-sum* when there are only two players 1 and 2 and $r^1_s + r^2_s = 0$.

When there is only one state (a repeated game) (Monderer & Shapley, 1996a) proved that fictitious play converges to Nash equilibria of the stage game (hence to stationary equilibria of the repeated game). Their proof uses extensively the multi-linearity of the (common) payoff function. Since we don't have this multi-linearity nor do we have multi-concavity or multi-quasi-concavity, we must use the structure of the problem, namely that stationary equilibria satisfy a dynamic programming principle (each player is optimizing at every state $1-\delta$ its current payoff plus δ the expectation of the continuation payoff). As such, our procedures will use fictitious play at the local static game in which the continuation payoff vector is fixed and update the continuation payoff using a Q-learning like rule. But, since learning can only occur if all states are visited infinitely often, the following assumption is needed.

Definition 2.3 (Ergodicity)**.** A stochastic game is ergodic if there is a finite time T such that for every s and s' there is a positive probability that the system starting from s is in s' after T steps for any actions taken.

Some final notations. We denote by $P_{ss'}(a)$ the probability to go to state s' starting from s with action profile $a \in A$. As players will be randomizing, let the functions $P_{ss'}$ and r^i_s be multi-linearly extended to mixed action profiles (i.e., $\Pi_{i\in I}\Delta(A^i)^S$) and are therefore I-linear. Finally, a stationary equilibrium payoff is an $S \times I$-vector that corresponds to a stationary Nash equilibrium.

3. Related Work

Fictitious Play Fictitious play (FP) is a procedure that was introduced in discrete time to play the same stage game repeatedly. It asks every player to play a best response to a prior (a mixed strategy profile) which is equal to the empirically past actions of the opponents. It was initially proposed by Brown (1951) and Robinson (1951) who proved that when the stage game is a zero-sum game and both players use FP, the empirical distribution of actions converges to the set of Nash equilibria of the stage game. A similar result has been obtained when the stage game is a potential game (Monderer & Shapley, 1996a), a $2 \times n$ game (Berger, 2005), or a mean-field game (Perrin et al., 2020). Similar convergence results have been obtained for some variants such as smooth FP, or vanishingly smooth FP, or stochastic FP (Benaïm & Faure, 2013; Fudenberg & Levine, 1995; 1998; Hofbauer & Sandholm, 2002) But FP as well as all no-regret algorithms fail to converge to Nash equilibra in all finite games (Shapley, 1964; Hart & Mas-Colell, 2003; Hofbauer & Sigmund, 1998; DeMichelis & Germano, 2000).

FP is much less studied when the stage games vary (stochastic games) and there are many possible extensions. Vrieze & Tijs (1982) used a FP like algorithm to compute the value of a stochastic game but it is not a learning procedure (as it assumes all the states are observed and updated at every stage and so it does not define a behavioral strategy). Perkins (2013) defined and studied another FP procedure in zero-sum and identical interest stochastic games relying crucially on two-timescale updates and with some restrictions on the discount factor (for zero-sum games) and the structure of equilibria (for identical interest stochastic games). More recently, Sayin et al. (2020) introduced another discrete-time variant of FP and deduced its convergence in zero-sum stochastic games from the convergence of an associated continuous-time best response dynamics explained below.

Best-response dynamics In continuous time, best-response dynamics (Harris, 1998; Matsui, 1992) is based on the same principle as fictitious play: each player adjusts its mixed action towards the best-response to the time-average mixed action of other players. For repeated –non-stochastic– games, it is the continuous-time counterpart of the discrete-time fictitious play as the stochastic approximations framework (Benaïm et al., 2005) allows to deduce from the convergence of the continuous time dynamics the convergence of the (discrete-time) FP. Recently Leslie et al. (2020) introduced an extension of the best-response dynamics to stochastis games with the continuation payoff updated at a slower pace in the spirit of Q-Learning (described below) and proved the convergence of this system to the set of stationary Nash equilibria of the underlying discounted stochastic game. However, Leslie et al. (2020) did not prove the convergence to stationary equilibria of the discrete-time counterpart of their dynamics: we will do it in this article for their dynamics and others in zero-sum and also in identical interest stochastic games. Sayin et al. (2020) defined an alternative dynamics close to the one of Leslie et al. (2020) and using a two time-scale stochastic approximation theory, they show convergence of their discrete time FP to stationary equilibria of the zero-sum discounted stochastic game. In contrast, we give a direct convergence proof of our discrete-time system in identical interest stochastic games. But the convergence of our discrete time FP procedure in zero-sum game will be deduced from the continuous time system, thanks to an elaborate stochastic approximation technique.

Q-learning Watkins (1989) introduced the Q-learning algorithm designed to control MDPs. It had a major impact and there are multiple generalizations, including offline Q-learning (Kumar et al., 2020), double Q-learning (Hasselt, 2010) or Q-learning with no-regret procedures (Kash et al., 2020). There is a wide range of applications, from robots control (Tai & Liu, 2016) to SAT solving (Kurin et al., 2020). Q-learning is a model-free algorithm, meaning that it does not require a complete specification of the environment such as the transition probability between states. A step proceeds as follows: starting from a state s_t, an action a_t is chosen and this results in a new (random) state s_{t+1} chosen by the environment while the learner gets an instantaneous payoff R_{t+1}. At every step, a Q-function Q_t defined on every state-action pair is updated towards $R_{t+1} + \delta \max_a Q_t(s_{t+1}, a)$.

Q-learning was generalized to multi-agent systems. One line of work comprises algorithms that solve at every step the stage game defined as follows: every player has actions of the current state, and payoffs are the payoff of the Q-function $Q_t(s_t, \cdot)$. Then the values of the Q-function are updated towards the values of the stage game. This leads to algorithms such as Nash-Q (Hu & Wellman, 1998; 2003) or Team-Q and Minimax-Q (Littman, 2001). For a complete survey, see (Busoniu et al., 2008) and references therein.

Combining Q-learning and fictitious play To extend FP to stochastic games, the challenge is to define and compute what is a best-response to empirical observations: given a strategy for every player, the total discounted payoff is not straightforward to compute and is non-linear with respect a player stationary strategy. To overcome this difficulty, Sayin et al. (2020) and Leslie et al. (2020) use (close but different) mechanisms similar to that of Q-learning to deal with multiple states: a Q-function (or a state-value function) defined on every state-action pair or on every state is updated during the play. The player can then consider a stage game that is built with this Q-function, which is linear with respect to its mixed actions, to play a best response. The Q-function is typically updated at a slower timescale. More precisely, the algorithm of Sayin et al. (2020) estimates $\hat{Q}_{i,s,k}(a)$ for every player i, state s, action a at time k. It is the expected payoff if players play action profile a starting from state s. Then, the procedure is to play the best-response against the belief on actions used by other players in current state, that is an element of $\arg\max_{a\in A^i} \hat{Q}_{i,s_k,k}(a, x_{s_k}^{-i})$ at time k where x_s^{-i} is the (uncorrelated) strategy of other players in state s believed by player i.

Leslie et al. (2020) introduced and studied a best-response dynamics in zero-sum stochastic games. Time is continuous, so this is not an online-learning algorithm. The proposed dynamics maintains a vector $u^i := \{u_s^i\}_{s\in S}$ for player i. It is the expected payoff starting from every state s (i.e., an estimate of the state-value function). It plays a role similar to that of the Q-function in Q-learning. Then player i plays a best-response to the stage game with payoffs composed of the instantaneous payoff and the expected later payoff, that is an element of $\arg\max_{a\in A^i}(1-\delta)r_s^i(a, x_s^{-i}) + \delta P_s(a, x_s^{-i}) \cdot u^i$.

(Leslie et al., 2020) are only concerned with two-player zero-sum stochastic games and only study a continuous-time system. We introduce a family of continuous-time dynamics that contains the dynamics of Leslie et al. (2020) as a particular case, as well as their discrete-time counterpart. We prove that our systems in both time regimes converge to stationary equilibria in identical interest and in zero-sum discounted stochastic games. See Appendix G for a summary of differences between our systems and the ones of Sayin et al. (2020) and Leslie et al. (2020).

Multi-Agent Reinforcement Learning This paper falls within the more general framework of multi-agent reinforcement learning (MARL), see Busoniu et al. (2008) for a survey. One interesting line of work is that of the complexity of MARL in the centralized case (i.e., with a central planner), see for instance (Liu et al., 2021) and references therein. However, Liu et al. (2021) rely on an oracle to choose correlated equilibria in an auxiliary game (similarly to Nash-Q) and as such can not be considered as a decentralized procedure.

4. Discrete Time : Fictitious Play Procedures

Procedures to play stochastic games Each asynchronous FP procedure, introduced in this subsection, is a behavioral strategy to play stochastic games for a player i, that is a function that provides a distribution of probability for the action a_n^i given the history of the play prior to n and the current state s_n. Formally, it is a mapping $\bigcup_{n\in\mathbb{N}}[(S\times A)^n \times S] \to \Delta(A^i)$. Such an extension of FP is called *asynchronous* because there is a unique current state (that every player observes) and actions are chosen by the players only for this state. This contrasts with another FP procedure that we also study and call *synchronous* FP. This not a behavioral strategy because there is no specific current state and players provide actions for all states at every stage, i.e., its a mapping $\bigcup_{n\in\mathbb{N}}(A^S)^n \to \Delta(A^i)^S$. It can be interpreted in various ways: either as a setting where the actual state is not known to the players, as a simulation of the system or as an algorithm, as in (Vrieze & Tijs, 1982).

Our discrete-time procedures are designed using two estimates per state: one is the empirical action that every player uses and the other one is the expected continuation payoff that a player estimates starting from this state. First, we define the two estimates, then proceed with a description of the action selection, and finally the updating rules.

Empirical actions We begin by exposing how the empirical action is computed for every state. Given a state $s \in S$ and a time step n, $s_n^\sharp$ denotes the number of times that s occurs between 0 and n i.e., $s_n^\sharp = \sharp\{k \mid 0 \le k \le n \wedge s_k = s\}$. Then the empirical action of player i in state s is defined in $\Delta(A^i)$ as:

$$x_{n+1,s}^i := \frac{1}{s_n^\sharp}\sum_{k=0}^{n} 1_{s_k=s} a_k^i = \frac{1_{s_n=s}a_n^i}{s_n^\sharp} + \frac{s_{n-1}^\sharp x_{n,s}^i}{s_n^\sharp} \quad (1)$$

with the convention that if $s_n^\sharp = 0$, then $x_{n+1,s}^i = x_{0,s}^i$ which is defined arbitrarily. Consequently, $x_{n+1,s}^i$ is equal to $x_{n,s}^i$ when s_n is not equal to s. Pure action a_k^i is seen as an element of the Euclidean space $\Delta(A^i)$.

The auxiliary Shapley game The second estimate is defined using the payoff of an auxiliary game. Given a continuation payoff vector $u \in \mathbb{R}^S$, we define (following (Shapley, 1953)) the auxiliary game parameterized by u as the one-shot game where the action set is A^i for every player i and the payoff function is $f_{s,u^i}^i(\cdot)$ where:

$$f_{s,u}^i(x) := (1-\delta)r_s^i(x) + \delta\sum_{s'\in S} P_{ss'}(x)u_{s'}$$

Fink (and Shapley) proved that stationary equilibria are the fixed point of an operator based on this auxiliary game.

Update steps α_n is the non-increasing sequence of positive update steps for the payoff estimates, $\sigma_n = \sum_{k=0}^n \alpha_k$, and starting values $u_{0,s}^i$ defined arbitrarily. We suppose that:

$$\sum_k \frac{\alpha_k}{\sigma_k} = \infty \qquad \text{(H1)}$$
$$0 < \alpha_n \le 1 \qquad \alpha_{n+1} \le \alpha_n$$

Payoff estimates Players estimate the continuation payoff in a vector $u_n^i \in \mathbb{R}^S$. Values of this vector are written $u_{n,s}^i$ for state s, at step n for player i. At every step n, the estimator is defined as:

$$\begin{aligned} u^i_{n+1,s} :=& \frac{1}{\sigma_n} \sum_{k=0}^{n} \alpha_k f^i_{s,u^i_k}(x_{k,s}) \\ =& \frac{\sigma_{n-1}}{\sigma_n} u^i_{n,s} + \frac{\alpha_n}{\sigma_n} f^i_{s,u^i_n}(x_{n,s}) \end{aligned}$$

Notice that in an identical interest stochastic game, r^i_s does not depend on i and as a consequence, $u^i_{n,s}$ does not depend on i either. The same holds for zero-sum games because the payoff of player 2 is the negative of that of player 1 and it is sufficient to follow player 1's payoff. As such, we omit the superscript for u^i in the rest of the paper.

Estimator $u_{n,s}$ can be seen as a mean where recent values of the expected payoffs $f_{s,u_n}(x_{n,s})$ are given less weight than oldest values. However, if the sequence $f_{s,u_n}(x_{n,s})$ is stationary, $u_{n,s}$ will ultimately converge to the same limit as $f_{s,u_n}(x_{n,s})$. A similar idea of a fast and a slow update rates is used in (Leslie et al., 2020; Perkins, 2013; Konda & Borkar, 1999; Sayin et al., 2020).

Remark: If $\alpha_n = 1$, then $\sigma_n = n$. Thus, the update rates of $x^i_{n,s}$ and $u_{n,s}$ are the same and the hypothesis (H1) holds. It also holds for $\alpha_n = \frac{1}{\log n}$ where $\sigma_n = \log \log n$. In this case, the update rate of $x^i_{n,s}$ is much faster than the one of $u_{n,s}$ and, as will be seen, this is the discrete-time analog of the continuous time dynamics in (Leslie et al., 2020).

Action selection We can now define the action selection of our FP procedures. It is an extension of the classical FP procedure. For repeated games, FP is defined as a behavioral strategy where at every stage, every player takes a best response against the empirical action of the opponents up to that stage. For stochastic games, we define fictitious play as a best response in the auxiliary Shapley game parameterized by a given continuation payoff u^i_n, that is for every n:

$$a^i_{n,s} \in \text{br}^i_{s,u_n}(x^{-i}_{n,s}) := \arg\max_{y \in A^i} f_{s,u_n}(y, x^{-i}_{n,s})$$

where $s = s_n$. When there are several best responses, our convergence results are independent on the selection rule.

Now we can define precisely our two FP procedures.

Asynchronous FP

$$\begin{cases} u_{n+1,s} - u_{n,s} = \dfrac{\alpha_n}{\sigma_n} \left(f_{s,u_n}(x_{n,s}) - u_{n,s} \right) \\ x^i_{n+1,s} - x^i_{n,s} = \dfrac{\mathbb{1}_{s_n = s}}{s^\sharp_n} \left(a^i_n - x^i_{n,s} \right) \\ a^i_n \in \text{br}^i_{s,u_n}(x^{-i}_{n,s}) \end{cases} \tag{AFP}$$

Estimates of continuation payoff $u_{n,s}$ are updated towards the estimated payoff in the auxiliary game $f^i_{s,u^i_n}(x_{n,s})$ for all states s at every step. Empirical actions $x^i_{n,s}$ are updated only for the current state (notice the indicator function) in the direction of the action played a^i_n. This is an incremental version of Eq. (1).

Therefore, this defines a behavioral strategy because the only information needed to update the system variables is the actions played in the current state, as the second equation shows.

Synchronous FP To get convergence in non ergodic stochastic games, we now define a version with synchronous updates on every state. This is an algorithm but not a behavioral strategy and thus, not a learning rule. Synchronous fictitious play is defined as follows:

$$\begin{cases} u_{n+1,s} - u_{n,s} = \dfrac{\alpha_n}{\sigma_n} \left(f_{s,u_n}(x_{n,s}) - u_{n,s} \right) \\ x^i_{n+1,s} - x^i_{n,s} = \dfrac{1}{n} \left(a^i_{n,s} - x^i_{n,s} \right) \\ a^i_{n,s} \in \text{br}^i_{s,u_n}(x^{-i}_{n,s}) \end{cases} \tag{SFP}$$

Contrary to AFP, an action is provided for every state at every step, as if the state was unknown. This allows to update $x^i_{n,s}$ and $u_{n,s}$ synchronously (but with at a rate that may be different).

Incremental updates In both FP procedures, $x^i_{n,s}$ and $u_{n,s}$ can be computed *via* incremental updates. This will enable us to make the link with the continuous version (Section 6). Moreover, it shows that for a machine implementation, the procedure only needs constant memory instead of storing the whole history.

Theorem 4.1 (Convergence of FP rules in identical interest stochastic games)**.** *Under H1, procedures SFP and AFP almost surely converge to the set of stationary Nash equilibria in identical interest ergodic discounted stochastic games. The convergence hold also in non ergodic games for SFP.*

As in (Monderer & Shapley, 1996b), our discrete-time proof for identical interest stochastic games is direct and is not derived from some associated continuous-time system.

Proof of Theorem 4.1 (sketch). The central idea is to show that the gap between $f_{s,u_n}(x_{n,s})$ and $u_{n,s}$ is lower-bounded by a sequence whose sum converge. This is possible because $f_{s,u_n}(x_{n,s})$ is mostly non-decreasing (but for the synchronization error of the players that optimize this function which is in $\frac{1}{n^2}$). This is similar to the proof continuous time (where the payoff function is a Lyapunov function of the system). Another key point is that $u_{n,s}$ moves towards f_{s,u_n} at a rate no faster than the updates of $x_{n,s}$. This lower bound is used to prove the convergence of $u_{n,s}$, and the convergence to the set of stationary Nash equilibria follows.

See Appendix A for the complete proof. □

5. Continuous Time: Best-Response Dynamics

This section extends and studies the best-response dynamics introduced and studied in zero-sum stochastic games by Leslie et al.. We generalize their updating rates and prove that all the extended dynamics converge to stationary equilibria in identical interest and in zero-sum stochastic games. These dynamics are the continuous-time counterpart of AFP and SFP as shown in the next section.

As in discrete-time, there are two sets of variables: $\{u_s^i, x_s^i\}_{s\in S, i\in I}$. These variables may have different update rates, and we suppose there is a function $\alpha : \mathbb{R}^+ \to \mathbb{R}^{+*}$ to express the update rates of variables u_s^i. Function α is continuous and non-increasing. We make the following additional assumption on α:

$$\int_0^t \alpha(y)\, dy \xrightarrow[t\to\infty]{} +\infty \qquad \text{(H2)}$$
$$\alpha(t) \geq 0 \text{ and } \alpha \text{ is non-increasing}$$

Synchronous Dynamics As in AFP, in the next dynamics, variables of all states are updated at the same time.

For $t \geq 0$ and every state s and player i, synchronous best-reply dynamics (SBRD) is defined as:

$$\begin{cases} \dot{u}_s^i(t) = \alpha(t)\left(f^i_{s,u^i(t)}(x_s(t)) - u_s^i(t)\right) \\ \dot{x}_s^i(t) \in \mathrm{br}^i_{s,u^i(t)}(x_s(t)) - x_s^i(t) \end{cases} \qquad \text{(SBRD)}$$

where $\mathrm{br}^i_{s,u^i(t)}(x_s(t)) := \mathrm{argmax}_{a\in A^i} f^i_{s,u^i(t)}(a, x_s^{-i}(t))$ (i.e., it is a best response to the auxiliary Shapley game). This action is used as an element of the Euclidean space $\Delta(A^i)$. Vector $u^i(t)$ denotes $\{u_s^i(t)\}_{s\in S}$.

Remark: This is a generalization of the definition of Leslie et al. who studied the case $\alpha(t) = \frac{1}{t+1}$. Replacing $f^i_{s,u^i(t)}(x_s(t))$ by the maximum over actions, that is $\max_{a\in A^i} f^i_{s,u^i(t)}(a, x_s^{-i}(t))$ is an alternative that would be closer to the system outlined by Sayin et al. and Q-learning in general. It could be an interesting system to study but as noted by Sayin et al., this would result in $u_s^i(t)$ to be different for two players even if the game is zero-sum or identical interest, which poses more theoretical challenges.

Differential inclusion SBRD classically admits a (typically non-unique) solution (Aubin & Cellina, 1984; Benaïm et al., 2005). Indeed, one can rewrite it as $\frac{dy}{dt} \in F(t, y)$ where y is a vector with every u_s^i, x_s^i and F is a closed set-valued map, with non-empty, convex values. Furthermore, as shown in Lemma B.1 of Appendix B, values are bounded, so the solution is defined on $\mathbb{R}^+$ (Aubin & Cellina, 1984, p. 97).

In identical interest games, $r_s^i = r_s$ for every player i. Therefore, for every s, u_s^i and f^i_{s,u^i} do not depend on i (when initial values are equal), hence we omit the superscript i in our statements. It is similar for zero-sum games.

Asynchronous Dynamics We now provide results regarding the convergence of asynchronous systems. In this system, the expected continuation payoff starting from state s is always updated at the same rate but the empirical action is not. It is defined as follows:

$$\begin{cases} \dot{u}_s(t) = \alpha(t)\left(f_{s,u(t)}(x_s(t)) - u_s(t)\right) \\ \dot{x}_s^i(t) \in \beta_s(t)\left(\mathrm{br}^i_{s,u(t)}(x_s^{-i}(t)) - x_s^i(t)\right) \\ \beta_s(t) \in [\beta_-, 1] \end{cases} \qquad \text{(ABRD)}$$

where $\beta_- \in (0, 1]$.

Value $\beta_s(t)$ is the update rate for state s at time t. If only one state was updated at every time point, then we would have $\beta_s(t)$ equal to 0 but in one state where it would be equal to 1. If the game is ergodic, then on average every state is reached a strictly positive proportion of the time, $> \beta_-$. Next section will show in the ergodic case, that this system is formally linked to the AFP procedure.

Theorem 5.1 (Convergence of ABRD and SBRD in identical interest stochastic games)**.** *Let $\{u_s, \beta_s, x_s^i\}_{s\in S, i\in I}$ be a solution of ABRD. Under H2, there is $\Phi \in \mathbb{R}^{|S|}$ such that:*

- *for all s, $f_{s,u(t)}(x_s(t)) \xrightarrow[t\to\infty]{} \Phi_s$ and $u(t) \xrightarrow[t\to\infty]{} \Phi$*
- *Φ is a stationary Nash equilibrium payoff*
- *$\{x_s(t)\}_{s\in S}$ converges to the set of stationary Nash equilibria with payoff Φ*

A sketch of the proof is provided below. A comprehensive proof with technical lemmas is provided in Appendix B.

Sketch of proof. We define, for $s \in S$:

$$\Gamma_s(t) := f_{s,u(t)}(x_s(t))$$
$$\Delta_s^i(t) := \max_{y\in A^i} f_{s,u(t)}(y, x_s^{-i}(t)) - f_{s,u(t)}(x_s(t)) \geq 0$$

We are going to lower bound $\Gamma_s(t) - u_s(t)$ for every s so as the differential of u_s is lower-bounded by an integrable function. This guarantees that, as u_s is bounded (see Lemma B.1), it converges. We then show that for every player i, $\Delta_s^i(t) \to 0$ and finish the proof of the theorems by studying convergence of $\Gamma_s(t)$ and the limit set of $x_s(t)$.

□

Theorem 5.2 (Convergence of ABRD in zero-sum stochastic games)**.** *Let $\{u_s, \beta_s, x_s^i\}_{s\in S, i\in I}$ be a solution of ABRD. There exists a constant $A > 0$ (which only depends on δ and r_s) such that if $\alpha^\star > \lim_{t\to\infty} \alpha(t)$, then, under H2:*

- *for all s, $\limsup_{t\to\infty} |f_{s,u(t)}(x_s(t)) - u(t)| \leq A\alpha^\star$*
- *$\{x_s(t)\}_{s\in S}$ converges to the set of stationary Nash $A\alpha^\star-$equilibria as $t \to \infty$.*

The proof is in Appendix C. Note that if $\alpha(t) \to 0$, then $\alpha^\star$ can be chosen arbitrarily close to 0 which is the case in (Leslie et al., 2020) ($\alpha(t) = t + 1$). Hence this is an extension of (Leslie et al., 2020).

6. Linking Continuous and Discrete Systems

This section uses the continuous time results of the previous section to deduce the convergence of the discrete-time procedures in zero-sum stochastic games (which is not covered by (Leslie et al., 2020)) but also in team stochastic games (defined below). Any identical interest stochastic game is a team stochastic game but not the reverse.

Zero-sum games We can discretize the continuous model using an extension of the stochastic approximation framework (see details in Appendix D and then using an algorithm with doubling trick (standard in RL procedures). Note that the doubling trick trigger $T(\alpha)$ can be computed, this is explained in Appendix D.

Algorithm 1 FP with Doubling Trick for Zero-Sum Games

```
$\alpha, x, u \leftarrow 1, x_0, u_0$
loop
    $x_{n+1,s_n} \leftarrow x_{n,s_n} + \frac{1}{n+1}\left(a_n - x_{n,s_n}\right)$
    $\forall s, u_{n+1,s} \leftarrow u_{n,s} + \frac{\alpha}{\sigma_n}(f_{s,u_n}(x_{n,s} - u_{n,s})$
    Choose $a^i_{n+1} \in \text{br}^i_{s,u_n}(x^{-i}_{n+1,s})$
    if $n > T(\alpha)$ then
        $\alpha \leftarrow \alpha/2$
    end if
end loop
```

Combining the stochastic approximation framework and Theorem 5.2 guarantees the convergence of this algorithm. A detailed proof is in Appendix D.

Theorem 6.1 (Convergence of FP with doubling trick in zero-sum stochastic games)**.** *Under H1, procedures SFP and AFP with the doubling trick as specified in Algorithm 1 almost surely converge to the set of stationary Nash equilibria in zero-sum ergodic discounted stochastic games. The convergence holds also in non ergodic games for SFP.*

Different Priors and Team Stochastic Games Holler (2020) proved that in exact global-potential stochastic games, players are divided into two categories: either they do not influence the transition or they have the same payoff function up to a constant. This last class is called *team stochastic games*. While the proof of convergence of FP in discrete-time is not straightforward in team games, it can be studied using the continuous time system and stochastic approximations techniques. Similarly, if players have different priors on continuations.

Theorem 6.2 (Convergence of FP with different priors in team stochastic games)**.** *If all players use a FP procedure as defined in Section 4 with priors on the continuations that may be different (i.e., $u^i_s(0)$ may not be equal to $u^j_s(0)$ for two players i and j and a state s), then the average actions $x_{s,n}$ and vectors u^i (for every player i) converge respectively to the set of stationary Nash equilibria and the corresponding continuation payoffs.*

Sketch of proof. We can define a Lyapunov function on the continuous-time system, yielding conditions on chain transitive sets. Details can be found in D.6. □

7. Conclusion

We defined a number of continuous and discrete time systems to learn stationary equilibria in stochastic games. They combine ideas from fictitious play and Q-learning and are extensions of a continuous-time system of (Leslie et al., 2020) who proved its convergence to stationary equilibria in zero-sum stochastic games. We prove their convergence to stationary equilibria in continuous time but also in discrete time; in zero sum but also in identical interest discounted stochastic games. An open problem is to show the convergence of the procedures of (Sayin et al., 2021) in identical interest and team stochastic game. The main difficulty relies on the fact that their updating rule does not preserve the identical interest objective along the trajectory.

Another interesting direction is the speed of convergence. As outlined in the proof, there are bounds for zero-sum stochastic games but none for identical interest ones. To the best of our knowledge, no results are known even in non-stochastic games (Monderer & Shapley, 1996a).

An interesting extension would be limiting average stochastic games. This could be achieved by increasing in AFP the discount factor δ_n from stage to stage to 1. Another nice extension would be to have a model free algorithm, that is an updating rule that does not use the knowledge of the probability distribution. A simple way to adapt AFP is by letting the players explore with small probability and replace in the up-dating rule of AFP the probability transition by its empirical estimation. When the game is ergodic, after some period $T(\varepsilon)$ the estimated transition probability will be close to the real transition with probability at least $1 - \varepsilon$. Consequently, under ergodicity and (H1), this modification of AFP converges to stationary equilibria in identical-interest and zero-sum stochastic games.

More tricky is to construct a learning procedure that converges to a stationary equilibrium when the players do not observe other players past actions but only their own past actions and the current state. Even when there is only one state (a repeated game) FP and all its variants fail because there is no way to form a belief about the opponents without observing their actions. A class of procedures that converge to Nash equilibria of the stage game in zero sum and identical interest repeated games are non-regret algorithms (Blum & Mansour, 2005; Hofbauer & Sandholm, 2002) with some exploration to be able to estimate what would the payoff be if a player has played differently (we are in a bandit setting).

But regret is not well defined in stochastic games even if players observe the past actions of the opponents (Mannor & Shimkin, 2003).

A last interesting question is: what happens if we let each player use simple Q-learning? This is very unstable in some repeated games (RG) (Wunder et al., 2010). In others RG, simulations in a repeated pricing game (a kind of repeated prisoner dilemma) show that Q-learning does not converge to the stationary Nash (Calvano et al., 2020) (i.e., not competitive pricing which would be defection at every stage) but to Pareto Nash equilibrium of the RG (to collusion, that is a cooperative equilibrium, similar to Tit for Tat). One may wonder if it is possible to construct a Q-learning like procedure which converges to Pareto optimal equilibria in every repeated/stochastic game. This was our motivating question because the Tit for Tat equilibrium in a RG is a stationary equilibrium in the auxiliary stochastic game where the current state is the last action profile in the RG.

8. Acknowledgements

We are very grateful to Guillaume Vigeral and Laurent Gourvès for their help throughout this work and also to Sylvain Sorin for comments on earlier versions of this article. We also thank the anonymous reviewers for their helpful suggestions.

References

Amir, R. Stochastic Games in Economics and Related Fields: An Overview. In Neyman, A. and Sorin, S. (eds.), *Stochastic Games and Applications*, NATO Science Series, pp. 455–470, Dordrecht, 2003. Springer Netherlands. ISBN 978-94-010-0189-2. doi: 10.1007/978-94-010-0189-2_30.

Aubin, J.-P. and Cellina, A. *Differential Inclusions: Set-Valued Maps and Viability Theory*, volume 264 of *Grundlehren Der Mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1984. ISBN 978-3-642-69514-8 978-3-642-69512-4. doi: 10.1007/978-3-642-69512-4.

Aumann, R. J. and Shapley, L. S. Long-Term Competition—A Game-Theoretic Analysis. In Megiddo, N. (ed.), *Essays in Game Theory: In Honor of Michael Maschler*, pp. 1–15. Springer, New York, NY, 1994. ISBN 978-1-4612-2648-2. doi: 10.1007/978-1-4612-2648-2_1.

Benaïm, M. and Faure, M. Consistency of Vanishingly Smooth Fictitious Play. *Mathematics of Operations Research*, 38(3):437–450, August 2013. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.1120.0568.

Benaïm, M., Hofbauer, J., and Sorin, S. Stochastic Approximations and Differential Inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, January 2005. ISSN 0363-0129, 1095-7138. doi: 10.1137/S0363012904439301.

Berger, U. Fictitious play in 2×n games. *Journal of Economic Theory*, 120(2):139–154, February 2005. ISSN 00220531. doi: 10.1016/j.jet.2004.02.003.

Blum, A. and Mansour, Y. From External to Internal Regret. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Auer, P., and Meir, R. (eds.), *Learning Theory*, volume 3559, pp. 621–636. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-26556-6 978-3-540-31892-7. doi: 10.1007/11503415_42.

Brown, G. W. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

Busoniu, L., Babuska, R., and De Schutter, B. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, March 2008. ISSN 1094-6977, 1558-2442. doi: 10.1109/TSMCC.2007.913919.

Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review*, 110(10):3267–3297, October 2020. ISSN 0002-8282. doi: 10.1257/aer.20190623.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge; New York, 2006. ISBN 978-0-511-19178-7 978-0-511-54692-1 978-0-511-18995-1 978-0-511-19059-9 978-0-511-19091-9 978-0-511-19131-2 978-0-521-84108-5.

Daskalakis, C., Skoulakis, S., and Zampetakis, M. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1466–1478, Virtual Italy, June 2021. ACM. ISBN 978-1-4503-8053-9. doi: 10.1145/3406325.3451125.

DeMichelis, S. and Germano, F. On the indices of zeros of nash fields. *Journal of Economic Theory*, 94(2):192–217, 2000. ISSN 0022-0531. doi: 10.1006/jeth.2000.2669.

Fink, A. M. Equilibrium in a stochastic n-person game. *Hiroshima Mathematical Journal*, 28(1), January 1964. ISSN 0018-2079. doi: 10.32917/hmj/1206139508.

Fudenberg, D. and Levine, D. K. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7):1065–1089, July 1995. ISSN 01651889. doi: 10.1016/0165-1889(94)00819-4.

Fudenberg, D. and Levine, D. K. *The Theory of Learning in Games*. Number 2 in MIT Press Series on Economic Learning and Social Evolution. MIT Press, Cambridge, Mass, 1998. ISBN 978-0-262-06194-0.

Fudenberg, D. and Maskin, E. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica : journal of the Econometric Society*, 54 (3):533–554, 1986. ISSN 00129682, 14680262.

Harris, C. On the Rate of Convergence of Continuous-Time Fictitious Play. *Games and Economic Behavior*, 22(2):238–259, February 1998. ISSN 08998256. doi: 10.1006/game.1997.0582.

Hart, S. and Mas-Colell, A. Uncoupled dynamics do not lead to nash equilibrium. *The American Economic Review*, 93(5):1830–1836, 2003. ISSN 00028282.

Hart, S. and Mas-Colell, A. *Simple Adaptive Strategies: From Regret-Matching to Uncoupled Dynamics*. Number v. 4 in World Scientific Series in Economic Theory. World Scientific, New Jersey, 2013. ISBN 978-981-4390-69-9.

Hasselt, H. Double q-learning. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Hofbauer, J. and Sandholm, W. H. On the Global Convergence of Stochastic Fictitious Play. *Econometrica*, 70(6): 2265–2294, 2002. ISSN 0012-9682.

Hofbauer, J. and Sigmund, K. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998. doi: 10.1017/CBO9781139173179.

Hofbauer, J., Sorin, S., ,Department of Mathematics, University College London, London WC1E 6BT, and ,Laboratoire d'Económetrie, Ecole Polytechnique, 1 rue Descartes, 75005 Paris. Best response dynamics for continuous zero–sum games. *Discrete & Continuous Dynamical Systems - B*, 6(1):215–224, 2006. ISSN 1553-524X. doi: 10.3934/dcdsb.2006.6.215.

Holler, J. E. *Learning Dynamics and Reinforcement in Stochastic Games*. PhD thesis, University of Michigan, 2020.

Hu, J. and Wellman, M. P. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 242–250, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

Hu, J. and Wellman, M. P. Nash q-learning for general-sum stochastic games. *The Journal of Machine Learning Research*, 4(null):1039–1069, December 2003. ISSN 1532-4435.

Kash, I. A., Sullins, M., and Hofmann, K. Combining no-regret and q-learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pp. 593–601, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-7518-4.

Konda, V. R. and Borkar, V. S. Actor-Critic–Type Learning Algorithms for Markov Decision Processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, January 1999. ISSN 0363-0129, 1095-7138. doi: 10.1137/S036301299731669X.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020.

Kurin, V., Godil, S., Whiteson, S., and Catanzaro, B. Can q-learning with graph networks learn a generalizable branching heuristic for a SAT solver? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9608–9621. Curran Associates, Inc., 2020.

Laraki, R., Renault, J., and Sorin, S. *Mathematical Foundations of Game Theory*. Universitext. Springer International Publishing, Cham, 2019. ISBN 978-3-030-26645-5 978-3-030-26646-2. doi: 10.1007/978-3-030-26646-2.

Leslie, D. S., Perkins, S., and Xu, Z. Best-response dynamics in zero-sum stochastic games. *Journal of Economic Theory*, 189:105095, September 2020. ISSN 00220531. doi: 10.1016/j.jet.2020.105095.

Littman, M. L. Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2(1): 55–66, April 2001. ISSN 13890417. doi: 10.1016/S1389-0417(01)00015-8.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A Sharp Analysis of Model-based Reinforcement Learning with Self-Play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Mannor, S. and Shimkin, N. The Empirical Bayes Envelope and Regret Minimization in Competitive Markov Decision Processes. *Mathematics of Operations Research*, 28 (2):327–345, May 2003. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.28.2.327.14483.

Matsui, A. Best response dynamics and socially stable strategies. *Journal of Economic Theory*, 57(2):343–362, August 1992. ISSN 0022-0531. doi: 10.1016/0022-0531(92) 90040-O.

Monderer, D. and Shapley, L. S. Fictitious Play Property for Games with Identical Interests. *Journal of Economic Theory*, 68(1):258–265, January 1996a. ISSN 00220531. doi: 10.1006/jeth.1996.0014.

Monderer, D. and Shapley, L. S. Potential Games. *Games and Economic Behavior*, 14(1):124–143, May 1996b. ISSN 0899-8256. doi: 10.1006/game.1996.0044.

Neyman, A. and Sorin, S. (eds.). *Stochastic Games and Applications*. Springer Netherlands, Dordrecht, 2003. ISBN 978-1-4020-1493-2 978-94-010-0189-2. doi: 10.1007/978-94-010-0189-2.

Perkins, S. *Advanced Stochastic Approximation Frameworks and Their Applications*. PhD thesis, University of Bristol, September 2013.

Perkins, S. and Leslie, D. S. Asynchronous Stochastic Approximation with Differential Inclusions. *Stochastic Systems*, 2(2):409–446, December 2012. ISSN 1946-5238, 1946-5238. doi: 10.1287/11-SSY056.

Perrin, S., Perolat, J., Lauriere, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13199–13213. Curran Associates, Inc., 2020.

Robinson, J. An Iterative Method of Solving a Game. *The Annals of Mathematics*, 54(2):296, September 1951. ISSN 0003486X. doi: 10.2307/1969530.

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

Sayin, M. O., Zhang, K., Leslie, D. S., Basar, T., and Ozdaglar, A. Decentralized Q-Learning in Zero-sum Markov Games. *arXiv:2106.02748 [cs, math]*, June 2021.

Shapley, L. S. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, October 1953. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.39.10.1095.

Shapley, L. S. Some Topics in Two-Person Games. In Dresher, M., Shapley, L. S., and Tucker, A. W. (eds.), *Advances in Game Theory. (AM-52), Volume 52*, pp. 1–28. Princeton University Press, 1964. doi: doi:10.1515/9781400882014-002.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

Tai, L. and Liu, M. A robot exploration strategy based on Q-learning network. In *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pp. 57–62, Angkor Wat, June 2016. IEEE. ISBN 978-1-4673-8959-4. doi: 10.1109/RCAR.2016.7784001.

Vrieze, O. J. and Tijs, S. H. Fictitious play applied to sequences of games and discounted stochastic games. *International Journal of Game Theory*, 11(2):71–85, June 1982. ISSN 0020-7276, 1432-1270. doi: 10.1007/BF01769064.

Watkins, C. J. and Dayan, P. Technical Note: Q-Learning. *Machine Learning*, 8(3/4):279–292, 1992. ISSN 08856125. doi: 10.1023/A:1022676722315.

Watkins, C. J. C. H. *Learning from Delayed Rewards*. PhD thesis, King's College, Oxford, 1989.

Wunder, M., Littman, M., and Babes, M. Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 1167–1174, Madison, WI, USA, June 2010. Omnipress. ISBN 978-1-60558-907-7.

Young, H. P. *Strategic Learning and Its Limits*. Oxford University Press, Oxford [England] ; New York, 2004. ISBN 978-0-19-926918-1.

A. Convergence of Discrete-Time Fictitious Play in Identical Interest Stochastic Games

In this section, we prove that systems SFP and AFP converge to a stationary Nash equilibrium. The proofs for the two systems are similar, except for the last part about the convergence of the empirical actions. Therefore, we write the first, identical part, only for AFP (the more complex system), and give the two proof in the last part.

Recall system AFP:

$$\begin{cases} u_{n+1,s} - u_{n,s} = \dfrac{\alpha_n}{\sigma_n}\left(f_{s,u_n}\left(x_{n,s}\right) - u_{n,s}\right) \\ x^i_{n+1,s} - x^i_{n,s} = 1_{s=s_n}\dfrac{a^i_n - x^i_{n,s}}{n} \\ a^i_n \in \text{br}^i_{s_n,u_n}(x^{-i}_{n,s}) \\ \sigma_n = \displaystyle\sum_{k=1}^{n} \alpha_k \end{cases} \tag{AFP}$$

Under these hypothesis on α_n:

$$\begin{aligned} \sum_k \frac{\alpha_k}{\sigma_k} &= \infty \\ \alpha_n &\leq 1 \\ \alpha_{n+1} &\leq \alpha_n \end{aligned} \tag{H1}$$

Note that this is satisfied for $\alpha_n = 1$ (single timescale, autonomous case) or $\alpha_n = \frac{1}{n}$ (u is updated significantly slower than x).

We are going to show that under H1, $u_{n,s}$ and $f_{s,u_n}\left(x_{n,s}\right)$ converge to the same equilibrium payoff for every s in an (ergodic for AFP), identical interest stochastic game. This implies that $x_{n,s}$ converges to a stationary Nash equilibrium.

Proof. Denote $\Gamma_{n,s} := f_{s,u_n}\left(x_{n,s}\right),\ w_n := \min_{s\in S}\Gamma_{n,s} - u_{n,s}$ and $s^-_n \in \arg\min_{s\in S}\Gamma_{n,s} - u_{n,s}$ so as $w_n = \Gamma_{n,s^-_n} - u_{n,s^-_n}$.

The energy of the system w_n converges We bound the changes in w_n:

$$\begin{aligned} w_{n+1} - w_n &= \Gamma_{n+1,s^-_{n+1}} - u_{n+1,s^-_{n+1}} - (\Gamma_{n,s^-_n} - u_{n,s^-_n}) \\ &\geq \Gamma_{n+1,s^-_{n+1}} - u_{n+1,s^-_{n+1}} - (\Gamma_{n,s^-_{n+1}} - u_{n,s^-_{n+1}}) \\ &= \Gamma_{n+1,s^-_{n+1}} - \Gamma_{n,s^-_{n+1}} - (u_{n+1,s^-_{n+1}} - u_{n,s^-_{n+1}}) \\ &= \Gamma_{n+1,s^-_{n+1}} - \Gamma_{n,s^-_{n+1}} - \frac{\alpha_k}{\sigma_k}\left(\Gamma_{n,s^-_{n+1}} - u_{n,s^-_{n+1}}\right) \end{aligned} \tag{2}$$

Now, there exists C (independent of n) such that $|\Gamma_{n,s^-_{n+1}} - u_{n,s^-_{n+1}}| - |\Gamma_{n,s^-_n} - u_{n,s^-_n}| < \frac{C}{n}$ (because for every s, $\Gamma_{n,s} - u_{n,s}$ changes of at most $\frac{C}{n}$ between n and $n+1$, so this is true for the minimum as well).

As a consequence, continuing (2):

$$\begin{aligned}
w_{n+1} - w_n &\geq \Gamma_{n+1,s^-_{n+1}} - \Gamma_{n,s^-_{n+1}} - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{n\sigma_n} \\
&\geq f_{s^-_{n+1},u_{n+1}}\left(x_{n+1,s^-_{n+1}}\right) - f_{s^-_{n+1},u_n}\left(x_{n+1,s^-_{n+1}}\right) \\
&\quad + f_{s^-_{n+1},u_n}\left(x_{n+1,s^-_{n+1}}\right) - f_{s^-_{n+1},u_n}\left(x_{n,s^-_{n+1}}\right) - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{n\sigma_n} \\
&\geq \delta \sum_{s' \in S} P_{s^-_{n+1},s'}(x_{n+1,s^-_{n+1}})\left(u_{n+1,s'} - u_{n,s'}\right) \\
&\quad + f_{s^-_{n+1},u_n}\left(x_{n+1,s^-_{n+1}}\right) - f_{s^-_{n+1},u_n}\left(x_{n,s^-_{n+1}}\right) - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{n\sigma_n} \\
&\geq \delta \frac{w_n}{\sigma_n \alpha_n} + f_{s^-_{n+1},u_n}\left(x_{n+1,s^-_{n+1}}\right) - f_{s^-_{n+1},u_n}\left(x_{n,s^-_{n+1}}\right) - \frac{\alpha_n w_n}{\sigma_n} - \frac{\alpha_n C}{n\sigma_n} \\
&\geq (\delta - 1)\frac{\alpha_n w_n}{\sigma_n} + f_{s^-_{n+1},u_n}\left(x_{n+1,s^-_{n+1}}\right) - f_{s^-_{n+1},u_n}\left(x_{n,s^-_{n+1}}\right) - \frac{\alpha_n C}{n\sigma_n}
\end{aligned}$$

The first order expansion of $f_{s^-_{n+1},u_n}\left(x_{n+1,s^-_{n+1}}\right)$ for AFP:

$$f_{s^-_{n+1},u_n}\left(x_{n+1,s^-_{n+1}}\right) = f_{s^-_{n+1},u_n}\left(x_{n,s^-_{n+1}}\right) + \sum_{i \in I} \frac{1_{s=s_n}}{n}\left(f_{s^-_{n+1},u_{n,s}}\left(a^i_n, x^{-i}_{n,s}\right) - f_{s^-_{n+1},u_{n,s}}\left(x_{n,s}\right)\right) + O\left(\frac{1}{n^2}\right) \quad (3)$$

The expansion (3) would be the same for SFP except for the indicator $1_{s=s_n}$ which would disappear.

In any case, the first order term is positive (because a^i_n is a best-response in the auxiliary game), therefore there exists $D > 0$ such that:

$$w_{n+1} - w_n \geq (\delta - 1)\frac{\alpha_n w_n}{\sigma_n} - \frac{D}{n^2} - \frac{\alpha_n C}{n\sigma_n}$$

Then, using Lemma F.1, for $n > m$:

$$w_n \geq w_m \Pi^n_{k=m}(1 + \frac{\delta - 1}{k}) - \sum_{k=m}^{n}\left[\frac{D}{k^2} + \frac{\alpha_k C}{k\sigma_k}\right] \geq E\Pi^n_{k=m}(1 + \frac{\delta - 1}{k}) - \sum_{k=m}^{\infty}\left[\frac{D}{k^2} + \frac{C}{k^2}\right]$$

for some $E > 0$ (independent of m because w_n is bounded). The last inequality is obtained using Lemma F.2. The right term goes to 0 as the rest of a convergent sum. Furthermore, the left term goes to 0 when n goes to ∞, so $w_n \to 0$.

The continuation payoffs u_n converge Sequence w_n can be lower bounded more precisely: $\sum_{k=m}^{\infty} \frac{D+C}{k^2} = \Omega\left(\frac{1}{m}\right)$ and $\Pi^n_{k=m}(1 + \frac{\delta-1}{k}) = \Omega\left(\left(\frac{m}{n}\right)^{\delta-1}\right)$, so with $m = [\sqrt{n}]$, $w_n \geq \Omega\left(\frac{1}{\sqrt{n}}\right) + \Omega\left(\frac{1}{n^{\frac{1-\delta}{2}}}\right) = \Omega\left(\frac{1}{n^{\frac{1-\delta}{2}}}\right)$.

Consequently, for every s, $u_{n+1,s} - u_{n,s} \geq \Omega\left(\frac{1}{n^{1+\frac{1-\delta}{2}}}\right)$, and as $u_{n,s}$ is bounded (again using Lemma F.2), it converges.

The payoff of the auxiliary game converges to the same limit Similarly, one can show that $f_{s,u_{n+1}}\left(x_{n+1,s}\right) - f_{s,u_n}\left(x_{n,s}\right)$ is lower bounded by $\Omega\left(\frac{1}{n^{1+\frac{1-\delta}{2}}}\right)$, so it converges, and it is the same limit as $u_{n,s}$ (otherwise $u_{n,s}$ could not be bounded).

The limit is an equilibrium payoff in AFP Using (3), (valid for every s), writing $\Delta_{s,n} := \sum_{i \in I} f_{s,u_{n,s}}\left(a^i_n, x^{-i}_{n,s}\right) - f_{s,u_{n,s}}\left(x_{n,s}\right)$:

$$f_{s,u_n}\left(x_{n+1,s}\right) - f_{s,u_n}\left(x_{n,s}\right) = \frac{1_{s=s_n}}{n}\Delta_{s,n} + O\left(\frac{1}{n^2}\right) \quad (4)$$

Then:

$$\begin{aligned} f_{s,u_{n+1}}\left(x_{n+1,s}\right) - f_{s,u_n}\left(x_{n+1,s}\right) &= \delta \sum_{s' \in S} P_{ss'}(x_{n+1,s})(u_{n+1,s'} - u_{n,s'}) \\ &= \delta \sum_{s' \in S} P_{ss'}(x_{n+1,s}) \frac{\alpha_n}{\sigma_n}(f_{s',u_n}\left(x_{n,s}\right) - u_{n,s'}) \\ &\geq \delta |S| P_{ss'}(x_{n+1,s}) \frac{\alpha_n}{\sigma_n} w_n \end{aligned} \tag{5}$$

Summing (4) and (5) gives:

$$f_{s,u_{n+1}}\left(x_{n+1,s}\right) - f_{s,u_n}\left(x_{n,s}\right) \geq \delta |S| \frac{\alpha_n}{\sigma_n} w_n + \frac{1_{s=s_n}}{n} \Delta_{s,n} + O\left(\frac{1}{n^2}\right) \tag{6}$$

However, $\Delta_{s,n} \geq 0$, so summing (6) over n gives that $\sum_n \frac{1_{s=s_n}}{n} \Delta_{s,n} < \infty$ with the same reasoning as above (because $\frac{\alpha_n}{\sigma_n} w_n = \Omega\left(\frac{1}{n^{1+(1-\delta)/2}}\right)$ and the terms in the left term cancel out).

Simple calculations yield that $\frac{1}{n}\sum_{k=1}^{n} 1_{s=s_k} \Delta_{s,k}$ goes to 0 as n goes to ∞. However, it is clear that changes in $\Delta_{s,n}$ are of the order of magnitude of the update steps, that is $\frac{1}{n}$. As a consequence, assuming that $\Delta_{s,n}$ does not go to 0, there exists a $A > 0$ such that for $\epsilon > 0$, if $\Delta_{s,n} \geq 3\epsilon$, then $\Delta_{s,n+m} \geq 3\epsilon - \sum_{k=1}^{m} \frac{A}{n+k} \geq 2\epsilon - A\log((n+m)/n)$ for n large enough (well known result of the harmonic series). But then, for $m = [n\exp(\epsilon/A) - 1)]$:

$$\frac{1}{n+m} \sum_{k=n}^{m+m} 1_{s=s_k} \Delta_{s,k} \geq \frac{1}{n+m} \sum_{k=n}^{n+m} 1_{s=s_k} \epsilon$$

Since the game is ergodic, with probability 1 when m goes to ∞ (and it goes to infinity when n goes to infinity), $\frac{1}{m}\sum_{k=n}^{n+m} 1_{s=s_k}$ is greater than a real β_- which depends only on the game (minimal frequency of visit of s using the law of large numbers).

$$\frac{1}{n+m} \sum_{k=n}^{m+m} 1_{s=s_k} \Delta_{s,k} \geq \frac{1}{n+m} m \beta_- \epsilon \geq \frac{n\exp(\epsilon/A) - 1) - 1}{n + n\exp(\epsilon/A) - 1)} \beta_- \epsilon \geq \frac{\exp(\epsilon/A) - 1) - \frac{1}{n}}{1 + \exp(\epsilon/A) - 1)} \beta_- \epsilon$$

This latest inequality is absurd, so $\Delta_{s,n}$ goes to 0 almost surely when n goes to infinity, proving that the limit of $u_{n,s}$ is an equilibrium payoff. Then, it is clear that $x_{n,s}$ converge towards the set of Nash equilibria almost surely (otherwise $f_{s,u_{n,s}}(x_{n,s})$ could not have the same limit as$u_{n,s}$.

Now, for SFP, the proof is similar but for the indicator function which disappears. As a consequence, it is not needed to use a β_- but we still prove that almost surely, $\Delta_{s,k}$ goes to 0. Therefore, we do not need the ergodicity hypothesis for this case.

□

B. Convergence of Best-Response Dynamics in Identical Interest Stochastic Games

In this section, we prove that the best-response dynamics converge in identical interest stochastic games. We first prove a boundedness lemma and then proceed with the convergence of every system in identical interest stochastic games.

Note that since we only deal with identical interest stochastic games, the superscript i in u_s^i can be omitted as all u_s^i are equals (see Section 5).

In what follows, let $\{u_s^i, x_s^i\}_{s \in S, i \in I}$ be a solution of (ABRD). Note that since (SBRD) is included in (ABRD), this is valid for solutions of (SBRD) as well (it is the case where $\beta_- = 1$).

We define:

$$\begin{aligned} \Gamma_s(t) :=& f_{s,u(t)}(x_s(t)) \\ \Delta_s^i(t) :=& \max_{y \in A^i} f_{s,u(t)}(y, x_s^{-i}(t)) - f_{s,u(t)}(x_s(t)) \\ =& \max_{y \in A^i} f_{s,u(t)}(y, x_s^{-i}(t)) - \Gamma_s(t) \end{aligned}$$

Lemma B.1. *Let $\{u_s^i, x_s^i\}_{s\in S, i\in I}$ a solution of (ABRD) or (SBRD). Then for all $s \in S$, functions u_s and $t \mapsto f_{s,u(t)}(x_s(t))$ are bounded.*

Proof. Let $M = \max_{s\in S, a\in A}\{|u_s(0)|, |\Gamma_s(0)|, |r_s(a)|\} + 1$.

Then $|u_s(0)| < M$ and $|\Gamma_s(0)| < M$ for every s. u_s and Γ_s are continuous, therefore if they are not bounded by M, there exists t minimal such that there exists $s \in S$ such that either:

- $u_s(t) = M$ and $|\Gamma_s(t)| < M$, therefore $\dot{u}_s(t) = \beta_t \alpha(t)(\Gamma_s(t) - u_s(t)) \leq 0$ for some β_t, therefore $u_s(t^-) \geq M$, which is absurd.
- $u_s(t) = -M$ and $|\Gamma_s(t)| < M$, therefore $\dot{u}_s(t) = \beta_t \alpha(t)(\Gamma_s(t) - u_s(t)) \geq 0$ for some β_t, therefore $u_s(t^-) \leq -M$, which is absurd.
- $\Gamma_s(t) = M$, therefore:
$$(1-\delta) r_s(x_s(t)) + \delta \sum_{s'\in S} P_{s,s'}(x_s(t)) u_{s'}(t) = M$$
But $r_s(x_s) < M$ and $u_{s'}(t) \leq M$ for all s',
therefore $\sum_{s'\in S} P_{s,s'}(x_s(t)) u_{s'}(t) \leq M$,
so $\Gamma_s(t) < M$ (because $0 < \delta < 1$), absurd.

□

Lemma B.2. *Function Γ_s is differentiable and its differential is:*

$$\frac{d\Gamma_s}{dt} = \delta \sum_{s'} P_{ss'}(x_s)\dot{u}_{s'} + \beta_s(t) \sum_i \Delta_s^i(t)$$

In the SBRD case, $\beta_s(t) = 1$.

Proof.
$$\frac{d\Gamma_s}{dt} = D_u(f_{s,\vec{u}(t)}(x_s(t)))(D_t u) + D_{x_s} f_{s,\vec{u}}(x_s)(D_t x_s)$$
where D_u is the partial differential in u.

$x_s \mapsto f_{s,u(t)}(x_s)$ is a n-linear map in x_s, therefore:

$$D_{x_s} f_{s,u(t)}(x_s)(D_t x_s) = \sum_i f_{s,u(t)}(\dot{x}_s^i, x_s^{-i})$$

$u \mapsto f_{s,u}(x_s(t))$ is a linear function in u, and:

$$D_u f_{s,u}(x_s(t)) = \delta \sum_{s'} P_{ss'}(x_s)\dot{u}_s$$

Therefore, $\frac{d\Gamma_s}{dt} = \delta \sum_{s'} P_{ss'}(x_s)\dot{u}_{s'} + \beta_s(t)\sum_i \Delta_s^i(t)$. □

Lemma B.3. *Function $\Delta_s^i(t)$ is Lipschitz.*

Proof. u is differentiable and its derivative is bounded by
$\sup_t |\Gamma_s(t) - u_s(t)|$, so u is $2M$-Lipschitz where M is a bound of the Γ_s and u_s. The derivative of x_s is also bounded, so it is also Lipschitz. As $f_{s,\cdot}$ is Lispschitz with respect to any parameter (it is multilinear). Therefore, for all y, $t \mapsto f_{s,u(t)}(y, x_s^{-i}(t))$ is Lispschitz with the same coefficient, so $\Delta_s^i(t)$ is also Lipschitz. □

Convergence of the synchronous and asynchronous system Let $s_-(t) \in \arg\min_{s \in S} (\Gamma_s(t) - u_s(t))$. This means that for every t we choose an arbitrary s that minimizes $\Gamma_s(t) - u_s(t)$. Note that as a consequence, $\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)$ is continuous because every $\Gamma_s(t) - u_s(t)$ is continuous.

Lemma B.4. *There exists $A \geq 0$ such that for every $s \in S$, $\Gamma_s(t) - u_s(t) \geq -A \exp(\int_1^t (\delta - 1)\alpha(t)\, dt)$*

Proof. By the previous lemma:

$$\begin{aligned}\frac{d\Gamma_s}{dt} &\geq \delta \sum_{s'} P_{ss'}(x_s)\alpha(t)\left(\Gamma_{s'}(t) - u_{s'}(t)\right)\\ &\geq \delta \sum_{s'} P_{ss'}(x_s)\alpha(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}\right)\\ &= \delta\alpha(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}\right)\end{aligned} \tag{7}$$

Moreover, for $h > 0$:

$$\begin{aligned}&\Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t))\\ &\geq \Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t+h)}(t) - u_{s_-(t+h)}(t))\\ &\geq h \min_{s\in S} \frac{d\Gamma_s}{dt} + o(h) + u_{s_-(t+h)}(t) - u_{s_-(t+h)}(t+h)\end{aligned} \tag{8}$$

For any s:

$$\begin{aligned}u_s(t) - u_s(t+h) &= -h\frac{du_s}{dt} + o(h)\\ &= -h\alpha(t)\left(\Gamma_s(t) - u_s(t)\right) + o(h)\end{aligned}$$

Now let us suppose that s is an accumulation point of $s_-(t+h)$ when h goes to 0. Then, as every $\Gamma_s(t) - u_s(t)$ is continuous, we have that $\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t) = \Gamma_s(t) - u_s(t)$ (else s can not be an accumulation point). So, the preceding equality can be rewritten as:

$$u_s(t) - u_s(t+h) = -h\alpha(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)\right) + o(h)$$

This is valid for every accumulation point of $s_-(t+h)$ (and is independent of s) and there is a finite number of such s, so we also have:

$$u_{s_-(t+h)}(t) - u_{s_-(t+h)}(t+h) = -h\alpha(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)\right) + o(h)$$

Now, from inequality (7), we have that:

$$h\min_{s\in S}\frac{d\Gamma_s}{dt} \geq h\delta\alpha(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}\right)$$

And these two last inequalities can be summed to get:

$$h\min_{s\in S}\frac{d\Gamma_s}{dt} + u_{s_-(t+h)}(t) - u_{s_-(t+h)}(t+h) + o(h) \geq h(\delta - 1)\alpha(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}\right) + o(h)$$

Going back to (8):

$$\begin{aligned}&\Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t))\\ &\qquad \geq h(\delta - 1)\alpha(t)\left(\Gamma_{s_-(t)}(t) - u_{s_-(t)}\right) + o(h)\end{aligned}$$

We now need a version of Grönwall Lemma that applies to this case, it is provided here for completeness:

Let $v(t) = \exp\left(\int_0^t (\delta - 1)\alpha\,(t)\,dt\right)$.

Then $\frac{dv}{dt} = (\delta - 1)\alpha\,(t)\,v(t)$, $v(0) = 1$, $v > 0$.

and $\frac{1}{v(t+h)} = \frac{1}{v(t)} - h(\delta - 1)\frac{\alpha(t)}{v(t)} + o(h)$

We now proceed with the classical proof of Grönwall Lemma:

$$\begin{aligned}
\frac{\Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h)}{v(t+h)} &\geq \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{v(t+h)} + h(\delta - 1)\alpha\,(t)\,\frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{v(t+h)} + o(h) \\
&\geq \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{v(t)} - h(\delta - 1)\alpha\,(t)\,\frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{v(t)} \\
&\qquad + h(\delta - 1)\alpha\,(t)\,\frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{v(t)} + o(h) \\
&\geq \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{v(t)} + o(h)
\end{aligned}$$

Therefore, $t \mapsto \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{v(t)}$ is increasing. We can conclude:

$$\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t) \geq \left(\Gamma_{s_-(0)}(0) - u_{s_-(0)}(0)\right) \exp\left(\int_0^t (\delta - 1)\alpha\,(t)\,dt\right)$$

□

Lemma B.5. *The gap between* $\Gamma_s(t)$ *and* $\max_{y \in A^i} f_{s,u(t)}(y, x_s^{-i}(t))$ *converges to 0:*

$$\forall s, \sum_i \Delta_s^i(t) \to 0$$

Proof. First, we show that $\forall i, s, \int_1^\infty \sum_{i \in I} \Delta_s^i(t)dt < +\infty$.

Using Lemma B.2: $\beta_s(t) \sum_i \Delta_s^i(t) = \frac{d\Gamma_s}{dt} - \delta \sum_{s'} P_{ss'}(x_s)\dot{u}_s$.

Therefore:

$$\forall T, \int_1^T \beta_s(t) \sum_i \Delta_s^i(t)dt = \int_1^T \frac{d\Gamma_s}{dt} - \delta \sum_{s'} \int_1^T P_{ss'}(x_s)\dot{u}_s$$

With the previous lemma:

$$\begin{aligned}
P_{ss'}(x_s)\dot{u}_s &= P_{ss'}(x_s)\alpha\,(t)\,(\Gamma_s(t) - u_s(t)) \\
&\geq -P_{ss'}(x_s)A\alpha\,(t)\exp\left(\int_1^t (\delta - 1)\alpha\,(t)\right)
\end{aligned}$$

Then, for all T:

$$\begin{aligned}
\beta_- \int_1^T \sum_i \Delta_s^i(t)dt &\leq \int_1^T \beta_s(t) \sum_i \Delta_s^i(t)dt \\
&\leq \Gamma_s(T) - \Gamma_s(1) + \delta \sum_{s'} P_{ss'}(x_s) \int_1^T A\alpha\,(t)\exp\left(\int_1^t \alpha\,(v)\,(\delta - 1)dv\right) \\
&= \Gamma_s(T) - \Gamma_s(1) + \delta \frac{A}{\delta - 1}\left(\exp\left(\int_1^T \alpha\,(v)\,(\delta - 1)dv\right) - 1\right) \\
&< \Gamma_s(T) - \Gamma_s(1) + \delta \frac{A}{1 - \delta}
\end{aligned}$$

Then, as $\Delta_s^i(t)$ is Lipschitz (Lemma B.3) and the limit of its integral is bounded and positive, $\Delta_s^i(t) \xrightarrow[t\to\infty]{} 0$. □

Lemma B.6 (Convergence of the synchronous and semi-asynchronous system)**.** *For all $s \in S$:*

$$\Gamma_s(t) = f_{s,u(t)}(x_s(t)) \xrightarrow[t\to\infty]{} \limsup \Gamma_s$$
$$\textit{and } u_s(t) \xrightarrow[t\to\infty]{} \limsup \Gamma_s$$

Proof.

$$\begin{aligned}\Gamma_s(t_2) &= \Gamma_s(t_1) + \int_{t_1}^{t_2} \frac{d\Gamma_s}{dt} dv \\ &\geq \Gamma_s(t_1) + \delta \int_{t_1}^{t_2} \alpha\left(v\right) \left(\Gamma_{s_-(v)}(v) - u_{s_-(v)}(v)\right) dv \\ &\geq \Gamma_s(t_1) - A\delta \int_{t_1}^{t_2} \alpha\left(v\right) \exp\left(\int_1^v (\delta - 1)\alpha\left(t\right) dt\right) dv \\ &\geq \Gamma_s(t_1) - A\frac{\delta}{1-\delta} \exp\left(\int_1^{t_1} (\delta - 1)\alpha\left(t\right) dt\right)\end{aligned}$$

So $\dfrac{A\delta}{1-\delta} \exp\left(\int_1^{t_1} (\delta - 1)\alpha\left(t\right) dt\right)$ goes to 0 when t_1 goes to $+\infty$ (thanks to hypothesis H2), therefore, it is sufficient to take t_1 big enough so that $\Gamma_s(t_1)$ is close to the $\limsup$ and the second term is small enough.

With a similar argument, u_s has a limit, and it is necessarily the same as Γ_s, otherwise u_s would be unbounded (because $\dot{u}_s = (\Gamma_s - u_s)/\alpha\left(t\right)$). □

Lemma B.7 (Convergence to the set of mixed stationary equilibria)**.** *$\{\lim \Gamma_s\}_{s\in S}$ is an equilibrium payoff of the δ discounted stochastic game. $\{x_s\}_{s\in S}$ converges to the set of mixed equilibria.*

Proof. Let $\tilde{x}$ be an accumulation point of the vector-valued function $x = \{x_s\}$. Then, from Lemma B.5:

$$\Delta_s^i(t) = f_{s,u(t)}(br^i_{s,u(t)}(x_s^{-i}(t)) - x_s^i(t), x_s^{-i}(t)) \to 0$$

So by continuity, for all s:

$$f_{s,\lim u}(br^i_{s,\lim u}(\tilde{x}_s^{-i}(t)) - \tilde{x}_s^i, \tilde{x}_s^{-i}) = 0$$

□

Proof of Theorem 5.1.
Lemma B.7 and B.6 prove the theorems for both the SBRD and ABRD systems. □

C. Convergence of Best-Response Dynamics in Zero-Sum Games

In this section, we prove that best-response dynamics converges in two players zero-sum games, therefore extending the result of Leslie et al.. The proof is very similar to that of (Leslie et al., 2020) but it is included below for completeness. Details can be found in the other paper.

Let $\{x_s, u\}_{s\in S}$ be a solution of $ABRD$ and let $\alpha^\star > \lim_{t\to+\infty} \alpha(t)$. Note that $\alpha^\star$ may be be arbitrarily close to 0 if the limit of α is 0, and indeed the case proven in (Leslie et al., 2020) is the case where $\alpha(t) = \frac{1}{t+1}$ and $\alpha^\star \to 0$.

We define the energy of the system, also known as the duality gap, as:

$$w_s\left(t\right) = \max_{a^1\in A^1} f_{s,u(t)}\left(a^1, x_s^2(t)\right) - \min_{a^2\in A^2} f_{s,u(t)}\left(x_s^1(t), a^2\right) \tag{9}$$

Lemma B.1 states that there exists a constant $M > 0$ such that $\|u_s\|_\infty \leq M$ and $\|f_{s,u(t)}\left(\right)\|_\infty \leq M$.

By definition x_s, u are differentiable almost everywhere. It is straightforward to see that it is also true for w_s.

We denote $v_{s,u(t)}$ the value of the auxiliary game in state s parameterized by $u(t)$.

Lemma C.1. *For every $s \in S$, there exists a time T such that for all $t \geq T$, $|f_{s,u(t)}\left(x_s(t)\right) - v_{s,u(t)}| \leq 4M\alpha^\star$.*

Proof. We define $D_u w_s$ and $D_{x_s} w_s$ as the partial derivatives of w_s when u and x_s are considered as parameter of w_s. With these notations, $\frac{dw_s}{dt} = \dot{u} \cdot D_u w_s + \dot{x_s} \cdot D_{x_s} w_s$.

On the one hand, using Lemma A.2 of (Leslie et al., 2020), $\dot{u} \cdot D_u w_s \leq 2\delta \max_{s' \in S} \dot{u_{s'}} \leq 4\delta M\alpha(t)$. On the other hand, using (Hofbauer et al., 2006), $\dot{x_s} \cdot D_{x_s} w_s \leq -\beta_- w_s\left(t\right)$.

Therefore, $\frac{dw_s}{dt} \leq -\beta_- w_s\left(t\right) + 4\delta M\alpha(t)$. Since α is decreasing, it is arbitrarily close to $\alpha^\star$ when t goes to ∞, so $w_s\left(t\right) \leq 4M\alpha^\star$ when t is big enough. Note that knowing α, a t satisfying this property can be computed. This will be used in the discrete time algorithm.

Since $|f_{s,u(t)}\left(x_s(t)\right) - v_{s,u(t)}| \leq w_s\left(t\right)$, this gives the desired result.

□

Define ϵ such as $\frac{(1-\delta)\epsilon}{16} = 4M\alpha^\star$ and and $t_1(\epsilon)$ as defined in Lemma C.1.

We define two distinguished states:

- $s_f(t) \in \arg\max_{s \in S} |f_{s,u(t)}\left(x_s(t)\right) - u_s(t)|$
- $s_v(t) \in \arg\max_{s \in S} |v_{s,u(t)} - u_s(t)|$

Lemma C.2. *If $t \geq t_1(\epsilon)$, $|u_{s_f(t)}(t) - f_{s_f,u(t)}\left(x_{s_f(t)}(t)\right)| \geq \epsilon$ and for an $s \in S$,*

$$\left||u_{s_f(t)}(t) - v_{s_f(t),u(t)}| - |u_s(t) - v_{s,u(t)}|\right| \leq \frac{(1-\delta)\epsilon}{8}$$

then:

$$\frac{d|u_s(t) - v_{s,u(t)}|}{dt} \leq -\frac{3(1-\delta)\alpha\left(t\right)\epsilon}{4}$$

Proof. First, using Lemma A.2 of (Leslie et al., 2020):

$$\frac{d|v_{s,u(t)}|}{dt} \leq \delta \max_{s \in S} |\dot{u}| = \delta\alpha\left(t\right)|f_{s_f,u(t)}\left(x_{s_f(t)}(t)\right) - u_{s_f(t)}(t)|$$

If $u_s(t) \geq v_{s,u(t)}$: Then $|u_{s_f(t)}(t) - v_{s_f(t),u(t)}| - u_s(t) + v_{s,u(t)} \leq \frac{(1-\delta)\epsilon}{8}$.

$$\begin{aligned}
\frac{du_s(t)}{dt} &= \alpha\left(t\right)\left(f_{s,u(t)}\left(x_s(t)\right) - u_s(t)\right) \\
&\leq \alpha\left(t\right)\left(f_{s,u(t)}\left(x_s(t)\right) + \frac{(1-\delta)\epsilon}{8} - v_{s,u(t)} - |u_{s_f(t)}(t) - v_{s_f(t),u(t)}|\right) \\
&\leq \alpha\left(t\right)\left(\frac{3(1-\delta)\epsilon}{16} - |u_{s_f(t)}(t) - v_{s_f(t),u(t)}|\right) \\
&\leq \alpha\left(t\right)\left(\frac{(1-\delta)\epsilon}{4} - |u_{s_f(t)}(t) - f_{s_f,u(t)}\left(x_{s_f(t)}(t)\right)|\right)
\end{aligned}$$

Summing with $v_{s,u(t)}$:

$$\begin{aligned}
\frac{du_s(t) - v_{s,u(t)}}{dt} &\leq \alpha\left(t\right)\left(\frac{(1-\delta)\epsilon}{4} + (\delta - 1)|u_{s_f(t)}(t) - f_{s_f,u(t)}\left(x_{s_f(t)}(t)\right)|\right) \\
&\leq \alpha\left(t\right)\left(\frac{(1-\delta)\epsilon}{4} - (1-\delta)\epsilon\right) \\
&\leq -\alpha\left(t\right)\left(\frac{3(1-\delta)\epsilon}{4}\right)
\end{aligned}$$

If $u_s(t) \leq v_{s,u(t)}$: similar calculations yield the same result.

□

We can now prove the two final lemma of this section.

Lemma C.3. *For all $s \in S$, $\limsup |u_s(t) - f_{s,u(t)}\left(x_s(t)\right)| \leq 2\epsilon$.*

Proof. We define $g(t) = \max\{|u_{s_f(t)}(t) - v_{s_f(t),u(t)}|, 2\epsilon\}$.

As a composition of maximum of locally Lipschitz function and as such is locally Lipschitz as well. (See Lemma B.4 of (Leslie et al., 2020) for detailed arguments.)

Now, if $|u_{s_f(t)}(t) - v_{s_f(t),u(t)}| \leq 2\epsilon$, then $\frac{dg}{dt} = 0$. If $u_{s_f(t)}(t) - v_{s_f(t),u(t)}| \geq 2\epsilon$, then, if t is greater than $t^1(\epsilon)$, $|u_{s_f(t)}(t) - f_{s_f,u(t)}\left(x_{s_f(t)}(t)\right)| \geq \epsilon$ (Lemma C.1). For $t \geq t^1(\epsilon)$, on a neighbourhood of t, every s that maximizes $|f_{s,u(t)}\left(x_s(t)\right) - u_s(t)|$ satisfies the condition of Lemma C.2, because $f_{s,u(t)}\left(x_s(t)\right)$ and $v_{s,u(t)}$ are close thanks to Lemma C.1. Therefore, Lemma C.2 can be used and $\frac{dg}{dt} \leq -\frac{3(1-\delta)\alpha(t)\epsilon}{4}$.

Using hypothesis H2, $g(t) \to 2\epsilon$, which gives the result.

□

Lemma C.4. *For all $s \in S$, $x_s(t)$ converge to the set of 2ϵ-Nash equilibria of the auxiliary game.*

Proof. The previous proof gives that $f_{s,u(t)}\left(x_s(t)\right)$ is 2ϵ close to $v_{s,u(t)}$, hence the result. □

D. Using Continuous Time Results for Discrete Time Fictitious Play with Stochastic Approximations

In this section, we describe how the stochastic approximation framework with differential inclusion (Benaïm et al., 2005) can be extended and used to prove result in discrete time in the autonomous case (i.e., α is constant).

D.1. Correlated Asynchronous Stochastic Approximation

An asynchronous system as defined in (Perkins & Leslie, 2012) is as follows. Assuming $y_n \in \mathbb{R}^k$, one defines a system where updated components of the vector at every step n are $S_n \subseteq K := [1 \dots k]$. We define $s^\sharp_n$ as the number of times util n that s occured:

$$s^\sharp_n = \sharp\{k \mid s \in S_k \wedge 0 \leq k \leq n\}$$

We now describe now a system where component $y_{s,n}$ is updated at rate $\gamma_{s^\sharp_n}$ if and only if $s \in S_n$, that is:

$$y_{s,n+1} - y_{s,n} - \gamma_{s^\sharp_n}(Y_{s,n} + d_{s,n}) \in 1_{s \in S_n} \gamma_{s^\sharp_n} F_s(y_n) \tag{10}$$

where variable $Y_{s,n}$ is a random noise with $\mathbb{E}[Y_{s,n}] = 0$ and $d_{s,n}$ goes to 0 when $n \to \infty$.

We define:

$$\overline{\gamma}_n = \max_{s \in S_n} \gamma_{s^\sharp_n}$$

$$M_{n+1} = \text{diag}\left\{1_{s \in I_n} \frac{\gamma_{s^\sharp_n}}{\overline{\gamma}_n} \mid s \in K\right\}$$

and we can rewrite (10) to:

$$y_{n+1} - y_n - \overline{\gamma}_n M_{n+1}(Y_n + d_n) \in \overline{\gamma}_n M_{n+1} F(y_n) \tag{11}$$

The continuous counterpart is defined as follows. For an $\epsilon > 0$, Ω^ϵ_k is the set of $k \times k$ diagonal matrices with coefficients between ϵ and 1:

$$\Omega^\epsilon_k := \{\text{diag}(\beta_1, \dots, \beta_k); \beta_i \in [\epsilon, 1], \forall i = 1, \dots, k\}$$

And the continuous system is:

$$\frac{dy}{dt} \in \overline{F}(y) := \Omega_k^\epsilon \cdot F(y) \tag{12}$$

where the multiplication is between sets (i.e., the resulting set is the multiplication of every pair of the initial sets).

Then, the limit set of solutions of (11) is internally chain transitive (see Definition D.2 below) for system (12) (Perkins & Leslie, 2012) under assumptions stated in Subsection D.2.

However, we need a modified version where the asynchronoucity can be correlated, meaning for instance that some components are updated synchronously or that updating may be done at the same time for a set of components. This is the case if x_s^i and u_s were updated at the same times for a specific state s (which is an extension of this current work, as explained in the conclusion) or for u_s which is always updated at every step in AFP. Therefore, we now suppose that every $S_n \in \mathcal{S} \subseteq K$. For instance, if the s component is updated at every step, it can be expressed with $\forall S' \in \mathcal{S}$, $s \in S'$. Then we define an alternative set of diagonal matrices for the continuous version: $\Omega_{k,\mathcal{S}}^\epsilon := \text{diag}(\text{conv}(\mathcal{S}) \bigcap [\epsilon, 1]^K)$ and the map $\overline{F}(y) := \Omega_{k,\mathcal{S}}^\epsilon \cdot F(y)$

Then we can link the internally chain transitive sets of differential inclusion $\frac{dy}{dt} \in \overline{F}(y)$ and limit sets of solutions of (11). As systems ABRD and SBRD can be written as $\overline{F}$ with a suitable $\mathcal{S}$ and F, making it possible to prove the rest of Theorem 4.1 using the convergence results of the continuous time systems of the previous section, see section D.4.

D.2. Formal Results

We start with the definition of Marchaud maps. They are used in most stochastic approximation theorems, even if the term is not always employed. In our systems, as the best-response map br is piecewise constant and the rest of the right hand side is continuous, right hand sides of the differential inclusions are Marchaud maps.

Definition D.1 (Marchaud map)**.** $F : \mathbb{R}^K \rightrightarrows \mathbb{R}^K$ is a Marchaud map if:

(i) F is a closed set-valued map, *i.e.* $\{(x, y) \in \mathbb{R}^K \times \mathbb{R}^K \mid y \in F(x)\}$ is closed.

(ii) for all $y \in \mathbb{R}^K$, $F(y)$ is a non-empty, compact, convex subset of $\mathbb{R}^K$

(iii) there exists $c > 0$ such that $\sup_{y \in \mathbb{R}^K z \in F(y)} ||z|| \leq c(1 + ||y||)$

We now need the definition of internally chain transitive sets, as stated in (Benaïm et al., 2005). They will later be used to characterize the limit sets of the discrete time systems.

Definition D.2 (Internally chain transitive)**.** A set A is internally chain transitive for a differential inclusion $\frac{dy}{dt} \in F(y)$ if it is compact and if for all $y, y' \in A$, $\epsilon > 0$ and $T > 0$ there exists an integer $n \in \mathbb{N}$, solutions $y_1, \ldots y_n$ to the differential inclusion and real numbers $t_1, t_2, \ldots, t_n$ greater than T such that:

- $y_i(s) \in A$ for $0 \leq s \leq t_i$
- $\|y_i(t_i) - y_{i+1}(0)\| \leq \epsilon$
- $\|y_1(0) - y\| \leq \epsilon$ and $\|y_n(t_n) - y'\| \leq \epsilon$

Definition D.3 (Asymptotic pseudo-trajectories)**.** A continuous function $z : \mathbb{R}^+ \to \mathbb{R}^m$ is an asymptotic pseudo-trajectory of a differential inclusion if $\lim_{t \to +\infty} \mathbf{D}(\Theta^t(z), S) = 0$ where $\Theta^t(z)(s) = z(t + s)$ (it is the translation operator), S is the set of all solutions of the differential inclusion and $\mathbf{D}$ is the distance between continuous functions defined as:

$$\mathbf{D}(f, g) := \sum_{k=1}^{\infty} \frac{1}{2^k} \min(\|f - g\|_{[-k,k]}, 1)$$

where $\| \cdot \|_{[-k,k]}$ is the supremum norm on the interval $[-k, k]$.

This two last definitions will be useful with Theorem 4.3 of (Benaïm et al., 2005) that establishes that the limit set of asymptotic pseudo-trajectories is internally chain transitive. What is left to prove is that an affine interpolation of the discrete time system is an asymptotic pseudo-trajectories. Below is the proof for the synchronous system SFP and the next section deals with semi-asynchronous and fully-asynchronous systems.

Lemma D.4. *The limit set of SFP is internally chain transitive with respect to SBRD for $\alpha(t) = 1$ for all t.*

Proof. Proposition 1.3 and Theorem 4.2 of (Benaïm et al., 2005) establish that the affine interpolation of sequences $x_{n,s}, u_n$ is a perturbed solution and then an asymptotic pseudo trajectory. Theorem 4.3 from the same article proves that the limit set is internally chain transitive. □

D.3. Correlated Asynchronous Stochastic Approximations

We now extend a theorem originally proven by Perkins & Leslie:

Theorem D.5 (Analog of Theorem 3.1 of (Perkins & Leslie, 2012))**.** *Suppose that:*

(i) $y_n \in C$ *for all* n *where* C *is compact*

(ii) The set valued application $F : C \rightrightarrows C$ *is Marchaud*

(iii) Sequence γ_n *is such that*

- *(a)* $\sum_n \gamma_n = \infty$ *and* $\gamma_n \xrightarrow[n\to\infty]{} 0$
- *(b) for* $x \in (0,1), \sup_n \gamma_{[xn]}/\gamma_n < A_x < \infty$ *where* $[\cdot]$ *is the floor function.*
- *(c) for all* n, $\gamma_n \geq \gamma_{n+1}$

(iv)

- *(a) For all* $y \in C$, $\mathcal{S}_n, \mathcal{S}_{n+1} \in \mathcal{S}$,

$$\mathbb{P}(S_{n+1} = \mathcal{S}_{n+1} | \mathcal{F}_n) = \mathbb{P}(S_{n+1} = \mathcal{S}_{n+1} | S_n = \mathcal{S}_n, y_n = y)$$

- *(b) The probability transition between* $\mathcal{S}_n$ *and* $\mathcal{S}_{n+1}$ *is Lipsichitz continuous in* x_n *and the Markov chain that* S_n *form is aperiodic, irreducible and for every* $s \in \mathcal{S}$*, there exists* $S \in \mathcal{S}$ *such that* $s \in S$*.*

(v) For all n, Y_{n+1} *and* S_{n+1} *are uncorrelated given* $\mathcal{F}_n$

(vi) For some $q \geq 2$, $\begin{cases} \sum_n \gamma_n^{1+q/2} < \infty \\ \sup_n \mathbb{E}(\|Y_n\|^q) < \infty \end{cases}$

(vii) $d_n \to 0$ *when* $n \to \infty$

Then with probability 1, affine interpolation $\overline{y}$ *is an asymptotic pseudo-trajectory to the differential inclusion,*

$$\frac{dy}{dt} \in \overline{F}(y)$$

where $\begin{cases} \overline{F}(y) := \Omega^{\epsilon}_{k,\sigma} \cdot F(y) \\ \Omega^{\epsilon}_k := \left\{ \mathit{diag}(\beta_1, \ldots, \beta_k) \middle| \begin{array}{r} \forall i \in \{1, \ldots, k\}, \\ \beta_i \in [\epsilon, 1] \end{array} \right\} \\ \epsilon > 0 \end{cases}$

However, at every step of the proof of Perkins and Leslie, we can take into account that $S_n \in \mathcal{S}$, therefore we do not need every matrix $\text{diag}([\epsilon,1]^K)$ in Ω^ϵ_k but only those that are also in $\text{diag}(\text{conv}_\epsilon(\mathcal{S}))$ where $\text{conv}(\mathcal{S})$ is the convex hull of $\mathcal{S}$ composed with $\max(\epsilon, \cdot)$ for every coordinate. Indeed, when update rates are manipulated, they are summed via integrals or floored by an $\epsilon > 0$. The resulting vectors belongs to $\text{conv}_\epsilon(\mathcal{S})$ at every step of the proof. Therefore, the conclusion of the theorem can be changed with $\Omega^\epsilon_{k,\mathcal{S}} := \text{diag}(\text{conv}(\mathcal{S}) \bigcap [\epsilon,1]^K)$. This makes it possible to use the Theorem in our asynchronous and semi-asynchronous cases.

The full proof is included in Section E.

D.4. Convergence of a Fictitious Play procedure in identical interest stochastic games

In this subsection, we first characterize the internally chain transitive sets of $SBRD$ and $ABRD$ before using this characterization to prove a second time the convergence of FP in identical interest stochastic games.

Lemma D.6 (Internally Chain Transitive Sets)**.** *If for all t, $\alpha(t) = 1$ and if L is internally chain transitive either for ABRD then*

$$L \subseteq \left\{ (x,u) \middle| \begin{array}{l} \forall s \in S\ \forall i \in I,\ f_{s,u}(x_s) = u_s \\ \wedge\ x_s^i \in \operatorname*{arg\,max}_{y^i \in A^i} f_{s,u}(y^i, x_s^{-i}) \end{array} \right\}$$

Proof.

We define:

$$A := \left\{ (x,u) \middle| \begin{array}{l} \forall s \in S\ \forall i \in I,\ f_{s,u}(x_s) = u_s \\ \wedge\ x_s^i \in \operatorname*{arg\,max}_{y^i \in A^i} f_{s,u}(y^i, x_s^{-i}) \end{array} \right\}$$
$$B := \{(x,u)\,|\forall s \in S\ f_{s,u}(x_s) \geq u_s\}$$

We first show that $L \subseteq B$. In order to do that, we take an element of L and show that any path starting from this element is brought towards B, leading to the fact that the element is necessarily already in B (by definition of internal chain transitivity).

Let $(x,u) \in L$ and suppose that $(x,u) \notin B$, that is:

$$-\zeta := \min_{s \in S} f_{s,u}(x_s) - u_s < 0$$

Then for the case of SBRD, for any $T > 0$, there exists $n \in \mathbb{N}$, solutions of SBRD $(x_1,u_1), \dots (x_n,u_n)$ and $t_1, \dots, t_n$ greater than T as in Definition D.2 for $\epsilon = \zeta/2$.

Then $\min_{s \in S} f_{s,u_1(0)}(x_{1,s}(0)) - u_{1,s}(0) \geq -\zeta - \zeta/2$.

Now we can use Lemma B.4 with $\alpha(t) = 1$, for all s:

$$f_{s,u_1(t_1)}(x_{1,s}(t_1)) - u_{1,s}(t_1) \geq (f_{s,u_1(t_1)}(x_{1,s}(t_1)) - u_{1,s}(t_1)) \exp((\delta-1)t_1) \geq (-\frac{3}{2}\zeta)\exp((\delta-1)T)$$

So for T big enough, then for all s:

$$f_{s,u_1(t_1)}(x_{1,s}(t_1)) - u_{1,s}(t_1) \geq -\zeta/4$$

Iteratively, we get:

$$f_{s,u_n(t_n)}(x_{n,s}(t_n)) - u_{n,s}(t_n) \geq -\zeta/4$$

which is contradictory to the fact that $\min_{s \in S} f_{s,u}(x_s) - u_s = -\zeta$.

For ABRD, we have the exact same proof.

So $L \subseteq B$.

We can now use a more classic argument to show that $L \subseteq A$ with a Lyapunov function now that the ambient space can be restricted to B. Let us define $V(x,u) := \sum\limits_{s \in S} f_{s,u}(x_s)$. Then, V is a Lyapunov function for set A with ambient space B. Indeed, on B, $\frac{du_s}{dt} \geq 0$, so $\frac{df_{s,u}(x_s)}{dt} \geq 0$ (with Lemma B.2). Therefore $\frac{df_{s,u}(x_s)}{dt} = 0$ for every s if and only if $(x,u) \in A$. Moreover, $V(A)$ has empty interior thanks to Sard's Theorem.

So we can use Proposition 3.27 of (Benaïm et al., 2005): it applies in case the Lyapunov function is defined on invariant set. So L is contained in A.

□

A second proof of Theorem 4.1 using continuous time Below, we prove a second time Theorem 4.1 in the autonomous case (that is, $\alpha_n = 1$) to show that our extension to the stochastic approximations framework is directly useful.

Proof of Theorem 4.1. For systems (AFP) we now need to apply Theorem D.5. Variable Y_n is 0 in our case because there is no noise. S_{n+1} is the next state variable and it has distribution $P_{S_n}(a_n)$. We check the assumptions:

- (i) is guaranteed because every variable of the system is bounded.
- (ii) is guaranteed because the best-response map is marchaud and the derivative of u is continuous.
- for (iii) and (vi) we use $\gamma(n) = 1/n$, so every assumption is trivial to verify.
- (iv) and (v) comes from the definition of a play and the ergodicity hypothesis on the game

Therefore the affine interpolation of a sequence of fictitious play for stochastic games under our assumption is an asymptotic pseudo-trajectory, which implies that its limit set is internally chain transitive by Theorem 4.3 of (Benaïm et al., 2005).

For system SFP Lemma D.4 states that the limit set is internally chain transitive.

Then Lemma D.6 concludes the proof: the limit set is internally chain transitive and consequently included in the set of equilibria.

□

D.5. Convergence of FP in zero-sum stochastic games

We consider SBRD and ABRD in zero-sum stochastic games in the autonomous case, that is for $\alpha(t) = \alpha^\star$

Proof of Theorem 6.1. The previous proof checks all the hypothesis necessary to apply the stochastic approximation framework we presented in this section. Therefore, a characterization of internally chain transitive sets will be sufficient to conclude.

In the proof of Lemma C.1, we showed that function $\max w_s(t) - \frac{1}{\beta_-} 4\delta M \alpha^\star$ is a Lyapunov function. Therefore, the set where the duality gap is lower or equal than $4\delta M\alpha^\star$ is a internally chain transitive set. Furthermore, in the proof of Lemma C.3, we defined a function g which is a Lyapunov function relative to the previous internally chain transitive set.

□

D.6. Different Priors and Team Games

In this subsection, we suppose that every player has its own u^i estimates. We are going to show that in this case, internally chain transitive sets are included into the set where the estimates u^i are equal (up to a constant) for every i.

Indeed, suppose that G is now a team game. Then every r^i_s can be written $r^i_s = r_s + M_i$ with the convention that $M_1 = 0$ and $r_s = r^1_s$. Then let us show that any internally chain transitive set L is included in $\{(x, u) \mid u^i_s = u^1_s + M_i \; \forall i, s\}$.

Define $V^i(x, u) = \arg\max_s |u^i_s - u^1_s - M_i|$

Let s that maximizes $|u^i_s - u^1_s - M_i|$ so that $V^i(x, u) = |u^i_s - u^1_s - M_i|$.

Then if $u^i_s > u^1_s - M_i$, then $V^i(x(t), u(t))$ can be differentiated for almost every t (using the same techniques as in Section B):

$$\begin{aligned}\frac{dV^i}{dt} &= \alpha(t)(f^i_{s,u^i}(x_s(t)) - f^1_{s,u^i}(x_s(t)) - u^i_s(t) + u^1_s(t) \\ &\le \alpha(t)((1-\delta)M_i + \delta V^i(x, u) + \delta M_i - u^i_s(t) + u^1_s(t) \quad \le \alpha(t)(\delta - 1)V^i(x, u)\end{aligned}$$

And similar calculations for the case $u^i_s \le u^1_s - M_i$ give the same results.

Therefore V^i is a Lyapunov function and $L \subseteq {V^i}^{-1}(\{0\})$, hence the result.

E. Proof of Theorem D.5

In this subsection, we show a proof of Theorem D.5. It is a modification of Theorem 3.1 of (Perkins & Leslie, 2012). In order to carry the proof, we first need a general theorem found in (Benaïm et al., 2005):

Theorem E.1 (Linear interpolation are asyptotic pseudo-trajectories)**.** *Consider the stochastic approximation process*

$$y_{n+1} - y_n \in \gamma_n \left[F(y_n) + Y_{n+1} + d_{n+1}\right] \tag{13}$$

under the assumptions:

(i) *For all* $T > 0$

$$\lim_{n\to\infty} \sup_k \left\{ \left\| \sum_{i=n}^{k-1} \gamma_{i+1} Y_{i+1} \right\| ; k = n+1, \dots, m(\tau_n + T) \right\} = 0 \tag{14}$$

where $\tau_0 = 0$*,* $\tau_n = \sum_{i=1}^{n}$
$gamma_i$ *and* $m(t) = \sup\{k \geq 0; t \geq \tau_k\}$*,*

(ii) $\tau_n \xrightarrow[n\to\infty]{} \infty$ *and* $\gamma_n \xrightarrow[n\to\infty]{} 0$

(iii) $\sup_n \|y_n\| = \mathcal{Y} < \infty$

(iv) F *is a Marchaud map*

(v) $d_n \to 0$ *as* $n \to \infty$ *and* $\sup_n \|d_n\| = d < \infty$

Then a linear interpolation of the iterative process $\{y_n\}_{n\in\mathcal{N}}$ *given by (13) is an asymptotic pseudo-trajectory of the differential inclusion*

$$\frac{dx}{dt} \in F(x) \tag{15}$$

Proof of Theorem D.5. We are going to use Theorem E.1 and the four conditions must be verified for stochastic process 11 so as its linear interpolation is an asymptotic pseudo-trajectory of 12.

To do this, we first define the discrete time system that Theorem E.1 will be applied to. We define $\tilde{M}_n := \mathrm{diag}(\max\{1_{s\in I_n} \frac{\gamma_{s^\sharp_n}}{\overline{\gamma}_n}, \epsilon\})$. Note that, consequently, $\tilde{M}_n \in \mathrm{diag}(\mathrm{conv}(\mathcal{S}) \bigcap [\epsilon, 1]^K) = \Omega^\epsilon_{k,\mathcal{S}}$. We select $f_n \in F(x_n)$ in the differential inclusion so as for every n, $y_{n+1} = y_n + \overline{\gamma}_{n+1} M_{n+1} \left[f_n + Y_{n+1} + d_{n+1}\right]$. Then define $\overline{Y}_{n+1} := f_n(M_{n+1} - \tilde{M}_{n+1}) + M_{n+1} V_{n+1}$, that is to say that $\overline{Y}_{n+1}$ is the noise Y_{n+1} plus the error induced by the fact that every state is updated at a minimum ϵ rate. Then we have $y_{n+1} = y_n + \overline{\gamma}_{n+1} \left[\tilde{M}_{n+1} f_n + \overline{Y}_{n+1} + \overline{d}_{n+1}\right]$.

So $y_{n+1} - y_n \in \overline{\gamma}_{n+1} \left(\Omega^\epsilon_{K,\mathcal{S}} \cdot F(y_n) + \overline{Y}_{n+1} + \overline{d}_{n+1}\right)$.

And now we verify assumptions of Theorem E.1:

(i) For $T > 0$:

$$\begin{aligned}
\sup_k &\left\{ \begin{matrix} \left\| \sum_{i=n}^{k-1} \overline{\gamma}_{i+1} \overline{Y}_{i+1} \right\| ; \\ k = n+1, \dots, \overline{m}(\overline{\tau}_n + T) \end{matrix} \right\} \\
&\leq \sup_k \left\{ \begin{matrix} \left\| \sum_{i=n}^{k-1} \overline{\gamma}_{i+1} M_{i+1} Y_{i+1} \right\| ; \\ k = n+1, \dots, \overline{m}(\overline{\tau}_n + T) \end{matrix} \right\} \\
&+ \sup_k \left\{ \begin{matrix} \left\| \sum_{i=n}^{k-1} \overline{\gamma}_{i+1} f_i (M_{i+1} - \tilde{M}_{i+1}) \right\| ; \\ k = n+1, \dots, \overline{m}(\overline{\tau}_n + T) \end{matrix} \right\}
\end{aligned}$$

The first part of the sum goes to 0 via classical Kushner-Clark condition and assumptions (iii) and (vi), the proof is detailled in Lemma 3.3 of (Perkins & Leslie, 2012). Regarding the second part, it is exactly Lemma 3.6 of (Perkins & Leslie, 2012) and this applies because of assumptions (iii), (iv) and (v).

(ii) This is assumption (iii).

(iii) This is assumption (i) of Theorem E.1.

(iv) The map is $\overline{F}(y) := \Omega^{\epsilon}_{k,\mathcal{S}} \cdot F(y)$ and it is Marchaud because F is Marchaud (assumption (ii)) and $\Omega^{\epsilon}_{k,\mathcal{S}}$ is compact (so every property of Definition D.1 holds).

(v) This is assumption (vii).

So Theorem E.1 applies and gives the desired result.

□

F. Technical Lemmas

Discrete-time Grönwall can be found in the literature with various assumptions. For the sake of completeness, we include here a version that matches the assumptions we have in our paper, with the associated proof. It is a differential version with error terms.

Lemma F.1 (Discrete-Time Grönwall). *Let $\{y_n\}$, $\{g_n\}$, $\{b_n\}$ sequences of real numbers such that $1 > 1 + g_n > 0$ for all n and:*

$$y_{n+1} - y_n \leq g_{n+1} y_n + b_{n+1}$$

Then $y_n \leq y_0 \Pi_{k=0}^{n}(1+g_k) + \sum_{k=0}^{n} b_k$.

Proof. We define $v_n := \frac{y_n - \sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)}$. We show that v_n is decreasing:

$$\begin{aligned}
v_{n+1} - v_n &= \frac{y_{n+1} - y_n}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{y_n}{\Pi_{k=0}^{n+1}(1+g_k)} - \frac{y_n}{\Pi_{k=0}^{n}(1+g_k)} - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{g_{n+1} y_n + b_{n+1}}{\Pi_{k=0}^{n}(1+g_k)} + \frac{y_n}{\Pi_{k=0}^{n}(1+g_k)} \left(\frac{1}{1+g_{n+1}} - 1 \right) - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{g_{n+1} y_n + b_{n+1}}{\Pi_{k=0}^{n}(1+g_k)} + \frac{y_n}{\Pi_{k=0}^{n}(1+g_k)} \frac{-g_{n+1}}{1+g_{n+1}} - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{b_{n+1}}{\Pi_{k=0}^{n+1}(1+g_k)} - \frac{\sum_{k=0}^{n+1} b_k}{\Pi_{k=0}^{n+1}(1+g_k)} + \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \\
&\leq \frac{\sum_{k=0}^{n} b_k}{\Pi_{k=0}^{n}(1+g_k)} \left(1 - \frac{1}{1+g_{n+1}} \right) \\
&\leq 0
\end{aligned}$$

And $v_0 = y_0$, hence the result.

□

Lemma F.2 (Bound on α_n). *Under hypothesis H1, for all n, $\frac{\alpha_n}{\sigma_n} \leq \frac{1}{n+1}$.*

Proof. By induction: for $n = 0$, $\frac{\alpha_n}{\sigma_n} = 1$. Now for $n+1$:

$$\frac{\alpha_{n+1}}{\sigma_{n+1}} \leq \frac{\alpha_n}{\sigma_n} \frac{\sigma_n}{\sigma_{n+1}} \leq \frac{1}{n+1} \left(1 - \frac{\alpha_{n+1}}{\sigma_{n+1}} \right)$$

As a consequence:

$$\frac{\alpha_{n+1}}{\sigma_{n+1}} \frac{n+2}{n+1} \leq \frac{1}{n+1}$$

And the result follows. □

G. Comparison with Existing Work

The table below summarizes the identical and different aspects of four systems: (asynchronous) fictitious play and best-response dynamics of our paper, and best-response dynamics of Leslie et al. (2020) and fictitious play of Sayin et al. (2020) which are the closest works to ours. We say that the timescales of actions and continuations are *different* when the ratio of the update rate of actions and the one of continuations goes to 0 when t goes to ∞.

System	Time	Classes of games with proven convergence to the set of equilibria	Timescales of actions and continuations
AFP	Discrete	Zero-sum (approximate convergence) and team games	Same or different timescales for team games, same timescale (or constant ratio) for zero-sum games
ABRD	Continuous	Identical-interest and zero-sum	Same or different timescales for identical-interest games, convergence with different timescales in zero-sum games, approximate convergence with same timescales in zero-sum games
Best-Response Dynamics of (Leslie et al., 2020)	Continuous	Zero-sum	Two timescales: update rate of $1/t$ for continuations and 1 for empirical actions
Fictitious Play of (Sayin et al., 2020)	Discrete	Zero-sum	Two timescales: strategies updated infinitely faster than Q-values

Figure 1. Comparison of recent adaptations of best-response dynamics and fictitious play in stochastic games