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

Lucas Baudin*[1]

[1]Université Paris-Dauphine-PSL, Paris, France, lucas.baudin@dauphine.eu

November 5, 2021

Abstract

This paper combines ideas from Q-learning [44] and fictitious play [8, 36, 30] to define three reinforcement learning procedures which converge to the set of stationary mixed Nash equilibria in identical interest discounted stochastic games. First, we analyse three continuous-time systems that generalize the best-response dynamics defined by Leslie et al. [26] for zero-sum discounted stochastic games. Under some assumptions depending on the system, the dynamics are shown to converge to the set of stationary equilibria in identical interest discounted stochastic games. Then, we introduce three analog discrete-time procedures in the spirit of Sayin et al. [37] and demonstrate their convergence to the set of stationary equilibria using our results in continuous time together with stochastic approximation techniques [5]. Some numerical experiments complement our theoretical findings.

1 Introduction

Learning equilibria of a static game is a subject that has been widely studied over the years [13, 16, 45, 10]. In standard game theory, the goal of learning might be prescriptive (how to design algorithms in multi-agent systems) or descriptive (which (kind of) equilibria is reached with a procedure that players might use). However, learning equilibria of stochastic games has comparatively been less developed, with noticeable exceptions [20, 27, 26, 37]. Stochastic games, introduced by Shapley [39], 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 [31] that is especially well suited for economic applications (see the survey by Amir [1] 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. [9] for a survey).

To learn stationary equilibria of stochastic games, this paper aims at combining a standard method of reinforcement learning on the one hand, with a standard game theory method on the other hand. A major trend in recent years is the advent of efficient reinforcement learning algorithms [40]. The optimization algorithm Q-learning [44] is one of the most successful model-free algorithms [23] with numerous extensions [17, 24]. On the game theory side, fictitious play [8, 36] is one of the most studied procedures to perform learning in games. Some recent proofs of the convergence of discrete time fictitious play rely on the theory of stochastic approximations that makes it possible to use results in continuous time [3, 5, 23, 33, 32]. The continuous-time counterpart of fictitious play is the best-response dynamics [29, 19] and we start this paper with a study of best-response dynamics in identical interest stochastic games.

*We acknowledge the valuable help of Rida Laraki, Guillaume Vigeral and Laurent Gourvès throughout the whole writing of this paper. We are also grateful to Sylvain Sorin for his remarks on earlier versions of this paper.

While adaptations of fictitious play for stochastic games were proposed earlier [42, 38], the idea of combining concepts from Q-learning and fictitious play emerged only recently with work of Perkins [32], Leslie et al. [26] on the one hand and Sayin et al. [37] on the other hand. In these papers, a mechanism similar to 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. These papers are dedicated to zero-sum games. 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. The present article extends these definitions to non zero-sum stochastic games and in particular prove convergence to equilibria for identical interest stochastic games [30], resulting in a decentralized algorithm for fully cooperative multi-agent reinforcement learning [9].

Contributions. Our contributions are as follows:

- We extend the definition of the best-response continuous-time dynamics of Leslie et al. [26] by loosening constraints on the update rates of the empirical actions and the expected payoffs. We define synchronous, asynchronous and semi-asynchronous versions of the procedure.
- We prove the convergence of the dynamics to the set of stationary equilibria for synchronous and semi-asynchronous versions in identical interest games for every discount rate. The convergence is also proven for the asynchronous version if the discount rate is small enough. We conjecture that it holds for all values of the discount rate and this is supported by simulations.
- We define procedures to play stochastic games in discrete time and in a fictitious play fashion. We establish a link between the limit sets of these discrete time algorithms and our continuous-time dynamical systems using the stochastic approximations framework. Then we deduce the convergence of the discrete-time systems in identical interest stochastic games using the results established for the continuous-time ones.

Outline Section 2 gives initial definitions and assumptions. The next section describes related work (which is complemented with Appendix D of the supplementary material). Then, continuous-time, best-response dynamics in identical interest stochastic games is defined in Section 4 and its convergence is studied. Section 5 describes several fictitious-play algorithms in discrete time for identical interest stochastic games. Extensions of these procedures and numerical simulations to support conjectures of Section 4 are in Section 6.

2 Preliminaries

Stochastic games We study dynamically interactive multi-agent systems based on the framework of stochastic games: two or more players can take actions over an infinite horizon. Their actions affect both the instantaneous payoff that they receive and the future state, which is the second determining factor of the total payoff. Therefore, compared to standard repeated games, 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. In this article, we focus on finite stochastic games: the state space, the action sets and the player set are finite.

Definition 1 (stochastic game)**.** *Stochastic games are tuples such as* $G = (S, I, (A^i)_{i\in I}, (r^i_s)_{i\in I, s\in S}, (P_s)_{s\in S})$ *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*,* $A := \prod_{i\in I} A^i$ *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 (whre* $\Delta(S)$ *is the set of probability distributions on* S*)*

The probability to go to state s' starting from s with action profile $a_s \in A$ is denoted by $P_{ss'}(a_s)$. Functions $P_{ss'}$ and r^i_s are linearly extended to mixed action profiles (i.e., $\prod_{i\in I}\Delta(A^i)^S$) and are therefore I-linear. For

an action profile $a_s \in A$ the action of a player i is denoted by a_s^i. The action profile obtained by replacing action of player i by b in action profile a_s is denoted by (b, a_s^{-i}). A strategy of player i is an element of $(A^i)^S$ and for $c \in (A^i)^S, a \in A^S$, (c, a^{-i}) is defined similarly. A strategy profile is an $|S|$-vector with action profile for every state (i.e., an element of A^S).

We are especially interested in one class of games: *identical interest stochastic games* where all players have the same stage reward function, i.e., for every state $s \in S$, there exists a function r_s such that for every player i, $r_s^i = r_s$ [30, 18].

We consider a sequence of play of a stochastic game in discrete time: 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, every player $i \in I$ chooses an action $a_n^i \in A^i$ and receives a stage reward of $r_{s_n}^i(a_n)$. 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 $(1-\delta)\sum_{k=0}^{\infty} \delta^k r_{s_k}^i(a_k)$ where $\delta \in (0,1)$ is the discount factor. It is well known that given a strategy profile, an expected total payoff can be defined, resulting in a $S \times I$ vector with an element for every initial state and player.

Definition 2 (Stationary Nash equilibrium)**.** *A stationary Nash equilibrium is a strategy profile $x \in \Pi_{i \in I} \Delta(A^i)^S$ such that no unilateral 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 $b \in (A^i)^S$.*

An equilibrium payoff is an $S \times I$-vector that corresponds to a strategy profile that is a stationary Nash equilibrium. There is a mixed stationary Nash equilibrium in every finite stochastic game [11].

For several results of this paper, stochastic games are supposed to be ergodic so that in any play, any state is visited infinitely often:

Definition 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.*

3 Related Work

Fictitious Play Fictitious play is a procedure to play repeated games. The central idea is that every player should play a best response to the empirically observed strategy of other players. It was initially proposed by Brown [8] and Robinson [36] to solve zero-sum (static) games (i.e., to determine which value every player can guarantee to itself). Numerous extensions have been studied for general-sum games and smooth best response [12, 13]. The convergence of such a procedure was proved in various cases such as zero-sum games, potential games [30], $2 \times n$ games [6], mean-field games [35], or for variants of the procedures such as smooth or vanishingly smooth fictitious play [4] or joint strategies with inertia [28]. It was less studied for stochastic games and there is no widely adopted definition yet. Vrieze and Tijs [42] used a fictitious play algorithm to compute the value of a stochastic game but it is not designed to be used during a play. Perkins [32] also defined a fictitious play procedure for identical interest and zero-sum games assuming that the player could solve Bellman equations. Perolat et al. [34] introduced actor critic fictitious-play for multi-stage games. Sayin et al. [37] recently introduced another variant of fictitious play for zero-sum stochastic games; it is detailed below and in Appendix D.

Best-response dynamics In continuous time, best-response dynamics [14] is based on the same principle as fictitious play: each player adjusts its mixed action towards the best-response to the current mixed action of other players. For normal form, non-stochastic, static games, it can be used as a continuous counterpart of fictitious play using the stochastic approximations framework [5]. In zero-sum stochastic games, Leslie et al. [26] defined an extension of the best-response dynamics with estimated payoff that are updated in the dynamics at a slower pace, see details below.

Q-learning Watkins [43] introduced the Q-learning algorithm designed to control Markov Decision Processes. It had a major impact and there are multiple generalizations, including offline Q-learning [24],

double Q-learning [17] or Q-learning with no-regret procedures [22]. There is a wide range of applications, from robots control [41] to SAT solving [25]. 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 [20, 21] or Team-Q and Minimax-Q [27]. For a complete survey, see [9] and references therein. Compared to these algorithms, our paper gives an algorithm that uses a model of the stochastic game (i.e., the transition probabilities and the payoff functions are known) but that does not require to solve intermediate games (which is computationally better).

Combining Q-learning and fictitious play To extend fictitious play 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. Sayin et al. [37] and Leslie et al. [26] use 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. [37] 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. [26] define a best-response dynamics for 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 $\arg\max_{a \in A^i} (1-\delta) r_s^i(a, x_s^{-i}) + \delta P_s(a, x_s^{-i}) \cdot u^i$.

Both papers [26, 37] are concerned with two plaers zero-sum stochastic games. They use proof techniques specific to thise kind of games. In our paper, our focus is on continuous-time dynamics and associated discrete-time algorithms that converge to the set of stationary mixed Nash equilibria in multiplayer identical interest stochastic games. See Appendix D for an extended comparison with these articles.

4 Continuous-Time Dynamics

In this section we study best-response dynamics in identical interest stochastic games based on a dynamics defined for zero-sum games by Leslie et al.. We extend this dynamics to identical interest games and with generalized update rates. We first look at a synchronous version and its convergence for ease of exposition. The results are then extended to the asynchronous and semi-asynchronous case.

Similarly to the dynamics of Leslie et al., 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 inverse of the update rates of variables u_s^i. Function α is continuous and non-decreasing. We make the following additional assumption on α: $\int_0^t \frac{1}{\alpha(y)} dy \xrightarrow[t \to \infty]{} +\infty$.

Auxiliary game Following Leslie et al. [26] and previous authors (for instance Shapley [39]), we define a so-called *auxiliary game* for every state s as a one-shot game parameterized by a vector $u := \{u_{s'}\}_{s' \in S}$ whose

action set is A and payoff of player i is, for any action profile $x_s \in A$, $f^i_{s,u}(x_s) = (1-\delta)r^i_s(x_s) + \delta \sum_{s' \in S} P_{ss'}(x_s)u_{s'}$. Vector u represents the continuation payoff believed by player i when it chooses an action, i.e., the expected payoffs starting from every state, hence the term $\delta \sum_{s' \in S} P_{ss'}(x)u_{s'}$ which is the discounted, expected payoff if players take action x.

Synchronicity and asynchronicity In our paper, we study three types of systems: synchronous, fully asynchronous and semi-asynchronous ones. In the synchronous kind, variables of all states are updated at the same time. There is no distinguished, current state. In semi-asynchronous systems, there is a current state, variables x^i_s are updated only if they are related to the current states but variables u^i_s are updated even if the current state is not s. In fully asynchronous systems, variables x^i_s and u^i_s are updated if and only if the current state is s.

Synchronous Dynamics In this dynamics, variables of all states are updated at the same time similarly to [42, 38]. This does not correspond to any play of the stochastic game. However, this can be seen either as an abstract view of a play or as an algorithm to compute an equilibrium. It is also mathematically more tractable, which is why we start with this case.

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

$$\begin{cases} \dot{u}^i_s(t) = \dfrac{f^i_{s,u^i(t)}(x_s(t)) - u^i_s(t)}{\alpha(t)} \\ \dot{x}^i_s(t) \in \mathrm{br}^i_{s,u^i(t)}(x_s(t)) - x^i_s(t) \end{cases} \tag{SBRD}$$

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

Remark: This is a generalization of the definition of Leslie et al. where $\alpha(t) = t+1$. Replacing $f^i_{s,u^i(t)}(x_s(t))$ by the maximum over actions, that is $\max_{a \in A^i} f^i_{s,u^i(t)}(a, x^{-i}_s(t))$ 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 $u^i_s(t)$ 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 [2, 5]. Indeed, one can rewrite it as $\frac{dy}{dt} \in F(t, y)$ where y is a vector with every u^i_s, x^i_s and F is a closed set-valued map, with non-empty, convex values. Furthermore, as shown in Lemma 1 of Appendix A, values are bounded, so the solution is defined on $\mathbb{R}^+$ [2, p. 97].

The rest of this section deals with identical interest games (i.e., $r^i_s = r_s$ for every player i). Therefore, for every s, u^i_s and f^i_{s,u^i} do not depend on i (when initial values are equal) and we omit the superscript i.

We aim at proving the following theorem that will later be transposed to the discrete-time case.

Theorem 1 (Convergence for identical interest stochastic games of SBRD)**.** *Let $\{u_s, x^i_s\}_{s \in S, i \in I}$ be a solution of SBRD. Then there exists $\Phi \in \mathbb{R}^{|S|}$ 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*

The proof proceeds as follows: first it is shown that the error term caused by wrong estimates of subsequent payoffs via the $\{u^i_s\}_{s \in S, i \in I}$ variable is bounded and then that this term does not make the convergence fail. Details are given at the end of the section and in Appendix A of the supplementary material.

Asynchronous Dynamics We now provide results regarding the convergence of semi-asynchronous systems and fully-asynchronous ones. In the semi-asynchronous one, the expected payoff starting from state s is always updated at the same rate but the empirical action is not, and for the fully asynchronous system, both the payoff estimates and the empirical action are updated at the same state-dependent rate.

The fully asynchronous system is defined as follows:

$$\begin{cases} \dot{u}_s(t) = \beta_s(t)\dfrac{f_{s,u(t)}(x_s(t)) - u_s(t)}{\alpha\left(\int_0^t \beta_s(y)dy\right)} \\ \dot{x}_s^i(t) \in \beta_s(t)\left(\mathrm{br}^i_{s,u(t)}(x_s^{-i}(t)) - x_s^i(t)\right) \\ \beta_s(t) \in [\beta_-, 1] \end{cases} \tag{ABRD}$$

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

The semi-asynchronous system (in the spirit of the system of Leslie et al. [26]) is:

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

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. However, our motivation to study this continuous time system is to prove the convergence of *discrete* time fictitious play in stochastic games where some variables are only updated for the current state. Using a theory developed by Perkins and Leslie, the rates can be supposed to be in $[\beta_-, 1]$ where $\beta_- \in (0,1]$. It represents an average on multiple time points in an ergodic stochastic game. This is a mathematically convenient way to use the ergodicity hypothesis.

Theorem 2 (Convergence for identical interest stochastic games)**.** *Let $\{u_s, \beta_s, x_s^i\}_{s\in S, i\in I}$ be a solution of SABRD. Then there exists $\Phi \in \mathbb{R}^{|S|}$ 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*

It also holds for solutions of ABRD when $\delta < \frac{1}{|S|}$.

A detailed sketch of the proof is provided below. A comprehensive proof with technical lemmas is provided in Appendix A. We conjecture it also holds for ABRD when $\delta \geq \frac{1}{|S|}$, see simulations in Section 6.

Proofs of Theorems 1 and 2 (sketch). We define, for $s \in S$:

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

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 1), it converges. We will then show that for every player i, $\Delta_s^i(t)ist \to 0$ and finish the proof of the Theorems by showing convergence of $\Gamma_s(t)$ and studying the limit set of $x_s(t)$.

Let $s \in S$. First, note that $\Delta_s^i(t) ist \geq 0$. Function Γ_s is differentiable:

$$\frac{d\Gamma_s}{dt} = \delta \sum_{s'} P_{ss'}(x_s)\dot{u}_{s'} + \beta_s(t) \sum_i \Delta_s^i(t) ist \tag{1}$$

where $\beta_s(t) = 1$ for SBRD and is already defined for SABRD and ABRD. See Lemma 2 in the supplementary material for details.

Lower bound of $\Gamma_s(t) - u_s(t)$ for SBRD and SABRD. The lower bound of $\Gamma_s(t) - u_s(t)$ is proven separately for SBRD and SABRD on the one hand and ABRD on the other hand. Until further notice, we suppose that $\{u_s, x_s\}_{s\in S}$ is **a solution of SABRD** (which includes the case SBRD).

Let $s_-(t) \in \arg\min_{s'\in S} (\Gamma_{s'}(t) - u_{s'}(t))$. Then, for any $s \in S$:

$$\begin{aligned}
\frac{d\Gamma_s}{dt} &\geq \delta \sum_{s'} P_{ss'}(x_s) \frac{\Gamma_{s'}(t) - u_{s'}(t)}{\alpha(t)} \\
&\geq \delta \sum_{s'} P_{ss'}(x_s) \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}}{\alpha(t)} \\
&= \delta \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}}{\alpha(t)}
\end{aligned} \tag{2}$$

Moreover, for $h > 0$:

$$\begin{aligned}
&\Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)) \\
&\geq \Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t+h)}(t) - u_{s_-(t+h)}(t)) \\
&\geq h \min_{s'\in S} \frac{d\Gamma_{s'}}{dt} + o(h) + u_{s_-(t+h)}(t) - u_{s_-(t+h)}(t+h)
\end{aligned} \tag{3}$$

Then, it can be shown that if s' is an accumulation point of $s_-(t+h)$ when h goes to 0, then:

$$u_{s'}(t) - u_{s'}(t+h) = -h \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)}{\alpha(t)} + o(h) \tag{4}$$

Since this is valid for every such s', combining (2), (3) and (4) gives:

$$\Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)) \geq h(\delta - 1) \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}}{\alpha(t)} + o(h)$$

Now we are going to apply a version of Grönwall Lemma (see details in Lemma 6) so we get:

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

Therefore, we can use this inequality in $\dot{u}_s(t)$:

$$\dot{u}_s(t) \geq -\frac{A}{\alpha(t)} \exp\left(\int_1^t (\delta - 1) \frac{1}{\alpha(t)} dt\right)$$

where $A > 0$. The right hand side term is integrable, and as u_s is bounded (see Lemma 1), it converges.

Lower bound of $\Gamma_s(t) - u_s(t)$ for ABRD. We now consider a solution of system ABRD. A similar reasoning can be done with function $\Psi(t) := \sum_{s\in S} (u_s(t) - \Gamma_s(t))_+$, see Lemma 8 for details. We can bound Ψ and in the end we can also bound $\dot{u}_s$

$$\dot{u}_s(t) \geq -A \frac{\exp\left(\int_0^t (\delta|S| - 1) \frac{\beta_-}{\alpha\left(\int_0^v \beta_s(w)dw\right)} dv\right)}{\alpha\left(\int_0^t \beta_s(v)dv\right)}$$

which is integrable. Therefore, u_s converges with the same arguments.

$\sum_{i\in I}\Delta_s^i(t)ist$ **goes to** 0. In either case, we show that $\Delta_s^i(t)ist \to 0$. First, we notice that in the SBRD and SABRD case:

$$\begin{aligned}\int_0^t \sum_{i\in I}\Delta_s^i(t)isvdv &\le \int_0^t \frac{\beta_s(u)}{\beta_-}\sum_{i\in I}\Delta_s^i(t)isvdv\\ &= \frac{1}{\beta_-}\left(\int_0^t \frac{d\Gamma_s}{dt}(v) - \delta\sum_{s'}P_{ss'}(x_s(v))\dot{u}_{s'}(v)dv\right)\\ &\le \frac{1}{\beta_-}\left(\Gamma_s(t)-\Gamma_s(0)\right) + \frac{A}{\beta_-(1-\delta)}\left(1-\exp\left(\int_1^t\frac{1-\delta}{\alpha(v)}dv\right)\right)\end{aligned}$$

So, this integral is bounded. However, as $\sum_{i\in I}\Delta_s^i(t)is\cdot$ is Lipschitz (see Lemma 3), we conclude that $\sum_{i\in I}\Delta_s^i(t)ist \xrightarrow[t\to\infty]{} 0$ (Lemma 5).

In the ABRD case, we have similar inequalities except for the argument of α which is $\int_0^t\beta_s(v)dv$. As it is bounded between $\beta_- t$ and t, it does not change much of the computations and the integral is bounded as well. See Lemma 9 for details.

Convergence of Γ_s. In the case of SBRD and SABRD, using (1), we can lower bound its derivative:

$$\frac{d\Gamma_s}{dt} \ge \delta\frac{\Gamma_{s_-(t)}(t)-u_{s_-(t)}(t)}{\alpha(t)} \ge -\delta A\frac{\exp\left(\int_0^t(\delta-1)\frac{1}{\alpha(u)}du\right)}{\alpha(t)}$$

This latest term being integrable and Γ_s being bounded, we conclude that Γ_s converges to its lim sup when t goes to $+\infty$. The limit is necessarily the same as u_s, otherwise u_s could not be bounded (see the comprehensive proof for details).

Limit set of $x_s(t)$. Let $\tilde{x}$ be an accumulation point of the vector-valued function $x = \{x_s\}$. Then, we previously showed:

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

So by continuity, for all s:

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

So $\tilde{x}$ belongs to the set of Nash equilibria.

□

5 Discrete-Time Procedures

One important consequence of the convergence of the best-response dynamics in identical interest stochastic games is that it makes it possible to study a discrete-time fictitious-play like procedure for these games. In particular, we are going to define an asynchronous procedure that can be used as an online control algorithm, which is a novelty compared to [26].

Procedures to play stochastic games A procedure to play stochastic games for a player i is a behavioral strategy, 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 $S\times\bigcup_{n\in\mathbb{N}}(S\times A)^n \to \Delta(A^i)$. This results in a process that we characterize as *asynchronous*: there is a unique current state (that every player is aware of) and actions are provided by players for this state. This is defined in contrast to *synchronous* systems where there is no specific current state and players provide actions for all states at every step, i.e., a mapping $\bigcup_{n\in\mathbb{N}}(A^S)^n \to \Delta(A^i)^S$. This may not be useful for online control applications but nevertheless, it can be interpreted in various ways: either as a simulation of the real system or as a pre-computation phase.

Fictitious play for identical interest games Similarly to the best-response dynamics, our discrete-time procedure is designed using two estimates per state: one is the empirical action that every player uses (identified with a mixed action) and the other is the expected total payoff that a player estimates starting from this state. First, we define the two estimates and then proceed with a description of the action selection.

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., $s_n^\sharp = \sharp\{k \mid 0 \leq k \leq n \wedge s_k = s\}$. Then the empirical action of player i in state s is defined in $\Delta(A^i)$ as:

$$x^i_{n+1,s} := \frac{1}{s_n^\sharp} \sum_{k=0}^{n} 1_{s_k=s} a^i_k = \frac{1_{s_n=s} a^i_n}{s_n^\sharp} + \frac{s_{n-1}^\sharp x^i_{n,s}}{s_n^\sharp}$$

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

Payoff estimates The second estimate $u^i_{n,s} \in \mathbb{R}^S$ is a vector of payoffs. Values of this vector are written $u^i_{n,s}$ for state s, at step n for player i. We also use the generic notation u^i for a vector of expected payoffs (with corresponding notation u^i_s for payoffs starting from s when the difference between s and n is not ambiguous). At every step n, the estimator is defined as:

$$u^i_{n+1,s} := \frac{1}{s_n^\sharp} \sum_{k=0}^{n} 1_{s_k=s} f^i_{s,u^i_k}(x_{k,s})$$

with, as previously, starting values $u^i_{0,s}$ defined arbitrarily.

Action selection We can now define the procedure. It is an extension of the standard fictitious play procedure. For repeated games, a fictitious play procedure is defined as a procedure where at every stage, every player takes a best response against the empirical action of the opponents up to that stage. Here, for stochastic games, fictitious play is defined as follows for every n:

$$a^i_{n,s} \in \mathrm{br}^i_{s,u^i_n}(x^{-i}_{n,s}) := \arg\max_{y \in A^i} f^i_{s,u^i_n}(y, x^{-i}_{n,s})$$

where $s = s_n$. Therefore, the action is a best response in the auxiliary game parameterized by u^i_n.

Incremental updates Both $x^i_{n,s}$ and $u^i_{n,s}$ can be computed *via* incremental updates. This will enable us to make the link with the continuous version. It also shows that, for a machine implementation, the procedure only needs constant memory instead of storing every value to calculate the average actions and payoffs. Thus, the system can be rewritten as follows, leading to asynchronous fictitious play (AFP):

$$\begin{cases} u^i_{n+1,s} - u^i_{n,s} = \dfrac{1_{s_n=s}}{s_n^\sharp}\left(f^i_{s,u^i_n}(x_{n,s}) - u^i_{n,s}\right) \\ x^i_{n+1,s} - x^i_{n,s} = \dfrac{1_{s_n=s}}{s_n^\sharp}\left(a^i_{n,s} - x^i_{n,s}\right) \end{cases} \tag{AFP}$$

Synchronous and semi-asynchronous fictitious play To get a simplified version of the dynamical system, we now define a version with synchronous updates on every state. The continuous counterpart is simpler as it does not require an indicator function to specify the current state.

If the stochastic game is played synchronously in every state (meaning that for every state s and player i there is a choice of action $a_s^i \in A_s^i$ at every time step), the empirical action of player i in state s is:

$$x_{n+1,s}^i = \frac{\sum_{k=0}^n a_{k,s}^i}{n} = \frac{a_{n,s}^i}{n} + \frac{(n-1)x_{n,s}^i}{n}$$

and the estimated payoff starting from s is updated as follows:

$$u_{n+1,s}^i = \frac{\sum_{k=0}^n f_{s,u_k^i}^i(x_{k,s})}{n} = \frac{f_{s,u_n^i}^i(x_{n,s})}{n} + \frac{(n-1)u_{n,s}^i}{n}$$

These updates can be written in an incremental fashion, leading to synchronous fictitious play:

$$\begin{cases} u_{n+1,s}^i - u_{n,s}^i = \dfrac{f_{s,u_n}^i(x_{n,s}) - u_{n,s}^i}{n} \\ x_{n+1,s}^i - x_{n,s}^i = \dfrac{a_{n,s}^i - x_{n,s}^i}{n} \end{cases} \tag{SFP}$$

An alternative to both SFP and AFP is semi-asynchronous fictitious play. It can be used during a standard asynchronous play of the stochastic game: $a_{n,s}^i$ are needed to update $x_{n,s}^i$ but not needed to update $u_{n,s}^i$, so only the updates on $x_{n,s}^i$ are asynchronous in SAFP:

$$\begin{cases} u_{n+1,s}^i - u_{n,s}^i = \dfrac{f_{s,u_n}^i(x_{n,s}) - u_{n,s}^i}{n} \\ x_{n+1,s}^i - x_{n,s}^i = 1_{s_n=s} \dfrac{a_n^i - x_{n,s}^i}{s_n^\sharp} \end{cases} \tag{SAFP}$$

In identical interest games, as noted above, vectors u_n^i are independent of player i as soon as players start with the same belief, i.e., for all $i, j \in I$, $u_0^i = u_0^j$, and $f_{s,u}^i(x) = f_{s,u}(x)$.

We now state the main convergence result for our definition of fictitious play for identical interest stochastic games.

Theorem 3 (Discrete stochastic fictitious play converges to an equilibrium)**.** *Procedures SFP and SAFP almost surely converge to the set of stationary mixed Nash equilibria for identical interest ergodic stochastic games and if $\delta < 1/|S|$, then the results also hold for AFP. It also holds for non ergodic games for SFP.*

The proof of this latest theorem is sketched in the rest of the section. It uses the stochastic approximation framework. We recall the classical theory and the asynchronous extension that we need to modify.

Stochastic approximations Stochastic approximations have long been used to study asymptotic behavior of discrete-time systems using their continuous-time counterpart [3, 5, 23]. In this framework, one typically assumes that there is a set valued function $F : \mathbb{R}^K \rightrightarrows \mathbb{R}^K$, a sequence of decreasing update steps $\{\gamma_n\} \in \mathbb{R}^{\mathbb{N}}$ and Y_{n+1} a noise difference random variable. Then the two following systems are related:

$$\frac{dy}{dt} \in F(y) \tag{5}$$

$$y_{n+1} - y_n - \gamma_{n+1} Y_{n+1} \in \gamma_{n+1} F(y_n) \tag{6}$$

Stochastic approximations theorems (see Appendix B for a precise statement) typically assert that the limit set of a solution of (6) is *internally chain transitive* for differential inclusion (5), meaning that two points of the limit set must be linked by a number of chained solutions of (5):

Definition 4 (Internally chain transitive)**.** *A set A is internally chain transitive (ICT) 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 $y_1, \dots y_n$ to the differential inclusion and real numbers $t_1, t_2, \dots, t_n$ 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* $\|y_n(t_n) - y'\| \leq \epsilon$

This framework links the synchronous systems: SFP and SBRD where $\alpha(t) = 1$ (see the next section for a discussion regarding other values of α). In our case, $y = \{u_s, x_s\}_{s \in S}$, there is no random noise and a vector of $F(y)$ is composed of an element of $\mathrm{br}^i_{s,u}(x_s^{-i}) i s u x_s^{-i} - x_s^i$ for lines corresponding to x_s^i and $f_{s,u}(x_s) - u_s$ for lines corresponding to u_s. This proves the part related to SFP of Theorem 3, see Appendix B.

Correlated Asynchronoucity To do similar proofs for systems SAFP and AFP, one needs asynchronous stochastic approximations: the standard stochastic approximation framework [5] is not sufficient to track every state and make every update rate depends on $s_n^\sharp$. We extend results from Perkins and Leslie [33] to the case of correlated asynchronoucity. Indeed, in SAFP, variables u_s are updated at every time step, independently of the current state and in AFP, variables u_s and x_s are updated at the same time. Therefore, as Perkins and Leslie [33] provided a theory where every variable was updated asynchronously with respect to the other ones. We extend this result, see Appendix B, and apply it to use systems ABRD and SABRD to prove Theorem 3 under the ergodicity hypothesis.

Proof of Theorem 3 (sketch). The first step of the proof is a generalization of the asynchronous stochastic approximation theorem of [33] to deal with correlated asynchronicity, this is done in Appendix C.

Then, systems are written in the form of (5) and (6). This is explained above.

Thus, it remains to show that the ICT sets of (5) are contained in the set of stationary Nash equilibria and their associated payoff. To do this, we use results of convergence of the previous section. Howevbrefer, there is no direct implication between the convergence to a set and the fact that this set is ICT. Therefore, the chain transitivity is proven in part using the original definition, and in part with a Lyapunov function. More precisely we want to show that any ICT set is included in:

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

First, we show that any element (x,u) of ICT sets are in B. Otherwise, we look at the chain between (x,u) and itself, and conclude to a paradox: any solution is arbitrarily close to B after a time T independently of the starting point. Then, relatively to B, we can define a function $V(x,u) := \sum_{s \in S} f_{s,u}(x_s)$ which is a Lyapunov function in B. This makes it possible to conclude that any ICT set is included in $V^{-1}(0)$ which is equal to A.

The whole proof is detailed in Appendix B of the supplementary material.

□

6 Extensions and Simulations

Generalization of fictitious play The procedure we defined in the previous section is a special case of a more general procedure defined as follows:

$$u^i_{n+1,s} := \left(\sum_{k=0}^{s^\sharp_n} \frac{1}{\alpha(k)} \right)^{-1} \sum_{k=0}^{n} \frac{1_{s_k=s}}{\alpha(s^\sharp_k)} f^i_{s,u^i_k}(x_{k,s})$$

where α is supposed non-decreasing. The former estimate corresponds to the case where α is the constant function equal to one.

This latest estimate would be the counterpart to the continuous system studied in the previous section. However, the link between the continuous and discrete systems is not straightforward. Therefore the convergence of the general version for identical interest games remains an open problem: the continuous system is non-autonomous and there is no general theory of stochastic approximations for such systems. An option could be to use two-timescale stochastic optimizations in the spirit of the work of Borkar [7] in a way similar to Sayin et al. [37]. However, there are regularities assumptions satisfied in zero-sum stochastic games that are not trivial to verify in our case and the continuous time systems would be different.

We now describe the simulations that we have conducted in order to assess the two conjectures that we made namely the case $\delta \geq 1/|S|$ for ABRD of Theorem 2 and $\alpha(k) \neq 1$ for Theorem 3. They were all run on a personal computer (Intel i5) in less than one hour.

Asynchronous continuous time system when $\delta \geq 1/|S|$ We ran a simulation of differential inclusion ABRD using a simple Euler method. Our extensive simulations suggest that the system always converges for any randomly generated game. If our conjectures were disproved, then we believe that it would occur in very intricate cases such as degenerated stochastic games or with more players, actions and states. It is also possible that the system does not converge when rates $\beta_s(t)$ are correlated, which is especially difficult to simulate with a randomization. For the sake of readability, our figures only contain the u_s, but actions converge quickly towards a pure action profile. There may be several equilibria, so the system does not always converge to the same one.

The random games that we study have 2 actions, 2 states and 2 players. We randomly chose an instance so as it is completely reproducible and plotted u_1, u_2, Δ_1, Δ_2 on Figure 1 where $\Delta_s := f_{s,u}(x) - u_s$ is the error in the Bellman equation ($\Delta_s = 0$ would mean that u_1 and u_2 are solutions of the Bellman equation). Other instances can be tried in the notebook in the supplementary material. This is the instance chosen as example:

$$P_1 = \begin{bmatrix} 0.09 & 0.05 \\ 0.95 & 0.79 \end{bmatrix} P_2 = \begin{bmatrix} 0.20 & 0.66 \\ 0.79 & 0.54 \end{bmatrix} r_1 = \begin{bmatrix} 0.65 & 0.37 \\ 0.00 & 0.73 \end{bmatrix} r_2 = \begin{bmatrix} 0.19 & 0.07 \\ 0.30 & 0.07 \end{bmatrix} \delta = 0.7$$

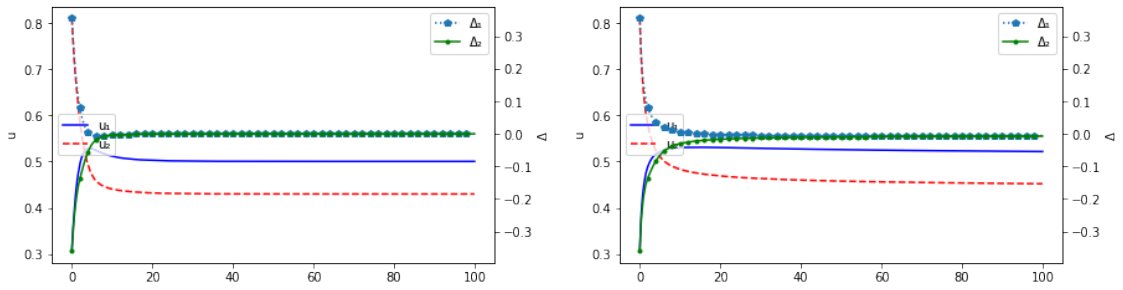

Figure 1: Simulation of the asynchronous best-response dynamics ABRD for $\alpha(t) = 1$ (top) and $\alpha(t) = t + 1$ (bottom)

The notebook of the simulation in the supplementary material and plots in Appendix E show that actions also converge. Vector (u_1, u_2) converges towards $(0.50, 0.42)$ which is the payoff of the stationary strategy that actions x^i_s converge to (see the notebook and Appendix E). This stationary strategy is a stationary

Nash equilibrium of the stochastic game: it is verified in the notebook by computing the discounted payoff of every pure strategy. Thus we conjecture that Theorem 2 holds for $\delta \geq 1/|S|$. Furthermore, the continuous time system with $\alpha(t) = 1$ converges, in the simulations, more rapidly than the one with $\alpha(t) = t + 1$.

Fictitious play with two timescales We ran our procedure ABRD on the previous instance with $\alpha(k) = 1$ and $\alpha(k) = k + 1$. Compared to ABRD with $\alpha(t) = 1$, fictitious play with $\alpha(k) = 1$ converges exponentially slower (which is predicted by the theory of stochastic approximations). Since fictitious play with $\alpha(k) = k+1$ converges exponentially slower compared to the version with $\alpha(k) = 1$, this results in a double exponential slowdown. Figure 2 shows simulations of this system with a logarithmic scale on the time coordinate and different time scales on every graph. Similarly to the continuous time systems, vector u converges to an equilibrium payoff which is $(0.50, 0.42)$ but other equilibria exist (see the notebook and Appendix E for details).

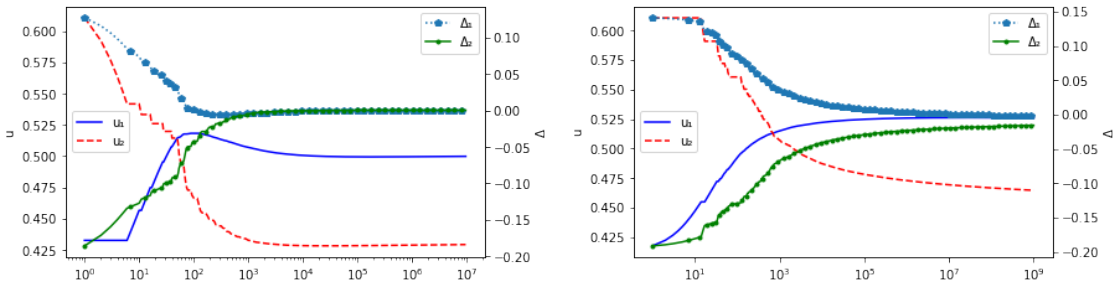

Figure 2: A fictitious play process AFP with $\alpha(k) = 1$ (top) and $\alpha(k) = k + 1$ (bottom)

7 Discussion

In this paper, we defined a number of continuous-time dynamics and discrete time systems. They are inspired by standard fictitious play and Q-learning procedures. We proved the convergence to stationary Nash equilibria for identical interest discounted stochastic games under assumptions depending on the system. We discuss below limitations and extensions of our work.

Extensions of our procedure There are a number of directions that deserve additional investigations. First, proving convergence for identical interest stochastic games of fictitious play for different values of α, and the case $\delta \geq 1/|S|$ for asynchronous systems. It would also be interesting to investigate whether these algorithms work in other types of games. To do so we expect that it will be necessary to switch to smooth best-responses in order to ensure general convergence (similarly to standard fictitious play [13]). Second, this version of fictitious play should be made more robust. In particular, it would be interesting to investigate whether different priors (i.e., different initial values for x and u^i) for every player perturb the algorithm. This could be useful in order to use the procedure in a fault-tolerant system.

Potential games The proof of convergence of fictitious play in identical interest (non-stochastic) games of [30] can be applied to potential games because these games have the same best-reply structure. An interesting question is to what extents there exists an analog of potential games for stochastic games [18] and whether our procedure can be used in those games as well.

Update rules and model-free algorithm In our paper, we suppose that every player knows the transition matrix, which is an assumption standard in game theory but less common in multi-agent systems. As noted in the related work section, other update rules that are closer to Q-learning would lead to a different

algorithm. It would be a step towards a model-free algorithm when players do not know the transition matrix. There are a number of challenges, including the fact that every state-action pair has to be visited an infinite number of times for the transition to be estimated. Comparing theoretically or experimentally the speed of convergence of model-free and model-based version of fictitious play (with respect to the transition and payoff matrices) would also be interesting.

References

[1] Rabah Amir. Stochastic Games in Economics and Related Fields: An Overview. In Abraham Neyman and Sylvain Sorin, editors, *Stochastic Games and Applications*, NATO Science Series, pages 455–470, Dordrecht, 2003. Springer Netherlands. ISBN 978-94-010-0189-2. doi: 10.1007/978-94-010-0189-2_30.

[2] Jean-Pierre Aubin and Arrigo Cellina. *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.

[3] Michel Benaim. A Dynamical System Approach to Stochastic Approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, March 1996. ISSN 0363-0129, 1095-7138. doi: 10.1137/S0363012993253534.

[4] Michel Benaïm and Mathieu Faure. 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.

[5] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. 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.

[6] Ulrich Berger. 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.

[7] Vivek S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, February 1997. ISSN 01676911. doi: 10.1016/S0167-6911(97)90015-3.

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

[9] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 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.

[10] Nicolò Cesa-Bianchi and Gábor Lugosi. *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.

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

[12] Drew Fudenberg and David K. Levine. 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.

[13] Drew Fudenberg and David K. Levine. *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.

[14] Christopher Harris. 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.

[15] Sergiu Hart and Andreu Mas-Colell. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 68(5):1127–1150, 2000. ISSN 0012-9682.

[16] Sergiu Hart and Andreu Mas-Colell. *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.

[17] Hado Hasselt. Double q-learning. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[18] John Edward Holler. *Learning Dynamics and Reinforcement in Stochastic Games*. PhD thesis, University of Michigan, 2020.

[19] Ed Hopkins. A Note on Best Response Dynamics. *Games and Economic Behavior*, 29(1-2):138–150, October 1999. ISSN 08998256. doi: 10.1006/game.1997.0636.

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

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

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

[23] Vijaymohan R. Konda and Vivek S. Borkar. 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.

[24] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc., 2020.

[25] Vitaly Kurin, Saad Godil, Shimon Whiteson, and Bryan Catanzaro. Can q-learning with graph networks learn a generalizable branching heuristic for a SAT solver? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9608–9621. Curran Associates, Inc., 2020.

[26] David S. Leslie, Steven Perkins, and Zibo Xu. 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.

[27] Michael L. Littman. 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.

[28] Jason R. Marden, GÜrdal Arslan, and Jeff S. Shamma. Joint Strategy Fictitious Play With Inertia for Potential Games. *IEEE Transactions on Automatic Control*, 54(2):208–220, February 2009. ISSN 0018-9286. doi: 10.1109/TAC.2008.2010885.

[29] Akihiko Matsui. 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.

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

[31] Abraham Neyman and Sylvain Sorin, editors. *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.

[32] Steven Perkins. *Advanced Stochastic Approximation Frameworks and Their Applications*. PhD thesis, University of Bristol, September 2013.

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

[34] Julien Perolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 919–928. PMLR, April 2018.

[35] Sarah Perrin, Julien Perolat, Mathieu Lauriere, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13199–13213. Curran Associates, Inc., 2020.

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

[37] Muhammed O. Sayin, Francesca Parise, and Asu Ozdaglar. Fictitious play in zero-sum stochastic games. *arXiv:2010.04223 [cs, math]*, October 2020.

[38] G. Schoenmakers, J. Flesch, and F. Thuijsman. Fictitious play in stochastic games. *Mathematical Methods of Operations Research*, 66(2):315–325, September 2007. ISSN 1432-2994, 1432-5217. doi: 10.1007/s00186-007-0158-9.

[39] L. S. Shapley. 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.

[40] Richard S. Sutton and Andrew G. Barto. *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.

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

[42] O. J. Vrieze and S. H. Tijs. 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.

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

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

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

A Convergence of the Continuous Time System

We first prove a boundedness lemma and then proceed with the convergence of every system in identical interest stochastic games. The first lemma can be proven for systems SBRD, SABRD and ABRD at once using a more general system that contains every other ones (for $t \geq 0$):

$$\begin{cases} \dot{u}_s^i(t) \in \left\{ \beta_s(t) \dfrac{f^i_{s,u^i(t)}(x_s(t)) - u_s^i(t)}{\alpha(t)}, \dfrac{f^i_{s,u^i(t)}(x_s(t)) - u_s^i(t)}{\alpha(t)} \right\} \\ \dot{x}_s^i(t) \in \beta_s(t) \left(\mathrm{br}^i_{s,u^i(t)}(x_s(t)) - x_s^i(t) \right) \\ \beta_s(t) \in [\beta_-, 1] \end{cases} \tag{7}$$

Since this system includes every other system studied in this article, the next lemmas are applicable to all of them.

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 4).

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 1. *Let $\{u_s^i, x_s^i\}_{s \in S, i \in I}$ a solution of (7). 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 $\dot{u}_s(t) = \beta_t(\Gamma_s(t) - u_s(t))/\alpha(t) \leq 0$ for some β_t, therefore $u_s(t^-) \geq M$, which is absurd.
- $u_s(t) = -M$ and $|\Gamma_s(t)| < M$, therefore $\dot{u}_s(t) = \beta_t(\Gamma_s(t) - u_s(t))/\alpha(t) \geq 0$ 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 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 3. *Function $\Delta_s^i(t) is\cdot$ 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 $\Delta_s^i(t) is\cdot$ is also Lipschitz. □

Convergence of the synchronous and semi-synchronous system In what follows $\{u_s^i, x_s^i\}_{s\in S, i\in I}$ is a solution of SABRD for identical interest stochastic games. In particular, this includes the synchronous case because a solution of SBRD is a solution of SABRD.

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

Proof. By the previous lemma:

$$\begin{aligned}
\frac{d\Gamma_s}{dt} &\geq \delta \sum_{s'} P_{ss'}(x_s)\frac{\Gamma_{s'}(t) - u_{s'}(t)}{\alpha(t)} \\
&\geq \delta \sum_{s'} P_{ss'}(x_s)\frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}}{\alpha(t)} \\
&= \delta \frac{\Gamma_{s_-(t)}(t) - u_{s_-(t)}}{\alpha(t)}
\end{aligned} \tag{8}$$

Moreover, for $h > 0$:

$$\begin{aligned}
&\Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t)}(t) - u_{s_-(t)}(t)) \\
&\geq \Gamma_{s_-(t+h)}(t+h) - u_{s_-(t+h)}(t+h) - (\Gamma_{s_-(t+h)}(t) - u_{s_-(t+h)}(t)) \\
&\geq h\min_{s\in S}\frac{d\Gamma_s}{dt} + o(h) + u_{s_-(t+h)}(t) - u_{s_-(t+h)}(t+h)
\end{aligned} \tag{9}$$

For any s:

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

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:

$$u_s(t) - u_s(t+h) = -h\frac{\Gamma_{s\_(t)}(t) - u_{s\_(t)}(t)}{\alpha(t)} + o(h)$$

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:

$$u_{s\_(t+h)}(t) - u_{s\_(t+h)}(t+h) = -h\frac{\Gamma_{s\_(t)}(t) - u_{s\_(t)}(t)}{\alpha(t)} + o(h)$$

Now, from inequality (8), we have that:

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

And these two last inequalities can be summed to get:

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

Going back to (9):

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

We now need a version of Grönwall Lemma that applies to this case, it is provided here for completeness:
Let $v(t) = \exp\left(\int_0^t (\delta-1)\frac{1}{\alpha(t)}dt\right)$.
Then $\frac{dv}{dt} = (\delta-1)\frac{1}{\alpha(t)}v(t)$, $v(0) = 1$, $v > 0$.
and $\frac{1}{v(t+h)} = \frac{1}{v(t)} - h(\delta-1)\frac{1}{\alpha(t)v(t)} + o(h)$

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

Therefore, $t \mapsto \frac{\Gamma_{s\_(t)}(t) - u_{s\_(t)}(t)}{v(t)}$ is increasing, and:

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

□

Lemma 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 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:

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

Then, for all T:

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

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

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

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

$$\text{and } u_s(t) \xrightarrow[t\to\infty]{} \limsup \Gamma_s$$

Proof.

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

So $\frac{A\delta}{1-\delta}\exp\left(\int_1^{t_1}(\delta-1)\frac{1}{\alpha(t)}dt\right)$ goes to 0 when t_1 goes to $+\infty$ (thanks to hypothesis (4)), therefore, it is sufficient to take t_1 big enough so that $\Gamma_s(t_1)$ is close to the lim sup 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 $\dot{u}_s = (\Gamma_s - u_s)/\alpha(t)$). □

Lemma 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 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 1 and SABRD part of Theorem 2.
Lemma 7 and 6 prove the theorems for both the SBRD and SABRD systems. □

Convergence of the asynchronous system We now suppose that $\delta \in (0, 1/|S|)$. In what follows $\{u_s^i, x_s^i\}_{s\in S, i\in I}$ is a solution of ABRD for identical interest stochastic games.

We use the same definitions for $\Delta_s^i(t)$ and $\Gamma_s(t)$:

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

We now define Ψ which measures how much the estimated payoffs u_s are over estimated compared to $\Gamma_s(t)$:

$$\begin{aligned}\Psi(t) &= \sum_{s\in S}(u_s(t) - \Gamma_s(t))_+\\ &= \sum_{s\in S} 1_{u_s-\Gamma_s(t)\geq 0}(u_s(t) - \Gamma_s(t))\end{aligned}$$

Lemma 8. *For a solution* $\{u_s, x_s^i, \beta_s\}_{i\in I, s\in S}$ *of ABRD:*

- $\Psi(t) \leq \Psi(1)\exp\left(\int_1^t(\delta|S|-1)\frac{\beta_-}{\alpha(u)}du\right)$
- *every* u_s *converges*
- *every* Γ_s *converges*
- *the limits of* u_s *and* Γ_s *are the same.*

Proof. First, for readability, we pose $1_s(t) := 1_{u_s-\Gamma_s(t)>0}$.

Ψ is continuous and differentiable almost everywhere (if there is an accumulation point where for a s, $u_s(t)-\Gamma_s(t)=0$, then $\frac{d}{dt}(u_s-\Gamma_s)=0$, so $1_{u_s(t)-\Gamma_s(t)\geq 0}(u_s-\Gamma_s)$ is differentiable and its derivative is 0). :

$$\begin{aligned}
\frac{d\Psi}{dt} &= \sum_{s\in S} 1_s(t)\left(\begin{array}{l}\beta_s(t)\dfrac{\Gamma_s(t)-u_s(t)}{\alpha\left(\int_0^t\beta_s(u)du\right)}-\beta_s(t)\displaystyle\sum_i\Delta_s^i(t)\\ -\delta\displaystyle\sum_{s'}P_{ss'}(x_s)\beta_{s'}(t)\dfrac{\Gamma_{s'}(t)-u_{s'}(t)}{\alpha\left(\int_0^t\beta_{s'}(u)du\right)}\end{array}\right)\\
&\leq \sum_{s\in S} 1_s(t)\left(\begin{array}{l}\beta_s(t)\dfrac{\Gamma_s(t)-u_s(t)}{\alpha\left(\int_0^t\beta_s(u)du\right)}\\ -\delta\displaystyle\sum_{s'}P_{ss'}(x_s)\beta_{s'}(t)\dfrac{\Gamma_{s'}(t)-u_{s'}(t)}{\alpha\left(\int_0^t\beta_{s'}(u)du\right)}\end{array}\right)\\
&= \sum_{s\in S}\left(-\delta\sum_{s'}1_{s'}(t)P_{s's}(x_{s'})\right)\beta_s(t)\frac{\Gamma_s(t)-u_s(t)}{\alpha\left(\int_0^t\beta_s(u)du\right)}
\end{aligned}$$

- If $u_s(t)-\Gamma_s(t)<0$, then the summed term is equal to $-\delta\sum_{s'}1_{s'}(t)P_{s's}(x_{s'})\beta_s(t)\frac{\Gamma_s(t)-u_s(t)}{\alpha\left(\int_0^t\beta_s(u)du\right)}\leq 0$. Since $(u_s(t)-\Gamma_s(t))_+=0$, the summed term is lower than:

$$-(1-\delta|S|)\frac{\beta_-}{\alpha(t)}(u_s(t)-\Gamma_s(t))_+$$

- If $u_s(t)-\Gamma_s(t)>0$, then the summed term is equal to $(1-\delta\sum_{s'}1_{s'}(t)P_{s's}(x_{s'}))\beta_s(t)\frac{\Gamma_s(t)-u_s(t)}{\alpha\left(\int_0^t\beta_s(u)du\right)}$
  - $\delta\sum_{s'}1_{s'}(t)P_{s's}(x_{s'})\leq\delta|S|<1$
  - so $1-\delta\sum_{s'}1_{s'}(t)P_{s's}(x_{s'})>0$.
  - so $(1-\delta\sum_{s'}1_{s'}(t)P_{s's}(x_{s'}))\beta_s(t)\frac{\Gamma_s(t)-u_s(t)}{\alpha\left(\int_0^t\beta_s(u)du\right)}<0$
  - but $\beta_s(t)\geq\beta_-$ and $\alpha\left(\int_0^t\beta_s(u)du\right)\leq\alpha(t)$,
    so $\frac{\beta_s(t)}{\alpha\left(\int_0^t\beta_s(u)du\right)}\geq\frac{\beta_-}{\alpha(t)}$
  - so the summed term is lower than:

$$(1-\delta|S|)\frac{\beta_-}{\alpha(t)}(\Gamma_s(t)-u_s(t))$$
$$=-(1-\delta|S|)\frac{\beta_-}{\alpha(t)}(u_s(t)-\Gamma_s(t))_+$$

Consequently:

$$\frac{d\Psi}{dt}\leq-(1-\delta|S|)\frac{\beta_-}{\alpha(t)}\Psi(t)$$

So by Grönwall Lemma (see Lemma 4 for the details of the application of Grönwall Lemma):

$$\Psi(t)\leq\Psi(1)\exp\left(\int_1^t-(1-\delta|S|)\frac{\beta_-}{\alpha(u)}du\right)$$

So Ψ goes to 0 when t goes to $+\infty$ because of Hypothesis (4).
Now we lower bound $\frac{du_s}{dt}$ to show convergence of u_s.

$$\begin{aligned}\frac{du_s}{dt} =& \beta_s(t)\frac{\Gamma_s(t)-u_s(t)}{\int \beta_s(u)du}\\ \geq& \beta_s(t)\frac{-\Psi(t)}{\alpha\left(\int \beta_s(u)du\right)}\\ \geq& -\psi(1)\frac{\exp\left(\int_1^t -(1-\delta|S|)\frac{\beta_-}{\alpha(u)}du\right)}{\alpha\left(\int \beta_s(u)du\right)}\\ \geq& -\psi(1)\frac{\exp\left(\int_1^t -(1-\delta|S|)\frac{\beta_-}{\alpha\left(\int \beta_s(u)du\right)}du\right)}{\alpha\left(\int \beta_s(u)du\right)}\end{aligned}$$

As $\delta|S|<1$, this last term is integrable, therefore, as u_s is bounded, this means that u_s converges. The same argument applies to Γ_s using Lemma 2, and the limits of u_s and Γ_s are necessarily the same (otherwise the derivative of u_s converge towards $\frac{\lim u_s - \lim \Gamma_s}{\alpha(t)}$ which would result in a diverging u_s). □

Lemma 9. *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. The proof proceeds similarly to Lemma 5. First, we show that $\forall i,s, \int_1^\infty \sum_{i\in I}\Delta_s^i(t)dt < +\infty$.

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

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

With the previous lemma:

$$\begin{aligned}P_{ss'}(x_s)\dot{u}_s &= P_{ss'}(x_s)\beta_{s'}(t)\frac{\Gamma_s(t)-u_s(t)}{\alpha\left(\int_0^t \beta_s(v)dv\right)}\\ &\geq -P_{ss'}(x_s)\frac{\Psi(t)}{\int_0^t \beta_s(v)dv}\end{aligned}$$

(because $\beta_s(t)\leq 1$ and $\Gamma_s(t)-u_s(t)\geq -(u_s(t)-\Gamma_s(t))_+$).

Then, for all T:

$$\begin{aligned}
\beta_- \int_1^T \sum_i \Delta_s^i(t)dt &\leq \int_1^T \sum_i \beta_s(t)\Delta_s^i(t)dt \\
&\leq \Gamma_s(T) - \Gamma_s(1) \\
&+ \delta \sum_{s'} P_{ss'}(x_s) \int_1^T \Psi(1) \frac{\exp\left(\int_1^t \frac{\delta|S|-1}{\alpha(v)} dv\right)}{\alpha\left(\int_0^t \beta_s(v)dv\right)} \\
&= \Gamma_s(T) - \Gamma_s(1) \\
&+ \delta \sum_{s'} \int_1^T \Psi(1) \frac{\exp\left(\int_1^t \frac{\delta|S|-1}{\alpha\left(\int_0^t \beta_s(w)dw\right)} dv\right)}{\alpha\left(\int_0^t \beta_s(v)dv\right)} \\
&= \Gamma_s(T) - \Gamma_s(1) \\
&+ \delta \frac{A}{\delta - 1} \left(\exp\left(\int_1^T \frac{\delta - 1}{\alpha\left(\int_0^t \beta_s(v)dv\right)} \right) - 1 \right) \\
&< \Gamma_s(T) - \Gamma_s(1) + \delta \frac{A}{1-\delta}
\end{aligned}$$

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

Proof of the part about ABRD of Theorem 2.
It is Lemma 8 and the same proof as Lemma 7.

□

B Stochastic Approximations

B.1 Correlated Asynchronous Stochastic Approximation

An asynchronous system as defined in [33] 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 $s_n^\sharp$ as the number of times util n that s occured:

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

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

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

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:

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

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

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_k^ϵ is the set of $k \times k$ diagonal matrices with coefficients between ϵ and 1:

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

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 6 below) for system (12) [33] under assumptions stated in Subsection B.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 for x_s^i and u_s which are updated at the same times for a specific state s in AFP or for u_s which is always updated at every step in SAFP. 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 SABRD and ABRD can be written as $\overline{F}$ with a suitable $\mathcal{S}$ and F, making it possible to prove the rest of Theorem 3 using the convergence results of the continuous time systems of the previous section, see section B.4.

B.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 5 (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 [5]. They will later be used to characterize the limit sets of the discrete time systems.

Definition 6 (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 7 (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 [5, Theorem 4.3] 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 10. *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 [5] 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. □

B.3 Correlated Asynchronous Stochastic Approximations

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

Theorem 4 (Analog of Theorem 3.1 of [33])**.** *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,

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

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

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 Subsection C.

B.4 Convergence of a Fictitious Play procedure

Lemma 11 (Internally Chain Transitive Sets)**.** *If for all t, $\alpha(t) = 1$ and if L is internally chain transitive either for SBRD or SABRD, or for ABRD and $\delta < 1/|S|$, then*

$$L \subseteq \left\{ (x, u) \middle| \begin{matrix} \forall s \in S \ \forall i \in I, \ f_{s,u}(x_s) = u_s \\ \wedge \ x^i_s \in \underset{y^i \in A^i}{\arg\max} f_{s,u}(y^i, x^{-i}_s) \end{matrix} \right\}$$

Proof.

We define:

$$A := \left\{ (x, u) \middle| \begin{matrix} \forall s \in S \ \forall i \in I, \ f_{s,u}(x_s) = u_s \\ \wedge \ x^i_s \in \underset{y^i \in A^i}{\arg\max} f_{s,u}(y^i, x^{-i}_s) \end{matrix} \right\}$$

$$B := \left\{ (x, u) \middle| \forall s \in S \ f_{s,u}(x_s) \geq u_s \right\}$$

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 6 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 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 SABRD, we have the exact same proof, and for ABRD, the coefficient in the exponential is $\delta|S| - 1$ but it is supposed to be negative, so the proof is the same as well.

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 $V(x,u) := \sum_{s\in S} f_{s,u}(x_s)$. 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 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 [5]: it applies in case the Lyapunov function is defined on invariant set. So L is contained in A.

□

Proof of Theorem 3. For systems (SAFP) and (AFP) we now need to apply Theorem 4. 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 [5].

For system SFP Lemma 10 states that the limit set is internally chain transitive.

Then Lemma 11 concludes the proof: the limit set is internally chain transitive and consequently included in the set of equilibria.

□

C Proof of Theorem 4

In this subsection, we show a proof of Theorem 4. It is a modification of Theorem 3.1 of [33]. In order to carry the proof, we first need a general theorem found in [5]:

Theorem 5 (Linear interpolation are asyptotic pseudo-trajectories)**.** *Consider the stochastic approximation process*

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

under the assumptions:

(i) For all $T > 0$

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

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 4. We are going to use Theorem 5 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 5 will be applied to. We define $\tilde{M}_n := \text{diag}(\max\{1_{s\in I_n}\frac{\gamma_{s_n^\sharp}}{\overline{\gamma}_n}, \epsilon\})$. Note that, consequently, $\tilde{M}_n \in \text{diag}(\text{conv}(\mathcal{S})\bigcap[\epsilon,1]^K) = \Omega^\epsilon_{k,\mathcal{S}}$. 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 5:

(i) For $T > 0$:

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

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 [33]. Regarding the second part, it is exactly Lemma 3.6 of [33] and this applies because of assumptions (iii), (iv) and (v).

(ii) This is assumption (iii).

(iii) This is assumption (i) of Theorem 5.

(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 5 holds).

(v) This is assumption (vii).

So Theorem 5 applies and gives the desired result.

□

D Extended Comparison with Existing Work

Fictitious play in non-stochastic normal form games Fictitious play was originally introduced by Brown and Robinson for zero-sum games. Its convergence was proven for several class of games (see 3 for details). Its continuous counterpart is the best-response dynamics and it was studied by numerous authors, including [14].

The studied games are not stochastic, so there is no state in the stochastic sense. Therefore there is only one strategy profile $x(t)$ that evolves with time t according to the following differential inclusion:

$$\frac{dx^i}{dt} \in \text{BR}^i(x^{-i}) \tag{16}$$

where $BR^i(x^{-i})$ is the best response of player i, i.e. the set of actions a that maximizes $r^i(a, x^{-i})$).

Benaïm, Hofbauer, and Sorin showed in [5] that the limit set of discrete time fictitious play can be studied using (16): the continuous time interpolation is a so called asymptotic pseudo-trajectory.

Benaïm and Faure exhibit a simple Lyapunov function for 16 in identical interest games (and more generally in potential games): $t \mapsto r^i(x(t))$ is increasing if r^i is concave (in the games defined in our paper, it is linear) in every variable.

This approach can not be used as-is for stochastic games: even identical interest stochastic games do not necessarily have concave value function in every player strategy across states. Therefore, it is necessary to rely on other potential functions or other proof techniques. This approach was carried out in [32] where the total discounted payoff of a strategy is computed at every step of the algorithm. Then, an immediate best-response is computed and played. Computing such functions across all states is potentially costly (it requires to solve the Markovian decision process). We described in our paper a procedure that does not need the whole Markovian decision process to be solved at every step.

Best-response dynamics in zero-sum stochastic games

Leslie, Perkins, and Xu introduced the best-response dynamic in a zero-sum stochastic game as follows for every s, i and $t \geq 1$:

$$\begin{cases} \dot{u}_s(t) = \dfrac{f_{s,u(t)}(x_s(t)) - u_s(t)}{t} \\ \dot{x}^i_s(t) \in \arg\max\limits_{a \in A^i} f_{s,u}(a, x^{-i}_s(t)) - x^i_s(t) \end{cases} \tag{17}$$

Compared to our paper, their work is dedicated to zero-sum stochastic games in continuous time whereas our paper deals with identical interest stochastic games and study discrete-time fictitious play using the continuous time best-response. Therefore, it is the first time, to the best of our knowlegde, that ideas from [26] are used in an algorithm to perform online learning.

Our paper is inspired from [26] and as such, share many similarities, including the different learning rates for u_s and x^i_s. In our continous time systems, the different learning rates are generalized and u_s are updated at rate $\alpha(t)$, so $\alpha(t) = t$ gives (17).

Furthermore, a state-dependent system is also defined in [26] as follows:

$$\begin{cases} \dot{u}_s(t) = \dfrac{f_{s,u(t)}(x_s(t)) - u_s(t)}{t} \\ \dot{x}^i_s(t) \in 1_{s=s(t)} \left(\arg\max\limits_{a \in A^i} f_{s,u}(a, x^{-i}_s(t)) - x^i_s(t) \right) \end{cases} \tag{18}$$

where $s(t)$ is the state at time t.

This corresponds to semi-asynchronous systems SABRD. However, in contrast to (18), in SABRD, the update rate is at least $\beta_- > 0$ because it stands for an average over a continuous period of time. This is a different way to use the ergodicity hypothesis and it is especially well suited for our article because our goal is to prove the convergence of the discrete time systems using continuous time systems. So, the average over time is consubstantial with stochastic approximations.

We also provide fully asynchronous systems ABRD where u_s is also updated in a stade dependent manner, which is not the case in [26].

Fictitious play in zero-sum games Sayin, Parise, and Ozdaglar introduced an algorithm that combines fictitious play and Q-learning in zero-sum stochastic games. It is defined with estimates of the state-action value function $\hat{Q}_{i,s,n}(a)$ for player i, state s, action a and step n of the algorithm and estimates of the other player startegies $\hat{\pi}_{1,s,n}$ (which is $x^i_{s,n}$) with our notations). This differ from our work in three fundamental directions. First, it is based on state-action value functions $\hat{Q}_{i,s,n}(a)$, whereas our work is built upon state value functions $u^i_{s,n}$. Second, the games considered are zero-sum (fully competitive) stochastic

games whereas we focus on identical interest (fully cooperative) stochastic games. Third, it is technically built upon different proof techniques as Sayin, Parise, and Ozdaglar also use stochastic approximations but the different timescale are not present in the continuous-time systems.

The update on the Q-function when the profile a is played and the current state is s at step n is (for other actions, the Q-function is unchanged):

$$\hat{Q}_{i,s,n}(a) = \hat{Q}_{i,s,n}(a) + \hat{\beta}_{s,n}\left(r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} \hat{v}_{i,n}(s') P_{ss'}(a) - \hat{Q}_{i,s,n}(a)\right)$$

where $\hat{v}_{i,n}(s') = \max_{a^i \in A^i} Q_{i,s,n}(a^i, x_{s,n}^{-i})$, $\hat{beta}_{s,n}$ is the update rate and other notations are those of our paper.

Since the estimate are per state-action and not only per state (as in our work) it is possible to use a model-free update rule where the transition is not needed:

$$\hat{Q}_{i,s,n}(a) = \hat{Q}_{i,s,n}(a) + \hat{\beta}_{s,n}\left(r^i(s,a) + \gamma v_{i,n}(s') - \hat{Q}_{i,s,n}(a)\right)$$

where s' is the next state.

Therefore, the state transition is implicitly estimated in the model-free version of the algorithm in the Q-function, as long as there is an infinite number of times when the system is in state s and every action a is played. A similar mechanism would be interesting for our procedure: we cannot directly use the same idea as our state-value estimate $u_{s,n}$ (which is the analog of $Q_{i,s,n}$) is not per action.

Fictitious play in multi-stage games Perolat, Piot, and Pietquin proposed a fictitious play process for multistage games [34]. In multi-stage games, states can be naturally ordered as a tree with an initial state and a final state, an assumption that we do not make. The fact that states are ordered as a tree is helpful to do proofs by inductions but covers a smaller class of games. There is nevertheless common features between our procedure and the procedure outlined in [34].

Technically, it involves the estimation of a state-action value function $Q_n^i(s,a)$ where i is a player, s the state, a the action and n the update step of the algorithm and the computation of a strategy where the probablity for player i to play a in state s at step n is $\pi_n^i(s,a)$.

Values $\pi_n^i(s,a)$ is updated towards a logit-choice best-response with regards to the current state-action value function $Q_n^i(s,a)$. This is a difference with our procedure which uses a plain best-response that are potentially non-unique, hence the need for differential inclusions (which are not needed for Perolat, Piot, and Pietquin). Our procedure is also simpler to implement (it is not necessary to use a logit function) and interesting from an epistemic point of view in game theory (players take a best response without randomness).

Values $Q_n^i(s,a)$ are updated towards an iterate of the Bellman operator (either using the current strategy or using the logit-choice best-response strategy). In contrast, our procedure updates state-values which is presumably a smaller set of estimates but require either the transition matrix or an estimation of these transitions.

In [34], other players strategy are not observed. In our paper, we model every player independently in a similar way as fictitious play of Brown and Robinson. This can be seen as a limitation of our work since it requires more information during the play. However, we believe that a more precise model of players can be interesting for two reasons. First, from a general perspective, it is interesting to study models where more information about the environment is known because it makes it possible to converge to sets that may not be attainable in more relaxed settings [15]. Second, we believe it can converge faster in the appropriate environment (especially with a large number of players), even if we lack theoretical or experimental evidence. It would be interesting to compare our algorithm to the one outlined in Perolat et al..

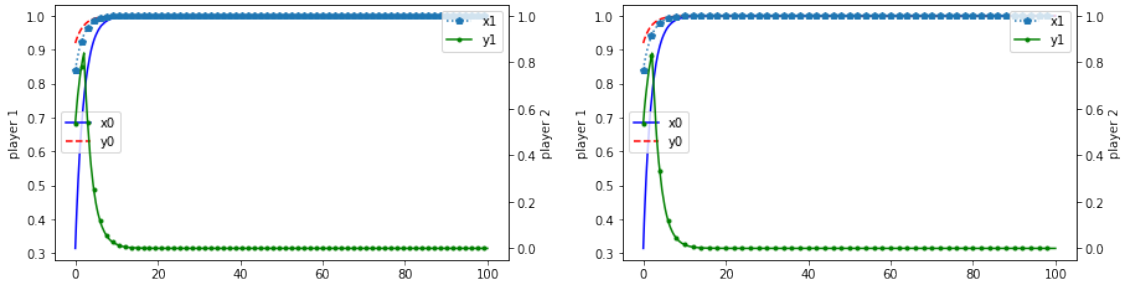

Figure 3: Evolution of actions in simulations of best-response dynamics in the case $\alpha(t) = 1$ (top) and $\alpha(t) = t$ (bottom)

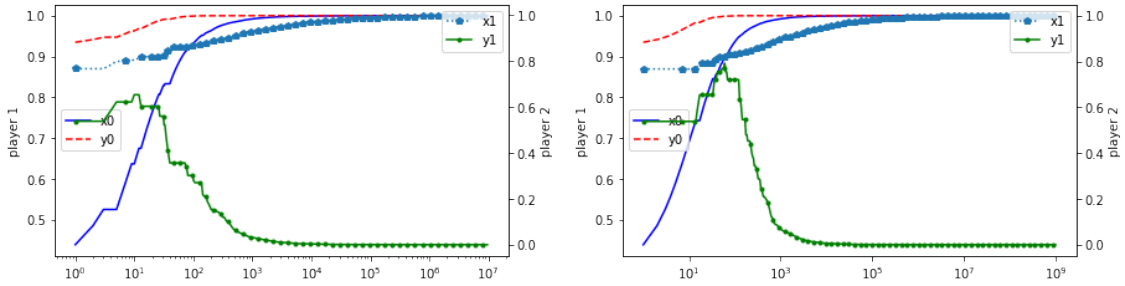

Figure 4: Evolution of actions in simulations of fictitious play in the case $\alpha(t) = 1$ (top) and $\alpha(t) = t$ (bottom)

E Supplementary Plots of Simulations

E.1 Continuous-time best response dynamic

Figure 3 shows the action variables x_s^i of one of the simulation we ran. A pure action is quickly approached. With more actions and states, it could be the case that actions approach several pure actions before converging to a final one. In our case, as is shown in the notebook of the supplementary material, this is a Nash equilbrium. Similar plots for fictitious play can be found in Figure 4

Note that the evolution of actions is similar in both case $\alpha(t) = 1$ and $\alpha(t) = t$, this is because the function α is involved in the u_s derivative but not directly in the expression of derivatives of x_s^i.