DOI: http://dx.doi.org/10.7551/978-0-262-31709-2-ch136

Comparing Reinforcement Learning and Evolutionary Based Adaptation in Population Games

Ana L. C. Bazzan

PPGC / UFRGS
Caixa Postal 15064, CEP 91501-970, Porto Alegre, RS, Brazil
bazzan@inf.ufrgs.br

Abstract

In evolutionary game theory, the main interest is normally on the investigation of how the distribution of strategies changes along time and whether an stable strategy arises. In this paper we compare the dynamics of two games in which three populations of agents interact: a three-player version of matching pennies and a game with several Nash equilibria. We do this comparison by three methods: continuous replicator dynamics, an evolutionary approach, and reinforcement learning. We show how the convergence depends on the nature of the underlying method used, as well as on the pace of adjustments by the agents.

Introduction

In game theory (GT), one traditional explanation of equilibrium is that it results from an introspective analysis by the players when the rules of the game, the rationality of players, and the payoff functions are all common knowledge. It is well-known that both conceptually and empirically this argument has many problems. Just to mention one of them, in games with more than one equilibria, even if one assumes that players are able to coordinate their expectations using some selection procedure, it is not clear how such a procedure comes to be common knowledge (Fudenberg and Levine, 1998). Moreover, in the real world, individuals, as for instance animals, are not necessarily rational as assumed by the GT. Thus, in the context of evolutionary GT (EGT), alternative explanations focus on equilibrium arising as a long-run outcome of a process in which populations of animals interact over time.

In the next section, we briefly review some of these approaches. In particular, we note that one of the approaches, the replicator dynamics (RD), presents some problems related to its justification among populations of animals. In fact, there has been a great deal of questioning about why should one care about using the RD, since neither economic agents nor artificial agents (and actually not even monkeys) are genetically programmed to "play" certain behaviors. Thus a justification for the replicator could be that there is an underlying model of learning (by the agents) that gives rise to the dynamics.

Some alternatives have been proposed in this line. Fudenberg and Levine (1998) refer to two interpretation that relate to learning. The first is social learning, a kind of "asking around" model in which players can learn from others in the population. The second is a kind of satisficing-rather-than-optimizing learning process in which the probability of a determined strategy is proportional to the payoff difference with the mean expected payoff.

This second variant has been explored, among others, by Börgers and Sarin (1997); Tuyls et al. (2006), which are based on some kind of reinforcement learning at individual level. In particular, in the stimulus-response based approach of Börgers and Sarin, the reinforcement is proportional to the realized payoff (in their formulation, necessarily positive). Thus each strategy's probability increases by a factor that is computed by the current probability multiplied by the difference between the strategy's expected payoff and the expected payoff of the player's current mixed strategy. In the limit, it is shown that the trajectories of the stochastic process converges to the continuous RD. This is valid in a stationary environment. However, as noted by Börgers and Sarin and Fudenberg and Levine, this does not imply that the RD and the stochastic process have the same asymptotic behavior when the play of *both* players follow a stimulus-response learning approach. We remark that Fudenberg and Levine (1998) specifically mention a two agent or two population game, but the same is true (even more serious), when it comes to more.

The reasons for this difference are manifold. First, Börgers and Sarin's main assumption is "...an appropriately constructed continuous time limit", i.e., a gradual (very small) adjustment is made by players between two iterations of the game. This implies that the discrete learning model evolves stochastically, whereas the equations of the RD are deterministic. Also, there is the fact that players may be stuck in suboptimal strategies because they are all using a learning mechanism, thus turning the problem non-stationary. These facts have as consequences that other popular dynamics in game-theory as, e.g., best response dynamics, which involve instantaneous adjustments to best replies,

have difference in the asymptotic behavior. For example, in the matching pennies game, "discrete time replicator dynamics will cycle along expanding trajectories, but will not get absorbed by any pure strategy outcome" (Börgers and Sarin, 1997). In the battle of the sexes game, a learning approach which is not based on the gradual adaptation (such as best response) may prevent convergence to an equilibrium while the gradual adjustment of RD permits such convergence.

In summary, there are advantages and disadvantages in using the discussed approaches and interpretations, i.e., the continuous, analytical variant of RD, and learning and adaptation approaches such as best response, genetic algorithms, and stimulus-response based models. Besides, as we verify, not all adaptation methods do replicate the RD.

In this paper we aim at applying different approaches and compare their performance. Specifically, we use three of them: the analytical equations of the RD, a genetic based evolutionary approach, and reinforcement learning (here Q-learning). We remark that, as discussed in the next section, some learning approaches are not appropriate for this problem as they either consider perfect monitoring (observation of other individuals' actions, as in Claus and Boutilier (1998)), or modeling of the opponent (as in fictitious play), or both. In our case this is not possible given the high number of individuals involved in the populations, and the unlikelihood of encounters happening frequently among the same individuals.

Further, we employ two games that are played by three populations of individuals and have been seen as metaphors for studying interactions in populations of individuals. The first is a three-person version of the matching pennies, due to Jordan (apud Fudenberg and Tirole (1991)). In this, as in the two-person original game, there is just a single Nash equilibrium (in mixed strategies). The second is due to B. O'Neill (apud Myerson (2002)), which pays a non-zero quantity to the players when exactly one of them select one of the two strategies. This game has eight Nash equilibria, four of which are in pure strategies, thus making the task of coordinating which equilibrium to select very hard for the individuals.

We are interested in the trajectory of a population of individuals with very little knowledge of the game. Indeed, they are only aware of the payoff or reward received, thus departing from the assumption of knowledge of payoff matrix and rationality being common knowledge, frequently made in GT. We show that, in the case of the three-player matching pennies, the evolutionary approach leads to the same result as the RD, namely to cycle, which means to the (unique) Nash equilibrium in mixed strategy. For the second game, it was possible to observe convergence to the Nash equilibria in pure strategies. In all cases, convergence depends on the rate of experimentation in the populations.

Background and Related Work

Evolutionary Game Theory

EGT investigates the relationship between individual and aggregate behaviors. Its inspiration comes from population genetics, where the focus is less on individual behavior and more on the aggregate population behavior. This shift from individual-level decision-making (eventually leading to user equilibrium), to dynamics of individuals interaction is in line with the increasing complexity in modern societies. There are many systems where we nowadays observe a tendency of a complex coupled decision-making process. Already in 1950, Nash saw this phenomenon, which he then called "mass-action interpretation". Later, this focus on equilibria was criticized by J. Maynard Smith: "An obvious weakness of the game-theoretic approach to evolution is that it places great emphasis on equilibrium states ..." (Smith, 1982). Besides, J. Maynard Smith also dealt with the shift from individual to population level. Even if he borrowed some definitions from standard GT when he introduced the concept of evolutionary stable strategy (ESS) as a way to understand conflicts among animals, he had already noticed that "there are many situations ... in which an individual is, in effect, competing not against an individual opponent but against the population as a whole... Such cases can be described as 'playing the field'..." (Smith, 1982).

Currently, this kind of modeling is called a population game, which models simultaneous interactions of a large number of simple individuals or agents distributed in a finite number of populations. Simple agents here mean that each has a (typically small) number of strategies to choose, causing a minor impact in other agents payoff. Despite this, the payoff of each agent is, as in the classical GT, conditioned by the distribution of strategies in each population. In EGT and population games, typically, one is not interested in constancy or equilibrium only. Rather, the major interest is on the dynamics of games. A population of decision-makers is considered, in which what is investigated is how the rate of the strategy profiles change as a response to the decisions made by all individuals in the population.

This idea that the composition of the population of individuals (and hence of strategies) in the next generations changes with time (in this case generations) suggests that we can see these individuals as replicators. The RD is based on gradual movement from worse to better strategies. One of the results of Börgers and Sarin is that in appropriately constructed continuous time limit, a stimulus-response based learning model converges to the continuous time version of the RD. They have proposed that such a continuous time limit is constructed so that each time interval sees many iterations of the game, and that the adjustments that the agents make between two iterations of the game are very small. This way the stochastic learning process becomes deterministic in the limit, thus replicating the system of differential equations which characterizes the RD.

However, as mentioned in the introduction, this result refers to arbitrary, finite points in time, and does not hold if infinite time is considered. When time tends to infinitely, the asymptotic behavior of the discrete time learning process can be different from the asymptotic behavior of the continuous time RD.

Additionally, the RD treats the player as a population (of strategies). By the construct of the continuous time of Börgers and Sarin, in each iteration, a random sample of the population is taken to play the game. Due to the law of large numbers, this sample represents the whole population. However, in the discrete learning process, at each time, only one strategy is played by each individual. Moreover, the outcome of each of these interactions affects the probabilities with which the strategies are used in the next time step.

These results have been extended in Tuyls et al. (2006). It was shown that the positive reinforcement model by Börgers and Sarin (1997) corresponds to the learning automata. Moreover, a similar dynamics was derived for Boltzmann action selection. The theoretical results were verified with experiments in 3 classes of 2×2 games.

These works suggest that other dynamics, e.g., based on less gradual adjustments may lead to different results in other games as well. We also remark that at least one of the games considered in in this paper is one in which not only the analytical computation of the RD is non-trivial, but also the fact that more populations and actions are involved contribute to results being less intuitive as the cases in Börgers and Sarin (1997); Tuyls et al. (2006).

Next, for sake of clarity, we briefly mention the Q-learning method, as well as discuss some related work on multiagent reinforcement learning (MARL).

Individual and Multiagent RL

Reinforcement learning (RL) by a single agent problems can be modeled as Markov decision processes (MDPs). An experience tuple $\langle s, a, s', r \rangle$ denotes the fact that the agent was in state s, performed action a and ended up in s' with reward r. Q-learning is a popular model-free algorithm in which the update rule for each experience tuple is given in Equation 1 where α is the learning rate and γ is the discount for future rewards.

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \, max_{a'} \, Q(s',a') - Q(s,a)\right) \tag{1}$$

When many individuals or agents learn simultaneously, the problems are well known. They arise mainly due to the fact that while one agent is trying to model the environment (other agents included), the others are doing the same and potentially changing the environment they share.

Besides, in MARL an issue is the exponential increase in the space of joint states and joint actions, if agent i *explicitly models the states and actions of other agents*. The decision on whether or not to include joint states and/or joint actions in the learning process of i is a key one as it has severe implications. In fact, most of the game-theoretic literature concentrates on games with few players and few actions because otherwise it is computationally prohibitive.

This paper deals with a large number of agents, thus approaches such as JAL (joint agent learners, Claus and Boutilier (1998)) cannot be used because an explicit model of other agents' actions, states, and rewards is necessary. After JAL, several approaches have been proposed for related as well as more general MARL problems. However, they cannot be used due to some restrictions: zero-sum game (e.g., Littman (1994)), few agents and/or few actions and/or assumption of perfect monitoring (e.g., Hu and Wellman (1998); Lauer and Riedmiller (2000); Kapetanakis and Kudenko (2002); Kuminov and Tennenholtz (2008)).

Methods

Formalization of Population Games

Population games are quite different from the games studied by the classical GT because population-wide interaction generally implies that the payoff to a given member of the population is not necessarily linear in the probabilities with which pure strategies are played. A population game can be defined as follows.

- (**populations**) $\mathcal{P} = \{1, ..., p\}$: society of $p \geq 1$ populations of individuals where $|p|$ is the number of populations
- (**strategies**) $\mathcal{S}^p = \{s_1^p, ..., s_m^p\}$: set of strategies available to agents in population p
- (**payoff function**) $\pi(s_i^p, \mathbf{q}^{-\mathbf{p}})$

Agents in each population p have m^p possible strategies. Let n_i^p be the number of individuals using strategy s_i^p. Then, the fraction of individuals using s_i^p is $x_i^p = \frac{n_i^p}{N^p}$, where N^p is the size of p. $\mathbf{q^p}$ is the m^p-dimensional vector of the x_i^p, for $i = 1, 2, ..., m^p$. As usual, $\mathbf{q}^{-\mathbf{p}}$ represents the set of $\mathbf{q^p}$'s when excluding the population p. The set of all $\mathbf{q^p}$'s is $\mathbf{q}$. Hence, the payoff of an agent of population p using strategy s_i^p while the rest of the populations play the profile $\mathbf{q}^{-\mathbf{p}}$ is $\pi(s_i^p, \mathbf{q}^{-\mathbf{p}})$. Consider a (large) population of agents that can use a set of pure strategies $\mathcal{S}^p$. A population profile is a vector σ that gives the probability $\sigma(s_i^p)$ with with strategy $s_i^p \in \mathcal{S}^p$ is played in p.

One important class within population games is the class of symmetric games, in which two random members of a *single* population meet and play the stage game, whose payoff matrix is symmetric. The reasoning behind these games is that members of a population cannot be distinguished, i.e., two meet randomly and each plays one role but these need not to be the same in each context. Thus the symmetry. However, there is no reason to restrict oneself to a symmetric modeling in other scenarios beyond population biology. For

	H	T
H	+1/+1/-1	-1/-1/-1
T	-1/+1/+1	+1/-1/+1

	H	T
H	+1/-1/+1	-1/+1/+1
T	-1/-1/-1	+1/+1/-1

Table 1: Payoff matrices for the three-player matching pennies game; payoffs are for player 1 / player 2 / player 3.

	x_2	y_2
x_1	0/0/0	**6/5/4**
y_1	**5/4/6**	0/0/0

	x_2	y_2
x_1	**4/6/5**	0/0/0
y_1	0/0/0	**0/0/0**

Table 2: Payoff matrices for the 3PEOY game; payoffs are for player 1 / player 2 / player 3 (the four Nash equilibria in pure strategies are indicated in boldface).

instance, in economics, a market can be composed of buyers and sellers and these may have asymmetric payoff functions and/or may have sets of actions whose cardinality is not the same. In asymmetric games, each agent belongs to one class determining the set of legal strategies.

Before we present the particular modeling of asymmetric population game, we introduce the concept of RD.

The previously mentioned idea that the composition of the population of individuals (and hence of strategies) in the next generations changes with time suggests that we can see these individuals as replicators. In the RD, the rate of use of a determined strategy is proportional to the payoff difference with the mean expected payoff, as in Eq. 2. As previously defined, the fraction of agents using s_i^p is $x_i^p = \frac{n_i^p}{N^p}$. The state of population p can be described as a vector $\mathbf{x^p} = (x_i^p, ..., x_m^p)$. We are interested in how the fraction of agents using each strategy changes with time, i.e., the derivative $\frac{dx_i^p}{dt}$ (henceforth denoted $\dot{x}_i^p$).

$$\dot{x}_i^p = (\pi(s_i^p, \mathbf{x^p}) - \bar{\pi}(\mathbf{x^p})) \times x_i^p \quad (2)$$

In Eq. 2, $\bar{\pi}(\mathbf{x^p})$ is the average payoff obtained by p:

$$\bar{\pi}(\mathbf{x^p}) = \sum_{i=1}^{m} x_i^p \pi(s_i^p, \mathbf{x^p})$$

Obviously, to analytically compute this average payoff, each individuals would have to know all the payoffs, which is quite unrealistic.

Two Scenarios for Three Player Games

In the three-population games considered here, to avoid confusion we use the term "player" with its classical interpretation, i.e., the decision-makers of the normal form game (NFG). Because this game is played by randomly matched individuals, one from each population, we call these individuals "agents". Thus player refers to a population of agents.

The just given description of the general population game is instantiated for our particular scenarios as follows.

Three Player Matching Pennies From the general definition of a population game given in the previous section, this is the particular instance for the three player matching pennies (henceforth 3PMP):

- (**populations**) $\mathcal{P} = \{1, 2, 3\}$
- (**strategies**) for each population $p \in \mathcal{P}$: $\mathcal{S}^1 = \mathcal{S}^2 = \mathcal{S}^3 = \{H, T\}$
- (**payoff function**) see Table 1

From the payoff matrix for this game (Table 1), one sees that as in the two player original game, there is a single Nash equilibrium, in mixed strategies. This is indicated in Table 3. In this table, columns 2–3 specify $\mathbf{x^1}$ (fraction of agents selecting each strategy H and T in population $p = 1$), columns 4–5 specify $\mathbf{x^2}$ of $p = 2$, and columns 6–7 specify $\mathbf{x^3}$ of $p = 3$. The last column gives the payoffs for the agents.

Three Player Coordination Game The second three player game that we consider is a kind of coordination game. As mentioned, it pays a non-zero quantity to the players when exactly one of them select one of the two strategies. Henceforth we denominate this game by 3PEOY (three players, where exactly one should select strategy y). This game has eight Nash equilibria, four of which are in pure strategies, thus making the task of coordinating over equilibrium selection very hard for the agents.

The corresponding population game is then defined:

- (**populations**) $\mathcal{P} = \{1, 2, 3\}$
- (**strategies**) for each population $p \in \mathcal{P}$: $\mathcal{S}^1 = \{x_1, y_1\}$, $\mathcal{S}^2 = \{x_2, y_2\}$, and $\mathcal{S}^3 = \{x_3, y_3\}$.
- (**payoff function**) see Table 2

profile	x_H^1	$x_T^1 = (1 - x_H^1)$	x_H^2	$x_T^2 = (1 - x_H^2)$	x_H^3	$x_T^3 = (1 - x_H^3)$	payoff
σ	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0/0/0

Table 3: Unique Nash equilibria for the 3PMP game.

profile	x_x^1	$x_y^1 = (1 - x_x^1)$	x_x^2	$x_y^2 = (1 - x_x^2)$	x_x^3	$x_y^3 = (1 - x_x^3)$	payoff
σ_a	1	0	1	0	0	1	4/6/5
σ_b	0	1	0	1	0	1	0/0/0
σ_c	0	1	1	0	1	0	5/4/6
σ_d	1	0	0	1	1	0	6/5/4
σ_e	1	0	≈ 0.44	≈ 0.56	≈ 0.54	≈ 0.45	≈ 2.63 / ≈ 2.73 / ≈ 2.22
σ_f	≈ 0.44	≈ 0.56	≈ 0.54	≈ 0.45	1	0	≈ 2.73 / ≈ 2.22 / ≈ 2.63
σ_g	≈ 0.54	≈ 0.45	1	0	≈ 0.44	≈ 0.56	≈ 2.22 / ≈ 2.63 / ≈ 2.73
σ_h	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	≈ 2.22 / ≈ 2.22 / ≈ 2.22

Table 4: Eight Nash equilibria for the 3PEOY game.

The eight Nash equilibria appear in Table 4. In this table, columns 2–3 specify $\mathbf{x^1}$ (fraction of agents selecting each strategy s_i^1 in population $p = 1$), columns 4–5 specify $\mathbf{x^2}$ of $p = 2$, and columns 6–7 specify $\mathbf{x^3}$ of $p = 3$. The last column gives the payoffs for the agents. For example, for the first equilibrium (profile σ_a), because $x_x^1 = 1$, $x_x^2 = 1$, and $x_x^3 = 0$, all agents in $p = 1$ selects action x_1, all agents in $p = 2$ select x_2 and all agents in $p = 3$ select y_3.

Regarding the mixed strategy profile σ_e (for example), all agents in $p = 1$ select action x_1 (because $x_y^1 = 0$), whereas in the other two populations nearly half of the agents select each action. Profiles σ_f to σ_h can be similarly interpreted.

In the classical GT interpretation of equilibrium, profiles σ_a, σ_b, σ_c, and σ_d would be Nash equilibria in pure strategies, while the other four equilibria would mean that agents randomize between at least two pure strategies. The EGT interpretation thought is as follows. If we consider that we are dealing with three populations of agents, we can think about the equilibria in terms of the percentage of individuals in one of the three populations that in fact select one of the actions available. This seems a more reasonable explanation for the concept of mixed strategies, given that, at each time, agents in fact only select an action.

The same reasoning obviously applies to the 3PMP game which has only a mixed strategy profile.

Moreover, it must also be noticed that in asymmetric games, all ESS are pure strategies (for a proof see, e.g., Webb (2007)). Thus, for the 3PEOY game, only σ_a to σ_d are candidates for ESS.

Note from Table 4 that in this game σ_b, though a Nash equilibrium, is not efficient (all agents receive zero), thus the learning models tested here should be able to recognize this.

Replicator Dynamics, Evolutionary Approach and Reinforcement Learning

As mentioned in the introduction, the continuous RD model, which is hard to justify, can be reproduced with some forms of learning. To compare the performance of these learning models, we have first formulated the continuous RD for our specific three-population game. The equations can be derived from Eq. 2.

In both the 3PMP and 3PEOY, we have checked which Nash equilibria are stable. In the 3PEOY in particular, the Nash equilibria that need to be investigated are, as mentioned, those in pure strategies, i.e., σ_a to σ_d from Table 4. To check which are the ESSs, it is necessary to analyze which are the stable rest points. In simple games, e.g., 2×2 or even symmetric 2×3 this can be done graphically. However, because our problem involves several variables the divergence operator was used.

Now we turn to the approaches based on evolution and reinforcement learning. In both cases, in each time step, agents from each population p play g games whose payoffs are given in Table 1 and Table 2.

For the evolutionary approach, mutations create genetically mutated versions of the agents. Here we use a mutation rate p_m: at each time step, each agent in the population receives a new strategy with probability p_m. We recall that, according to Börgers and Sarin (1997), it is expected that if the adjustment is not gradual, there may be no convergence to the behavior of the continuous RD. The sum of the payoffs obtained by playing these g games is then the fitness of the agent. After these g games are played, the populations of agents are reproduced: In each population p, the fittest agents have a higher probability of being selected. Then each individual suffers mutation with probability p_m, which means that its strategy is changed to another one randomly selected.

For the reinforcement learning, agents learn using individual Q-learning (Eq. 1), thus assessing the value of each strategy by means of Q values. For action selection, ε-greedy was used, i.e., action selection is random with probability ε, otherwise it is greedy. In line with the just mentioned issue of gradual adjustments, and from Tuyls et al. (2006), we know that the value of ε is key to reproduce the behavior of the continuous RD.

Experiments and Results

In this section we discuss the numerical simulations of the evolutionary and learning based approaches and compare them with the continuous RD, from which we know the Nash equilibria, and the candidates to be ESSs, for both games 3PMP and 3PEOY.

We are interested in investigating issues such as what happens if each population p starts using a given profile σ^p in games that have more than one equilibrium. To which extent the rate p_m shifts this pattern? For instance, if the population starts using any σ^p, what happens if it is close to (but not actually at) σ^*? Will it tend to evolve towards σ^* or move away? If it reaches σ^*, how long has it taken? What happens if there are multiple equilibria?

The main parameters, as well as the values that were used in the simulations are: $\mathcal{P} = \{0, 1, 2\}$, $N^0 = N^1 = N^2 = 300$, $g = 10,000$, $\alpha = 0.5$; ε, Δ (number of time steps) and p_m were varied. In all cases, at step 0, agents select strategies from a uniform distribution of probability.

The next subsections discuss the evolutionary and the reinforcement learning approaches respectively.

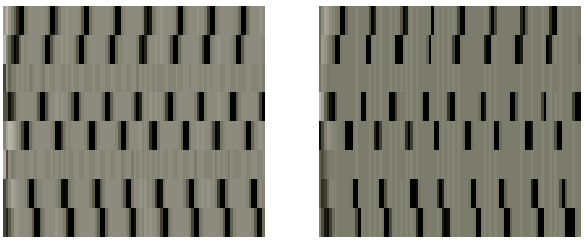
Adaptation with Evolutionary Approach

In this case, because more than two variables (strategies) are involved, it is not possible to show typical (2d) RD-like plots that depict the trajectory of these variables. Therefore, as an alternative to show the dynamics, we use heatmaps.

We start with the 3PMP game. In the plots that appear in Figure 1 (which were reduced due to lack of space), heatmaps are used to convey the idea of the intensity of the selection of each of the 8 joint actions (represented in the vertical axis) along time (horizontal axis), with $\Delta = 1000$ time steps. These 8 joint actions are those that appear in Table 1. Due to an internal coding used, the 8 joint actions are labeled such that 0 and 1 mean the selection of first strategy (H) and second strategy (T) respectively.

In each triplet (joint action), the first digit indicates the action of $p = 3$, the second digit is for the action of $p = 2$, and the third digit is for $p = 1$.

In the heatmaps, to render the figure cleaner we just use shades of gray color (instead of hot colors as usual). In any case, the darker the shade, the more intense one joint action is selected. Thus we should expect that the Nash equilibria correspond to the darker strips.



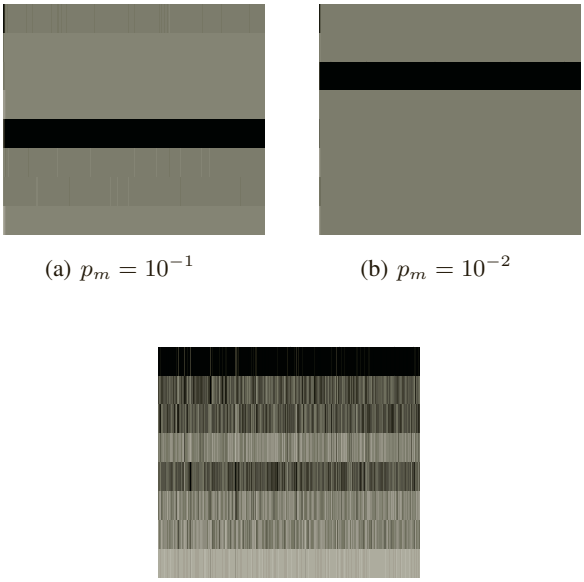
(a) $p_m = 10^{-1}$ (b) $p_m = 10^{-2}$

Figure 1: 3PMP: Evolution of Dynamics for Different values of p_m, evolutionary approach.

In Figure 1 it is possible to see that there is no convergence to any of the pure strategy. Rather, agents cycle among 6 of the joint actions. This happens for all value of p_m tested. The two joint actions that pay -1 to all agents are quickly discarded. Remember that a different pattern occurs when fictitious play was used, according to Börgers and Sarin (1997).

Regarding the second game, 3PEOY, plots appear in Figure 2. Again, there are 8 joint actions. These are labeled such that 0 and 1 mean the selection of first strategy (x) and second strategy (y) respectively.

In particular, the four Nash equilibria (σ_a, σ_b, σ_c, and σ_d) in pure strategies of these game are represented as: `100`, `111`, `001` and `010` respectively.



(a) $p_m = 10^{-1}$ (b) $p_m = 10^{-2}$

(c) $p_m = 0.5$

Figure 2: 3PEOY: Evolution of Dynamics for Different values of p_m, evolutionary approach.

In Figure 2 we show how the selection evolves along time, for two values of p_m. Figure 2(a) is for $p_m = 10^{-1}$, i.e., changes in strategy occur with this rate. In these particular plot we show clear convergence to σ_a (100). However, convergence is also possible to other joint actions, as, e.g. 010 and 001, but not to 111 because this is a non efficient Nash equilibria. We do not show all plots as they are similar to Figure 2(a).

When p_m is lowered as in Figure 2(b), the pattern is the same, i.e., clear convergence to either σ_a, σ_c or σ_d.

When p_m is set higher, this pattern does not occur, i.e., there is no clear convergence to any of the Nash equilibria. An extreme case with $p_m = 0.5$ is depicted in Figure 2(c), where one sees that the performance is poor due to the high rate of changes by the agents. Interestingly, frequently, the Pareto inefficient Nash equilibria σ_b is selected, which means payoff of zero to each agent.

Reinforcement Learning

We now turn to the individual learning using Q-learning. Experiments were run with different values of α and change in ε. It seems that α has much less influence in the result than ε. Thus we concentrate on $\alpha = 0.5$.

For both games, here we show plots for ε starting at 1.0 with various decay rates each time step.

For the 3PMP game, to render the picture more clear, plots in Figure 3 depict the evolution within time of the percentage of H being selected by agents in population $p = 1$ only as others are similar. One can see that agents shift from H to T and vice-versa. This happens no matter if the decay of ε is fast (Figure 3(a)) or more slow (Figure 3(b)).

For the 3PEOY game, due to a more clear convergence pattern, it is possible to plot not only the percentage of selection of action x by agents in the first population, but also for $p = 2$ and $p = 3$. Figure 4 depicts them. Figure 4(a) refers to a faster decay of ε, 0.9 at each time step, and shows convergence to σ_c. However we note that similar patterns occur leading to σ_a and σ_d (but not σ_b). Figure 4(b) is for decay of 0.99. In this particular plot, convergence is to σ_d.

How agents have converged to a given profile is better seen examining the trajectories of the probabilities to play each strategy, for each population. Due to the number of variables, it is not possible to plot them all together. Thus we opted to show the behavior of selected variables in a pairwise fashion, namely x_1 x x_2, x_1 x x_3, x_2 x x_3 (Figure 5). In this plot, for each of the three pairs, the x-axis represents the probability with which the first component of the pair selects strategy x, while the y-axis represents the probability with which the second component of the pair selects x.

It is possible to see that although these percentages all start at 0.5, they all converge to 1.0 or close. In the particular plot convergence was to $x_1 = 1$, $x_2 = 0$, and $x_3 = 1$ and hence this corresponds to the same pattern as in Figure 4(b), σ_d.

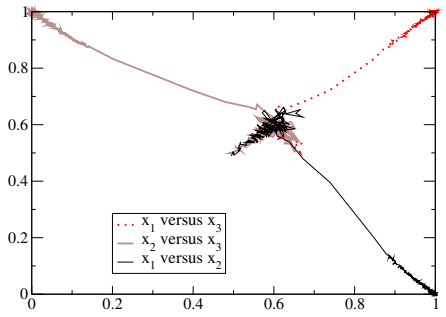


Figure 5: Trajectories: x_1 x x_2, x_1 x x_3, and x_2 x x_3 (all start at $x_i = 0.5$).

In short, some conclusions can be drawn from these simulations. First, simultaneous learning by the agents does not always lead to the Nash equilibrium, much less to the ESS computed for the corresponding RD of the static NFG.

If an evolutionary approach is used, depending on the p_m rate, any of the Nash equilibria may establish, or agents may be stuck at a sub-optimal state.

This is in line with the result in Börgers and Sarin (1997), which prescribes gradual adjustments. Profiles that are dominated do not establish.

Conclusion

In this paper, two three-population games were used to illustrate the patterns of convergence when different methods are used for replicating the dynamics prescribed by analytical, continuous methods such as the RD. It was seen that the extent of match between the continuous RD and discrete learning methods (whether an evolutionary approach or Q-learning) depends on the pace of the adjustment.

Also, compared to results in Tuyls et al. (2006), an extra population in the matching pennies game has slowed down the convergence as it takes more time for agents to learn in the presence of three populations.

Future work will consider information broadcast to agents, in order to have a further way to model action selection and try to improve the coordination task, as well as investigate issues regarding correlated equilibria.

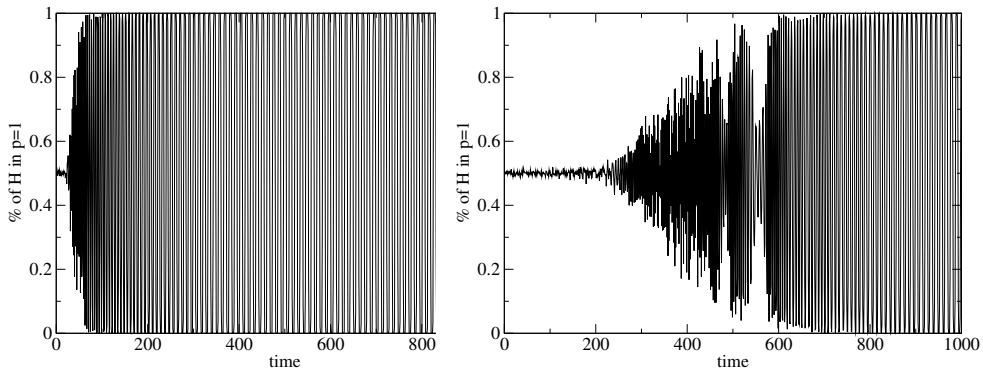
Acknowledgements

The author and this work are partially supported by CNPq.

References

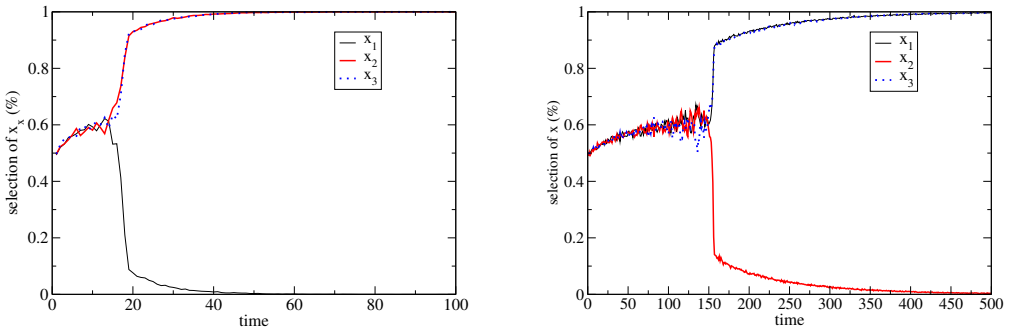
Börgers, T. and Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14.

Claus, C. and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of*



(a) Selection of H along time, decay of 0.90 (b) Selection of H along time, decay of 0.99

Figure 3: Evolution of the selection of the first player in the 3PMP game, QL.



(a) Selection of H along time, decay of 0.90 (b) Selection of H along time, decay of 0.99

Figure 4: Evolution of the selection of the three players in the 3PEOY game, QL.

the Fifteenth National Conference on Artificial Intelligence, pages 746–752.

Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games*. The MIT Press.

Fudenberg, D. and Tirole, J. (1991). *Game Theory*. MIT Press Books. The MIT Press.

Hu, J. and Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proc. 15th International Conf. on Machine Learning*, pages 242–250. Morgan Kaufmann.

Kapetanakis, S. and Kudenko, D. (2002). Reinforcement learning of coordination in cooperative multi-agent systems. In *AAAI/IAAI*, pages 326–331.

Kuminov, D. and Tennenholtz, M. (2008). As safe as it gets: Near-optimal learning in multi-stage games with imperfect monitoring. In *Proceeding of the ECAI 2008*, pages 438–442, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Lauer, M. and Riedmiller, M. (2000). An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proc. 17th International Conference on Machine Learning*, pages 535–542. Morgan Kaufmann, San Francisco, CA.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning, ML*, pages 157–163, New Brunswick, NJ. Morgan Kaufmann.

Myerson, R. (2002). *Game theory*. Harvard Univ. Press, Cambridge, MA, 5. print edition.

Nash, J. (1950). *Non-Cooperative Games*. PhD thesis, Priceton University.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK.

Tuyls, K., Hoen, P. J., and Vanschoenwinkel, B. (2006). An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multiagent Systems*, 12(1):115–153.

Webb, J. N. (2007). *Game Theory – Decisions, Interaction and Evolution*. Springer, London.