TRACKING REINFORCEMENT LEARNING AGENT PLAYING A FINITE STATE MARKOV GAME WITH A FLUID APPROXIMATION

YANN KERZREHO
UNIVERSITÉ PARIS-DAUPHINE
SUPERVISORS: PIERRE CARDALIAGUET, YANNICK VIOSSAT

Contents

1. Introduction

In recent years, the study of multi-agent reinforcement learning (MARL) has developed, particularly in environments where multiple agents interact over time in structured settings, such as Markov games (Littman 1994). These systems are appearing in fields ranging from robotic coordination to economic markets. However, analysing the long-term behaviour of learning agents in such environments remains a challenge due to the high variance and non-stationarity of the learning process. In this paper, we show how a fluid approximation can be used to study the asymptotic behaviour of agents. To do this, a stochastic process defining the behaviour of the agents is approximated by a process limit as the learning rate of the agent decreases and the number of iterations increases. This limit defines a relatively simple ODE, and can be used to study agent behaviour analytically. We apply it to a framework where the agents are Q learners with softmax exploration, showing that the system admits a deterministic ODE approximation capturing the expected evolution of the agents' strategies.

This paper aims to extend the idea of Banchio and Mantegazza 2022 to MARLs where the Markov game has a finite number of states. That last paper only derived a fluid limit for a stateless game using Kurtz 1970, this paper proposes a new theorem for averaging algorithm learning around the invariant distribution of game states. To our knowledge, this advance is original and paves the way for further research into algorithmic collusion, the field studying the emergence of cooperation between algorithms on marketplaces.

Section 2 introduces the main theorem of the paper and its proof. Section 3 explains how to apply it to RL agents playing a Markov game, section 4 is an example of 2 Q-tables playing a prisoner's dilemma, the canonical example of cooperation.

Date: 2025.

2. Main Theorem and its Proof

2.1. **Set up.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where $(X_n)_{n\in\mathbb{N}}$ is a process in $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ and $(G_n)_{n\in\mathbb{N}}$ a process in $(E, \mathcal{E})$ where E is finite. Let also assume that there is a function f and a probability kernel $P_\cdot$ such that,

$$f : \mathbb{R}^d \times E \to \mathbb{R}^d \quad \text{measurable s.t.} \quad X_{n+1} = X_n + f(X_n, G_n),$$
$$\mathbb{P}(G_{n+1} = g | G_n, X_n) = P_{X_n}(G_n, g) \quad \forall g \in E.$$

Let $(X_n, G_n)_{n\in\mathbb{N}}$ be a discrete time homogeneous Markov chain, we are interested in the fluid approximation of X in $\mathbb{R}^d$. Let us also define $\left((X_n^N, G_n^N)_{n\in\mathbb{N}}\right)_{N\in\mathbb{N}}$ s.t. $X_{n+1}^N = X_n^N + \frac{1}{N} f(X_n^N, G_n^N)$ and G_{n+1}^N with $P_{X_n^N}$ as kernel, a sequence of markov chain. Then we have,

$$X_n^N = X_0 + \frac{1}{N}\sum_{k=0}^{n-1} f(X_k^N, G_k^N).$$

We want to show the convergence between the X_n^N and the ODE defined by $y(0) = X_0$ and $y'(t) = \beta(y(t))$ where the derivative of fluid limit is defined as,

$$\beta : x \in \mathbb{R}^d \mapsto \sum_{g\in E} \pi_x(g) f(x, g) \in \mathbb{R}^d \tag{1}$$

where π_x is the ergodic probability measure of the chain given by the kernel P_x.

Assumption 1. *The following assumptions are made,*

- *For any $t \in \mathbb{N}$, there is a set $D \subset \mathbb{R}^d$ for which the process $(X_n^N)_{n\in\{1,\dots tN\}} \in D$ for all n and $N \in \mathbb{N}$.*
- *Kernels P_x are uniformly respecting the Doeblin condition: their exist an integer k, a positive c and a probability measure q on E s.t. ${P_x}^k(i,j) \geq c\, q(j)$ for all $x \in D$ and for all $i, j \in \mathcal{S}$.*
- *The transition matrix P_x is Lipschitz in D.*
- *$f : \mathbb{R}^d \times E \to \mathbb{R}^d$ restricted on D is Lipschitz in its first argument and bounded by $\|f\|_\infty = \|f|_D\|_\infty$*

From the set of assumptions we can deduce the following proposition,

Proposition 1. *When the set of assumption 1 holds true, then $x \mapsto \pi_x$ where π_x is the unique invariant measure of the Markov kernel P_x, is Lipschitz continuous on D. It also immediately implies that β defined as (1) is Lipschitz continuous on D.*

The proof, which is not specific to the paper and is fairly straightforward, can be found in the appendix.

2.2. **Theorem and its proof.**

Theorem 1. *Let consider a Markov chain $(X_n, G_n)_{n\in\mathbb{N}}$ defined as above and respecting the set of assumptions 1, then for any $t \in \mathbb{N}$ and $y_n^N = X_0 + \frac{1}{N}\sum_{k=0}^{n-1} \beta(y_k)$ where β is defined as (1),*

$$\sup_{0\leq n\leq tN} \mathbb{E}\left\|X_n^N - y_n^N\right\| = O\left(N^{-1/3}\right)$$

Proof. **Step 1.** For $N, M \in \mathbb{N}$ s.t. $tN/M \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left\|X_{tN}^N - y_{tN}^N\right\| &= \frac{1}{N}\mathbb{E}\left\|\sum_{n=0}^{tN-1} f(X_n^N, G_n^N) - \beta(y_n)\right\| \\
&\leq \frac{1}{N}\mathbb{E}\left\|\sum_{n=0}^{tN-1} f(X_n^N, G_n^N) - \beta(X_n^N)\right\| + \frac{1}{N}\mathbb{E}\left\|\sum_{n=0}^{tN-1} \beta(X_n^N) - \beta(y_n)\right\| \\
&\leq \frac{1}{N}\sum_{k=0}^{M-1} \mathbb{E}\left\|\sum_{n=ktN/M}^{(k+1)tN/M-1} f(X_n^N, G_n^N) - \beta(X_n^N)\right\| + \frac{1}{N}\mathbb{E}\left\|\sum_{n=0}^{tN-1} \beta(X_n^N) - \beta(y_n)\right\|
\end{aligned}$$

The last term can be controlled by Grönwall Lemma. Each term of the sum over k can be decomposed as follows, for simplicity we change indices to $k = 0$,

$$\mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_n^N, G_n^N) - \beta(X_n^N)\right\| \leq \mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_n^N, G_n^N) - f(X_n^N, \tilde{G}_n^N)\right\|$$
$$+ \sum_{n=0}^{tN/M-1} \mathbb{E}\left\|\beta(X_0^N) - \beta(X_n^N) + f(X_n^N, \tilde{G}_n^N) - f(X_0^N, \tilde{G}_n^N)\right\| + \mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_0^N, \tilde{G}_n^N) - \beta(X_0^N)\right\|$$

where $(\tilde{G}_n^N)_{n\geq k}$ is the Markov chain defined by the kernel $P_{X_0^N}$ and $\tilde{G}_0^N = G_0^N$. Moreover $G_{n+1}^N = h(X_n^N, G_n^N, \xi_{n+1})$ and $\tilde{G}_{n+1}^N = h(X_0^N, \tilde{G}_n^N, \xi_{n+1})$, with $(\xi_n)_{n\in\mathbb{N}} \sim \mathcal{U}[0,1]$ i.i.d., $\varphi : \{1, \dots \#E\} \subset \mathbb{N} \to E$ a bijection and,

$$h(x, g, \xi) = \begin{cases} \varphi(1) & \text{if } \xi \in \big[0, P_x\big(g, \varphi(1)\big)\big] = I_{x,g}^1 \\ \varphi(i) & \text{if } \xi \in \left]\sum_{j=1}^{i-1} P_x\big(g, \varphi(j)\big),\ \sum_{j=1}^{i} P_x\big(g, \varphi(j)\big)\right] = I_{x,g}^i, \text{ for } i > 1 \end{cases}$$

where $\bigsqcup_{i=1}^{\#E} I_{x,g}^i = [0,1]$ and with λ the Lebesgue measure, $\lambda(I_{x,g}^i) = P_x\big(g, \varphi(i)\big)$.

Step 2. Bounding $\mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_n^N, G_n^N) - f(X_n^N, \tilde{G}_n^N)\right\|$

$$\mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_n^N, G_n^N) - f(X_n^N, \tilde{G}_n^N)\right\| \leq \sum_{n=0}^{tN/M-1} \mathbb{E}\left\|\sum_{g\in E} f(X_n^N, g)\big(\mathbb{1}\{G_n^N = g\} - \mathbb{1}\{\tilde{G}_n^N = g\}\big)\right\|$$
$$\leq \|f\|_\infty \sum_{n=0}^{tN/M-1} \mathbb{E}\left[\sum_{g\in E}\left|\mathbb{1}\{G_n^N = g\} - \mathbb{1}\{\tilde{G}_n^N = g\}\right|\right] \leq \|f\|_\infty \sum_{n=0}^{tN/M-1} \mathbb{P}\left(G_n^N \neq \tilde{G}_n^N\right)$$

By Lemma 1 bellow, there is a positive C such that,

$$\|f\|_\infty \sum_{n=0}^{tN/M-1} \mathbb{P}\left(G_n^N \neq \tilde{G}_n^N\right) \leq \|f\|_\infty \sum_{n=0}^{tN/M-1} 1 - \left(1 - C\frac{M}{N}\right)_+^M \leq \|f\|_\infty \frac{tN}{M}\left(1 - \left(1 - C\frac{M}{N}\right)_+^M\right)$$

Step 3. Bounding $\mathbb{E}\sum_{n=0}^{tN/M-1}\left\|\beta(X_n^N) - \beta(X_0^N) + f(X_n^N, \tilde{G}_n^N) - f(X_0^N, \tilde{G}_n^N)\right\|$

As for any $n \in \mathbb{N}$,

$$\left\|X_{n+1}^N - X_n^N\right\| \leq \frac{1}{N}\|f\|_\infty$$

We have that,

$$\left\|X_L^N - X_0^N\right\| \leq \sum_{l=0}^{L-1}\left\|X_{l+1}^N - X_l^N\right\| \leq \frac{L}{N}\|f\|_\infty$$

Then,

$$\mathbb{E}\sum_{n=0}^{tN/M-1}\left\|\beta(X_0^N) - \beta(X_n^N) + f(X_n^N, \tilde{G}_n^N) - f(X_0^N, \tilde{G}_n^N)\right\| = O\left(\frac{t^2 N}{M^2}\right)$$

Step 4. Bounding $\mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_0^N, \tilde{G}_n^N) - \beta(X_0^N)\right\|$

$$\mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_0^N, \tilde{G}_n^N) - \beta(X_0^N)\right\| \leq \sum_{g\in E}\mathbb{E}\left\|\sum_{n=0}^{tN/M-1} f(X_0^N, g)\mathbb{1}\{\tilde{G}_n^N = g\} - f(X_0^N, g)\pi_{X_0^N}(g)\right\|$$
$$\leq \sum_{g\in E}\|f\|_\infty \mathbb{E}\left|\sum_{n=0}^{tN/M-1} \mathbb{1}\{\tilde{G}_n^N = g\} - \pi_{X_0^N}(g)\right|$$

By the Lemma 2 bellow if $tN/M \to \infty$ the order of the sum is $\left(\frac{tN}{M}\right)^{1/2}$.

Step 5. Aggregating all previous steps we have,

$$\mathbb{E}\left\|X_{tN}^N - y_{tN}^N\right\| \leq O\left(1-(1-M/N)_+^{\ M}\right) + O\left(1/M\right) + O\left((M/N)^{1/2}\right) + \frac{1}{M}\mathbb{E}\left\|\sum_{n=0}^{tN-1}\beta(X_n^N) - \beta(y_n)\right\|$$

By using $1-(1-M/N)^M = 1-\exp(M\ln(1-M/N)) = 1-\exp\left(-M^2/N+o(M^2/N)\right) \sim M^2/N$, we can find a M optimizing the convergence rate. Let $M=N^\alpha$, then the order is $N^{2\alpha-1}+N^{-\alpha}+N^{\alpha/2-1/2}$. For $\alpha = 1/3$ the final order is $N^{-1/3}$.

Step 6. Let κ the error defined as,

$$\kappa(N) = \mathbb{E}\left\|X_{tN}^N - y_{tN}^N\right\| - \frac{1}{M}\mathbb{E}\left\|\sum_{n=0}^{tN-1}\beta(X_n^N) - \beta(y_n)\right\|$$

an application that allow to bound the error of order $N^{-1/3}$. We denote L_β the Lipschitz constant of β, then we have for N large enough and for all $n \in \{0,\dots tN\}$,

$$\begin{aligned}\mathbb{E}\left\|X_n^N - y_n^N\right\| &= \kappa(N) + \frac{1}{N}\sum_{k=0}^{n-1}\mathbb{E}\left\|\beta(X_k^N) - \beta(y_k^N)\right\| \\ &\leq \kappa(N) + L_\beta\frac{1}{N}\sum_{k=0}^{n-1}\mathbb{E}\left\|X_k^N - y_k^N\right\|\end{aligned}$$

By using Grönwall lemma: $u_n \leq B + \sum_{k=0}^{n-1}\alpha_k u_k \implies u_n \leq B\exp\left\{\sum_{k=0}^{n-1}\alpha_k\right\}$, we have for all $n \in \{0,\dots tN\}$,

$$\begin{aligned}\mathbb{E}\left\|X_n^N - y_n^N\right\| &\leq \kappa(N)\exp\left\{\sum_{k=0}^{n-1}L_\beta/N\right\} \\ &\leq \kappa(N)e^{tL_\beta} = O\left(N^{-1/3}\right)\end{aligned}$$

As we have the inequality for all $n \in \{0,\dots tN\}$ we pass to the sup to get the final result. □

Lemma 1. *Let the set of assumptions 1 hold true. Define $\varphi : \{1,\dots \#E\} \subset \mathbb{N} \to E$ a bijection and,*

$$h(x,g,\xi) = \begin{cases}\varphi(1) & \text{if } \xi \in \left[0, P_x\big(g,\varphi(1)\big)\right] = I_{x,g}^1 \\ \varphi(i) & \text{if } \xi \in \left]\sum_{j=1}^{i-1}P_x\big(g,\varphi(j)\big),\ \sum_{j=1}^{i}P_x\big(g,\varphi(j)\big)\right] = I_{x,g}^i\end{cases}$$

where $\bigsqcup_{i=1}^{\#E} I_{x,g}^i = [0,1]$ and with λ the Lebesgue measure, $\lambda(I_{x,g}^i) = P_x\big(g,\varphi(i)\big)$.

With $(\xi_n)_{n\in\mathbb{N}} \sim \mathcal{U}[0,1]$ i.i.d. we define two Markov chains induced by the same sequence of random variable, $G_{n+1}^N = h(X_n^N, G_n^N, \xi_{n+1})$ et $\tilde{G}_{n+1}^N = h(X_0^N, \tilde{G}_n^N, \xi_{n+1})$. Then there is $C>0$ such that for all $n \in \{0,...M\}$

$$\mathbb{P}\left(G_n^N \neq \tilde{G}_n^N\right) \leq 1 - \left(1 - C\frac{M}{N}\right)_+^{\ M}$$

Proof.

$$\mathbb{P}\left(G_{n+1}^N\tilde{G}_{n+1}^N\right) \geq \mathbb{P}\left(G_n^N = \tilde{G}_n^N\right)\mathbb{P}\left(G_{n+1}^N = \tilde{G}_{n+1}^N \mid G_n^N = \tilde{G}_n^N\right)$$

We are looking for a lower bound of $\mathbb{P}\left(G_{n+1}^N = \tilde{G}_{n+1}^N \mid G_n^N = \tilde{G}_n^N\right)$ to induce an upper bound of $\mathbb{P}\left(G_{n+1}^N \neq \tilde{G}_{n+1}^N\right)$. With,

$$\mathbb{P}(G_{n+1}^N = g | G_n^N, X_n^N) = P_{X_n^N}\big(g, G_n^N\big),$$

Assuming $G_n^N = \tilde{G}_n^N$, we have that $G_{n+1}^N = \tilde{G}_{n+1}^N$ if and only if there exist a $i \in \{1, \dots \#E\}$ such that the random variable ξ_{n+1} is in both $I^i_{X_n^N, G_n^N}$ and $I^i_{X_0^N, \tilde{G}_n^N}$, then we have,

$$\mathbb{P}\left(G_{n+1}^N = \tilde{G}_{n+1}^N \mid G_n^N = \tilde{G}_n^N\right) = \lambda\left(\bigcup_{i=1}^{\#E}\left(I^i_{X_n^N, G_n^N} \cap I^i_{X_0^N, \tilde{G}_n^N}\right)\right)$$
$$= \sum_i^{\#E} \lambda\left(I^i_{X_n^N, G_n^N} \cap I^i_{X_0^N, \tilde{G}_n^N}\right)$$

With P L_P-Lipschitz in X, $\|X_n^N - X_m^N\| \leq \frac{|m-n|}{N}\|f\|_\infty$ and,

$$\sum_{j=1}^{i} P_{X_n^N}\big(g, \varphi(j)\big) \in B\left(\sum_{j=1}^{i} P_{X_0^N}\big(g, \varphi(j)\big),\ L_P\|f\|_\infty \frac{n-k}{N} i\right)$$

With $n - k \leq M$ and $i \leq \#E$, we get,

$$\lambda\left(I^i_{X_n^N, G_n^N} \cap I^i_{X_0^N, \tilde{G}_n^N}\right) \geq P_{X_0^N}\big(g, \varphi(i)\big) - 2L_P\|f\|_\infty \frac{M}{N}(\#E)$$

And finally, there is a $C < 0$ such that,

$$\mathbb{P}\left(G_{n+1}^N = \tilde{G}_{n+1}^N \mid G_n^N = \tilde{G}_n^N\right) = \sum_{i=1}^{\#E} \lambda\left(I^i_{X_n^N, G_n^N} \cap I^i_{X_0^N, \tilde{G}_n^N}\right)$$
$$\geq \left(1 - 2L_P\|f\|_\infty \frac{M}{N}(\#E)^2\right)_+ = \left(1 - C\frac{M}{N}\right)_+$$

With this lower bound and $G_0^N = \tilde{G}_0^N$, an induction can be used to get,

$$\mathbb{P}\left(G_{n+1}^N = \tilde{G}_{n+1}^N\right) \geq \left(1 - C\frac{M}{N}\right)_+ \mathbb{P}\left(G_n^N = \tilde{G}_n^N\right)$$
$$\mathbb{P}\left(G_n^N = \tilde{G}_n^N\right) \geq \left(1 - C\frac{M}{N}\right)_+^n \geq \left(1 - C\frac{M}{N}\right)_+^M$$

Then we can conclude. □

Lemma 2. *$(G_n)_{n\in\mathbb{N}}$ is a Markov chain in $(E, \mathcal{P}(E))$ with E finite. Its transition matrix is denoted P, and define μ^n as the distribution of G_n. If the chain respects the Doeblin condition: there exist an integer k, a positive c and a probability measure q on E s.t. $P^k(i,j) \geq c\, q(j)$ then,*

$$\|\mu^n - \pi\|_{TV} = \frac{1}{2}\sum_{g\in E} |\mu^n(g) - \pi(g)| \leq (1-c)^{\lfloor n/k \rfloor}$$

$$\mathbb{E}\left|\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{G_n = g\} - \pi(g)\right| = O(N^{-1/2}) \quad \textit{for any } g \in E.$$

The proof of the first point can be found in Giné et al. 1997. The proof of the second point can be found in the appendix and mainly uses Jensen's inequality. It allows the expectation to be bounded by an error term and the square root of the variance, which can simply be bounded under the assumptions made.

Corollary 1. *Under the assumptions of Theorem 1, the fluid-scaled process $\left(X_{tN}^N\right)_{t\geq 0}$ converges uniformly in expectation on any finite time interval to the unique solution $y(t)$ of the ordinary differential equation defined by,*

$$y(0) = X_0, \quad y'(t) = \beta(y(t)).$$

In particular, for any $T > 0$,

$$\sup_{0\leq t\leq T} \mathbb{E}\left\|X_{tN}^N - y(t)\right\| = O\left(N^{-1/3}\right).$$

The proof of the corollary is given in the appendix and uses a classical Euler scheme of order N^{-1} and the main theorem.

3. Application to a Markov Game played by an RL-agent

3.1. **Framework.** We will define here the formal framework of an algorithm playing a Markov game defined as the tuple $\mathcal{G} = (\mathcal{I}, \mathcal{S}, (A^i)_{i\in I}, T, (R^i)_{i\in I})$, where $\mathcal{I}$ is the set of N players, $\mathcal{S}$ a finite state space, A^i a finite set of player i's actions, with $\mathcal{A} = \times_{i\in I} A^i$. The map $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the probability transition function that gives the probabilities of reaching another state from a state given the actions played. The function $R^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function of player i.

At each iteration k of the game, each agent chooses an action a_k^i according to s_k, then the game transitions to state s_{k+1} according to T, and each agent i is rewarded by $R^i(s_k, a_k)$. Taking the formalism of Banchio and Mantegazza 2022, Let us define a *reinforcer*, an algorithm, *e.g.* a Q-table, that can play a Markov game.

Definition 1. *A reinforcer indexed by i is a tuple (X^i, π^i) with,*

- *X^i a d_i-dimensional stochastic process in $\mathbb{R}^{d_i}$ following,*

$$X^i_{k+1} = X^i_k + f^i\left(a_k, r^i_k, X^i_k, s_k, s_{k+1}\right)$$

- *$\pi^i : \mathbb{R}^{d_i} \times S \to \Delta(A^i)$ a policy that maps X^i and state to the probability of playing each action.*

where $a_k \in \mathcal{A}$ is the action played by players, r^i_k the reward of player i, and $s_k \in \mathcal{S}$ the state of the game, all at iteration k. When all players are reinforcers, a single vector $X_k = (X^1_k, \dots, X^N_k)$ of dimension d can be used as a synthesis of all players, then a single function $f : \mathcal{A} \times \mathbb{R}^N \times \mathbb{R}^d \times \mathcal{S}^2 \to \mathbb{R}^d$ and a single $\pi : \mathbb{R}^d \times \mathcal{S} \to \Delta(\mathcal{A})$ define the various stochastic process updates and policies.

3.2. **Application of the Theorem.** We can define $G_k = \left(s_k, a_k, s_{k+1}\right) \in E$ as the wrapped game, which includes all the information needed to update the algorithm at iteration k. Here $E = \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is finite. For simplicity, r^i_k is defined by the current state and joint actions, so the update of the process $(X_n)_{n\geq 0}$ at iteration $k+1$ is entirely known from $\left(X_k, G_k\right)$.

Conversely, G_{k+1} is defined by: s_{k+1} which is in G_k; a_{k+1}, which is determined by s_{k+1}, X_{k+1} and π; s_{k+2}, which is determined by a_{k+1}, s_{k+1} and the application T. Therefore, the law of G_{k+1} is given by the couple (X_k, G_k). We clearly have,

$$\begin{aligned} &f : \mathbb{R}^d \times E \to \mathbb{R}^d \quad \text{measurable s.t.} \quad X_{n+1} = X_n + f(X_n, G_n) \\ &\mathbb{P}(G_{n+1} = g | G_n, X_n) = P_{X_n}(G_n, g), \quad \forall g \in E. \end{aligned}$$

Then $(X_n, G_n)_{n\in\mathbb{N}}$ is a discrete time homogeneous Markov chain. Assuming that the set of assumptions 1 holds, we can apply the main theorem.

Intuitively, applying this theorem to a Markov game enables us to determine the limiting behaviour of agents when their learning speed is divided as much as the number of times they learn is increased. Overall, the order of magnitude of learning remains the same, but it is averaged across all possible game configurations. In very practical terms, to simulate convergence towards the ODE, all you have to do is to multiply the number of iterations by N and divide the learning speed parameter by N.

3.3. **Q-table with a soft-max policy.** We will showcase an algorithm that validates the assumptions. A Q-table assigns to each possible pair of action and state $(a^i, s) \in A^i \times \mathcal{S}$ a Q-value, which is updated with the Bellman equation to target the discounted rewards for playing action a^i in state s. The Q-values at iteration k for the couple (a^i, s) will be denoted $X^i_k(a^i, s)$.

When we equipped Q-tables with a policy, they can be seen as *reinforcers* such that for all (a^i, s),

$$X^i_{k+1}(a^i, s) = \begin{cases} X^i_k(a^i, s) + \alpha\left(r^i_k + \gamma \max\limits_{b\in A^i}\left(X^i_k(b, s_{k+1})\right) - X^i_k(a^i, s)\right) & \text{if } (a^i, s) = (a^i_k, s_k) \\ X^i_k(a^i, s) \text{ otherwise.} \end{cases} \tag{2}$$

where γ is the discount factor. Let's define the policy π^i made with a soft-max of temperature τ and with uniformly random exploration of rate ε, such that $\forall a^i \in A^i$,

$$\pi^i(X^i_k, s_k)(a^i) = (1-\varepsilon)\frac{\exp(X^i_k(a^i, s_k)/\tau)}{\sum_{b\in A_i}\exp(X^i_k(b, s_k)/\tau)} + \frac{\varepsilon}{\#A^i} \tag{3}$$

The verification of the set of assumptions is simple and can be found in the annex. Just note that the assumption of the uniform Doeblin condition is fulfilled thanks to the ε-rate of uniform exploration. It allows the wrapped game to explore in 2 iterations all the elements in E.

4. Q-table playing a Prisoner's Dilemma with memory

4.1. **Algorithms and Game.** In this subsection, two Q-tables (2) equipped with soft-max policies (3) playing a Prisoner's Dilemma are introduced. For both players A and B, $A^A = A^B = \{C, D\}$ and the gain matrix of this game will be parameterized by $g \in [1, 2]$ as below,

	C	D
C	(2g, 2g)	(g, 2 + g)
D	(2+g, g)	(2, 2)

Table 1. Gain matrix of the game.

The state transitions of the game are deterministic, with the next state determined by the current actions of the players. For example, if Player A plays C and Player B plays D, the next state will be CD. Then $\mathcal{S} = \{CC, CD, DC, DD\}$. Each agent thus has 8 Q-values, denoted $X_k^i \in \mathbb{R}^8$. The initial state s_0 will be uniformly distributed over $\mathcal{S}$.

An example simulation is shown in Figure 1, where each column represents the agent's Q values and each row represents a set of training parameters. The fluid approximation is also shown for comparison.

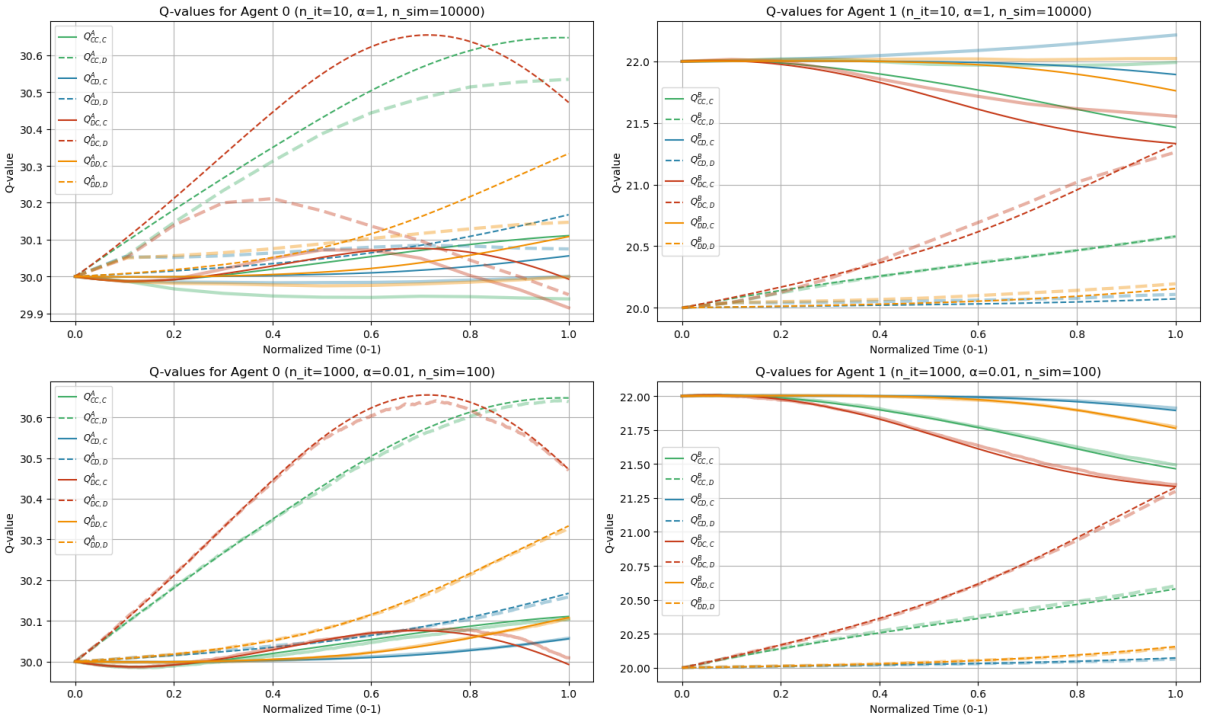

Figure 1. Fluid approximation (thin and opaque) and Simulation mean (large and translucent) Comparison with $g = 1.5, \tau = 0.5, \varepsilon = 0.1, \gamma = 0.9$. n_it the number of iteration of the game simulated and n_sim the number of simulation used to compute the mean.

References

[1] Martino Banchio and Giacomo Mantegazza. "Artificial intelligence and spontaneous collusion". In: *arXiv preprint arXiv:2202.05946* (2022).

[2] Evarist Giné et al. "Lectures on finite Markov chains". In: *Lectures on probability theory and statistics: École d'été de Probabilités de Saint-Flour XXVI-1996* (1997), pp. 301–413.

[3] Thomas G Kurtz. "Solutions of ordinary differential equations as limits of pure jump Markov processes". In: *Journal of applied Probability* 7.1 (1970), pp. 49–58.

[4] Michael L Littman. "Markov games as a framework for multi-agent reinforcement learning". In: *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.

Appendix A. Additional Proofs

A.1. **Proposition 1.** When the set of assumption 1 holds true, then $x \mapsto \pi_x$ where π_x is the unique invariant measure of the Markov kernel P_x, is Lipschitz continuous on D. It also immediately implies that β defined as (1) is Lipschitz continuous on D.

Proof. Let $x, y \in D$. Since E is finite and P_x satisfies a uniform Doeblin condition, there exists an integer $k \geq 1$ and a constant $\theta \in (0,1)$ such that for all probability measures μ, ν on E,

$$\|\mu P_x^k - \nu P_x^k\|_{\mathrm{TV}} \leq \theta \|\mu - \nu\|_{\mathrm{TV}}.$$

Moreover, since $\pi_x P_x = \pi_x$, it holds that

$$\|\pi_x - \pi_y\|_{\mathrm{TV}} = \|\pi_x P_x^k - \pi_y P_y^k\|_{\mathrm{TV}} \leq \|\pi_x P_x^k - \pi_x P_y^k\|_{\mathrm{TV}} + \|\pi_x P_y^k - \pi_y P_y^k\|_{\mathrm{TV}}.$$

Since the second term can be bounded with the contraction property, we still have to bound the first term $\|\pi_x P_x^k - \pi_x P_y^k\|_{\mathrm{TV}}$. Let us prove by induction that for all $k \in \mathbb{N}$,

$$\|P_x^k - P_y^k\|_{\mathrm{op}} \leq k L_P \|x - y\|,$$

where $\|\cdot\|_{\mathrm{op}} = \sup_{\mu \in \mathcal{P}(E)} \|\mu(\cdot)\|_{\mathrm{TV}}$ and L_P is the Lipschitz constant of $x \mapsto P_x$ on D such that,

$$\|P_x - P_y\|_{\mathrm{op}} \leq L_P \|x - y\|.$$

Assume the property holds for some $k \geq 1$. Then,

$$\begin{aligned}
\|P_x^{k+1} - P_y^{k+1}\|_{\mathrm{op}} &= \|P_x P_x^k - P_y P_y^k\|_{\mathrm{op}} \leq \|P_x (P_x^k - P_y^k)\|_{\mathrm{op}} + \|(P_x - P_y) P_y^k\|_{\mathrm{op}} \\
&\leq \|P_x\|_{\mathrm{op}} \cdot \|P_x^k - P_y^k\|_{\mathrm{op}} + \|P_x - P_y\|_{\mathrm{op}} \cdot \|P_y^k\|_{\mathrm{op}} \\
&\leq k L_P \|x - y\| + L_P \|x - y\| = (k+1) L_P \|x - y\|.
\end{aligned}$$

Therefore, using $\pi_x \in \mathcal{P}(E)$,

$$\|\pi_x P_x^k - \pi_x P_y^k\|_{\mathrm{TV}} \leq \|P_x^k - P_y^k\|_{\mathrm{op}} \leq k L_P \|x - y\|.$$

Use the contraction property of P_y^k,

$$\|\pi_x - \pi_y\|_{\mathrm{TV}} \leq k L_P \|x - y\| + \theta \|\pi_x - \pi_y\|_{\mathrm{TV}}.$$

Which gives,

$$\|\pi_x - \pi_y\|_{\mathrm{TV}} \leq \frac{k L_P}{1 - \theta} \|x - y\|.$$

□

A.2. **Lemma 2.** $(G_n)_{n \in \mathbb{N}}$ is a Markov chain in $(E, \mathcal{P}(E))$ with E finite. Its transition matrix is denoted P, and define μ^n as the distribution of G_n. If the chain respects the Doeblin condition: there exist an integer k, a positive c and a probability measure q on E s.t. $P^k(i,j) \geq c\, q(j)$ then,

$$\|\mu^n - \pi\|_{TV} = \frac{1}{2} \sum_{g \in E} |\mu^n(g) - \pi(g)| \leq (1 - c)^{\lfloor n/k \rfloor}$$

$$\mathbb{E}\left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{G_n = g\} - \pi(g) \right| = O(N^{-1/2}) \quad \text{for any } g \in E.$$

Proof. The first point is classical results, we refer to Giné et al. 1997 first section. For any $g \in E$ Let us define $S_N(g) = \sum_{n=1}^{N} \mathbb{1}\{G_n = g\}$. We have $\mathbb{P}(G_n = g) = \pi(g) + r_n$ with $|r_n| \leq 2(1-c)^{\lfloor n/k \rfloor}$

$$\mathbb{E}\left| \frac{1}{N} S_N(g) - \pi(g) \right| \leq \mathbb{E}\left| \frac{1}{N} S_N(g) - \frac{1}{N} \mathbb{E}[S_N(g)] \right| + \frac{1}{N} \sum_{n=1}^{N} 2(1-c)^{\lfloor n/k \rfloor}$$

We have $\frac{1}{N} \sum_{n=1}^{N} 2(1-c)^{\lfloor n/k \rfloor} \leq \frac{k}{cN}$ and by the Jensen inequality we have,

$$\mathbb{E}\left| \frac{1}{N} S_N(g) - \frac{1}{N} \mathbb{E}\big[S_N(g)\big] \right| \leq \frac{1}{N} \sqrt{\mathrm{Var}\big(S_N(g)\big)}$$

Let us compute $\mathrm{Var}(S_N)$,

$$\mathrm{Var}(\mathbb{1}\{G_n = g\}) = \mathbb{P}(G_n = g)[1 - \mathbb{P}(G_n = g)] = \pi(g)[1 - \pi(g) - 2r_n] + r_n[1 - r_n]$$

Thus,

$$\sum_{n=1}^{N} \mathrm{Var}(\mathbb{1}\{G_n = g\}) = N\pi(g)[1-\pi(g)] - 2\pi(g)\sum_{n=1}^{N} r_n + \sum_{n=1}^{N} r_n(1-r_n)$$
$$\leq N\pi(g)[1-\pi(g)] + O(1)$$

And with $i < j$,

$$\mathrm{Cov}(\mathbb{1}\{G_i = g\}, \mathbb{1}\{G_j = g\}) = \mathbb{P}(G_i = g, G_j = g) - \mathbb{P}(G_i = g)\mathbb{P}(G_j = g)$$
$$= \mathbb{P}(G_i = g)[P^{j-i}(g,g) - \mathbb{P}(G_j = g)] \leq (\pi(g) + r_i)\big[2(1-c)^{\lfloor (j-i)/k \rfloor} - r_j\big]$$

Thus with $\pi(g) + |r_i| < 2$ and $2(1-c)^{\lfloor (j-i)/k \rfloor} - r_j \leq 4(1-c)^{\lfloor (j-i)/k \rfloor}$

$$2\sum_{i<j\leq N} \mathrm{Cov}(\mathbb{1}\{G_i = g\}, \mathbb{1}\{G_j = g\}) \leq 2(\pi(g)+1)\sum_{i<j\leq N} 4(1-c)^{\lfloor (j-i)/k \rfloor} \leq 16\frac{kN}{c}$$

The variance is then bounded,

$$\mathrm{Var}\big(S_N(g)\big) = \sum_{n=1}^{N} \mathrm{Var}(\mathbb{1}\{G_n = g\}) + 2\sum_{i<j\leq N} \mathrm{Cov}(\mathbb{1}\{G_i = g\}, \mathbb{1}\{G_j = g\})$$
$$\leq N\pi(g)[1-\pi(g)] + 16\frac{kN}{c} + O(1)$$

Then we conclude with,

$$\mathrm{Var}\big(S_N(g)\big) = O(N)$$

□

A.3. **Corollary 1.** Under the assumptions of Theorem 1, the fluid-scaled process $\left(X^N_{tN}\right)_{t\geq 0}$ converges uniformly in expectation on any finite time interval to the unique solution $y(t)$ of the ordinary differential equation defined by,

$$y(0) = X_0, \quad y'(t) = \beta(y(t)).$$

In particular, for any $T > 0$,

$$\sup_{0\leq t\leq T} \mathbb{E}\left\|X^N_{tN} - y(t)\right\| = O\left(N^{-1/3}\right).$$

Proof. Recall that we defined the Euler approximation by

$$y^N_0 = X_0, \quad y^N_{n+1} = y^N_n + \frac{1}{N}\beta(y^N_n).$$

By the triangular inequality, for any $T > 0$ we have

$$\mathbb{E}\sup_{0\leq t\leq T} \|X^N_{tN} - y(t)\| \leq \mathbb{E}\sup_{0\leq t\leq T} \|X^N_{tN} - y^N_{tN}\| + \sup_{0\leq t\leq T} \|y^N_{tN} - y(t)\|.$$

From our main theorem, we have

$$\sup_{0\leq t\leq T} \mathbb{E}\|X^N_{tN} - y^N_{tN}\| = O\left(N^{-1/3}\right).$$

On the other hand, by standard results for the convergence of the Euler scheme, since β is Lipschitz, we have

$$\sup_{0\leq t\leq T} \|y^N_{tN} - y(t)\| = O\left(\frac{1}{N}\right).$$

Combining the two estimates yields

$$\sup_{0\leq t\leq T} \mathbb{E}\|X^N_{tN} - y(t)\| = O\left(N^{-1/3}\right),$$

□

Appendix B. Validation of the Assumptions for the Q-table

- For any $t \in \mathbb{N}$, there is a set $D \subset \mathbb{R}^d$ for which the process $(X_n^N)_{n\in\{1,...tN\}} \in D$ for all n and $N \in \mathbb{N}$.

We have that $\|X_{k+1}^N - X_k^N\|_\infty \leq \frac{1}{N}\alpha(3 + \gamma\|X_k^N\|_\infty)$, then by the Grönwall Lemma,

$$\begin{aligned}\|X_{tN}^N\|_\infty &\leq \|X_0\|_\infty + \sum_{k=0}^{tN-1} \|X_{k+1}^N - X_k^N\|_\infty \leq \|X_0\|_\infty + \sum_{k=0}^{tN-1} \frac{1}{N}\alpha(3 + \gamma\|X_k^N\|_\infty) \\ &\leq \|X_0\|_\infty + 3t\alpha e^{t\alpha}\end{aligned}$$

And thus we can bound the process in a compact,

$$X_n^N \in \overline{B}_{\|.\|_\infty}(0, \|X_0\|_\infty + 3t\alpha e^{t\alpha}) = D, \quad \forall n \in \{1, ...tN\}$$

- Kernels P_x are uniformly respecting the Doeblin condition: their exist an integer k, a positive c and a probability measure q on E s.t. ${P_x}^k(i,j) \geq c\, q(j)$ for all $x \in D$ and for all $i, j \in \mathcal{S}$.

For any $x \in D$ and $s \in \mathcal{S}$, $\pi(x,s) \geq \varepsilon/\#\mathcal{S}$, then for all $i, j \in \mathcal{S}$, ${P_x}^2(i,j) \geq (\varepsilon/\#\mathcal{S})^2$. So it is true for $c = \varepsilon^2/\#\mathcal{S}$ and q uniform.

- $f : \mathbb{R}^d \times E \to \mathbb{R}^d$ restricted on D is Lipschitz in its first argument and bounded by $\|f\|_\infty = \|f|_D\|_\infty$

Is immediate with the process being bounded and by using $\|.\|_\infty$.

- The transition matrix P_x is Lipschitz in D.

We have for all $x, y \in D$ and for all $i, j \in \mathcal{S}$,

$$\begin{aligned}P_x(i,j) &= \mathbb{P}\big(G_{n+1} = j | X_{n+1} = x + f(x,i), G_n = i\big) \\ &= \sum_{a\in A} \pi(x + f(x,i), s_n)(a) \cdot T(s_n, a)(j)\end{aligned}$$

Moreover the policy π is L_π-Lipsichtz on D as it is uniformly derivable in a compact. Then,

$$\begin{aligned}\big|P_x(i,j) - P_y(i,j)\big| &\leq \sum_{a\in A} T(s_n,a)(j)\big|\pi(x + f(x,i), s_n)(a) - \pi(y + f(y,i), s_n)(a)\big| \\ &\leq \sum_{a\in A} T(s_n,a)(j) L_\pi \big|x + f(x,i) - y - f(y,i)\big| \\ &\leq (\#A) L_\pi (1 + L_f)\big|x - y\big|\end{aligned}$$