State-Policy Dynamics in Evolutionary Games

Ilaria Brunetti, Yezekael Hayel, Eitan Altman

State Policy Dynamics in Evolutionary Games

Ilaria Brunetti · Yezekael Hayel · Eitan Altman

Abstract Standard Evolutionary Game Theory framework is a useful tool to study large interacting systems and to understand the strategic behavior of individuals in such complex systems. Adding an individual state to model local feature of each player in this context, allows one to study a wider range of problems in various application areas as networking, biology, etc. In this paper, we introduce such an extension of evolutionary game framework and particularly, we focus on the dynamical aspects of this system. Precisely, we study the coupled dynamics of the policies and the individual states inside a population of interacting individuals. We first define a general model by coupling replicator dynamics and continuous-time Markov Decision Processes and we then consider a particular case of a two policies and two states evolutionary game. We first obtain a system of combined dynamics and we show that the rest-points of this system are equilibria profiles of our evolutionary game with individual state dynamics. Second, by assuming two different time scales between states and policies dynamics, we can compute explicitly the equilibria. Then, by transforming our evolutionary game with individual states into a standard evolutionary game, we obtain an equilibrium profile which is equivalent, in terms of occupation measures and expected fitness to the previous one. All our results are illustrated with numerical analysis.

This work has been partially supported by the European Commission within the framework of the CONGAS project FP7-ICT-2011-8-317672, see www.congas-project.eu.

I. Brunetti
INRIA Sophia Antipolis and LIA/CERI, University of Avignon, 84911, Avignon, France
E-mail: ilariabrun@gmail.com

Y. Hayel
LIA/CERI, University of Avignon, 84911, Avignon, France
E-mail: yezekael.hayel@univ-avignon.fr

E. Altman
INRIA Sophia Antipolis and LINCS - Laboratory of Information, Network and Communication Sciences
E-mail: eitan.altman@inria.fr

Keywords Evolutionary Game Theory · Dynamic Processes · Replicator Dynamics · Singular Perturbation

1 Introduction

Evolutionary Game Theory (EGT) has been first introduced by J.Maynard Smith [1] to model the evolution of species in biology. It makes use of Game Theory tools to describe the dynamics of populations sizes as a result of a competition between them, where players are repeatedly and randomly matched through pairwise interactions. While in classical GT players are supposed to be rational individuals which interact and choose their strategies in order to maximize the individual fitness function, in EGT there is no rationality assumption. All players in a population are supposed to use some action (or behavior type) and the utility is interpreted as a Darwinian fitness depending on the behavior of others and thus on the population's profile, i.e. on the frequencies of the strategies in the whole population. Strategies with higher fitness are supposed to spread within the population.

A key notion in EGT is the *Evolutionarily Stable Strategy* (ESS), which is an action such that, if adopted by all the players, is robust against deviations of a (possibly small) fraction of the population. From a biological point of view it can be seen as a generalization of Darwin's idea of survival of the fittest, while from a GT perspective it is a refinement of Nash Equilibrium, which satisfies a stability property. In order to explain how a population reaches a stable situation, one needs to introduce another fundamental concept of EGT, the replicator dynamics, which serves to highlight the role of selection from a dynamic perspective. It is formalized by a system of ordinary differential equations and it establishes that the evolution of the size of the populations depends on the fitness they get in interactions. An action will spread if its fitness is larger than the fitness averaged over the various strategies used in the whole population. The *folk theorem of evolutionary games* allows to establish a strict connection between the stable points of the Replicator Dynamics and Nash Equilibria [2].

Evolutionary game dynamics are important foundations for understanding behavior of individuals' strategies in a population game. One main feature of population dynamics, is the relationship with learning algorithms in games. For example, in [3], the authors analyze a simple reinforcement learning model (the "Cross' model") discussing its relationship to the replicator dynamics (in fact, their learning model converges to the asymmetric, continuous time version of the replicator dynamics) and propose a discussion on why this learning process is suitable for economic agents. In [4], the author examines the convergence of fitness and strategies in Erev and Roth's model of reinforcement learning. This learning mechanism is one of the most simple as each individual needs only his own fitness to update his action. Also for this learning procedure, the author shows that its long-run behavior is governed by equations related to Maynard Smith's version of the replicator dynamic. A learning procedure

describes how each individual adapts his action based on several information like his own fitness, average fitness, historical actions of the others, probabilistic beliefs on the other actions, etc. This question of level of information under learning processes in games is widely studied and deeply analyzed in [5]. Many different learning algorithms/dynamics are proposed in the literature, like Brown-Nash-Von Neumann, Logistic dynamics, etc. All these dynamics can be generalized to the notion of revision protocols which define a general rule (compare and innovate, target and innovate, compare and non-innovate, etc) followed by individuals [6]. Then, depending on the rule, we get one of the well-known dynamics. There are two main observations that motivate our paper's analysis: first, the population dynamics are usually studied for stability purposes of the equilibrium. In this paper, we consider the dynamics from a control perspective. Second, taking into account individual local dynamics, intrinsically related to the global strategies dynamics, to the best of our knowledge, has never been studied from our point of view: mixing Markov Decision Process (MDP) and Evolutionary Dynamics. Note that similar game theoretic frameworks have been proposed in [7] and [8] but they do not study the dynamical aspects of the strategies of the game. In [9], the authors consider an evolutionary stochastic game framework by assuming each player is playing a best-response to the induced stationary population strategy. Then, the authors introduce the replicator dynamics into their context. We show in our paper, that we can reach such game against the stationary population state by considering a two time scale behavior. Considering different velocities in controlling dynamical systems is a common assumption in automatic control application like robotic control in [10]. Moreover, Evolutionary game dynamics has been proposed in [11] in order to introduce novel extremum seeking controllers. The analysis of the dynamical system is based on the singular perturbation method. The authors consider coupled state-action dynamics but their multi-population model differs largely from ours in several points. First, they consider a multi-population model in which all agents of the same population maximize their common cost function. Second, the agent dynamics is based on an extremum seeking dynamics as the cost functions of each population are unknown. Finally, the authors consider a singular perturbation method assuming that the speed of state dynamics is negligible compared to that of the decision dynamics, which is the opposite point of view of our model.

We introduce in this paper a particular population game in which each individual controls a Continuous Time Markov Decision Process (CT-MDP). This mathematical framework consisting in mixing CT-MDP and evolutionary games, yields interesting insights related to the dynamics of this complex system with several decision makers in interaction. Indeed, the replicator dynamics becomes naturally coupled to the Markov dynamics of each individual. In this context, we define the equilibrium population profile and study the properties of the rest-points of these coupled state policy dynamics.

1.1 Motivations and applications

The scenario proposed in this paper finds its first motivation in the study of optimal power control policies in wireless networks [12]. MDP is a suitable mathematical structure to study optimal problems in stochastic environment. When considering interactions between several decision makers, competitive MDP and stochastic games are useful tools. Our framework aims to enlarge this family of models, by considering evolutionary perspective in the game and then creating a link with dynamical systems. In fact, evolutionary games can be interpreted as a dynamical system through the Replicator Dynamics equations. Thus, since previous works only analyzed the evolutionary stability concept, we propose in this paper a study of the dynamical aspect of this framework. The usefulness of the paper is related to the applications that can be studied based on it. In Information and Communications Technology, our framework finds many application domains like social networks, crowd sourcing and Internet of Things (IoT). For example, emerging applications in engineering such as crowd-sourcing and (mis)information propagation involve a large population of heterogeneous users or agents in a complex network who strategically make dynamic decisions. These agents interact with each other in a complex environment, in which each agent makes strategic and dynamic decisions in response to the agents it interacts with. In all these applications, the action set of each agent depends on a local state. For example, in social networks, each agent may decide to add/remove friends/news based on his own current status. His decision impacts his own status dynamics but also the interaction with other agents. In IoT, a sensor has to determine when to upload his information to the fusion center. This decision impacts his battery level but also the communication quality as collisions may occur for example. As pointed out in several references cited above, the Replicator Dynamics equations are related to several learning algorithms that can be implemented in such sensors or actuators in IoT. Then, by studying these equations, we can understand the convergence behaviour of decentralized algorithms that can be used in such applications. Finally, we would like to mention that our framework is totally coherent with the ideas developed in [13], quoting: *From an engineering point of view, one of the main benefits of multi-agent learning (highly linked to the Evolutionary dynamics like the RD) is its potential applicability as a design methodology for distributed control, which is a branch of control theory that deals with design and analysis of multiple controllers that operate together to satisfy certain design requirements.*

1.2 Contributions of the paper

We first introduce a general framework, in which we associate a state to each player, and we suppose that this state determines the set of available actions. We consider deterministic stationary policies and we suppose that the choice of a policy determines the fitness of the player and it impacts the

evolution of the state. We define the interdependent dynamics of states and policies and we introduce the State Policy coupled Dynamics (SPcD). We then analyze a particular simple case in order to solve the system and we establish the relation between the equilibria of our system of differential equations and the equilibria of the game. We assume that the processes of states and policies move with different velocities: this assumption allows us to solve the system and then to find the equilibria of our game with two different methods: the singular perturbation method and a matrix approach. The main contributions of this paper are listed below.

- We propose a rigorous model of population policies dynamics that takes into account the individual state dynamics.
- We give some general results about the convergence of the coupled dynamical system to an equilibrium of the population game.
- In a particular setting of population game, we give a deep analysis of the convergence of the coupled dynamics to the equilibrium and we propose two different approaches of the problem.

The paper is organized as follows: in Section 2 we briefly present standard EGT main definitions and results. In Section 3 we introduce our evolutionary game framework that takes into account an individual state dynamics coupled to the policies ones. A complete characterization of the coupled dynamical system is performed in section 4, in the particular case of a two states and two actions game. An analysis of this complete setting is done considering singular perturbations techniques in subsection 5.1, and another method, based on rewriting the problem as an equivalent matrix game is described and solved in subsection 5.2. Finally, the solutions obtained are compared in subsection 5.3 and some applications in network systems are proposed in section 6. Then, we give in subsection 6.3 a numerical example of our problem, illustrating the singular perturbation solution and we conclude the paper in section 7 by proposing some perspectives of our framework.

2 Standard Evolutionary Game Theory

2.1 Evolutionarily Stable Strategy

Consider an infinitely large population of players, where individuals are repeatedly matched at random to play a symmetric normal form game, i.e. a two players game in which players dispose of the same set of actions and they have the same fitness function. Let $\mathcal{A} := \{1, \dots, K\}$ be the finite set of actions and let $\Delta = \{p \in \mathbb{R}_+^K | \sum_{i \in \mathcal{A}} p_i = 1\}$ be the set of strategies, that are probability measures over the action space. Note that an action $k \in \mathcal{A}$ can be represented through the unit vector $e_k = (0, \dots, 0_{k-1}, 1, 0_{k+1}, \dots, 0) \in \Delta$, $k = 1, \dots, K$. We define by $F(p, q) := \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}} p_i q_j J(i, j)$ the expected fitness of an individual choosing strategy p against an opponent choosing strategy q, with $p, q \in \Delta$. For all actions $i, j \in \mathcal{A}$, $J(i, j)$ is the immediate fitness

obtained by individual playing action i against an individual playing action j. For symmetric games, Nash equilibrium is defined as follows.

Definition 1 Strategy $q \in \Delta$ is a Nash equilibrium if:

$$F(q,q) \geq F(p,q), \quad \forall p \in \Delta.$$

If, for $p \neq q$ it holds with strict inequality, q is called a strict Nash equilibrium.

In EGT, the main concept of equilibrium is Evolutionarily Stable Strategy (ESS), which is a strategy that, if adopted by the whole population, is resistant against mutations of a small fraction of individuals (*mutants*) in the population.

Definition 2 Strategy $q \in \Delta$ is an ESS if $\forall p \in \Delta,\ p \neq q$, there exists some $\epsilon_p > 0$ such that:

$$\forall \epsilon \in (0, \epsilon_p) \qquad F(q, \epsilon p + (1-\epsilon)q) > F(p, \epsilon p + (1-\epsilon)q). \tag{1}$$

The following proposition allows to characterize an ESS through its stability properties.

Proposition 1 ([14] Proposition 2.1) *Strategy $q \in \Delta$ is an ESS if and only if it satisfies the following conditions:*

- *Nash Condition:* $F(q,q) \geq F(p,q) \qquad \forall p \in \Delta$,
- *Stability Condition:* $F(q,q) = F(p,q) \Rightarrow F(q,p) > F(p,p) \qquad \forall p \neq q$.

It immediately follows that any strict Nash equilibrium is an ESS, while the converse is not true. Finally, there is the following necessary and sufficient condition for a strategy q to be an ESS.

Proposition 2 ([15] Theorem 6.4.1) *The strategy $q \in \Delta$ is an ESS if and only if:*

$$F(q,p) > F(p,p),$$

for all $p \neq q$ in some neighborhood of q in Δ.

In population games [16], the notion of evolutionary stability is associated to the *population state*, defined by a vector $\mathbf{x} = (x_1, \ldots, x_K)$, where x_i is the fraction of individuals in the population playing action $i \in \mathcal{A}$ (i.e. choosing strategy e_i). Note that, as $\sum_{i=1}^{K} x_i = 1$, then $x \in \Delta$ and it is formally equivalent to a mixed strategy. If $q \in \Delta$ and $x \in \Delta$, the expected fitness $F(q,x)$ of a population game is thus thought of as the average expected fitness of a group of individuals such that a fraction q_i of the group uses pure action $i \in \mathcal{A}$, against a population in state x. When the fitness function F is linear in the population state, then evolutionarily stable state and ESS coincide: in the following, we will make no distinction between these two interpretations.

2.2 Replicator Dynamics

We suppose that players play actions in $\mathcal{A}$. The population profile at time t is given by a vector $x(t) \in \Delta$. The replicator dynamics, first defined in [16], describes how the distribution of pure actions evolves in time depending on interactions between individuals. Replicator dynamics of action $i \in \mathcal{A}$ is expressed by the following equation:

$$\dot{x}_i(t) = x_i(t)(F_i(x(t)) - \bar{F}(x(t))), \tag{2}$$

where by $F_i(x(t)) := F(e_i, x(t))$ we denote the immediate expected fitness of an individual playing action $i \in \mathcal{A}$ in a population in state $x(t)$ and $\bar{F}(x(t)) = \sum_{i=1}^{K} x_i(t)F_i(x(t))$ is the average immediate expected fitness of the population.

The replicator equation has numerous properties and there is a close relationship between its rest points, which are the equilibrium points of the ODE (2), and the equilibria of a game. The *folk theorem of evolutionary game theory* [17] states that:

1. any Nash equilibrium profile is a rest point of the replicator equation;
2. if a Nash equilibrium profile is strict then it's asymptotically stable;
3. if a rest point is the limit of an interior orbit for $t \to \infty$, then it is a Nash equilibrium profile;
4. any stable rest point of the replicator dynamics is a Nash equilibrium profile,

where a rest point x^* is stable if, for every neighborhood U_{x^*} of x^* there exists a neighborhood V_{x^*} of x^* such that $x(0) \in V_{x^*}$ implies $x(t) \in U_{x^*}$, $\forall t \geq 0$; an orbit is interior if it is such that $x(t) \in int\Delta := \{\mathbf{x} \in \Delta | x_i > 0, \forall i = 1, \ldots, K\}$; x^* is said to be attracting if it has a neighborhood U_{x^*} such that $x(t) \to x^*$ for $t \to \infty$ holds for $\forall x \in U_{x^*}$ and it is asymptotically stable if it is both stable and attracting. Any ESS is an asymptotically stable rest point and an interior ESS is globally stable, but the converse does not hold in general. In [15], it is proven that a state $\mathbf{x}^*$ is evolutionarily stable if and only if the function $\Gamma(\mathbf{x}) := \prod_i x_i^{x_i^*}$, with $\mathbf{x} \in \Delta$ is a strict local Lyapunov function for the replicator dynamics.

3 Individual States and Policies in Evolutionary Games: a general model

3.1 The individual state dynamical model

We consider a population game in which each individual is controlled by a CT-MDP [18]. In this type of controlled Markov process, the action of the decision maker determines the transition rate of the system. Let $\mathcal{S}$ be the finite state space of each player, with $|\mathcal{S}| = N$ and $\mathcal{A}$ the finite set of actions, with $|\mathcal{A}| = K$. Let us first describe the state dynamics and the Markov process associated to each individual.

Let $\mathcal{R}_s(s', a)$ be the transition rate from state s' to state s given action a, which satisfies $\mathcal{R}_s(s', a) \geq 0$ for all $s' \in \mathcal{S}$, $s' \neq s$, and $a \in \mathcal{A}$. These transition rates are *conservative*, i.e.

$$\sum_s \mathcal{R}_s(s', a) = 0, \quad \forall s' \in \mathcal{S}, \forall a \in \mathcal{A}.$$

Also the transition rates are *stable*, i.e.

$$\sup_{a \in \mathcal{A}} \mathcal{R}_{s'}(a) < \infty, \quad \forall s' \in \mathcal{S}.$$

with $\mathcal{R}_{s'}(a) := -\mathcal{R}_{s'}(s', a) \geq 0$. In our context, the set of actions is finite, which guarantees the stability of the transition rates. Note that the transition rate of an individual depends only on the action of the individual, and not of the others.

Remark 1 In order to establish the relationship between the discrete time MDP and the CT-MDP, it is possible to consider the discrete time MDP embedded at transition epochs of each event, as proposed in [18]. But, as we focus here on the continuous time replicator dynamics, we suppose that each individual controls his transition rates.

We define a randomized Markov policy as a mapping between state space, action space and time into the unit interval. For each state $s' \in \mathcal{S}$ and time t, a randomized Markov policy $u_t(a|s')$ determines the probability to play action a in state s' at time t. Only one action is used at each time, depending on the time variable t and the current state s'. Then, for a given state s' and time t, a randomized Markov policy $u_t(.|s')$ determines a probability distribution over $\mathcal{A}$. We denote by $\mathcal{U}$ the set of all randomized Markov policies. A randomized Markov policy $u_t(.|s')$ is said to be randomized stationary if $u_t(.|s') = u(.|s')$ for each state s' and time t. We denote by $\mathcal{U}_S$ the set of all randomized stationary policies. If a randomized stationary policy $u(.|s')$ determines, for all state s', a specific Dirac measure over the action set $\mathcal{A}$, the policy is said to be deterministic stationary. Finally, we denote by $\mathcal{U}_D$ the set of all deterministic stationary policies. In the following, a randomized Markov policy, a randomized stationary policy and a deterministic stationary policy are simply referred to respectively as a Markov policy, a stationary policy and a deterministic policy. By definition, we have the following relationship: $\mathcal{U}_D \subset \mathcal{U}_S \subset \mathcal{U}$.

In standard evolutionary games, each individual plays a strategy, whereas in our framework, individuals choose a deterministic policy in the finite set $\mathcal{U}_D = \{u_1, \ldots, u_D\}$. The choice of a policy determines the action played in each state and also the time spent by each individual in each state. Indeed, for any state s' and action a, the sojourn time in state s' is a random variable which follows an exponential distribution with parameter $\mathcal{R}_{s'}(a) = \sum_{s \neq s'} \mathcal{R}_s(s', a)$. Then, under a given deterministic policy $u_l \in U_D$, as there is a unique action $a = u_j(s')$ associated to each state s', i.e. $u_j(a|s') = 1$, the time spent in any

state s' for any individual choosing this policy, follows an exponential distribution with parameter $\mathcal{R}_{s'}(a) = \mathcal{R}_{s'}(u_j(s'))$. Then, the fraction of individuals in a given state depends on the distribution of policies over the population. This fraction allows to define the reward obtained for each individual at each pairwise interaction. For any deterministic policy $u_j \in \mathcal{U}_D$ and any state $s' \in \mathcal{S}$, the average time an individual playing policy u_j spends in state s' is given by $1/\mathcal{R}_{s'}(u_j(s'))$. Note that general sojourn time distribution can be also considered, but we keep in the rest of the paper the exponential assumption in order to obtain closed-form solutions of the equilibrium Markovian policy of our game.

Assumption 1 *Under any deterministic policy, the stochastic process of the individual states forms an ergodic continuous time Markov chain.*

3.2 Population of decision makers

Let us consider a fixed population of M decision makers, where each one has his own CT-MDP as described before, with $\mathcal{S}$ and $\mathcal{A}$ respectively the finite set of states and the finite set of actions. Let Δ be the set of distributions over $\mathcal{A}$. We define the fraction of individuals (decision makers) that are in state $s \in \mathcal{S}$ at time t as $\mathbf{w}_s^M(t) := \frac{1}{M}\sum_{l=1}^{M} \mathbb{1}_{\{s^l(t)=s\}}$, where $\mathbb{1}$ is the indicator function, i.e. $\mathbb{1}_{\{s^l(t)=s\}} = 1$ if the state of player l at time t is s, and it equals zero otherwise. For each state s, we denote by $Y_s^l(t)$ the probability that a given individual is in state s at time t under a deterministic policy $u \in \mathcal{U}_D$, i.e. $Y_s^l(t) = \mathbb{P}_{s_0}^u(s^l(t) = s)$, where s_0 is the initial state. Then, from Assumption 1, $\mathbb{P}_{s_0}^u$ doesn't depend on the initial state and, when the size of the population grows to infinity, from the law of large numbers, $Y_s^l(t)$ can be approximated by the fraction of individuals in state s at time t, $\mathbf{w}_s(t) = \lim_{M\to\infty} \mathbf{w}_s^M(t)$. The individual state dynamics, thus corresponds to the dynamics of the fraction of individuals in state s in the global population. We further suppose that the individual dynamics also depends on the policies and that, for any state s, there exists a Lipschitz function h_s which describes the individual state dynamics as follows:

$$\dot{w}_s(t) = h_s(\mathbf{w}(t), \mathbf{q}(t)) \qquad \forall s \in \mathcal{S}, \tag{3}$$

where $\mathbf{w}(t) = (w_1(t) \ldots, w_N(t))$ is the vector of state distribution, i.e. $w_i(t)$ is the fraction of individuals in state i at time t, and $\mathbf{q}(t) = (q_1(t) \ldots, q_D(t))$ is a distribution over the deterministic policies in the population, i.e. $q_j(t)$ corresponds to the fraction of individuals playing a deterministic policy $u_j \in \mathcal{U}_D$ at time t, where $|\mathcal{U}_D| = D$.

3.3 State Policy Coupled Dynamics

Based on imitation/learning evolutionary process, we assume that the fraction of individuals choosing each deterministic policy is evolving over time as a

dynamical process. Without specifying any revision protocol, we define the dynamics of deterministic policies through a set of Lipschitz continuous functions $G := \{g_1, \dots, g_D\}$, such that:

$$\dot{q}_j(t) = g_j(\mathbf{w}(t), \mathbf{q}(t)) \qquad \forall u_j \in \mathcal{U}_D. \tag{4}$$

Then, the dynamical evolution of states and policies fractions inside the population is represented by the following system of $N + D$ differential equations:

$$(S) : \begin{cases} \dot{w}_{s_1} = h_{s_1}(\mathbf{w}(t), \mathbf{q}(t)) \\ \vdots \\ \dot{w}_{s_N}(t) = h_{s_N}(\mathbf{w}(t), \mathbf{w}(t)) \\ \dot{q}_1(t) = g_1(\mathbf{w}(t), \mathbf{q}(t)) \\ \vdots \\ \dot{q}_D(t) = g_D(\mathbf{w}(t), \mathbf{q}(t)) \end{cases} \tag{5}$$

We refer to system (5) as the State-Policy coupled Dynamics (SPcD). A rest point of the SPcD is a state-policy distribution pair $(\mathbf{w}^*, \mathbf{q}^*)$ satisfying:

$$\begin{cases} h_{s_1}(\mathbf{w}^*, \mathbf{q}^*) = 0 \\ \vdots \\ h_{s_N}(\mathbf{w}^*, \mathbf{q}^*) = 0 \\ g_1(\mathbf{w}^*, \mathbf{q}^*) = 0 \\ \vdots \\ g_D(\mathbf{w}^*, \mathbf{q}^*) = 0 \end{cases}$$

In standard evolutionary game theory, an equilibrium is related to the rest points of the replicator dynamics (see the folk theorem of evolutionary game theory stated in previous section). In order to investigate this relation in our setting, we define the fitness function and the equilibria of our game. We denote by $F_i(\mathbf{w}, \mathbf{q})$ the immediate expected fitness of an individual choosing policy $u_i \in \mathcal{U}_D$ in a population whose state is defined by $(\mathbf{w}, \mathbf{q})$, and by $\bar{F}(\mathbf{w}, \mathbf{q})$ the average fitness in such population. Functions F and $\bar{F}$ are assumed to be continuous and linear in $\mathbf{q}$.

Definition 3 A state-policy distribution pair $(\mathbf{w}, \mathbf{q})$ is an equilibrium for the state-policy evolutionary game if $\forall u_i \in \mathcal{U}_D$ we have that:

$$F_i(\mathbf{w}, \mathbf{q}) \geq F_j(\mathbf{w}, \mathbf{q}) \qquad \forall u_j \neq u_i,\ u_j \in \mathcal{U}_D.$$

Note that, as in standard game theory, an equilibrium pair $(\mathbf{w}, \mathbf{q})$ satisfies the indifference principle, i.e. $F_i(\mathbf{w}, \mathbf{q}) = F_j(\mathbf{w}, \mathbf{q})$, for policies that are used by individuals in the population. Given a distribution vector of policies used, we denote by $supp(\mathbf{q}) := \{u_i \in \mathcal{U}_D | q_i > 0\}$. We now specify the evolution of the share of individuals $q_i(t)$ using deterministic policy $u_i \in \mathcal{U}_D$ at time t by

introducing the policy based replicator dynamics (PbRD). The PbRD is given by the following equation:

$$\dot{q}_i = g_i(\mathbf{w}(t), \mathbf{q}(t)) := q_i(F_i(\mathbf{w}, \mathbf{q}) - \bar{F}(\mathbf{w}, \mathbf{q})), \quad \forall u_i \in \{u_1, \ldots, u_D\}. \tag{6}$$

We observe that any unused policy u_i remains unused for ever as $\dot{q}_i = 0$. This property is specific of the replicator dynamics which are based on imitations and not on mutations [16]. Considering the PbRD equation, we obtain the following relationships between rest points of (5) and the equilibria of the state-policy game.

Proposition 3 *A rest point which is the limit of an interior orbit of the SPcD (S) is an equilibrium profile of the state-policy evolutionary game.*

Proof Trivially, if $(\mathbf{w}^*, \mathbf{q}^*)$ is internal rest point of the SPcD determined by system (S), then $F_i(\mathbf{w}, \mathbf{q}^*) = \bar{F}(\mathbf{w}, \mathbf{q}^*)\ \forall u_i$, and thus the couple $(\mathbf{w}^*, \mathbf{q}^*)$ satisfies the indifference principle, i.e. $\forall u_i, u_j \in \mathcal{U}_\mathcal{D}$, $F_i(\mathbf{w}^*, \mathbf{q}^*) = F_j(\mathbf{w}^*, \mathbf{q}^*) = \bar{F}(\mathbf{w}^*, \mathbf{q}^*)$, so $\xi^* = (\mathbf{w}^*, \mathbf{q}^*)$ is an equilibrium profile profile of the state-policy evolutionary game.

Remark 2 Note that the converse does not necessarily hold. Any equilibrium policy distribution $\mathbf{q}^*$ is a rest point of the PbRD, but the corresponding $\mathbf{w}^*$ is not necessarily a rest point of the individual state dynamics.

Proposition 4 *Any stable rest point of the SPcD (S) is an equilibrium profile of the state-policy game.*

Proof Suppose that $(\mathbf{w}^*, \mathbf{q}^*)$ is a stable rest point but not an equilibrium. There exists a policy u_i used in the population, i.e. $u_i \in supp(\mathbf{q}^*)$, such that $F_i(\mathbf{w}^*, \mathbf{q}^*) > \bar{F}(\mathbf{w}^*, \mathbf{q}^*)$, and, from the continuity of the fitness function, there exists a neighborhood U of $(\mathbf{w}^*, \mathbf{q}^*)$ such that $\forall (\mathbf{w}, \mathbf{q}) \in U$, with $(\mathbf{w}, \mathbf{q}) \neq (\mathbf{w}^*, \mathbf{q}^*)$, $F_i(\mathbf{w}, \mathbf{q}) > \bar{F}(\mathbf{w}^*, \mathbf{q}^*)$. This implies that, for this state-policy pair, the component q_i increases exponentially in the PbRD, which contradicts the stability of $(\mathbf{w}^*, \mathbf{q}^*)$. This completes the proof.

3.3.1 Two time scales behavior

We assume here that the states and the policies dynamics move with different velocities. The individual state dynamics are supposed to move very fast compared to the slow updating policies processes. This two time scale assumption, allows us to consider the singular perturbation method [19] to find the rest points of the system (5).We introduce the parameter $\epsilon > 0$ and we rewrite the states and policies system as follows:

$$\begin{cases} \epsilon \dot{w}_{s_1}(t) = h_{s_1}(\mathbf{w}(t), \mathbf{q}(t)) \\ \vdots \\ \epsilon \dot{w}_{s_N}(t) = h_{s_N}(\mathbf{w}(t), \mathbf{q}(t)) \\ \dot{q}_1(t) = g_1(\mathbf{p}(t), \mathbf{q}(t)) \\ \vdots \\ \dot{q}_D(t) = g_D(\mathbf{p}(t), \mathbf{q}(t)) \end{cases} \tag{7}$$

The parameter ϵ is a small positive scalar which serves to represent the different timescales of the two processes, where the velocity of the state process, $\dot{w}_i = h_i(\mathbf{w}, \mathbf{q})/\epsilon$, is fast when ϵ is small. When $\epsilon \to 0$ the states dynamics may rapidly converge to its steady-state and by the singular perturbation theory, under certain assumption, one can solve the reduced model and find a good approximation of the solution of the original system (5).

3.4 Continuous Time Markov Decision Evolutionary Game

Another possible technique to solve the problem consists in assuming that the distribution of the states is stationary and then defining a matrix game where a player chooses a policy in the set $\mathcal{U}_D$ instead of an action. By following the approach presented in [7], where the authors define a Markov Decision Evolutionary Game (MDEG), we can define here an analogous continuous time MDEG (CT-MDEG) as follows. We first define the fitness matrix, representing the fitness of the raw player, depending on the policies chosen:

$$\mathcal{H} = \begin{array}{c} \\ u_1 \\ \dots \\ u_i \\ \dots \\ u_D \end{array} \begin{array}{c} \begin{array}{ccccc} u_1 & \dots & u_j & \dots & u_D \end{array} \\ \begin{bmatrix} F(u_1,u_1) & \ddots & F(u_1,u_j) & \ddots & F(u_1,u_D) \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ F(u_i,u_1) & \ddots & F(u_i,u_j) & \ddots & F(u_i,u_D) \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ F(u_D,u_1) & \ddots & F(u_D,u_j) & \ddots & F(u_D,u_D) \end{bmatrix} \end{array}, \tag{8}$$

where $F(u_i, u_j)$ is the immediate average fitness of an individual playing pure policy u_i against an individual using u_j, $i, j \in x, y$, when the distribution over the individual states is stationary. The fitness of the column player is given by the transposed matrix $\mathcal{H}^T$. Let $\mathcal{R}_{u_i} = [\mathcal{R}_s(s', u_i(s))]_{s,s'}$ be the $N \times N$ *transition rate matrix* (or *infinitesimal generator*) associated to the deterministic policy $u_i \in \mathcal{U}_D$, and let $\pi(u_i)$ be the eigenvector satisfying:

$$\pi(u_i)\mathcal{R}_{u_i} = 0$$

The ergodicity Assumption 1 assures the existence of $\pi(u_i)$. Let $J(s, a; s', a')$ be the immediate fitness that a player gets when it is in state s and plays action a against an individual in state s' using action a'. We can thus express the immediate average fitness $F(u_i, u_j)$ as:

$$F(u_i, u_j) = \sum_{s,s' \in \mathcal{S}} \pi_s(u_i) J(s, u_i(s); s', u_j(s')) \pi_{s'}(u_j). \tag{9}$$

where $\pi_s(u_i)$ is the time ratio spent in state s under policy u_i.

We define the vector of distributions of the pure policies at time t, when the distributions of the individual states are stationary, $\delta(t) := (\delta_1(t), \dots, \delta_D(t)) \in [0, 1]^D$. Let $F_i(\delta)$ be the immediate expected fitness of an individual choosing

policy $u_i \in \mathcal{U}_D$ in a population whose policies distributions are given by the vector δ, and by $\bar{F}(\delta)$ the average fitness in such population. Thus, δ is an equilibrium for the CT-MDEG if $\forall u_i \in supp(\delta) = \{u_i \in \mathcal{U}_D | \delta_i > 0\}$, $F_i(\delta) \geq F_j(\delta)$, $\forall u_j \neq u_i, \in supp(\delta)$. The dynamics of the proportion of individuals adopting pure policy $u_i \in \mathcal{U}_D$ (when the distribution of the states is assumed to be stationary), are given by:

$$\dot{\delta}_i(t) = \delta_i(t) \left(F_i(\delta(t)) - \bar{F}(\delta(t)) \right). \tag{10}$$

As we proved in the general case, if the distribution of the individual states is stationary, then any interior and any stable rest point of the replicator dynamics (10) are equilibria of the CT-MDEG.

Corollary 1 *1. If the trajectory of $\delta(t)$ converges to an interior rest point δ^*, then δ^* is an equilibrium distribution for the CT-MDEG.*
2. If the trajectory of $\delta(t)$ converges to a stable rest point δ^*, then δ^* is an equilibrium distribution for the CT-MDEG.

Proof The proofs of the two corollaries follows the same line of the proofs of Proposition 3 and Proposition 4.

1. Trivially, if δ^* is internal rest point, then $F_i(\delta^*) = \bar{F}(\delta^*)$ and thus $\forall u_i, u_j \in \mathcal{U}_{\mathcal{D}}$, $F_i(\delta^*) = F_j(\delta^*) = \bar{F}(\delta^*)$, so δ^* is an equilibrium profile.
2. Suppose that δ^* is a stable rest point but not an equilibrium. There exists a policy u_i used in the population, i.e. $u_i \in supp(\delta^*)$, such that $F_i(\delta^*) > \bar{F}(\delta^*)$, and, from the continuity of the fitness function, there exists a neighborhood U of δ^* such that $\delta \in U$, with $\delta \neq \delta^*$, $F_i(\delta) > \bar{F}(\delta^*)$. This implies that, for this state-policy pair, the component δ_i increases exponentially, which contradicts the stability of δ^*. This completes the proof.

In the next section, we present a complete analysis and characterization of the equilibrium policy in the case of two states and two strategies, considering the coupled dynamical system. We then demonstrate that, if the system under study can be described in a simple manner, a closed-form solution can be obtained. Otherwise, singular perturbation based algorithms can be used to simulate the system given by equations (7).

4 Complete characterization with two states and two strategies

4.1 Individual state and its dynamics

In this section we suppose that each player can be in one of two possible states, $\mathcal{S} = \{1, 0\}$; every individual goes through a cycle that starts at state 1 and moves to states 0 after some random time at a rate that depends on its policy. After some exponentially distributed time it returns to state 1 and so on. At each pairwise interaction, the set of available actions of a player depends on its state: in state 1, $\mathcal{A}_1 = \{x, y\}$, whereas in state 0 an individual can only use y and thus $\mathcal{A}_0 = \{0\}$.

We consider the set of deterministic policies $\mathcal{U}_D := \{u_x, u_y\}$. Let u_x (resp. u_y) be the deterministic policy which consists in always playing action x (resp. y) in state 1. In state 0, an individual always plays y. Each player chooses one deterministic policy and we denote by $q_x(t)$ the fraction of individuals in the population that play the deterministic policy u_x at time t. The policy chosen impacts the fitness of the player interacting with another individual and also the time he spends in state 1. We define by μ_i the rate of decay from state 1 to state 0 when using policy u_i, $i \in \{x, y\}$, where $\mu_x > \mu_y$, and by μ the rate of change from state 0 to state 1.

As stated in Section 3.2, since the population considered is large, from the law of large numbers, the individual state dynamics can be approximated by the population state dynamics. Let $w_1(t)$ denote the probability that any individual is in state 1 at time t. We define the dynamics of $w_1(t)$ as follows:

$$\dot{w}_1(t) = -\mu_x w_1(t) q_x(t) - \mu_y w_1(t)(1 - q_x(t)) + \mu(1 - w_1(t)). \tag{11}$$

The first (resp. the second) term on the right side of the equation indicates that if an individual is in state 1 and chooses policy u_x (resp. u_y) with probability q_x (resp. $1 - q_x$), he leaves state 1 at a rate μ_x (resp. μ_y). The last term indicates that at a rate μ an individual in state 0 goes to state 1.

4.2 Individual fitness

At each pairwise interaction, the immediate fitness obtained by an individual depends on his current action and the current action of its opponent, as represented by the following fitness matrix:

$$A := \begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{array}, \tag{12}$$

where x and y are the available actions and the matrix entry A_{ij} indicates the fitness respectively of the first (row) player, and the fitness on the second player is given by the transposed matrix A^T. The expected fitness of a player interacting at time t, depends on the population profile at time t, which is now expressed by the couple $\xi(t) := (w_1(t), q_x(t)) \in [0, 1] \times [0, 1]$.

We denote by $J_x(\xi(t))$ (resp. $J_y(\xi(t))$) the expected fitness of an individual playing action x (resp. y) against a population whose profile is $\xi(t)$. By considering payoff matrix (12), we obtain the following expressions:

$$J_x(\xi(t)) := w_1(t)(q_x(t)a + (1 - q_x(t))b) + (1 - w_1(t))b,$$

$$J_y(\xi(t)) := w_1(t)(q_x(t)c + (1 - q_x(t))d) + (1 - w_1(t))d.$$

We can now define the expected fitness of an individual choosing deterministic policy $u_i \in \mathcal{U}_D$ at time t, denoted by $F_i(\xi(t))$, $i = x, y$. The expected

fitness $F_i(\xi(t))$ depends on the population profile $\xi(t)$ and on the individual's state and the policy chosen, which determines the action played in each state. As we are dealing with a large system, from the law of large numbers, we can assume that the probability that any individual is in state 1 at time t is given by $w_1(t)$, as explained in Section 3.2. Then, an individual choosing policy u_x, will be in state 1 (resp. 0) at time t with probability $w_1(t)$ (resp. $1 - w_1(t)$), and then he will get an immediate expected fitness $J_x(\xi(t))$ (resp. $J_y(\xi(t))$). Then the average immediate fitness of an individual choosing policy u_x at time t is given by:

$$F_x(\xi(t)) = w_1(t)J_x(\xi(t)) + (1 - w_1(t))J_y(\xi(t)). \tag{13}$$

If an individual chooses u_y, then in both states he plays pure action y, which leads to:

$$F_y(\xi(t)) = w_1(t)J_y(\xi(t)) + (1 - w_1(t))J_y(\xi(t)) = J_y(\xi(t)). \tag{14}$$

The average expected fitness in the whole population, whose profile at time t is $\xi(t) = (w_1(t), q_x(t))$ is then defined as:

$$\bar{F}(\xi(t)) = q_x(t)F_x(\xi(t)) + (1 - q_x(t))F_y(\xi(t)). \tag{15}$$

In this context, an *equilibrium profile* is given by the pair $\xi^* = (w_1^*, q_x^*)$ such that $\forall u_i \in supp(q^*)$, $F_i(\xi^*) \geq F_j(\xi^*)$, $\forall j \neq i$, $i, j \in \{x, y\}$. An equilibrium profile is thus $\xi^* = (w_1^*, q_x^*)$ stable in the sense of robustness against a deviation of the fraction of individuals playing the deterministic policy u_x. In other words, this definition says that no individuals have an interest to change its deterministic policy, considering this population profile.

4.3 Policy Based Replicator Dynamics

The PbRD can be rewritten as:

$$\begin{aligned} \dot{q}_x(t) &= g(w_1(t), q_x(t)) := q_x(t)(F_x(\xi(t)) - \bar{F}(\xi(t))) \\ &= q_x(t)[F_x(\xi(t)) - (q_x(t)F_x(\xi(t)) - (1 - q_x(t))F_y(\xi(t))] \\ &= q_x(t)(1 - q_x(t))(F_x(\xi(t)) - F_y(\xi(t))). \end{aligned} \tag{16}$$

We can investigate the dynamics of actions in this framework, where the fitness is a function of the population profile depending on policies and states. If we pick one random individual in the population at time t, the probability that he plays pure action x, denoted by $q(t)$, is given by the product $q_x(t)w_1(t)$. By carrying out the expression of $\dot{q}_x(t)$, we get the following equation for the

growth rate of the fraction of individuals playing action x in the population at time t:

$$\dot{q}(t) = \dot{q}_x(t)w_1(t) + q_x(t)\dot{w}_1(t) = q_x(t)[w_1(t)(F_x(\xi(t)) - \bar{F}(\xi(t))) + \dot{w}_1(t)],$$
$$= \frac{q(t)}{w_1(t)}[w_1(t)(F_x(\xi(t)) - \bar{F}(\xi(t))) + \dot{w}_1(t)].$$

We thus obtain:

$$\frac{\dot{q}(t)}{q(t)} = (F_x(\xi(t)) - \bar{F}(\xi(t))) + \frac{\dot{w}_1(t)}{w_1(t)}. \tag{17}$$

Equation (17) shows how the evolution of states impacts the dynamics of actions in our context. We observe that the growth rate of action x is increasing in the growth rate of state 1 and a sufficiently high growth rate of state 1 can lead to a growing rate of action x even if policy u_x is non-optimal.

The SPcD system, which combines the dynamics of the individual state and the dynamics of the policies used in the population, simplifies to:

$$(S)\begin{cases} \dot{w}_1 = h(\xi(t)) \\ \dot{q}_x = g(\xi(t)) \end{cases}$$

where $\xi(t) = (w_1(t), q_x(t))$ is the population profile. Functions h and g are continuously differentiable in ξ, (i.e. the partial derivatives $\partial h/\partial w_1$, $\partial h/\partial q_x$, $\partial g/\partial w_1$, $\partial g/\partial q_x$ are continuous), and thus they are locally Lipschitz continuous with respect to ξ in the compact space $[0,1]^2$, which guarantees the existence of a solution of the system (S). Since in Proposition 3 and Proposition 4 we established the relation between the rest points of the SPcD and the equilibria of the state-policy game, in the next section we solve the system (S) by applying the singular perturbation method.

5 Approximation techniques

In this section we show in details the two approximation techniques briefly introduced in the general case in Section 3, in the particular case of two states and two actions.

5.1 Singular perturbations approach

As introduced in Section 3, we assume that the state and the policy processes move with different velocities. We thus introduce a small parameter $\epsilon > 0$, such that:

$$\epsilon\dot{w}_1 := h(w_1, q_x).$$

We then rewrite the system of the two coupled differential equations as follows:

$$(S_\epsilon)\begin{cases} \epsilon\dot{w}_1 = h(w_1, q_x), \\ \dot{q}_x = g(w_1, q_x). \end{cases}$$

Considering this two-time scale version (S_ϵ) of the initial system (S), we can approximate the solution of (S) using the standard Singular Perturbation Model [19].

When $\epsilon \to 0$, we can consider the *quasi-steady-state-model* [19] by first solving in w_1 the transcendental equation $0 = h(w_1, q_x)$ and then rewriting the differential equation $\dot{q}$ as a function of the obtained roots. As the latter equation has a unique real solution $w_1^*(q_x)$, our system is in *normal form* . This allows us to solve the second differential equation called the quasi-steady-state equation:

$$\dot{q}_x = g(w_1^*(q_x), q_x). \tag{18}$$

If the Assumption 3.2 defined in [19] is satisfied, the reduced model is a good approximation of the original system. As defined in this book, this approximation is good in the sense that a strong stability property is verified for the reduced system. Particularly, the trajectory of the approximation system is $O(\epsilon)$ of the initial system, and hence also for the rest points. This assumption on strict negativity of eigenvalues real parts simplifies in our case to the following condition: $\frac{\partial h}{\partial w_1}(w_1, q_x) < 0$. We thus verify that, in our case, $\frac{\partial h}{\partial w_1}(w_1, q_x) < 0$, which guarantees that we can apply the singular perturbation method to solve (S_ϵ).

The two-time scale behavior of $w_1(t)$ and $q_x(t)$ has a geometric interpretation, as trajectories in $\mathbb{R}^2$. If we define the manifold sets $M_\epsilon := \{\varphi \quad \text{s.t.} \quad w_1 = \varphi(q_x, \epsilon) \quad \text{and} \quad \epsilon = h(\varphi(q_x, \epsilon), q_x)\}$, it is possible to rewrite the problem in terms of invariant manifolds. When $\epsilon = 0$, the manifold M_0 corresponds to the expression of the quasi steady state model. When the condition $\frac{\partial h}{\partial w_1}(w_1, q_x) < 0$ is satisfied, we have that the equilibrium manifold M_0 is stable (attractive). In particular, the existence of a conditionally stable manifold M_0 for $\epsilon = 0$ implies the existence of an invariant manifold M_ϵ satisfying the following convergence for all $\epsilon \in [0, \epsilon^*]$:

$$\varphi(\epsilon, q_x) \to \varphi(0, q_x), \quad \text{and} \quad M_\epsilon \to M_0 \quad \text{as} \quad \epsilon \to 0.$$

The positive constant ϵ^* is determined such that the following manifold condition is satisfied:

$$\epsilon \frac{\partial \varphi}{\partial x} g(\varphi(q_x, \epsilon), q_x) = h(\varphi(q_x, \epsilon), q_x),$$

for all q_x and $\epsilon \in [0, \epsilon^*]$. The attractiveness of the slow manifold M_0 is illustrated in the numerical illustrations section. Let us now compute the solution of the approximate system (S_0). We thus suppose that the distribution of the individual states is stationary (expressed by Equation (11)). By imposing $\dot{w}_1 = 0$, we obtain the following slow manifold $M_0 := \{\varphi \quad \text{s.t.} \quad w_1 = \varphi(q_x, 0) \quad \text{and} \quad 0 = h(q_x, \varphi(q_x, 0))\}$:

$$\varphi(q_x, 0) = \frac{\mu}{\mu + \mu_x q_x + \mu_y (1 - q_x)} := \varphi_1(q_x). \tag{19}$$

The PbRE (16) can now be rewritten as:

$$\dot{q}_x(t) = q_x(t)(1 - q_x(t)) \left[F_x(w_1^*(q_x(t)), q_x(t)) - F_y(w_1^*(q_x(t)), q_x(t))\right]. \quad (20)$$

Proposition 5 *For ϵ sufficiently small, the solution of the system (S_ϵ) can be approximated by the solution of S_0, which is given by the population profile $\xi^* = (w_1^*, q_x^*)$, such that:*

$$w_1^* = \frac{\mu - s^*(\mu_x - \mu_y)}{\mu + \mu_y} \quad \textit{and} \quad q_x^* = \frac{s^*(\mu + \mu_y)}{\mu - s^*(\mu_x - \mu_y)}, \quad (21)$$

where s^* is the equilibrium of the standard replicator dynamics (2) when considering payoff matrix (12):

$$s^* = \frac{d - b}{\gamma} \quad \textit{with} \quad \gamma = a - b - c + d.$$

Proof Let us first study the equation $\dot{q}_x = 0$ before substituting the stationary equation of the state dynamics. Solving this equation is equivalent to finding the population profile $\xi = (w_1^*, q_x)$ such that:

$$F_x(w_1^*, q_x) = F_y(w_1^*, q_x).$$

By explicitig the expressions of the fitness, after some manipulations we get the equivalent equation:

$$\begin{aligned} &w_1^* a w_1^* q_x + w_1^* b(1 - w_1^* q_x) + (1 - w_1^*) c w_1^* q_x \\ &+ (1 - w_1^*) d(1 - w_1^* q_x) = c w_1^* q_x + d(1 - w_1^* q_x). \end{aligned} \quad (22)$$

Then, after some algebra, we get:

$$w_1^* q_x = \frac{d - b}{\gamma} := s^*.$$

The stationary condition of the first differential equation (11) leads the following relation between w_1 and q_x:

$$w_1 = w_1^*(q_x) = \frac{\mu}{\mu + \mu_x q_x + \mu_y(1 - q_x)},$$

then we have to solve: $w_1^*(q_x) q_x = s^*$. This last equation is equivalent to:

$$\frac{\mu q_x}{\mu + \mu_x q_x + \mu_y(1 - q_x)} = s^*.$$

After some simple manipulations we obtain:

$$q_x = \frac{s^*(\mu + \mu_y)}{\mu - s^*(\mu_x - \mu_y)} := q_x^*.$$

Finally, as we have that $w_1^*(q_x^*) = \frac{s^*}{q_x^*}$ which leads to: $w_1^* = \frac{\mu - s^*(\mu_x - \mu_y)}{\mu + \mu_y}$.

Note that the rest point q_x^* of the PbRE (16) verifies:

$$q_x^* w_1^*(q_x^*) = s^*.$$

We thus obtain that the probability that any individual picked out randomly in the population, is playing action x at the equilibrium, is equal to s^*. This value is the mixed equilibrium of the standard matrix game given by matrix A. It means that, if we consider a state dependent action (instead of policy) game, the equilibrium is obtained under conditional probability over the state.

We have the following necessary and sufficient condition under which the solution obtained is a strict interior point.

Lemma 1 *The solution q_x^* obtained in proposition (5) is a strict interior point if and only if:*

$$\mu > \mu_x \frac{s^*}{1 - s^*}.$$

Proof The solution obtained in Proposition 5 is:

$$q_x^* = \frac{s^*(\mu + \mu_y)}{\mu - s^*(\mu_x - \mu_y)}.$$

This solution is a strict interior point if and only if:

$$0 < q_x^* < 1.$$

First, let's look at the positivity condition $q_x^* > 0$. This is equivalent to:

$$0 < q_x^* \iff \mu > s^*(\mu_x - \mu_y).$$

After some basic algebra, the second condition becomes:

$$q_x^* < 1 \iff \mu > \mu_x \frac{s^*}{1 - s^*}.$$

We have clearly that for all $s^* \in]0, 1[$, μ_x and μ_y:

$$\frac{s^*}{1 - s^*}\mu_x > s^*\mu_x > s^*(\mu_x - \mu_y).$$

Then if $\mu > \mu_x \frac{s^*}{1-s^*}$ the solution is a strict interior point, and the converse is true. This concludes the proof.

Note that this condition does not depend on the rate μ_y. In the next section, we present an alternative method based on rewriting our game problem into a matrix game considering only pure policies.

5.2 CT-MDEG approach

In alternative to the singular perturbation method, we can define the matrix game presented in Section 3.4. In this two states-two actions game, matrix $\mathcal{H}$ has size 2×2 and we can define the stationary probabilities to be respectively in states 1 and in state 0 when playing action $i \in \{x, y\}$ as follows:

$$\pi_1(i) = \frac{\frac{1}{\mu_i}}{\frac{1}{\mu} + \frac{1}{\mu_i}} = \frac{\mu}{\mu + \mu_i},$$

$$\pi_0(i) = \frac{\frac{1}{\mu}}{\frac{1}{\mu} + \frac{1}{\mu_i}} = \frac{\mu_i}{\mu + \mu_i}.$$

The fitness $F(u_i, u_j)$ can thus be expressed as:

$$F(u_i, u_j) = \sum_{s,s' \in \mathcal{S}} \pi_s(u_i(s)) J(u_i(s), u_j(s')) \pi_{s'}(u_j(s')), \quad u_i, u_j \in \{u_x, u_y\}$$

where $J(u_i(s), u_j(s'))$ is the immediate fitness of a player using action $u_i(s) \in \{x, y\}$ against an opponent playing $u_j(s') \in \{x, y\}$, . If we consider the immediate payoff matrix (12), we obtain:

$$\begin{aligned} &F(u_y, u_y) = d, \\ &F(u_x, u_y) = \pi_1(x)b + \pi_0(x)d, \\ &F(u_y, u_x) = \pi_1(x)c + \pi_0(x)d, \\ &F(u_x, u_x) = \pi_1(x)\left[\pi_1(x)a + \pi_0(x)b\right] + \pi_0(x)\left[\pi_1(x)c + \pi_0(x)d\right]. \end{aligned} \tag{23}$$

We can now define the replicator dynamics of the CT-MDEG:

$$\begin{aligned} \dot{\delta}_x(t) &= \delta_x(t)(1-\delta_x(t))(F(u_x, \delta_x(t)) - F(u_y, \delta_x(t)) \\ &= \delta_x(t)(1-\delta_x(t))\left[F(u_x, u_y) - F(u_y, u_y) + \delta_x(t)(F(u_x, u_x) - F(u_y, u_x)\right. \\ &\quad \left. + F(u_y, u_y) - F(u_x, u_y))\right]. \end{aligned} \tag{24}$$

where $\delta_x(t)$ is the probability that an individual chooses policy u_x at time t. We obtain the following result.

Proposition 6 *If the distribution of the individual states is stationary, the equilibrium policy of the game can be computed by considering the CT-MDEG, which leads to the equilibrium:*

$$\delta_x^* = \frac{s^*}{\pi_1(x)}, \tag{25}$$

under the condition $0 \leq s^* \leq \pi_1(x)$.

Proof The equilibrium of the policy game can be computed by imposing the indifference principle, which leads to:

$$\delta_x^* = \frac{F(u_y,u_y) - F(u_x,u_y)}{F(u_x,u_x) - F(u_y,u_x) + F(u_y,u_y) - F(u_x,u_y)}.$$

By substituting the values of the fitnesses (23) into the latter equation and by carrying out the values of the time ratios $\pi_1(x)$ and $\pi_0(x)$, we get:

$$\begin{aligned}\delta_x^* &= \frac{\frac{\mu(d-b)}{\mu+\mu_x}}{\frac{\mu^2 a+\mu\mu_x b+\mu_x\mu c+\mu_x^2 d}{(\mu+\mu_x)^2} + d - \frac{\mu b+2\mu_x d+\mu c}{(\mu+\mu_x)}} \\ &= \frac{\frac{\mu(d-b)}{\mu+\mu_x}}{\frac{\mu^2 a+\mu\mu_x b+\mu_x\mu c+\mu_x^2 d+d(\mu+\mu_x)^2-(\mu+\mu_x)(\mu b+2\mu_x d+\mu c)}{(\mu+\mu_x)^2}}.\end{aligned}$$

After some algebra:

$$\delta_x^* = \frac{\mu(d-b)}{\mu+\mu_x} \cdot \frac{(\mu+\mu_x)^2}{\mu^2(a+d-b-c)} = \frac{s^*}{\pi_1(x)}.$$

In order for δ_x^* to be an admissible equilibrium, it must satisfy $\delta_x^* \in [0,1]$, which completes the proof.

5.3 Relation between the equilibria

In section 5.1, we adopted the singular perturbation method, which allows to obtain the equilibrium profile $\xi^* = (w_1^*, q^*)$. In section 5.2, we first assumed that the distribution over the individual states is stationary and we then rewrote the game as an evolutionary game where players choose pure policies instead of actions. This approach leads to the equilibrium distribution δ_x^*. We now compare these two equilibria.

Corollary 2 *We have the following relation between the equilibria obtained with the two different approximation techniques of the state-policy game presented:*

$$w_1^* q_x^* = \pi_1(x)\delta_x^*.$$

Proof It straightforwardly follows from the proofs of Proposition 5 and Proposition 6, since we obtain respectively that: $\pi_1(x)\delta_x^* = s^*$ and $w_1^* q_x^* = s^*$.

Remark 3 Note that $\pi_1(x)$ is the conditional stationary probability of an individual to be in state 1 under policy u_x, while w_1^* is the probability that an individual randomly selected in the population whose profile is ξ^* is in state 1 at the equilibrium, which implies that $w_1^* > \pi_1(x)$. We can mathematically prove this inequality as:

$$\begin{aligned}w_1^* - \pi_1(x) &= \frac{\mu - s^*(\mu_x-\mu_y)}{\mu+\mu_y} - \frac{\mu}{\mu+\mu_x} \\ &= \frac{(\mu_x-\mu_y)(\mu - s^*(\mu+\mu_x)}{(\mu+\mu_x)(\mu+\mu_y)}.\end{aligned}$$

The denominator is always positive, $\mu_x > \mu_y$ by definition, and, from the proof of Lemma 1, $(1 - s^*)\mu > s^*\mu_x$, so the numerator is also positive. This implies that the strict inequality $w_1^* > \pi_1(x)$ always holds. We thus have that $q_x^* < \delta_x^*$.

We now compare the two equilibria in terms of average fitness obtained by the population, i.e. $F(\delta_x^*, \delta_x^*)$ and $\bar{F}(\xi^*)$, where $\xi^* = (w_1^*, q_x^*)$.

Proposition 7 *The average fitness in the population at the two equilibria obtained with the two approaches are equal, i.e.*

$$\bar{F}(\xi^*) = F(\delta_x^*, \delta_x^*) = s^*c + (1 - s^*)d.$$

Proof Considering the first approach based on the singular perturbations method, we have:

$$\bar{F}(\xi^*) = q_x^* F_x(\xi^*) + (1 - q_x^*)F_y(\xi^*) = F_y(\xi^*) + q_x^*(F_x(\xi^*) - F_y(\xi^*))$$

At the equilibrium state, we have $F_x(\xi^*) = F_y(\xi^*)$ and thus $\bar{F}(\xi^*) = F_x(\xi^*) = F_y(\xi^*) = J_y(\xi^*)$. Then, we get the average expected fitness of the population at the equilibrium with the first approach is:

$$\bar{F}(\xi^*) = w_1^*(q_x^* c + (1 - q_x^*)d) + (1 - w_1^*)d.$$

Since $q_x^* w_1^* = s^*$, then:

$$\bar{F}(\xi^*) = s^*c + (1 - s^*)d.$$

Considering the second method of rewriting the game into a matrix game, we obtain the following equilibrium profile: $\delta_x^* = \frac{s^*}{\pi_1(x)}$. The average fitness of the population in this case is:

$$F(\delta_x^*, \delta_x^*) = \delta_x^* F(u_x, \delta_x^*) + (1 - \delta_x^*)J(u_y, \delta_x^*) = F(u_y, \delta_x^*) + \delta_x^*(F(u_x, \delta_x^*) - F(u_y, \delta_x^*)).$$

At the ESS, we have the following equality $F(u_x, \delta_x^*) = F(u_y, \delta_x^*)$ and then the average fitness of the population becomes simply:

$$F(\delta_x^*, \delta_x^*) = F(u_y, \delta_x^*) = \delta_x^* F(u_y, u_x) + (1 - \delta_x^*)F(u_y, u_y).$$

Then, the average fitness of the population is: $F(\delta_x^*, \delta_x^*) = \delta_x^*(\pi_1(x)c + \pi_0(x)d) + (1 - \delta_x^*)d$. We have that $\delta_x^* \pi_1(x) = s^*$ which leads to: $F(\delta_x^*, \delta_x^*) = s^*c + (1 - s^*)d$.

Finally, we prove that the two mixed strategies obtained with the two approaches are in the same equivalent class in terms of time ratios in individual states, which means that the time ratio spent in each state is the same for the two equilibria. Since we are considering large populations of players, we can think of q_x and δ_x as the probabilities that a player chooses policy u_x at the equilibrium. Then, in the first case, $w_1(q_x)$ can be interpreted as the time ratio spent in state 1 by an individual choosing u_x with probability q_x*. If we consider the matrix game and we denote by $\bar{\pi}_1(\delta_x)$ the time ratio that an

individual playing deterministic policy u_x with probability δ_x spends in state 1, we that:

$$\bar{\pi}_1(\delta_x) := \delta_x \pi_1(x) + (1-\delta_x)\pi_1(y) = \delta_x \frac{\mu}{\mu+\mu_x} + (1-\delta_x)\frac{\mu}{\mu+\mu_y}. \quad (26)$$

In the following proposition we prove that the time ratios in state 1 at the equilibria, obtained respectively with the singular perturbation technique and with the matrix approximation technique, are equal.

Proposition 8 *The two different equilibria obtained with the two approximation techniques yield to the same time ratio in state* 1*, i.e.*

$$\bar{\pi}_1(\delta_x^*) = w_1(q_x^*).$$

Proof We first rewrite δ_x^* as a function of the immediate payoffs $\{a, b, c, d\}$:

$$\delta_x^* = \frac{(\mu_x+\mu)(d-b)}{\mu\gamma}.$$

where $\gamma := a - b - c + d$. We substitute it in (26), and we get:

$$\bar{\pi}_1(\delta_x^*) = \frac{\mu\gamma + (\mu_x+\mu_y)}{\mu\gamma(\mu+\mu_y)}.$$

Analogously, we substitute the expression of s^* in q_x^* in Proposition 5, and we obtain:

$$w_1(q_x^*) = \frac{\mu\gamma + (\mu_x+\mu_y)}{\mu\gamma(\mu+\mu_y)},$$

which proves that $w_1(q_x^*) = \bar{\pi}_1(\delta_x^*)$.

Remark 4 The previous results show that, when solving the two approximated games, we obtain two different equilibrium distributions of pure policies which yield the same average fitness and the same time ratios.

6 Applications in Network Systems

6.1 Energy Control in Wireless Network

The two-states two-actions model can be applied to describe a particular case of a problem that arises in dynamic power control in mobile networks, which has been presented in [12]. The idea of the problem is to consider that the battery life is a very critical issue in wireless systems, and then, defining optimal transmission policies based on battery levels is very important. Moreover, this energy management problem is even more important when interactions occurs between the devices and then complicate the analysis of such control systems. Then, we consider a system in which the action of each device impacts the lifetime of his battery or its battery level, and also impacts

its transmission rate. A large number of mobiles transmit packets occasionally. Each transmitter can be in Full (F) or Almost empty (A) battery state. When a mobile is in F state it can choose to transmit packets using high (h) or low (l) power, whereas if it is in state A, it can only transmit packets using l power. In general, several mobiles try to join a common receiver at the same time and interferences occur between the received signals. We suppose that transmissions are sparse so that the probability that more than two mobiles transmit simultaneously is negligible. We assume also that a transmission is successful either if the mobile is the only one transmitting during a slot or if it transmits at higher power than all the others. **Therefore, the fitness of the player can be obtained by analyzing the number of successful transmission over the common channel. A closed-form expression of the fitness is described in [12]. A feedback mechanism, like packet acknowledgment can be used for each player to compute its fitness.** The time spent in state F depends on the action chosen by the mobile. Then the state of the mobile changes to the other battery state A. After an exponentially distributed time, its battery state becomes empty. We assume that the battery is immediately recharged, so that the mobile goes back to state F. When transmitting at high power, the mobile's battery is consumed faster, and thus the transition rate from F to A is faster. Then, considering this framework, the state space corresponds to $\mathcal{S} := \{A, F\}$, the action space is $\mathcal{A} := \{h, l\} = \mathcal{A}_F$ and the restricted action space for state A is $\mathcal{A}_A := \{l\}$. The set of deterministic policies $\mathcal{U}_D := \{u_h, u_l\}$ is composed of the policy u_h such that $u_h(A) = l$ and $u_h(F) = h$; and the policy u_l such that $u_l(A) = l$ and $u_l(F) = l$. Then, the system (S) of coupled dynamics describe the time evolution of the fraction of mobiles in each state A and F, and at the same time the fraction of mobiles using policy u_h and u_l. By assuming that the state dynamic is highly faster than the policy dynamic (the change of policy has to be reimplemented into the mobiles by manufacturer or designers), then our analysis describes the equilibrium situation with corresponds to the long term evolution of this system.

6.2 Network Formation Games

Another application of the proposed model can be found in network formation games [20]. We consider a large number of nodes where each node is in one of two possible states: Infected or Susceptible, so that $\mathcal{S} = \{I, S\}$. Nodes interact through pairwise interactions, during which, both nodes exchange contents. If a node is in state S it determines the type of unidirectional link to the node he is interacting with. The type of link can be charged at a price (p) or free (f); if a node is in the infected state (state I), it can only create free links. Pay connection is safer, so that when a link is not a free one, the probability for a node to be infected is lower, independently of the choice of the other node to pay or not and also independent of the state of the other node. After some random time in I state, a node becomes susceptible again.

The fitness of a player, a node in this context, should depend on the evolution of its own state. Then, the fitness is easily computed based on local information which is known by each player. This application into networks formation games could ask more assumptions on the model, especially if the transition rate depends on the state of the opponent. In this case we should define a more general game framework considering interactive MDPs, like anonymous sequential games [21]. This generalized framework has a highly more complicated internal structure. We thus let its analysis as an extension for future works.

It has to be noted that the singular perturbation approach, proposed in section V.A is valid for this application, by considering a more complicated dynamics of individual state, which depends on the action also of the opponent.

6.3 Numerical Illustrations

We illustrate here the theoretical results obtained in previous sections with numerical solutions and simulations. We consider a first numerical example with the following transition rates: $\mu = 10$, $\mu_x = 1.5$ and $\mu_y = 1$. The fitnesses of the matrix game are: $a = -0.3$, $c = 0$, $b = 1$ and $d = 0.5$. Those values yield to the following equilibrium of the standard evolutionary game $s^* = \frac{5}{8} = 0.625$.

We plot on figure 1 the trajectories of the system (S_ϵ) of the coupled differential equations for different initial conditions and for $\epsilon = 0.01$. We simulate a discrete time version of the differential equations. We plot also the invariant manifold M_0 and we observe that it is an attractor of the trajectories.

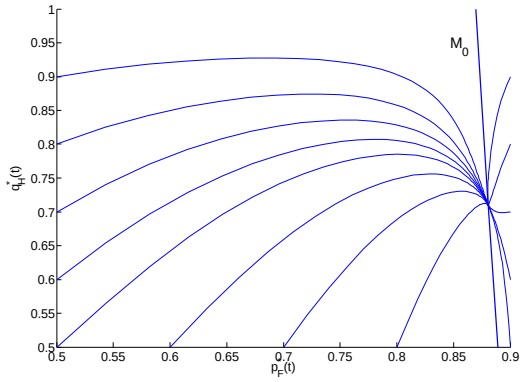

Fig. 1 Trajectories of the system (S_ϵ) from different starting points and the slow manifold M_0 with $\epsilon = 0.01$.

Proposition 5 gives the following solution of the system (S_0), by considering the singular perturbation method based on the steady-state model:

$$q_x^* = 0.7097, \quad \text{and} \quad w_1^* = 0.8807.$$

This couple corresponds exactly to the attractor of the trajectories on figure 1 and then our simulation validates the result of this proposition. The matrix game approach gives the following equilibrium:

$$\delta_x^* = \frac{s^*(\mu + \mu_x)}{\mu} = 0.71875 > q_x^*,$$

which verifies the Proposition 2. The replicator dynamics equation given by equation (24) for this matrix game is depicted on figure 2.

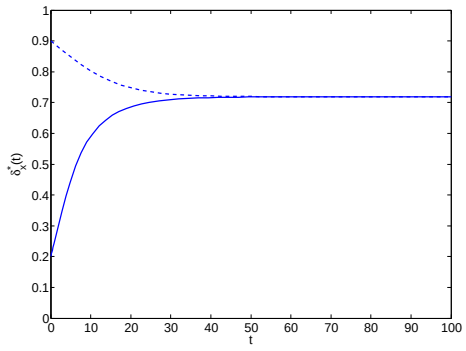

Fig. 2 Convergence of the replicator dynamics equation (24) to the ESS $\delta_x^* = 0.71875$ starting from $q_x(0) = 0.2$ and $q_x(0) = 0.9$.

7 Conclusions and perspectives

In this paper, we considered a particular type of evolutionary game in which the action of an individual not only determines its immediate fitness but it also impacts the transition rates of its embedded Markov process. The aim of this paper is to describe the coupled dynamics of the controlled dynamics between individual states and the dynamics of policies inside the global population of individuals in pairwise interaction. This latter dynamics is assumed to follow the well-known replicator dynamics. Once we introduced these combined dynamics, we proved that any stable rest point corresponds to an equilibrium profile of the evolutionary game. We proposed two methods to obtain the rest points under the assumption that the two dynamics evolve at different velocities. We gave a complete characterization of the equilibrium profiles for a two states two actions setting, and we showed that these equilibrium profiles are comparable in terms of occupation measures and fitness. Finally, we illustrated our framework with two application scenarios in network systems.

8 Notation

- $\mathcal{A}$: finite set of pure actions, with $|\mathcal{A}| = K$;
- $\Delta := \{\mathbf{p} \in \mathbb{R}_+^K | \sum_{i \in \mathcal{A}} p_i = 1\}$: set of strategies;
- $\mathcal{S}$: set of states, with $|\mathcal{S}| = N$;
- $\mathcal{U}$: set of Markov policies;
- $\mathcal{U}_S$: set of stationary policies;
- $\mathcal{U}_D$: set of deterministic policies;
- $\mathcal{R}_s(s', a)$: transition rate from state $s' \in \mathcal{S}$ to state $s \in \mathcal{S}$ given action $a \in \mathcal{A}$;
- $\pi_s(u_i)$: time ratio spent in state s under policy u_i;
- w_s: fraction of individual in the population in state s; $\mathbf{w} = (w_1, \ldots, w_N)$: distribution over all states in the population;
- q_i: fraction of individuals choosing deterministic policy u_i;
- $\xi = (w_1, q_x)$: population profile;
- $J(s, a; s', a')$: immediate fitness that a player get when in state s plays action a in an interaction with an individual in state s' playing a';
- $F(.,.)$: fitness of an individual
- $\bar{F}(.)$: average expected fitness of a population
- $F_i(.)$: expected fitness of an individual choosing deterministic policy $u_i \in \mathcal{U}_D$
- $J_i(.)$: expected fitness of an individual playing action $i \in \mathcal{A}$

References

1. J. M. Smith and G. R. Price, The logic of animal conflict, Nature, vol. 246, pp. 15-18 (1973).
2. J. Hofbauer, P. Schuster, and K. Sigmund, A note on evolutionarily stable strategies and game dynamics, Journal of Theoretical Biology, vol. 81, pp. 609-612 (1979).
3. T. Borgers and R. Sarin, Learning through reinforcement and replicator dynamics, Journal of Economic Theory, vol. 77 (1997).
4. A. Beggs, On the convergence of reinforcement learning, Journal of Economic Theory, vol. 122 (2005).
5. D. Fudenberg and D. Levine, Theory of Learning in Games, MIT press, Cambridge (1998).
6. W. H. Sandholm, Population Games and Evolutionary Dynamics. MIT Press, Cambridge (2010).
7. E. Altman and Y. Hayel, Markov decision evolutionary games, IEEE Transactions on Automatic Control, vol. 55 (2010).
8. B. Jovanovic and R. Rosenthal, Anonymous sequential games, Journal of Mathematical Economics, vol. 17, pp. 77-87 (1988).
9. J. Flesch, T. Parthasarathy, F. Thuijsman, and P. Uyttendaele, Evolutionary stochastic games, Dynamic Games and Applications, vol. 3, pp. 207-219 (2013).
10. P. M. S. Gros, D. Buccieri and D. Bonvin, Two-time-scale control scheme is illustrated via the simulation of a flying robotic structure, in in IFAC Conference on Nonlinear Model Predictive Control (2005).
11. J. Poveda and N. Quijano, Extremum seeking for multi-population games, in 52th IEEE Conference on Decision and Control (2013).
12. E. Altman and Y. Hayel, Stochastic evolutionary game approach to energy management in a distributed aloha network, in IEEE INFOCOM (2008).
13. S. Mannor and J. Shamma, Multi-agent learning for engineers, Artificial Intelligence, vol. 171, pp. 417-422 (2007).

14. J. Weibull, Evolutionary Game Theory. MIT Press, Cambridge (1995).
15. J. Hofbauer and K. Sigmund, Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge UK, (1998).
16. P. Taylor and L. B. Jonker, Evolutionary stable strategies and game dynamics, Mathematical Biosciences, vol. 40, pp. 145-156 (1978).
17. J. Hofbauer and K. Sigmund, Evolutionary game dynamics, American Mathematical Society, vol. 40, pp. 479-519 (2003).
18. X. Guo and O. Hernandez-Lerma, Continuous Time Markov Decision Process. Springer, Berlin (2009).
19. P. Kokotovic, H. Khalil, and J. O'reilly, Singular perturbation methods in control, SIAM (1986).
20. M. Jackson, A Survey of Network Formation Models: Stability and Efficiency. Group Formation in Economics: Networks, Clubs, and Coalitions book, Cambridge University Press, Cambridge UK (2005).
21. P. Wiecek and E. Altman, Stationary anonymous sequential games with undiscounted rewards, Journal of Optimization Theory and Applications (2014).