A Multi-Agent Reinforcement Learning Approach to Promote Cooperation in Evolutionary Games on Networks with Environmental Feedback

Zhang, Tuo; Gupta, Harsh; Suprabhat, Kumar; Stella, Leonardo

DOI:
10.1109/CDC49753.2023.10383787

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Zhang, T, Gupta, H, Suprabhat, K & Stella, L 2024, A Multi-Agent Reinforcement Learning Approach to Promote Cooperation in Evolutionary Games on Networks with Environmental Feedback. in *2023 62nd IEEE Conference on Decision and Control, CDC 2023.* Proceedings of the IEEE Conference on Decision and Control, Institute of Electrical and Electronics Engineers (IEEE), pp. 2196-2201, 62nd IEEE Conference on Decision and Control, CDC 2023, Singapore, Singapore, 13/12/23. https://doi.org/10.1109/CDC49753.2023.10383787

A Multi-Agent Reinforcement Learning Approach to Promote Cooperation in Evolutionary Games on Networks with Environmental Feedback

Tuo Zhang, Harsh Gupta, Kumar Suprabhat and Leonardo Stella

***Abstract*— A prominent feature of biological organization in many species of social animals is the ability to achieve cooperation. However, despite its predominance in natural evolution, cooperative behaviors come at a cost, typically in the form of *do ut des* mechanisms (e.g., reciprocal altruism in vampire bats) with given thresholds for sharing resources or communication efforts. In this paper, we investigate the conditions of cooperation through the evolutionary dynamics of the prisoner's dilemma (PD) game as well as the learning dynamics resulting from the corresponding multi-agent reinforcement learning (MARL) model. In both cases, the interactions in the population are captured by a regular network and the impact of the players' actions is reflected through the evolution of an environmental resource, which also acts as a feedback on the dynamics. The following is a list of contributions: i) we provide a full characterization of the stability properties of the networked feedback-evolving PD game; ii) we determine a set of threshold values below which cooperation is promoted; iii) we develop the corresponding cross-learning model, which is a stateless MARL model, and we show that this model is equivalent to the networked PD game with environmental feedback.**

I. Introduction

Since the pioneering work by Smith and Price on evolutionary game theory in 1973 and 1982 [1], [2], the combined efforts by the research community have led to a more precise understanding of the dynamics of natural evolution in terms of the strategic interactions in a population of decision-makers. Scrupulous attention has been given to the conditions for cooperation to thrive in a population of selfish decision-makers. Indeed, the prominent work by Axelrod and Hamilton was one of the first and more prominent works to address this topic within the context of the Prisoner's Dilemma (PD) game [3].

Understanding the conditions for cooperation to thrive and prosper has always fascinated scholars, especially when they rely on the animal kingdom for inspiration. Examples of cooperative behaviors are frequent in nature, such as reciprocal altruism in clouds of vampire bats [4] and collective decision-making in honeybee swarms [5]. However, especially in the latter case, cooperation is achieved when the resource sharing is limited to a certain threshold, determining a watershed on the amount of interactions in the swarm [6].

In the vast body of literature on social dilemmas, the majority of early studies considered well-mixed populations, namely, where every member interacts with everyone else. However, this is a significant limiting assumption, as it is not able to capture a finer level of interactions. To overcome this limitation, the contribution in [7] and in [8] were some of the first attempts to investigate structured populations in the PD game. The main difference with the classical formulation of the game dynamics is that players are now edges of a network and communicate via a number of edges. This results in the death of players with lower fitness in favor of those with higher fitness. The means for it to happen is given via a number of update rules, of which the most popular are: Birth-Death (BD), Death-Birth (DB) and Imitation (IM) [9], [10].

More recently, the introduction of game-environment feedback mechanisms have sparked growing interest in the research community [11], [12]. The intuition is that the evolution of the frequencies of strategies is coupled with the evolution of an environmental resource which in turn acts as a feedback on the population dynamics. This approach has led to interesting applications for its ability to capture the complexity of many real systems in a range of disciplines, including social sciences, economics and biology [13]–[16]. In the same line of research, we previously investigated irrational behaviors via prospect theory [17] and the impact of the interactions on cooperation [18].

Furthermore, motivated by the increasing attention received by machine learning (ML), we have studied the overlap between evolutionary dynamics and multi-agent reinforcement learning (MARL). Gaining qualitative insights into the learning dynamics of MARL is a persistent challenge. In the past decade, evolutionary dynamics have developed a range of MARL algorithms, providing a range of approaches in different circumstances [20]. The formal link between evolutionary dynamics and MARL was first established by Börgers and Sarin in 1997 [21]. They demonstrated a formal relationship between cross learning [19] and replicator dynamics in formal games where the set of available actions is discrete. The works that followed, e.g., see [22] and [23], extended this link to formal games with continuous action spaces. Other works also demonstrated the applicability of this approach to stochastic games with discrete actions [24]–[26]. More recently, research by Perolat *et al.* has shown that replicator dynamics can be used to develop algorithms under specific design choices [27]. Using this approach, reinforcement learning agents have mastered the Stratego game, to name an example.

Highlights of contributions. The contribution of this paper is threefold. First, we provide a full characterization of the stability properties of the networked feedback-evolving PD game under any choice of the system parameters. The network topology is captured by a regular network of degree k. Second, we extend previous works in the literature by giving a precise equivalent of the thresholds for the most common update rules, namely, Birth-Death (BD), Death-Birth (DB) and Imitation (IM). Third, motivated by the popularity of

MARL, we develop the MARL model corresponding to the networked PD game with environmental feedback where we introduce the network topology and the environmental resource as elements of novelty. We call the proposed model *networked resource-evolving cross learning*, in line with the stateless MARL model called cross learning.

This paper is organized as follows. After the notation, we introduce our model in Section II. In Section III, we carry out the stability analysis of this model to fully characterize the stability properties of this system. In Section IV, we formulate the corresponding MARL model in the stateless setting. In Section V, we provide two sets of simulations to corroborate the theoretical results. Finally, in Section VI, we draw conclusions and discuss future research.

II. Networked Feedback-Evolving PD Model

In this section, we introduce the networked PD game model. The evolution of the frequencies of the strategies are given by x_k and $(1-x_k)$ (because of the conservation of mass law) representing the portion of the population choosing to cooperate and to defect, respectively, subject to a regular network of degree k. The evolution of these frequencies is mutually dependent on an environmental resource n. This resource evolves over time as a result of the population dynamics and, in turn, acts as a feedback onto the population dynamics via the payoff matrix $A(n)$.

In line with the framework originating in [11], let the environment-dependent payoff matrix be defined as:

$$\begin{aligned} A(n) &= (1-n)\begin{bmatrix} T & P \\ R & S \end{bmatrix} + n\begin{bmatrix} R & S \\ T & P \end{bmatrix} \\ &= \begin{bmatrix} T-n\delta_{TR} & P-n\delta_{PS} \\ R+n\delta_{TR} & S+n\delta_{PS} \end{bmatrix}, \end{aligned} \tag{1}$$

where $\delta_{TR} = T - R$ and $\delta_{PS} = P - S$, and the parameters obey the following inequality $T > R > P > S$. The above matrix can be given a physical interpretation by considering the embedded symmetry of the two parts of the matrix: when the value of the environmental resource is large, players are incentivized to defect (as each player can attain a portion of the resource), shifting the matrix towards a PD game (where defection is the Nash equilibrium). On the other hand, when the resource is scarce, players will find it more rewarding to cooperate (shifting the matrix towards the Harmony game, i.e., cooperation is the Nash equilibrium) [11].

For the evolutionary game dynamics on a regular graph of degree k can be described by the following transformation of the payoff matrix A: $[a_{ij}] \to [a_{ij} + b_{ij}(k)]$, where a_{ij} represents the (i,j) entry of the original payoff matrix A and $b_{ij}(k)$ is the players' interactions parameter whose value depends on the degree of the network k and on the update rule chosen. In [9], the authors derive three update rules:

- In the Birth-Death (BD) rule, a node is selected with a probability proportional to its fitness, and one of its k neighbors at random is replaced by the offspring:

$$b_{ij}(k) = \frac{a_{ii} + a_{ij} - a_{ji} - a_{jj}}{k-2}.$$

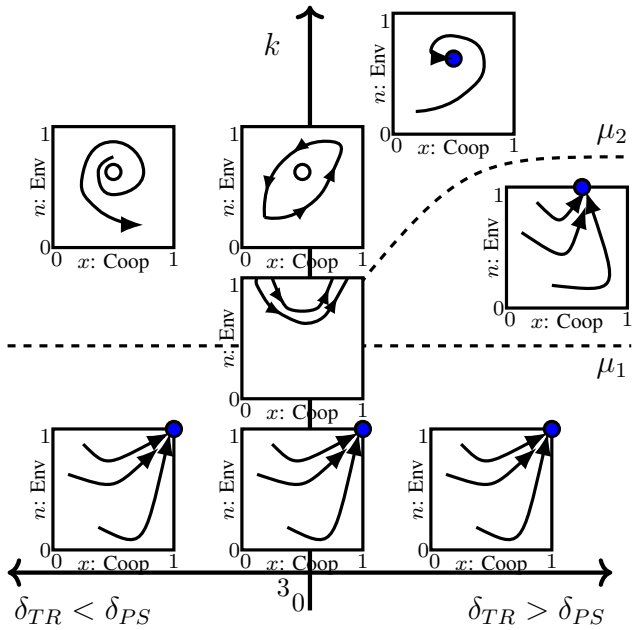

Fig. 1. Qualitative classification of the system dynamics in terms of the possible combinations of payoffs and network thresholds for the DB and IM update rules (hence, the general terms μ_1 and μ_2). Locally stable and unstable points are denoted by a blue circle and an empty circle, respectively.

- In the Death-Birth (DB) rule, a node is selected at random, and one of its k neighbors replaces it with its offspring with a probability proportional to their fitness:

$$b_{ij}(k) = \frac{(k+1)a_{ii} + a_{ij} - a_{ji} - (k+1)a_{jj}}{(k+1)(k-2)}.$$

- The Imitation (IM) rule, a node is randomly chosen to update its strategy by imitating one of its k neighbours proportionally to their fitness:

$$b_{ij}(k) = \frac{(k+3)a_{ii} + 3a_{ij} - 3a_{ji} - (k+3)a_{jj}}{(k+3)(k-2)}.$$

Note that regardless of the update rule chosen, $b_{ii}(k) = 0$. Under weak selection and pair approximation (see [7], [9]), the networked PD game model with environmental feedback resulting from the replicator equation is [18]:

$$\begin{aligned} \epsilon\dot{x} &= x(1-x)[(\delta_{PS} + (\delta_{TR} - \delta_{PS})x)(1-2n) + b_{12}(k)], \\ \dot{n} &= n(1-n)[(1+\lambda)x - 1], \end{aligned} \tag{2}$$

where the term $n(1-n)$ is used to ensure that the state of the environment is within the domain $[0,\ 1]$. Parameter ϵ denotes the rate at which the population dynamics change the environment and parameter $\lambda > 0$ represents the ratio between enhancement and degradation effects in the environment: when $\lambda < 1$, the degradation effect is stronger than the enhancement effect and vice versa when $\lambda > 1$; when $\lambda = 1$, the two effects are balanced.

III. Stability Analysis

In this section, we investigate the stability of system (2). We provide an initial result in the following lemma.

Lemma 1: Consider system (2). This system has 7 fixed points, as listed in Table I.

Proof: The proof is straightforward and we refer the reader to [18]. ∎

Remark. A physical interpretation of these equilibria is to have full cooperation or full defection in a depleted or replete environments, or the case of mixed populations in an intermediate environment.

Now, we investigate the stability of the fixed point stated in the above lemma. To this end, we consider the three update rules introduced in the previous section and formulate the corresponding thresholds. To this end, let the threshold for the BD, DB and IM rules be, respectively:

$$\mu_{BD}:=1-\delta_{PS}/\delta_{TR}, \tag{3}$$

$$\mu_{DB1}:=\frac{T-P+\sqrt{(P-T)^2-4\delta_{TR}(\delta_{PS}-\delta_{TR})}}{2\delta_{TR}}, \tag{4}$$

$$\mu_{DB2}:=\frac{T-P-S\lambda+R\lambda+\sqrt{\alpha}}{2(\delta_{TR}+\delta_{PS}\lambda)}, \tag{5}$$

$$\mu_{IM1}:=\frac{-(T-2R+P)+\sqrt{\gamma}}{2\delta_{TR}}, \tag{6}$$

$$\mu_{IM2}:=\frac{-(T-2R+P+2p\lambda-s\lambda-R\lambda)+\sqrt{\beta}}{2(\delta_{TR}+\delta_{PS}\lambda)}, \tag{7}$$

where $\alpha=(P-T+S\lambda-R\lambda)^2-4(\delta_{TR}+\delta_{PS}\lambda)(\delta_{TR}-\delta_{PS})(\lambda-1)$, $\beta=(T-2R+P+2P\lambda-S\lambda-R\lambda)^2-4(\delta_{TR}+\delta_{PS}\lambda)(3P+3R-S-5T)(1-\lambda)$, $\gamma=(T-2R+P)^2-4\delta_{TR}(3P+3R-S-5T)$.

A major difference with the standard PD game is that the DB and IM rules have now two thresholds. Figure 1 depicts the system dynamics for the DB and IM rules.

Theorem 1: Consider system (2) and the BD update rule. All fixed points are unstable regardless of the value of k. The system exhibits closed periodic orbits centred at the fixed point VII.

Proof: To study the stability of this system, we derive the Jacobian:

$$J(x,n)=\begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix},$$

where $J_{11}=(1-2x)[(\delta_{PS}+(\delta_{TR}-\delta_{PS})x)(1-2n)+b_{12}]+x(1-x)(1-2n)(\delta_{TR}-\delta_{PS})$, $J_{12}=-2x(1-x)[\delta_{PS}+(\delta_{TR}-\delta_{PS})x]$, $J_{21}=n(1-n)(1+\lambda)$ and $J_{22}=(1-2n)[(1+\lambda)x-1]$. Linearizing the Jacobian about point VII yields:

$$J(x,n)=\begin{bmatrix} 0 & \frac{-2\lambda(\delta_{TR}+\delta_{PS}\lambda)}{(1+\lambda)^3} \\ \frac{1+\lambda}{4} & 0 \end{bmatrix}.$$

The eigenvalues of the above Jacobian have no real part, indicating that the point VII is a center. A similar conclusion can be drawn for the other fixed points. ∎

Theorem 2: Consider system (2) under the DB update rule. The stability of this system can be fully characterized by parameters $\delta_{PS}, \delta_{TR}, k$ as in the following.

Case 1: $\delta_{TR}>\delta_{PS}$.

a) When $k<\mu_{DB1}$, the trajectories converge to point IV.
b) When $\mu_{DB2}>k>\mu_{DB1}$, they converge to point VI.
c) When $k>\mu_{DB2}$, they converge to point VII.

Case 2: $\delta_{TR}<\delta_{PS}$.

a) When $k<\mu_{DB1}$, the only stable point is point IV.
b) When $k>\mu_{DB1}$, all fixed points are unstable.

Case 3: $\delta_{TR}=\delta_{PS}$.

a) When $k<\mu_{DB1}$, the only stable point is point IV.
b) When $k>\mu_{DB1}$, the trajectories are closed periodic orbits centred at point VII.

Proof: The proof is divided into three parts. In the first part, we analyze whether the fixed points exist in the given domain. In the second part, we study the stability of each point. In the third part, the results of the first two parts are combined to determine the stability property of the system. This proof only contains three points: IV, VI, and VII. For the sake of conciseness, we restrict the proof to **Case 1**, i.e., $\delta_{TR}>\delta_{PS}$. This proof relies on an assumption $k\geq 3$, which must be true when we use the DB update rule.

Part 1. Only points VI and VII need to be analysed to see if they are in the domain of available values. For point VI, the condition for its existence is

$$0\leq -(b_{12}(k)-\delta_{PS})/(\delta_{PS}-\delta_{TR})\leq 1.$$

As $T>R>P>S$ holds and $\delta_{PS}<\delta_{TR}$. The equation above yields $\delta_{PS}\leq b_{12}(k)\leq\delta_{TR}$. As for point VII, the condition for its existence is

$$0\leq (b_{12}(k)+\delta_{TR}+b_{12}(k)\lambda+\delta_{PS}\lambda)/2(\delta_{TR}+\delta_{PS}\lambda)\leq 1,$$

for which we have $b_{12}(k)\leq(\delta_{TR}+\delta_{PS}\lambda)/(1+\lambda)$.

Part 2. To study the stability of this system, we derive the Jacobian as in the proof of Theorem 1. Linearizing about point IV yields:

$$J(x,n)=\begin{bmatrix} \delta_{TR}-b_{12}(k) & 0 \\ 0 & -\lambda \end{bmatrix},$$

which is asymptotically stable when $b_{12}(k)\geq\delta_{TR}$ and unstable otherwise (since one of the two eigenvalues would be positive). Linearizing the Jacobian about point VI yields:

$$J(x,n)=\begin{bmatrix} \frac{-(\delta_{PS}-b_{12})(\delta_{TR}-b_{12})}{\delta_{TR}-\delta_{PS}} & \frac{-b_{12}(\delta_{PS}-b_{12})(\delta_{TR}-b_{12})}{(\delta_{TR}-\delta_{PS})^2} \\ 0 & \frac{\delta_{TR}+\delta_{PS}\lambda-b_{12}\lambda-b_{12}}{\delta_{TR}-\delta_{PS}} \end{bmatrix},$$

where b_{12} in place of $b_{12}(k)$ for conciseness. This is asymptotically stable when

$$\frac{\delta_{TR}+\delta_{PS}\lambda-b_{12}(k)\lambda-b_{12}(k)}{\delta TR-\delta_{PS}}\leq 0,$$

which translates into $b_{12}(k)\geq(\delta_{TR}+\delta_{PS}\lambda)/(1+\lambda)$, and unstable otherwise, since one of the two eigenvalues would be positive. The other eigenvalue $\frac{-(\delta_{PS}-b12(k))(\delta_{TR}-b12(k))}{\delta_{TR}-\delta_{PS}}$ is always less than or equal to 0 when point VI exists. Finally, linearizing the Jacobian about point VII yields:

$$J(x,n)=\begin{bmatrix} \frac{\lambda b_{12}(\delta_{PS}-\delta_{TR})}{(1+\lambda)(\delta_{TR}+\delta_{PS}\lambda))} & \frac{-2\lambda(\delta_{TR}+\delta_{PS}\lambda)}{(1+\lambda)^3} \\ \frac{((\delta_{TR}+\delta_{PS}\lambda)^2-b_{12}^2(1+\lambda)^2)(1+\lambda)}{4(\delta_{TR}+\delta_{PS}\lambda)^2} & 0 \end{bmatrix},$$

TABLE I

LIST OF ALL THE FIXED POINTS FOR SYSTEM (2).

#	x	n
I	0	0
II	1	0
III	0	1
IV	1	1
V	$(b_{12}(k)+\delta_{PS})/(\delta_{PS}-\delta_{TR})$	0
VI	$-(b_{12}(k)-\delta_{PS})/(\delta_{PS}-\delta_{TR})$	1
VII	$1/(\lambda+1)$	$(b_{12}(k)+\delta_{TR}+b_{12}\lambda+\delta_{PS}\lambda)/2(\delta_{TR}+\delta_{PS}\lambda)$

from which it is evident that the top-left element is always less than 0 since $\delta_{PS} < \delta_{TR}$. Similarly, the top-right element is always less than 0. The bottom-left element is greater than 0 if $b_{12}(k) \leq (\delta_{TR}+\delta_{PS}\lambda)/(1+\lambda)$ which is the condition of existence of point VII in the given domain. Therefore, according to the three equation above, point VII is asymptotically stable when it exists.

Part 3. By combining the results obtained in Part 1 and 2, it can be shown that point VII exists and is asymptotically stable when $b_{12}(k) \leq (\delta_{TR}+\delta_{PS}\lambda)/(1+\lambda)$, which leads to $k > \mu_{DB2}$. Similarly, point VI is asymptotically stable when $\delta TR > b_{12}(k) \geq (\delta_{TR}+\delta_{PS}\lambda)/(1+\lambda)$, which leads to $\mu_{DB1} < k < \mu_{DB2}$. Finally, point IV is asymptotically stable when $\delta_{TR} \leq b_{12}(k)$, which leads to $k < \mu_{DB1}$. This concludes the proof. ■

Theorem 3: Consider system (2) under the IM update rule. The stability of this system can be fully characterized by parameters $\delta_{PS}, \delta_{TR}, k$ as in the following.

Case 1: $\delta_{TR} > \delta_{PS}$.

a) When $k < \mu_{IM1}$, the trajectories converge to point IV.
b) When $\mu_{IM2} > k > \mu_{IM1}$, they converge to point VI.
c) When $k > \mu_{IM2}$, they converge to point VII.

Case 2: $\delta_{TR} > \delta_{PS}$.

a) When $k < \mu_{IM1}$, the only stable point is point IV.
b) When $k > \mu_{IM1}$, all fixed points are unstable.

Case 3: $\delta_{TR} = \delta_{PS}$.

a) When $k < \mu_{IM1}$, the only stable point is point IV.
b) When $k > \mu_{IM1}$, the trajectories are closed periodic orbits centred at point VII.

Proof: The proof of Theorem 3 is analogous to the proof of Theorem 2, and it only differs on the calculation of $b_{12}(k)$, resulting in different threshold values. ■

Remark. We can provide a physical interpretation of the above results. In the case of low connectivity, namely, when the node degree k is small, cooperation is the stable strategy in a replete environment. As the connectivity of the network increases above a certain threshold, the value parameters in the payoff matrix determine different behaviors. Specifically, if the sum of the anti-diagonal of $A(n)$ is greater than the sum of the diagonal of $A(n)$, the population dynamics converge to an equilibrium in mixed strategies. It is worth noting that the internal fixed point exists only if $(x^\star, n^\star) \leq [0,1]^2$, i.e., the values of x, n must be between 0 and 1.

IV. MULTI-AGENT REINFORCEMENT LEARNING

In this section, we first introduce a common state-less model called cross learning and show the link between this model and evolutionary dynamics. Then, we extend this model to the networked PD game with environmental feedback. We call our proposed approach networked resource-evolving cross learning (NRE cross learning).

Cross learning, a specific type of finite action-set learning automata, is one of the most basic stateless reinforcement learning algorithms which uses an approach that consists of policy iteration. The basic idea is to start with a random policy to explore the environment and learn from the actions. The policy is then updated based on the reinforcement signal obtained by the environment, which enables the agent to learn the optimal policy and maximize the expected reward. At the start of an epoch t, the agent randomly selects an action $a(t)$ from the set of available actions $\mathcal{A}$. This choice is based on the probability vector $\pi(t)$, which is referred to as strategy. After the action $a(t)$ is selected, the environment provides a reinforcement signal $r(t)$ in the form of a reward. Using this reward $r(t)$, the automaton updates the policy $\pi(t)$ to the new policy $\pi(t+1)$. The update rule is given below.

$$\pi_i(t+1) \leftarrow \pi_i(t) + \begin{cases} \alpha r(t)(1-\pi_i(t)) & \text{if } a(t)=i, \\ -\alpha r(t)\pi_i(t), & \text{otherwise,} \end{cases} \tag{8}$$

where the reward signal is normalized, i.e., $r(t) \in [0,1]$, in order to guarantee policy validity. Parameter α represents the step size of the learning. We now present the following lemma to prove the equivalence between evolutionary dynamics and cross learning in the two-player case, where a population represents a player and the corrsponding probability of choosing one of the two actions [20].

Lemma 2: Consider a cross learning model in a PD game between two players. If $\alpha \to 0$, then the trajectory of cross learning converges to the trajectory of the PD game resulting from the replicator dynamics.

Proof: Consider the replicator equation, given two players x and y, and the corresponding payoff matrices A and B. The change in strategy of player x over time can then be written as:

$$\begin{aligned} \frac{dx_i}{dt} &= x_i[\textstyle\sum_j a_{ij}y_j - \sum_i x_i \sum_j a_{ij}y_j] \\ &= x_i[(Ay)_i - x^\top Ay]. \end{aligned} \tag{9}$$

Similarly, the change of player y over time can then be written as:

$$\frac{dy_i}{dt} = y_i[(Bx)_i - y^\top Bx]. \tag{10}$$

Then, we consider a learning system using the same matrices A and B, and let $\pi := [\pi_1, \pi_2]^\top$ and $\sigma := [\sigma_1, \sigma_2]^\top$ denote the policies of two agents. We now calculate the expected change in the policy of player 1 as:

$$\begin{aligned}
\mathbb{E}(\Delta\pi_i(t)) &= \pi_i(t+1) - \pi_i(t) \\
&= \pi_i(t)\alpha[E(r_i(t))(1-\pi_i(t)) - \textstyle\sum_{j\neq i}(r_j(t)\pi_j(t))] \\
&= \pi_i(t)\alpha[E(r_i(t)) - \textstyle\sum(r_j(t)\pi_j(t))] \\
&= \pi_i(t)\alpha[(A\sigma)_i - \textstyle\sum((A\sigma)_j\pi_j)] \\
&= \pi_i(t)\alpha[(A\sigma)_i - \pi^\top A\sigma]. \qquad (11)
\end{aligned}$$

Similarly, the expected change in the policy of the other player can be derived as:

$$\mathbb{E}(\Delta\sigma_i(t)) = \sigma_i(t)\alpha[(B\sigma)_i - \sigma^\top B\pi]. \tag{12}$$

If we treat the distribution of actions in the policy and the population distribution equally, it can be seen that equations (9)-(10) are equivalent to equations (11)-(12) but scaled by α. This concludes the proof. ■

To model the environmental feedback, we add a resource factor n to adjust the reward $r(i)$ of the environment when action i is taken. Meanwhile, every time the environment receives a stimulus from the agent's action, this resource factor changes as a result. To ensure that the agent can learn the best action in a competitive setting, the agent plays the game with a mirror copy of itself, which imitates the current policy of the agent in each game they play. The new environment is given below:

$$R(t) = [a_{ij}(t) + b_{ij}(t,k)], \tag{13}$$

$$n(t+1) \leftarrow n(t) + n(t)(1-n(t))[(1+\lambda)a(t) - 1], \tag{14}$$

where $R(t)$ represents the reward matrix. We denote the actions of each player by $a(t)$, using 1 and 0 for cooperation and defection, respectively.

Starting with the result from Lemma 2, we can calculate the expected change of the policy in the new environment:

$$\mathbb{E}(\Delta\pi_i(t)) = \pi_i(t)\alpha[(R(t)\sigma)_i - \pi^\top R(t)\sigma]. \tag{15}$$

In view of the identical policy adherence of agent 2 to that of agent 1 at each step, upon incorporating the equation (13), we can deduce the following:

$$\mathbb{E}(\Delta\pi_1(t)) = \pi_1(t)\alpha[(A\sigma)_1 - \pi^\top A\sigma + b_{12}(k)], \tag{16}$$

which is same as population dynamics for system (2) scaled by α. Analogously, the expected change of the environment factor n can be calculated as:

$$\mathbb{E}(\Delta n(t)) = n(t)(1-n(t))[(1+\lambda)\pi_1(t) - 1], \tag{17}$$

which is the equivalent to the environmental feedback in the second equation of system (2). The NRE cross learning algorithm is shown in Algorithm 1.

Algorithm 1: NRE Cross Learning

Input: reward matrix A, node degree k, step size α, change rate ϵ, ratio λ, time horizon T, initial (π, n)
Output: updated π', updated n'

1 : **for** step $t = 1 \ldots, T$ **do**
2 : $act1$ = randomly choose action from (0,1) with probability ($1-\pi$,π).
3 : $act2$ = randomly choose action from (0,1) with probability ($1-\pi$,π).
4 : r = **Reward**(A, k, n, $act1$, $act2$) by (13)
5 : π' = **UpdatePolicy**(r, α, π, ϵ) by (8)
6 : n' = **UpdateEnvironment**($act1$, λ) by (14)
7 : **end for**
8 : **return** π', n'

V. Numerical Analysis

In this section, we present two sets of simulations to corroborate the theoretical results. In all three examples, we set $\epsilon = 1$, $\lambda = 2$ and keep them constant.

In the first set of simulations, we consider system (2) under the DB update rule.In the first example, we set $k = 3 < \mu_{DB1} = 3.21$. According to Theorem 2, the only asymptotically stable point is $(x^\star, n^\star) = (1, 1)$, regardless of the values of R, S, T, P. For the sake of completeness, since $\delta_{TR} > \delta_{PS}$ the system exhibits the behavior shown in Fig 1, bottom-right. As it can be seen in Fig. 2, all the trajectories converge to this equilibrium point.

In the second example, we set $k = 4$ and $k = 6$ to show the system dynamics in **Case 1b)** and **Case 1c)** of Theorem 2. In the first case, we have that $\mu_1 < k < \mu_2$, and all trajectories converge to the fixed point VI, mixed populations in a replete environment as shown in Fig 3 (left). In the second case we have $k > \mu_2$, and the only stable equilibrium is the internal fixed point, i.e., point VII. Figure 3 (right) depicts this case.

In the second set of simulations, we turn our attention to the proposed NRE cross learning algorithm. In the third example, we set $k = 3$ and $k = 4$ corresponding to **Case**

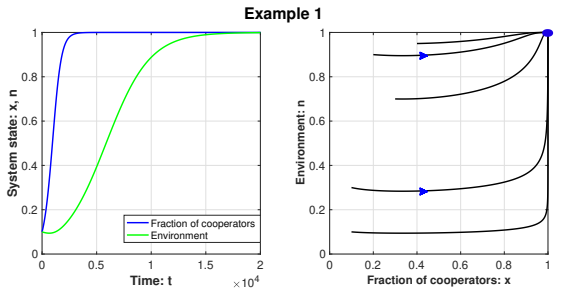

Fig. 2. Example 1. The first plot (left) shows the evolution of cooperators and the environment over time. The second plot (right) shows the phase plane dynamics in the x-n plane for system (2): the blue arrows indicate the direction of the dynamics and the blue circle denotes the stable point.

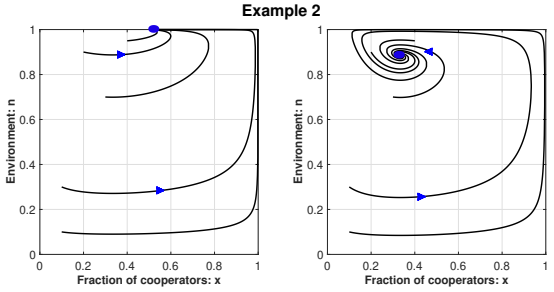

Fig. 3. Example 2. Both plots shows the phase plane dynamics in the x-n plane, when $k = 4$ (left) and $k = 6$ (right).

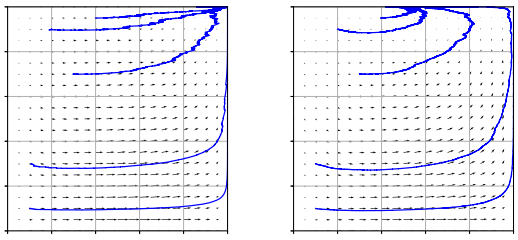

Fig. 4. Example 3. Both plots shows the phase plane dynamics in the x-n plane, when $k = 3$ (left) and $k = 4$ (right). In both cases, $\delta_{TR} > \delta_{PS}$. The arrows show the direction of the policy trajectories.

1a) and **Case 1b)** of Theorem 2. Figure 4 (left) and (right) depict the situation where the policy trajectories converge to the asymptotically stable point $(x^\star, n^\star) = (1, 1)$, and the situation where the policy trajectories converge to point VI, respectively. This shows that the policy trace of our proposed multi-agent reinforcement learning model approximates the corresponding dynamical system.

VI. Conclusion

The scope of this work is to study the impact of structured population in feedback-evolving games via evolutionary game dynamics and multi-agent reinforcement learning. Indeed, we have investigated the impact of players' interactions in the form of a regular network. Motivated by the interest of the control community in multi-agent reinforcement learning, we have developed a stateless model that is able to capture the dynamics of the networked PD game. We envision two paths for future directions of research: i) on the evolutionary game theory path, we will focus on multi-regular graphs and the considerable complexity that this extension brings to the system dynamics; ii) on the machine learning path, we will focus on the extension of our proposed model to consider tabular Q-learning for real-world problems.

References

[1] J. Maynard Smith and G. Price, "The Logic of Animal Conflict," *Nature*, vol. 246, no. 5427, pp. 15–18, 1973.

[2] J. Maynard Smith, *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.

[3] R. Axelrod and W. D. Hamilton, "The Evolution of Cooperation," *Science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[4] G. G. Carter and G. S. Wilkinson, "Food Sharing in Vampire Bats: Reciprocal Help predicts Donations more than Relatedness or Harassment", *Proceedings of the Royal Society B*, vol. 280, 2013.

[5] N. F. Britton, N. R. Franks, S. C. Pratt and T. D. Seeley, "Deciding on a New Home: How Do Honeybees Agree?", *Proceedings of the Royal Society B*, vol. 269, no. 1498, pp. 1383–1388, 2002.

[6] L. Stella and D. Bauso, "Bio-Inspired Evolutionary Game Dynamics in Symmetric and Asymmetric Models," *IEEE Control Systems Letters (L-CSS)*, vol. 2, no. 3, pp. 405–410, 2018.

[7] H. Matsuda, N. Ogita, A. Sasaki and K. Satō, "Statistical Mechanics of Population: the Lattice Lotka-Volterra Model," *Progress of Theoretical Physics*, vol. 88, no. 6, pp. 1035–1049, 1992.

[8] E. Lieberman, C. Hauert and M. A. Nowak, "Evolutionary Dynamics on Graphs," *Nature*, vol. 233, pp. 312–316, 2005.

[9] H. Ohtsuki, C. Hauert, E. Lieberman and M. A. Nowak, "A Simple Rule for the Evolution of Cooperation on Graphs and Social Networks," *Nature*, vol. 441, no. 7092, pp. 502–505.

[10] H. Ohtsuki and M. A. Nowak, "The Replicator Equation on Graphs," *Journal of Theoretical Biology*, vol. 243, no. 1, pp. 86–97.

[11] J. Weitz, C. Eksin, K. Paarporn, S. Brown and W. Ratcliff, "An Oscillating Tragedy of the Commons in Replicator Dynamics with Game-Environment Feedback," *Proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. E7518–E7525, 2016.

[12] A. Tilman, J. Plotkin and E. Akçay, "Evolutionary Games with Environmental Feedbacks," *Nature Communications*, vol. 11, no. 1, 2020.

[13] J. Lee, Y. Iwasa, U. Dieckmann and K. Sigmund, "Social Evolution leads to Persistent Corruption," *Proceedings of the National Academy of Sciences*, vol. 116, no. 27, pp. 13276–13281, 2019. Available: 10.1073/pnas.1900078116.

[14] S. Estrela *et al.*, "Environmentally Mediated Social Dilemmas," *Trends in Ecology & Evolution*, vol. 34, no. 1, pp. 6–18, 2019.

[15] W. Baar and D. Bauso, "Environmental Feedback incorporated on a Collective Decision Making Model," *Proceedings of the 21st IFAC World Congress*, 2020, vol. 53, no. 2, pp. 2832–2837.

[16] L. Stella, D. Bauso and P. Colaneri, "Mean-field Game for Collective Decision-making in Honeybees via Switched Systems," *IEEE Transactions on Automatic Control*, 2021.

[17] L. Stella and D. Bauso, "The Impact of Irrational Behaviors in the Optional Prisoner's Dilemma with Game-environment Feedback," *International Journal of Robust and Nonlinear Control*, 2021.

[18] L. Stella, W. Baar and D. Bauso, "Lower Network Degrees Promote Cooperation in the Prisoner's Dilemma With Environmental Feedback," *IEEE Control Systems Letters*, vol. 6, pp. 2725–2730, 2022.

[19] J. G. Cross, "A Stochastic Learning Model of Economic Behavior," *The Quarterly Journal of Economics*, vol. 87, pp. 239–266, 1973.

[20] D. Bloembergen, K. Tuyls, D. Hennes and M. Kaisers, "Evolutionary Dynamics of Multi-agent Learning: a Survey," *Journal of Artificial Intelligence Research*, vol. 53, 2015.

[21] T. Börgers and R. Sarin, "Learning through Reinforcement and Replicator Dynamics," *Journal of Economic Theory*, vol. 77, 1997.

[22] K. Tuyls and R. Westra, "Replicator Dynamics in Discrete and Continuous Strategy Space," *Multi-agent Systems: Simulation and Applications*, pp. 215—240, 2009.

[23] A. Galstyan, "Continuous Strategy Replicator Dynamics for Multi-agent Q-Learning," *Autonomous agents and multi-agent systems*, vol. 26, pp. 37–53, 2013.

[24] P. Vrancx, K. Tuyls, R. Westra and A. Nowé, "Switching Dynamics of Multi-Agent Learning," *International Conference on Autonomous Agents and Multiagent Systems*, 2008, vol. 26, pp. 307–313.

[25] D. Hennes, K. Tuyls and M. Rauterberg, "State-coupled Replicator Dynamics," *International Conference on Autonomous Agents and Multiagent Systems*, 2009, pp. 789–796.

[26] D. Hennes, M. Kaisers and K. Tuyls, "RESQ-learning in Stochastic Games," *Adaptive and Learning Agents Workshop (AAMAS)*, 2010, p. 8.

[27] P. Julien *et al.*, "Mastering the Game of Stratego with Model-free Multiagent Reinforcement Learning," *Science*, vol. 378, pp. 990–996, 2022.