Smooth Fictitious Play in Stochastic Games with Perturbed Payoffs and Unknown Transitions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent extensions to dynamic games (Leslie et al. [2020], Sayin et al. [2021], Baudin and Laraki [2022]) of the well-known fictitious play learning procedure in static games were proved to globally converge to stationary Nash equilibria in two important classes of dynamic games (zero-sum and identical-interest discounted stochastic games). However, those decentralized algorithms need the players to know exactly the model (the transition probabilities and their payoffs at every stage). To overcome these strong assumptions, our paper introduces regularizations of the systems in Leslie et al. [2020], Baudin and Laraki [2022] to construct a family of new decentralized learning algorithms which are model-free (players don't know the transitions and their payoffs are perturbed at every stage). Our procedures can be seen as extensions to stochastic games of the classical smooth fictitious play learning procedures in static games (where the players best responses are regularized, thanks to a smooth strictly concave perturbation of their payoff functions). We prove the convergence of our family of procedures to stationary regularized Nash equilibria in zero-sum and identical-interest discounted stochastic games. The proof uses the continuous smooth best-response dynamics counterparts, and stochastic approximation methods. When there is only one player, our problem is an instance of Reinforcement Learning and our procedures are proved to globally converge to the optimal stationary policy of the regularized MDP. In that sense, they can be seen as an alternative to the well known Q-learning procedure.

1 Introduction

Fictitious play (FP) (Brown [1951], Robinson [1951]) is the oldest and most famous independent (i.e., decentralized) learning algorithm in game theory. It is a simple behavioural strategy that agents may use to repeatedly play a fixed normal form game G. Namely, at every repetition of G, each player best replies to the empirical distribution of the uncorrelated past actions of the opponents.

When G is a finite game, it has been shown that the empirical distribution of the players' actions under FP globally converges to the set of mixed Nash equilibria, when G is zero-sum (Brown [1951], Robinson [1951]) or identical-interest (Monderer and Shapley [1996]). The FP procedure was extensively studied since then with numerous additional convergence results (Miyasawa [1961], Berger [2005], van Strien and Sparrow [2011]) and a number of generalizations (Leslie and Collins [2006], Benaïm and Faure [2013] to cite a few).

Unfortunately, FP has no guarantee in terms of regret. Indeed, it may well lead to the worst possible payoff in a zero-sum game if the opponent best-replies to it at every stage. Fortunately, the smooth version of FP (call it SFP), also named stochastic or regularized fictitious play, is proved to have the no-regret property (Fudenberg and Levine [1995], Benaïm and Faure [2013]) and to converge to regularized Nash Equilibria in the classes mentioned so far (Hofbauer and Sandholm [2002],

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Berger [2005], Heliou et al. [2017], even when G is non atomic or with continuous action sets, see Hadikhanloo et al. [2021], Perrin et al. [2020]). In the online optimization and learning community, fictitious play for two player games (or correlated FP for more than two players, where a player best responds to the correlated empirical past actions of the opponents) is precisely Follow The Leader (FTL), and Smooth Fictitious Play is Follow The Regularized Leader (FTRL) (see Shalev-Shwartz [2011], Cesa-Bianchi and Lugosi [2001], Belmega et al. [2018], Kwon and Mertikopoulos [2017], Giannou et al. [2021a,b]).

How can FP and SFP be extended to learn Nash equilibria in zero-sum and identical-interest games when the stage game is not fixed but changes with time (such as the stochastic game of Shapley [1953])?

The difficulty is that, in stochastic games (SG), a state variable evolves from a period to the next one with a probability depending on the current state and actions players take. Therefore, every player, when choosing its action at any given stage, must strike a balance between its instantaneous reward (determined by the current state and the current profile of actions) and its continuation future rewards, which depend in particular on the next state. Extending the result of Shapley [1953] in the zero-sum case, Fink [1964] proved that when there are finitely many players and actions (the framework of our paper), any discounted stochastic game (DSG) admits a stationary Nash equilibrium in mixed strategies (e.g. a decentralized randomized policy that depends only on the state variable).

Stationary Nash are the simplest possible equilibria that a dynamic game can have and it is our objective to construct independent learning algorithms that are provable to globally converge to them in finite zero-sum and identical-interest DSG. This is a hard problem because the payoff function of a player in a DSG is not own-payoff concave or quasi-concave when the players are restricted to their stationary strategies and thus, no gradient based method is guaranteed to converge, even to a local Nash equilibrium (Daskalakis et al. [2021]). Only recently some positive results have been obtained. That is, Leslie et al. [2020], Sayin et al. [2021], Baudin and Laraki [2022] combined the Fictitious Play behavioral strategy with a Q-learning like updating rule to design a family of decentralized rules such that the sequence of independent empirical distributions of the players converge to the set of stationary Nash equilibria in the two major classes of zero-sum and identical-interest ergodic DSG.

These latest results are based on ideas of the well-known and comparatively much better understood framework of Reinforcement Learning (RL). In this setting, there is only one player (a Markov Decision Process). The advent of efficient and model-free algorithms in this context such as Q-learning (Watkins [1989]) has had a lot of impacts on RL with numerous extensions, including offline Q-learning (Kumar et al. [2020]), double Q-learning (Hasselt [2010]) and a wide range of applications (Tai and Liu [2016], Kurin et al. [2020]). However, the convergence of Q-learning does not extend to the multiagent setting (see Wunder et al. [2010], Kianercy and Galstyan [2012]).

Unfortunately, the algorithms in Leslie et al. [2020] and Baudin and Laraki [2022] need the players to know precisely the model from the beginning (i.e., the transition and payoff functions). To avoid this drawback, our paper introduces a robust regularized version of their rules which combines Smooth FP, Q-learning-like rule (i.e., updates at every step an estimate of the continuation payoff) and empirical estimates of the unknown parameters. This leads us to a family of model-free independent learning algorithms, where the payoffs and the transitions are unknown to the players, and where the stage payoffs are imperfectly observed (e.g. randomly perturbed with a zero-mean noise). Our algorithms are proved to converge to the set of regularized stationary Nash equilibria in zero-sum and identical-interest ergodic DSG (those are a subset of stationary ε-Nash equilibria, and as ε goes to zero (corresponding to a vanishing SFP), they refine the set of stationary Nash equilibria).

Another efficient class of algorithms in RL are the Projected Gradient Methods (PGM) and their stochastic version (Williams [1992], Sutton and Barto [2018]). Very recently, PGM and Stochastic PGM (SPGM) have been studied by Daskalakis et al. [2020] in zero-sum stochastic games to prove that when the players play independently a PGM (or a SPGM) with different time scales,[1] one can approach a best stationary equilibrium iterate[2] in a finite time $T(\varepsilon)$, whenever the stochastic game is episodic.[3] Leonardos et al. [2022] proved a similar convergence result in episodic identical-interest

[1] Daskalakis et al. [2020] proved that their convergence result fails if the players have the same time scale.

[2] I.e., there is $t \in \{1, ..., T(\varepsilon)\}$ such that (in expectation in case of Stochastic PGM), the players play an ϵ stationary Nash equilibrium at time t.

[3] A SG is episodic if, at every stage, there is a positive probability that the game stops.

DSG, as soon as all the players use a PGM or SPGM.[4] Those are extremely interesting and promising convergence results. Note however that they don't cover our framework because they assume the game to have a positive probability to stop (implying that their trajectories terminate almost surely in finite time) while in our ergodic setting, all the trajectories are infinite and the game never terminates. So, our results and those of Daskalakis et al. [2020], Leonardos et al. [2022] are non-comparable and complementary: they prove the existence of a best iterate approximation, we prove a time-average convergence, we use different algorithms and tools, and orthogonal assumptions (episodic vs ergodic).

The counterpart of FP in continuous time is best-response dynamics (Matsui [1992], Harris [1998]) and modern proofs for the convergence of SFP rely on the convergence of the continuous model combined with stochastic approximation techniques (see for instance Benaïm et al. [2005], Benaïm and Faure [2013], Hadikhanloo et al. [2021]). We follow a similar approach: we prove the convergence to regularized stationary Nash equilibria of an associated smooth continuous-time dynamics and, using some advanced stochastic approximation tools, deduce the convergence of our discrete-time rules. This is an important technical difference with Daskalakis et al. [2020], Leonardos et al. [2022] who can prove their results directly in discrete time, with an explicit finite convergence bound $T(\varepsilon)$. Our stochastic approximation methods do not provide us with a convergence rate, which is an open problem for SFP and FP even in the classical setting.

Contributions

- We introduce a family of independent learning algorithms in stochastic games that are model-free (unknown transitions and imperfect observation of the stage payoffs);
- We identify the corresponding smooth continuous-time dynamics, and show it globally converges to regularized stationary Nash equilibria in identical-interest and zero-sum DSG;
- From the continuous-time convergence, we deduce that in our two classes of DSG, the uncorrelated empirical frequencies of actions generated by our discrete-time algorithms almost surely globally converge to the set of regularized stationary Nash equilibria.

Outline Section 2 gives the main definitions and notations of the paper. Section 3 introduces the smooth fictitious play procedure together with the continuation payoff updating rule. In order to prove the convergence of the discrete time algorithm, and also for its own sake, a smooth best-response dynamics with a simple continuation payoff up-dating rule dynamics is described in Section 4 followed with sketches of proofs of both our discrete-time and continuous-time systems. Section 5 describes in more details the related work. The appendix contains the detailed proofs.

2 Preliminaries

Strategic situations where several agents interact, get rewards and modify an environment can be modeled as Stochastic Games (SG). In our settings, the number of players, actions and possible states are finite. We consider players that are interested in the so-called discounted reward on an infinite horizon, that is players strike a balance between instantaneous rewards and future ones.

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

Main Restrictions We are interested in two classes of games: *zero-sum* stochastic games are two-players SG where $r_s^1 = -r_s^2$ for every s and *team* stochastic games are such that the payoff functions of the players differ only by a constant (there is $r_s(\cdot) : A \to \mathbb{R}$ such that for every i and s,

[4] Leonardos et al. [2022] proved their results in the larger class of potential stochastic games (they called Markov Potential Games). In such a class of games, the players are divided into two categories: either they do not influence the transition or they have the same payoff function up to a constant (see Holler [2020], Leonardos et al. [2022]). This last class is slightly larger than identical-interest SG is called *Team Stochastic Games* in Holler [2020]. All our convergence results extend to team stochastic games. The class of potential SG where the transitions are independent of the player's actions is easy to solve, just let each player uses a smooth FP myopically state per state independently.

$r_s^i(\cdot) = r_s(\cdot) + c^i$ for a constant c^i). A special case is *identical-interest* SG where all payoff functions are equal ($r_s^i(\cdot) = r_s(\cdot)$ for all i). A SG is *ergodic* if there exists $T \in \mathbb{N}$ such that for any sequence of actions of length T, the probability to reach any state s' starting from any state s is positive.

Game Form A stochastic game is played as follows: starting from an initial state s_0, at every step $n \in \mathbb{N}$, every player i choose an action a_n^i given the history of play and the current state s_n. The next state s_{n+1} is drawn from distribution $P_{s_n}(a_n)$.

Discounted Payoffs We suppose that every player i is interested in maximizing its discounted payoff, that is the expectancy of $\sum_{k\in\mathbb{N}} \delta^k r_{s_k}^i(a_k)$ where $\delta \in (0,1)$ is the discount factor.

Strategies A behavioral strategy σ^i for player i is a mapping associating with each stage $n \in \mathbb{N}$, history $h_n \in (S \times A)^n$ and current state s, a mixed action $x_n^i = \sigma^i(n, h_n, s)$ in $\Delta(A^i)$. The behavioral strategy is pure if its image is always in A^i. A stationary strategy of player i is the simplest of behavioral strategies. It depends only on the current state s but not on the period n nor on the past history h_n. As such, a stationary strategy can be identified with an element of $\Delta(A^i)^S$ (a mixed action per state interpreted as: whenever the state is s, i plays randomly according to x_s^i). The set of stationary strategy profiles is $\Pi_{i\in I}\Delta(A^i)^S$. Set $X = \Pi_{i\in I}\Delta\left(A^i\right)$ so a stationary profile is an element of X^S. For $y^i \in (\Delta(A^i))^S$ and $x \in \Pi_{i\in I}\Delta(A^i)^S$, we denote by (y^i, x^{-i}) the stationary profile where i changes its strategy from x^i to y^i. A stationary profile $x \in X^S$ is a Nash equilibrium if and only if no player has a profitable behavioral deviation. Fink [1964] proved the existence of stationary Nash equilibria in every finite DSG and that it is sufficient to check pure stationary deviations.

Regularizer In this paper, we are interested in exploratory algorithms, which may classically be generated by the use of some steep concave regularizer. This regularizer is added to the payoff functions r_s^i and can be given several interpretations (see Fudenberg and Levine [1998], Hofbauer and Sandholm [2002] for details): it models the uncertainty of the payoff caused by the "trembling hand" of players or is a way to generate strict incentives to explore all the actions. Formally player i maximizes a perturbation of its payoff function $r_s^i + \epsilon h^i$ under the following hypotheses:

$$\begin{gathered} h^i : X \to \mathbb{R}^+, \text{ strictly concave in } x^i,\ C^1 \text{ on the interior,} \\ \lim_{x^i \to \partial\Delta(A^i)} \|\nabla_{x^i} h^i(x)\| = +\infty \text{ and } \epsilon > 0 \end{gathered} \tag{H1}$$

In this paper we study the convergence of some discrete and continuous time systems to regularized Stationary Nash equilibria, parameterized by the regularizers $(h^i)_{i\in I}$ and parameter ϵ.

Definition 2.1. Regularized Stationary Nash Equilibria of a DSG are the stationary Nash equilibria of the DSG with the perturbed payoff functions $r_s^i + \epsilon h^i$, $i \in I$.

Remark. In identical-interest (resp. zero-sum) SG, we suppose all players take the same (resp. the opposite) regularizer h so as the regularized payoff functions remain identical. In this context, one can take for example a separable regularizer function which is equal to the sum (resp. the difference) of concave functions depending only on x^i.

Remark. If in a profile, all actions are β-optimal with respect to discounted payoff based on functions $r_s^i + \epsilon h^i$ (*i.e.,* deviations are at most β profitable), then this profile is called a β-regularized equilibria.

3 Smooth Fictitious Play in Stochastic Games (Discrete-Time Algorithms)

Fictitious Play in Repeated Games Introduced by Robinson [1951] and Brown [1951], fictitious play is a decentralized behavioral strategy to repeatedly play a fixed normal form game. At every step n, every player i chooses an action a_{n+1}^i that is a best response to the past empirical average action of other players: a_{n+1}^i must maximize $r^i\left(\cdot, x_n^{-i}\right)$ where $x_n^{-i} = \frac{1}{n+1}\sum_{k=0}^n a_k^{-i}$.

A famous variation of this procedure is smooth fictitious play (Fudenberg and Levine [1995]) where players choose their action according to a regularized payoff function. Formally, a player i draws an action according to a distribution that maximizes $r^i\left(\cdot, x_n^{-i}\right) + \epsilon h^i\left(\cdot, x_n^{-i}\right)$ (with h^i and ϵ defined in the previous section). An interesting property of such a procedure is that with suitable property on ϵ,

it has no *regret* (up to ϵ), meaning that if played unilaterally by a player i, other players can not trick i into using a suboptimal action. This is due to the randomness of the action choice: the distribution assigns positive probability to every action, so a player's behavior remains unpredictable.

In this section, we extend the smooth FP procedure to stochastic games. This builds upon the recent extension of FP in Leslie et al. [2020], Sayin et al. [2021], Baudin and Laraki [2022] (see Section 5 for more details). Indeed, it is not easy to derive the convergence to a regularized equilibrium by applying directly the definition of smooth fictitious play to the discounted stochastic game in which the players are restricted to play in stationary strategies, because the payoff function in this game is non-linear (nor is it concave, or quasi-concave) with respect to a player stationary strategy. To overcome this difficulty, and following the idea in Leslie et al. [2020], we update two sets of variables for every state: one concerns the uncorrelated empirical actions and the other is an estimate of the continuation payoffs (*i.e.,* payoffs that players can anticipate to achieve if that state is reached). These continuation payoffs are used as a parameter in an auxiliary game, often called the Shapley operator.

Auxiliary game Following numerous authors including Shapley [1953], we define a so-called auxiliary game parameterized by a family of vector $u^i \in \mathbb{R}^S$ (one for every player i) and a state s. It is a one-shot game every player i has the same action set as in G but gets a one-time payoff of:

$$f^i_{s,u^i}\left(a_s\right) := (1-\delta) r^i_s\left(a_s\right) + \delta \sum_{s' \in S} P_{ss'}(a_s) u^i_{s'} \tag{1}$$

Smooth Best-Response While playing the stochastic games, players maintain a set of continuation payoffs u^i that are used to choose their actions. Given that the current state is s, for a player i, its action is drawn from the distribution that is the smooth best-response in the auxiliary game with respect to empirical action profile x_s, that is:

$$\mathrm{sbr}^i_{s,u^i}\left(x_s\right) := \arg \max_{y^i \in \Delta(A^i)} f^i_{s,u^i}\left(y^i, x_s^{-i}\right) + \epsilon h^i\left(y^i, x_s^{-i}\right) \tag{2}$$

This is well and uniquely-defined because of the strict concavity of h^i.

Example. If the regularizer h^i is taken to be the Shannon entropy, that is:

$$h^i\left(x_s\right) = -\sum_{i \in I} \sum_{a^i \in A^i} x^i_s(a^i) \log\left(x^i_s(a^i)\right)$$

Then the smooth best-response function is the logit function:

$$\mathrm{sbr}^i_{s,u^i}\left(x_s\right)\left(a^i\right) := \frac{\exp(\epsilon^{-1} f^i_{s,u^i}\left(a^i, x_s^{-i}\right))}{\sum_{b^i \in A^i} \exp(\epsilon^{-1} f^i_{s,u^i}\left(b^i, x_s^{-i}\right))}$$

Smooth fictitious play with known transitions and deterministic payoff To extend smooth fictitious play to stochastic games, we use two sets of variables. The $x_{s,n}$ variable is the distribution of empirical actions for every state s prior to step n. The other variable $u^i_{s,n+1}$ can be interpreted as the continuation payoffs at s and is defined as the time average of the regularized payoffs up to step n. This leads to the following system:

$$\begin{cases} u^i_{s,n+1} = \dfrac{1}{n+1} \displaystyle\sum_{k=0}^{n} \left(f^i_{s,u^i_k}\left(x_{s,k}\right) + \epsilon h^i\left(x_{s,k}\right)\right) \\ x_{s,n} = \dfrac{1}{s^\sharp_n} \displaystyle\sum_{k=0}^{n} 1_{s=s_k} a_k \\ a^i_{n+1} \sim \mathrm{sbr}^i_{s,u^i_n}\left(x_{s,n}\right) \end{cases}$$

where $s^\sharp_n$ is the number of times s was reached, that is $\sum_{k=0}^{n} 1_{s=s_k}$ and every action a_k is embedded into the Euclidean space containing $\Delta\left(A^i\right)$. The interpretation is simple: at each stage t, each player best replies to the belief that the other players will play according to the past uncorrelated empirical distribution of actions, and that the future continuation payoffs are equal to the time average of the past estimated perturbed payoffs, calculated using the past empirical frequencies of actions.

This system can be generalized and rewritten in an incremental fashion. It is our first main system:

$$\begin{cases} u^i_{s,n+1} - u^i_{s,n} = \dfrac{\beta}{n+1}\left(f^i_{s,u^i_n}\left(x_{s,n}\right) + \epsilon h^i\left(x_{s,n}\right) - u^i_{s,n}\right) \\ x_{s,n+1} - x_{s,n} = \dfrac{1_{s=s_n}}{s^\sharp_n}\left(\mathrm{sbr}^i_{s,u^i_n}\left(x_{s,n}\right) - x_{s,n}\right) \end{cases} \tag{SFP}$$

where $\beta = 1$ corresponds to the equation above.

Remark. If there is only one state (e.g the classical repeated game sitting), this procedure is exactly standard smooth fictitious play since the smooth best-response does not depend on the value of $u^i_{s,n}$.

Smooth fictitious play with unknown transitions and perturbed payoff We suppose now that payoff functions are not known, and that each stage payoff is observed with some zero-mean noise which follows a distribution that may depend on the history, the current state and actions taken by the other players. Therefore, at step n, player i gets a random reward R^n_i that is drawn according to a distribution determined by actions a_n and current state s_n whose expectancy is $r^i_{s_n}\left(a_n\right)$ and bounded variance conditionally on the history (see Appendix B for details). We also suppose that transitions are not known. Therefore, both transitions and expected payoff may be empirically estimated as follows:

$$\begin{cases} \hat{P}_{ss',n}\left(a\right) = \dfrac{\sum_{k=0}^n 1_{s_k=s\wedge a_k=a} 1_{s_{k+1}=s'}}{\sum_{k=0}^n 1_{s_k=s\wedge a_k=a}} \\ \hat{r}^i_{s,n}\left(a\right) = \dfrac{\sum_{k=0}^n 1_{s_k=s\wedge a_k=a} R^k_i}{\sum_{k=0}^n 1_{s_k=s\wedge a_k=a}} \end{cases} \tag{MFP.1}$$

Consequently, we define the estimated auxiliary payoff using these two estimators:

$$\hat{f}^i_{s,u^i}\left(x_s\right) := (1-\delta)\hat{r}^i_{s,n}\left(x_s\right) + \delta \sum_{s'\in S} \hat{P}_{ss',n}\left(x_s\right) u^i_{s'} \tag{MFP.2}$$

Now we define a model-free version of smooth fictitious play, similar to SFP but using estimators:

$$\begin{cases} u^i_{s,n+1} - u^i_{s,n} = \dfrac{\beta}{n+1}\left(\hat{f}^i_{s,u^i_n}\left(x_{s,n}\right) + \epsilon h^i\left(x_{s,n}\right) - u^i_{s,n}\right) \\ x_{s,n+1} - x_{s,n} = \dfrac{1_{s=s_n}}{s^\sharp_n}\left(\mathrm{sbr}^i_{s,u^i_n}\left(x_{s,n}\right) - x_{s,n}\right) \end{cases} \tag{MFP}$$

where $\mathrm{sbr}^i_{s,\cdot}$ is defined relatively to $\hat{f}^i_{s,\cdot}$.

Theorem 3.1 (Convergence in identical-interest ergodic DSG)**.** *In a identical-interest ergodic discounted stochastic game, if all players follow SFP (resp. MFP), their empirical actions $x_{s,n}$ converge almost surely to the set of regularized stationary Nash equilibria and their expected vector of continuation payoffs $u^i_{s,n}$ converges to the optimal continuation payoff of limiting equilibrium set.*

This theorem means that even if the trajectories of SFP (resp. MFP) cycle between several stationary equilibria, they all share the same optimal continuation payoff vector.

Theorem 3.2 (Convergence in zero-sum ergodic DSG)**.** *In a zero-sum ergodic discounted stochastic game, if all the players follow SFP (resp. MFP) with the same initial values, their empirical actions $x_{s,n}$ converge almost surely to the set of $D\beta$-regularized Nash equilibria (where $D > 0$ is a constant that only depends on G) and their expected vector of continuation payoffs $u^i_{s,n}$ converges to the corresponding continuation payoff.*

To prove these results, we are going to define in the next section an associated smooth best-response continuous time counterpart to our independent learning algorithms. The proof of Theorems 3.1 and 3.2 are sketched at the end of Section 4 and fully detailed in Appendix B.

Remark 1. Theorem 3.2 suggests that it is possible to use a doubling-trick mechanism to converge to the set of 0-regularized stationary Nash equilibria. Indeed, players can compute the duality gap (see Appendix A for a definition) and decide to reduce the update rate β every time it is below a certain threshold. This is a standard trick to achieve no-regret in reinforcement learning.

Remark 2. Our rules suppose that each player observes the other player's past actions. This is an important difference with the algorithms developed in Daskalakis et al. [2020], Leonardos et al. [2022] where a player observes only its own actions. On the other hand, they do not prove a time-average convergence of their trajectories, while we do, but they prove a best iterate convergence that we do not prove.

4 Smooth Best-Response in Stochastic Games (Continuous-Time Dynamics)

Continuous counterpart of discrete time systems The continuous-time counterpart of fictitious play is best-response dynamics of Matsui [1992], Harris [1998]. Together with the theory of stochastic approximations (Benaïm et al. [2005]), proofs of convergence of best-response are key ingredients to modern proofs of convergence of fictitious play and smooth FP in repeated games (Benaïm et al. [2005], Benaïm and Faure [2013]) and more recently in proving the convergence of FP in stochastic games (Sayin et al. [2021], Baudin and Laraki [2022]).

In this section, we define the continuous-time counterpart to our discrete-time algorithms. This leads us to a regularized version of the continuous-time best-response dynamics in Leslie et al. [2020], Baudin and Laraki [2022]. Similarly to the previous section, our continuous time system has two sets of variables $u_s^i(t)$ which represents the belief aabout the continuation payoffs and $x_s(t)$ which corresponds to belief about the players actions. Our main continuous-time system is:

$$\begin{cases} \dot{u}_s^i = \beta(t)\left(f^i_{s,u^i(t)}\left(x_s(t)\right) + \epsilon h^i\left(x_s(t)\right) - u_s^i(t)\right) \\ \dot{x}_s = \alpha_s(t)\left(\mathrm{sbr}^i_{s,u^i(t)}\left(x_s(t)\right) - x_s(t)\right) \\ \alpha_s(t) \in [\alpha_-, 1] \end{cases} \tag{SBRD}$$

Remark 3. In zero-sum or identical-interest games, if players have the same initial conditions, then we can omit the superscript i in u_s^i as they are equal.

Update rates Profile $x_s(t)$ evolve towards the smooth best-response at a rate of $\alpha_s(t)$. Variable $\alpha_s(t)$ corresponds to the frequency at which a state s is visited. The fact that it is bounded below by $\alpha_- > 0$ is a mathematically convenient way to exploit the ergodicity of the stochastic game.

Remark 4. Note that this is different from the model of continuous-time stochastic game outlined for instance by Neyman [2017]. In this paper, we are interested in using the theory of stochastic approximation, therefore discrete-time and continuous-time system are related through an exponential change of variable in time. As a consequence, we can consider that the state occupation is averaged, which is not the case in some other studies in stochastic games.

Regarding continuation payoffs, they evolve towards perturbed payoffs at a rate of $\beta(t)$, which is typically oblivious, that is only determined by time t and state s independently of the value of other variables. We make the following assumptions that guarantee that the update rates are not too small:

$$\begin{gathered} \beta(t) \geq 0 \text{ and } \beta \text{ is decreasing} \\ \int_0^\infty \beta(u)du = +\infty \end{gathered} \tag{H2}$$

Remark 5. With $\beta(t) = \frac{1}{t+1}$, the system is close to the one outlined in Leslie et al. [2020]. In this case, continuation payoffs are updated with a *slower* timescale. Indeed, having two timescales may be useful to establish convergence or apply stochastic approximation theorems. In our paper, we do not make such an assumption, allowing the possibility that the two variables are on a single timescale.

Model-Free System In order to also study the convergence of the model-free version of our procedure, we also define the smooth best-response dynamics when the model is progressively learned. The estimators are defined as follows:

$$\begin{cases} \dot{\hat{P}}_{ss'}(b) = \alpha_s(t) a(t)(b)\left(P_{ss'}(b) - \hat{P}_{ss'}(b)(t)\right) \\ \dot{\hat{r}}_s^i(b) = \alpha_s(t) a(t)(b)\left(r_s^i(b) - \hat{r}_s^i(b)(t)\right) \\ a^i(t) = \mathrm{sbr}^i_{s,u^i(t)}\left(x_s(t)\right) \end{cases} \tag{3}$$

where $a(t) := \Pi_{i\in I} a^i(t)$ is the profile of selected actions.

Then, $\hat{f}^i_{s,\cdot}$ is defined as in (MFP.2) and we obtain a system MBRD by using estimators in SBRD.

Solutions Systems SBRD and MBRD are differential inclusions. A general theory can be found in Aubin and Cellina [1984]. Here, the right-hand side of each system forms a closed, set-valued map (because there are several possible values for $\alpha_s(t)$) whose images are convex and uniformly bounded. Under these assumptions, it is known that such systems admit (typically non unique) solutions.

Theorem 4.1. *Under hypothesis H2, MBRD converges to the set of regularized Nash equilibria in identical-interest stochastic games. Furthermore, if $\beta^\star \geq \limsup_{t\to\infty} \beta(t)$, then MBRD converges to the set of $D\beta^\star$-regularized stationary Nash equilibria in zero-sum games, where D is a positive constant that depends only on the game G. In particular if $\beta(t)$ goes to 0 then MBRD converges to the set of regularized stationary Nash equilibria.*

The convergence of MBRD is helpful to characterize the limit set of discrete-time smooth fictitious play systems. Indeed, they are contained in the internally chain transitive sets (see Appendix B for a definition) of MBRD using the theory of stochastic approximations. Below, we sketch the proof of continuous-time Theorem 4.1. The discrete-time counterpart Theorem 3.1 and Theorem 3.2 can be deduced in a similar fashion as Baudin and Laraki [2022], although there are several technical subtleties. Complete proofs of these results are in Appendix A.

Sketch of the proof of Theorem 4.1. The proofs for both class of games proceed quite differently: this is not surprising since there are no (to the best of our knowledge) unified convergence proof of even simple FP in potential and zero-sum games.

For identical-interest stochastic games, the key point is showing that the gap between (estimated) auxiliary payoffs $f^i_{s,u^i(t)}(x_s(t))$ and the auxiliary values $u^i_s(t)$ is narrowing. Technically, the difference is bounded below by an integrable function, which implies that $u^i_s(t)$ converges and then that $f^i_{s,u^i(t)}(x_s(t))$ converges and their limits are necessarily equal. Then, a study of the behavior of actions shows that they belong to the set of regularized equilibria, otherwise $f^i_{s,u^i(t)}(x_s(t))$ could not converge (based on Lipschitzian properties of all these quantities).

Regarding zero-sum stochastic games, the first part of the proof studies a rather standard quantity called the duality gap. It goes to 0 (or at least near 0), which implies that the value of the auxiliary game is mostly reached by the auxiliary payoffs $f^i_{s,u^i(t)}(x_s(t))$. Then, comparing relative speed of auxiliary values in all states leads to the convergence of these values and of other variables.

□

5 More on the Related Works

Single-Player Markov Decision Process Watkins [1989] introduced Q-learning to solve Markov Decision Process with guarantees when the environment is stationary. The central idea is to estimate the so-called Q-values for every state-action pairs that represents the continuation payoff using the following scheme:

$$Q(s,a) \leftarrow Q(s,a) + \beta(r_s(a) + \delta \max_{a'} Q(s',a') - Q(s,a)) \tag{4}$$

where s is the current state, a the action played, s' the next state. Q-values for other state-action pairs are left unchanged. This simple rule does not need any information about the transitions.

Since then, numerous extensions have been introduced, for instance double Q-learning (Hasselt [2010]) or Q-learning with deep learning (van Hasselt et al. [2016]). However, in multiagent systems, from the one player's point of view the environment is not stationary because it comprises other players that are also using some non stationary learning rule. Therefore, there is no guarantee and indeed it may not converge to the set of Nash equilibria (Wunder et al. [2010], Calvano et al. [2020]. Our paper is a way to overcome this difficulty with a different updating rule for estimating the continuation payoffs. This updating rule is closer to Expected Sarsa (Sutton and Barto [2018]) where the continuation values are moved towards the *expectation* of future payoffs.

Nash Q-learning Hu and Wellman [2003] is another way to learn Nash equilibria based on Q-learning, supposing that players have access to a Nash equilibrium of the auxiliary game parameterized by the Q-values. In our paper, we do not suppose that the players have access to such an oracle.

Repeated games When there is only one state, this is the classical framework in which the players repeatedly play a fixed normal form game. Leslie and Collins [2005] studies Q-learning in this context. On the FP and SFP side, most of the literature assumes that the players know precisely the parameters of the problems. We only assume that the stage rewards are imperfectly observed. Hence, our results also improve upon the classic framework.

Fictitious Play Algorithms in Stochastic Games Brown [1951] and Robinson [1951] introduced fictitious play to learn the value of repeated fixed zero-sum game in a turn based fashion (see also Berger [2007]. It was later studied in the usual simultaneous updating by Fudenberg and Levine [1998] with numerous extensions, including smooth fictitious play. However, none of these procedures converge to equilibria in the repetition of the zero-sum normal form game Γ induced by the stationary strategies of our DSG, mainly because the payoff functions of Γ, on a player own-strategy, are neither linear, nor concave Daskalakis et al. [2021] and so new idea are necessary to overcome this difficulty.

Vrieze and Tijs [1982] studied fictitious play in the context of varying stage games. However, similarly to the original fictitious play of Robinson and Robinson, it is designed as a way to compute the value of a zero-sum game, and it is not a behavioral strategy (i.e., there is no current state, the goal is to compute the minmax of a series of matrices). More recently, Leslie et al. [2020] proposed a best-response dynamics in continuous time which converges to stationary Nash equilibria in zero-sum DSG. The convergence was extended to discrete-time procedures and to identical-interest discounted stochastic games by Baudin and Laraki [2022]. These discrete-time and continuous-time systems are close to our systems, but they only work in model-based settings with no perturbation, whereas ours is suited for model-free settings and uses smooth best-responses. Sayin et al. [2021] introduced a fictitious play algorithm for zero-sum stochastic games in discrete time with a continuation up-dating rule close to Q-learning: both players update a Q-table with an entry per state-action pair whereas our continuation payoff are only indexed by the state. It works in a model-free setting, but with exact best-responses and it was proved to converge only in zero-sum SG.

Smooth or Regularized Learning Learning using regularizers is a widespread technique in machine learning, for instance with "follow the perturbed leader" (Cesa-Bianchi and Lugosi [2006]) or so-called stochastic fictitious play (Fudenberg and Levine [1995]). It makes it possible to have no-regret (Perchet [2014], Shalev-Shwartz [2011]) and from another point of view it is a substitute to the simpler ϵ-greedy exploration scheme in reinforcement learning (Sutton and Barto [2018]).

Local vs Global Our algorithms lead to all the players globally optimizing. Other algorithms such as COMA (Foerster et al. [2018]) lead to local optima, and so belong to another line of work.

6 Conclusion

We defined a number of decentralized continuous and discrete time systems with exploration to learn its own payoff function, the transition probability function, and which converge to regularized (and so ε-approximate) stationary Nash equilibria in discounted stochastic games. This is an area that has not been widely studied and as such, a number of questions remain to be answered–either regarding our systems or regarding other recent decentralized learning algorithms in stochastic games.

First, the theory of stochastic approximations gives no clue about the speed of convergence of the corresponding discrete-time algorithms. Moreover, even in continuous time, this question remains open even for potential games Harris [1998]. Therefore, an interesting question in both the stochastic and the repeated game setting is, which guarantee FP and SFP have regarding the rate of convergence?

Second, and inspired by the convergence of vanishing fictitious play to equilibria in classical repeated games (see Benaïm and Faure [2013], Hadikhanloo et al. [2021]), it would be of interest to study a vanishing version of our algorithm with a parameter $\varepsilon(t)$ that goes to zero in a way that guarantees the convergence to (exact) stationary Nash equilibria of the DSG.

Third, convergenve of FP or SFP in repeated game extends to infinite action games and to non atomic games. It would be of interest to relaxe our finiteness assumption, at least for the action sets.

A challenging open problem is the design of independent learning algorithms that converge to stationary equilibria in ergodic zero-sum and identical-interest DSG without the knowledge of the other player's past actions. The Projected Gradient Methods of Daskalakis et al. [2020], Leonardos et al. [2022] have this minimal information property but their results are proved only episodic SG.

Last but not least, obtaining a last iterate convergence instead of a time average convergence (as in our paper) or best iterate convergence (as in Daskalakis et al. [2020], Leonardos et al. [2022]) is another interesting direction.

References

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.

Lucas Baudin and Rida Laraki. Fictitious Play and Best-Response Dynamics in Identical-Interest and Zero-Sum Stochastic Games. *https://arxiv.org/pdf/2111.04317.pdf, accepted in ICML 2022*, May 2022.

E. Veronica Belmega, Panayotis Mertikopoulos, Romain Negrel, and Luca Sanguinetti. Online convex optimization and no-regret learning: Algorithms, guarantees and applications. *arXiv:1804.04529 [cs, math, stat]*, April 2018.

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.

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.

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.

Ulrich Berger. Brown's original fictitious play. *Journal of Economic Theory*, 135(1):572–578, July 2007. ISSN 00220531. doi: 10.1016/j.jet.2005.12.010.

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

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

Nicolò Cesa-Bianchi and Gábor Lugosi. Potential-Based Algorithms in Online Prediction and Game Theory. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, Lecture Notes in Computer Science, pages 48–64, Berlin, Heidelberg, 2001. Springer. ISBN 978-3-540-44581-4. doi: 10.1007/3-540-44581-1_4.

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.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive 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 5527–5540. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/3b2acfe2e38102074656ed938abf4ac3-Paper.pdf`.

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

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.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

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.

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.

Angeliki Giannou, Vlatakis-Gkaragkounis Emmanouil-Vasileios, and Mertikopoulos Panayotis. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *COLT '21: Proceedings of the 34th Annual Conference on Learning Theory*, 2021a.

Angeliki Giannou, Vlatakis-Gkaragkounis Emmanouil-Vasileios, and Mertikopoulos Panayotis. The convergence rate of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021b.

Saeed Hadikhanloo, Rida Laraki, Panayotis Mertikopoulos, and Sylvain Sorin. Learning in nonatomic games, part I: Finite action spaces and population games. *Journal of Dynamics and Games, to appear*, 2021. doi: https://arxiv.org/pdf/2107.01595.pdf.

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.

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.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/39ae2ed11b14a4ccb41d35e9d1ba5d11-Paper.pdf`.

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

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

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.

Ardeshir Kianercy and Aram Galstyan. Dynamics of Boltzmann Q learning in two-player two-action games. *Physical Review E*, 85(4):041145, April 2012. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.85.041145.

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.

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.

Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics & Games*, 4(2):125–148, 2017. ISSN 2164-6074. doi: 10.3934/jdg.2017008.

S. Leonardos, Panageas I., Overman W., and Piliouras G. Global convergence of multi-agent policy gradient in markov potential games. In *ICLR-22: 10th International Conference on Learning Representations, to appear*, 2022. doi: https://arxiv.org/pdf/2106.01969.pdf.

David S Leslie and E J Collins. Individual Q-Learning in Normal Form Games. *SIAM Journal on Control and Optimization*, 44(2):20, 2005. doi: 10.1137/S0363012903437976.

David S. Leslie and E. J. Collins. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, August 2006. ISSN 0899-8256. doi: 10.1016/j.geb.2005.08.005.

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.

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.

K. Miyasawa. On the convergence of learning processes in a 2x2 non-zero-person game. Levine's Working Paper Archive 419, David K. Levine, 1961.

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.

Abraham Neyman. Continuous-time stochastic games. *Games and Economic Behavior*, 104:92–130, July 2017. ISSN 08998256. doi: 10.1016/j.geb.2017.02.004.

Vianney Perchet. Approachability, regret and calibration: Implications and equivalences. *Journal of Dynamics & Games*, 1(2):181–254, 2014. ISSN 2164-6074. doi: 10.3934/jdg.2014.1.181.

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.

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.

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

Muhammed O. Sayin, Francesca Parise, and Asu Ozdaglar. Fictitious play in zero-sum stochastic games. *arXiv:2010.04223 [cs, math]*, December 2021.

Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000018.

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.

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.

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.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2094–2100, Phoenix, Arizona, 2016. AAAI Press.

Sebastian van Strien and Colin Sparrow. Fictitious play in 3×3 games: Chaos and dithering behaviour. *Games and Economic Behavior*, 73(1):262–286, September 2011. ISSN 0899-8256. doi: 10.1016/j.geb.2010.12.004.

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.

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

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, may 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL `https://doi.org/10.1007/BF00992696`.

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

7 Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See main restrictions in Section 2 and conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See theorem statements and hypothesis
   (b) Did you include complete proofs of all theoretical results? [Yes] See sketch of proof in the paper and complete detailed proofs in the Appendices A and B
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]