Paying the Price of Learning Independently in Route Choice

Ana L.C. Bazzan

PPGC / Instituto de Informática – UFRGS
Caixa Postal 15.064, 91.501-970, Porto Alegre, RS, Brazil
bazzan@inf.ufrgs.br

Abstract. In evolutionary game theory, one is normally interested in the investigation about how the distribution of strategies changes along time. Equilibrium-based methods are not appropriate for open, dynamic systems, as for instance those in which individual drivers learn to select routes. In this paper we model route choice in which many agents adapt simultaneously. We investigate the dynamics with a continuous method (replicator dynamics), and with learning methods (social and individual). We show how the convergence to one of the Nash equilibria depends on the underlying learning dynamics selected, as well as on the pace of adjustments by the driver agents.

1 Introduction and Related Work

The number of large metropolitan areas with more than ten million inhabitants is increasing rapidly, with the number of these so-called mega-cities now at more than 20. This increase has strong consequences to traffic and transportation. According to the keynote speaker of the IEEE 2011 forum on integrated sustainable transportation systems, Martin Wachs, *mobility is perhaps the single greatest global force in the quest for equality of opportunity* because it plays a role in offering improved access to other services. Congestion is mentioned as one of the major problems in various parts of the world, leading to a significant decrease in the quality of life, especially in mega-cities of countries experiencing booming economies. Mobility is severely impacted with 4.8 billion hours of travel delay that put the cost of urban congestion in USA alone at 114 billion dollars (www.its.dot.org). Moreover environmental costs must be considered.

One important part in the whole effort around intelligent transportation systems (ITS), in which artificial intelligence (AI) plays a role, relates to efficient management of the traffic. This in turn can be achieved, among others, by an efficient assignment of routes, especially if one thinks that in urban scenarios a considerable important part of the network remains sub-utilized whereas jams occur in a small portion of the network.

The conventional process of modeling assignment is macroscopic, based on equilibrium. In the traffic engineering literature, the Nash equilibrium is referred as Wardrop equilibrium [1]. This is convenient as it has sound mathematical grounds. Static traffic assignment models have mathematical properties such as

L. Correia, L.P. Reis, and J. Cascalho (Eds.): EPIA 2013, LNAI 8154, pp. 42–53, 2013.

existence and uniqueness of equilibrium. Each link in the transportation network can be described by the so-called volume-delay functions expressing the average or steady-state travel time on a link. However, equilibrium in traffic networks are often inefficient at collective level. Important results regarding this relate to the price of anarchy [2–4] and Braess paradox [5]. Given these results, one may try to improve the load balancing by routing drivers. However, whereas it is possible to do this in particle systems (e.g., data packages in communication networks), this is not the case in social systems.

Other drawbacks of equilibrium-based methods are that they assume stationarity, and rationality plus common knowledge. Regarding the former, it is not clear what equilibria can say about changes in the environment. Regarding rationality and 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.

Therefore, evolutionary game theory (EGT) offers alternative explanations to rationality and stationarity by focusing on equilibrium arising as a long-run outcome of a process in which populations of bounded rational agents interact over time. See for instance [6] for an extensive discussion about the various approaches based on learning in games. However, we note that one of the EGT approaches, the replicator dynamics (RD), presents some problems regarding its justification in decentralized or multiagent encounters. Thus an explanation 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 [6] refer to two interpretations 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 satisfying-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 [7, 8]. In particular, in the stimulus-response based approach of Börgers and Sarin, the reinforcement is proportional to the realized payoff. In the limit, it is shown that the trajectories of the stochastic process converge to the continuous RD. This is valid in a stationary environment. However, 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 [6] specifically mention a two agent or two population game, but the same is true (even more seriously), 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. More specifically, the RD treats the player as a population (of strategies). By the construct of the continuous time of [7], 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 agent. Moreover, the outcome of each of these interactions affects the probabilities with which the strategies are used in the next time step. 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 involves instantaneous adjustments to best replies, depict difference in the asymptotic behavior. Theoretical results regarding an extension by [8] were verified with experiments in 3 classes of 2×2 games. The differences in behavior of the continuous/discrete models were verified, with matching pennies converging to a pure strategy (thus, not a Nash equilibrium), while the RD cycles. This suggests that other dynamics, e.g., based on less gradual adjustments may lead to different results in other games as well.

In summary, there are advantages and disadvantages in using the discussed approaches and interpretations, i.e., the continuous, analytical variant of RD, and learning approaches such as best response, social learning, and stimulus-response based models.

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 kind of social learning where dissatisfied agents ask their peers, and individual Q-learning. We remark that other multiagent reinforcement learning approaches are not appropriate for this problem as they either consider perfect monitoring (observation of other agents' actions, as in [9]), or modeling of the opponent (as in fictitious play), or both. [1] In our case this is not possible given the high number of agents involved and the unlikelihood of encounters happening frequently among the same agents.

Here, the scenario is an asymmetric population game that models traffic assignmen. Populations with different sets of actions interact and the payoff matrix of the stage game is also asymmetric. We note that, most of the literature refers to homogeneous population, two actions, rendering the game symmetric. We are interested in the trajectory of a population of decision-makers 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.

In the next section we introduce our method and the formalization of route choice as population games. Experiments and their analysis follow in Section 3. The last section concludes the paper and discusses some future directions.

[1] Since a comprehensive review on learning in repeated games is not possible here, we refer the reader to [10, 11] and references therein for a discussion on the assumptions made in the proposed methods.

2 Methods

2.1 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 agents, 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**) $\pi(s_i^p, \mathbf{q^{-p}})$

Agents in each p have m^p possible strategies. Let n_i^p be the number of individuals using strategy s_i^p. Then, the fraction of agents 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, $\mathbf{q^{-p}}$ 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 $\mathbf{q^{-p}}$ is $\pi(s_i^p, \mathbf{q^{-p}})$.

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 which strategy $s_i^p \in \mathcal{S}^p$ is played in p.

One important class within population games is that 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 contest. Thus the symmetry. However, there is no reason to use a symmetric modeling in other scenarios beyond population biology. For 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 the route choice game discussed here, asymmetric games correspond to multi-commodity flow (more than one origin-destination pair).

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 agents (and hence of strategies) in the next generations changes with time suggests that we can see these agents as replicators. Moreover, an evolutionary stable strategy (ESS) may not even exist, given that the set of ESSs is a possibly empty subset of the set of Nash equilibria computed for the normal form game (NFG). 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. 1.

$$\dot{x}_i^p = (\pi(s_i^p, \mathbf{x^p}) - \bar{\pi}(\mathbf{x^p})) \times x_i^p \tag{1}$$

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,

Table 1. Available routes in the three populations

route	description	length
G0	$B1 \rightarrow C1 \rightarrow C3 \rightarrow E3 \rightarrow E5$	7
S0	$B1 \rightarrow F1 \rightarrow F2 \rightarrow E2 \rightarrow E5$	9
B0	$B1 \rightarrow F1 \rightarrow F4 \rightarrow E4 \rightarrow E5$	9
G1	$A2 \rightarrow A3 \rightarrow E3 \rightarrow E5$	7
S1	$A2 \rightarrow A5 \rightarrow E5$	10
B1	$A2 \rightarrow A6 \rightarrow F6 \rightarrow F4 \rightarrow E4 \rightarrow E5$	13
G2	$D5 \rightarrow D3 \rightarrow E3 \rightarrow E5$	5
S2	$D5 \rightarrow D4 \rightarrow C4 \rightarrow C5 \rightarrow E5$	5

Table 2. Payoff matrices for the three-player traffic game; payoffs are for player 0 / player 1 / player 2 (the three Nash equilibria in pure strategies are indicated in boldface)

	G2		
	G1	S1	B1
G0	1/1/4	**5/6/7**	5/1/7
S0	3/4/6	4/6/8	4/1/8
B0	5/5/7	**5/6/8**	4/0/9

	S2		
	G1	S1	B1
G0	4/4/8	7/4/6	7/1/8
S0	4/6/8	5/4/6	5/1/8
B0	**5/7/8**	5/4/6	4/0/8

i.e., the derivative $\frac{dx_i^p}{dt}$ (henceforth denoted $\dot{x}_i^p$). In Eq. 1, $\bar{\pi}(\mathbf{x^p})$ 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 agent would have to know all the payoffs, which is quite unrealistic in real scenarios.

Henceforth, in order to illustrate the approach and introduce the evaluation scenario, we refer to a specific instance. However this has some important properties: non-symmetry and presence of several equilibria.

In the three-population game 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 way the three populations interact determines their reward functions. For the sake of illustration, it is assumed that the three populations of agents use a road network Γ to go from their respective origins, and that there is a single destination (typical morning peak).

Each agent in each p has some alternative routes or strategies $\mathcal{S}^p$. These are named after the following reasoning: G means greedy selection (G is the most preferred because this route yields the highest utility *if not shared with other populations*); S means second preferred alternative; and B means border route

Table 3. Five Nash equilibria for the three-player traffic game

	G0	S0	B0	G1	S1	B1	G2	S2	
profile	x_0^0	x_1^0	$(1-x_0^0-x_1^0)$	x_0^1	x_1^1	$(1-x_0^1-x_1^1)$	x_0^2	$(1-x_0^2)$	payoff
σ_a	1	0	0	0	1	0	1	0	5/6/7
σ_b	0	0	1	0	0	1	1	0	5/6/8
σ_c	0	0	1	1	0	0	0	1	5/7/8
σ_d	0	0	1	$\frac{2}{3}$	$\frac{1}{3}$	0	$\frac{3}{4}$	$\frac{1}{4}$	$5/5.5/\approx 7.3$
σ_e	≈ 0.474	0	≈ 0.526	≈ 0.386	≈ 0.614	0	≈ 0.352	≈ 0.648	$5/\approx 4.7/\approx 6.8$

(a route that uses the periphery of Γ). Populations $p = 0$ and $p = 1$ have strategies $\mathcal{S}^0 = \{G0, S0, B0\}$ and $\mathcal{S}^1 = \{G1, S1, B1\}$. $p = 2$ has only two strategies: $\mathcal{S}^2 = \{G2, S2\}$. Combining all these sets, there are 18 possible assignments of routes.

Each agent selects a route and the payoff obtained is a function of the delay on the route taken. The delay in each route is the sum of delays on each link in the route. These are given by a volume-delay function (VDF). The VDF used in the present paper considers the number of agents using each link. Specifically, it adds 1 unit each time an agent uses a given link. This way, a link has cost 0 if no agent uses it; cost 1 if one agent uses it; and so forth. The only exception is a bottleneck link b that belongs to the greedy routes G0, G1, and G2. Link b does not accommodate all agents. Thus if too many agents use it at the same time, each receives a penalty of 1 unit. Hence, considering the 18 combinations of routes that appear in Table 1, costs depend on length of routes and how routes share the Γ. The maximum cost is incurred by agents in $p = 1$ when the following combination of route choices is made: B0 / B1 / G2. This cost is 13. In order to deal with maximization (of payoff) rather than cost minimization, costs are transformed in payoffs. The highest cost of 13 is transformed in reward zero and so on. Payoffs computed this way are given in Table 2. Note that strategy B1 is dominated or $p = 1$, thus the learning models tested here must be able to recognize this.

Values in Table 2 represent an arbitrary assignment of utility of the three players involved, based on the topology of Γ as explained. The utility function $u(.)$ that underlies Table 2 is however equivalent to any other $\hat{u}(.)$ if $\hat{u}(.)$ represents identical preferences of the players, and $u(.)$ and $\hat{u}(.)$ differ by a linear transformation of the form $\hat{u}(.) = A \times u(.) + B$, $A > 0$. Of course equivalence here refers to the solution concept, i.e., a qualitative, not quantitative concept. Equivalent game models will make the same prediction or prescription.

For the three-population game whose payoffs are given in Table 2, there are five Nash equilibria. All appear in Table 3. In this table, columns 2–4 specify $\mathbf{x^0}$ (fraction of agents selecting each strategy s_i^0 in population $p = 0$), columns 5–7 specify $\mathbf{x^1}$ and the last two columns specify $\mathbf{x^2}$. For example, for the first equilibrium (profile σ_a), because $x_0^0 = 1$, $x_1^1 = 1$, and $x_0^2 = 1$, all agents in $p = 0$ select action G0, all agents in $p = 1$ select S1 and all agents in $p = 2$ select G2. Regarding the mixed strategy profile σ_d, all agents in $p = 0$ select action B0 (because $x_0^0 = x_1^0 = 0$), whereas in $p = 1$, $\frac{2}{3}$ of agents select G1 and $\frac{1}{3}$ select S1. In $p = 2$, $\frac{3}{4}$ of agents select G2 and $\frac{1}{4}$ select S2. Profiles σ_b, σ_c, and σ_e can be similarly interpreted.

It must also be noticed that in asymmetric games, all ESS are pure strategies (for a proof see, e.g., [12]). Thus only σ_a, σ_b, and σ_c are candidates for ESS. Besides, clearly, among σ_a, σ_b, and σ_c, the first two are (weakly) Pareto inefficient because σ_c is an outcome that make all agents better off.

2.2 Replicator Dynamics, Social and Individual Learning

As mentioned, 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 first formulate the continuous RD for our specific three-population game. The equation for $\dot{x}_0^0$ (others are similar), derived from Eq. 1, is: $\dot{x}_0^0 = x_0^0(-x_0^2x_0^1 - 2x_0^2 - 4x_0^1 + 3 + x_0^0x_0^2x_0^1 + 2x_0^2x_0^0 + 4x_0^0x_0^1 - 3x_0^0 + x_0^2x_1^0 + 2x_1^0x_0^1 - x_1^0 - x_1^1 + x_0^0x_1^1 + x_1^0x_1^1)$

The three Nash equilibria that need to be investigated are those in pure strategies (σ_a, σ_b, and σ_c). We have analytically checked that only σ_c is an ESS (by means of the divergence operator, to find out where all derivatives are negative). This way, it was verified that the only Nash equilibria where all derivatives are negative is σ_c.

Now we turn to the learning models. In both models reported below, in each time step, agents from each population p play g games in which payoffs are as in Table 2.

For the social learning, we use one of the possibilities mentioned in [6], which is based on an ask-around strategy. This of course involves at least some sampling of other agents' rewards. However, it does not involve sophisticated modeling as perfect monitoring. It works as follows: when dissatisfied with their own rewards, some agents ask around in their social network and eventually change strategy. To replicate this behavior, we use an ask-around rate p_a: at each time step, with probability p_a each agent in the population p copies the strategy of a better performing acquaintance. The higher p_a, the more "anxious" is the agent (i.e., the faster it gets dissatisfied). We recall that, according to [7], it is expected that if the adjustment is not gradual, there may be no convergence to the behavior of the continuous RD.

For the individual learning, no ask-around strategy is used. Rather, agents learn using individual Q-learning (Eq. 2), thus assessing the value of each strategy by means of Q values.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\ (r + \gamma\, max_{a'}\, Q(s',a') - Q(s,a)) \qquad (2)$$

For action selection, ε-greedy was used. In line with the just mentioned issue of gradual adjustments, and from [8], we know that the value of ε is key to reproduce the behavior of the continuous RD.

3 Experiments and Results

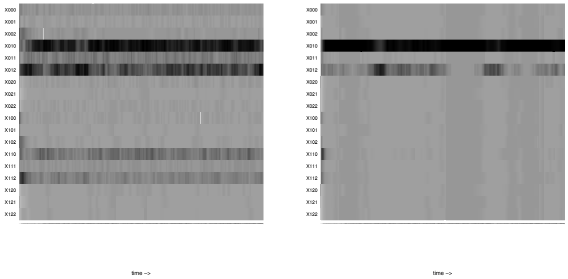
In this section we discuss the numerical simulations of the learning based approaches and compare them with the continuous RD, from which we know the

Nash equilibria, and that the only ESS is profile σ_c. With the learning models, 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_a shifts this pattern?

The main parameters of the model, 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_a were varied. In all cases, at step 0, agents select strategies from a uniform distribution of probability.

We first discuss the social learning. Because five 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. In the plots that appear in Figures 1 to 3, heatmaps are used to convey the idea of the intensity of the selection of each of the 18 joint actions (represented in the vertical axis) along time (horizontal axis), with $\Delta = 1000$. Due to an internal coding used, the 18 joint actions are labeled such that 0, 1 and 2 mean the selection of G, S, or B respectively. The order of the triplet is such that the first digit indicates the action of $p = 2$, the second digit is for the action of $p = 1$, and the third digit is for $p = 0$. In particular, the three Nash equilibria (σ_a, σ_b, and σ_c) are represented as `010` (G2-S1-G0), `012` (G2-S1-B0), and `102` (S2-G1-B0).

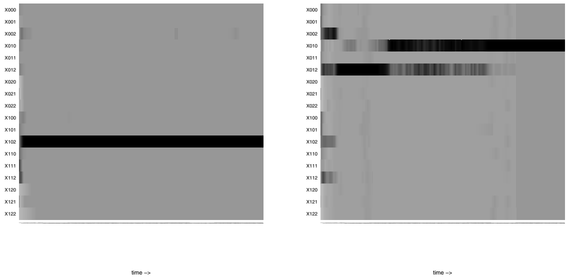
In the heatmaps, to render them cleaner we just use shades of gray color (instead of hot colors as usual). In any case, the darker the shade, the more intense one joint action is selected. Thus we should expect that the three Nash equilibria correspond to the darker strips.



(a) $p_a = 10^{-1}$ (b) $p_a = 10^{-3}$

Fig. 1. Evolution of dynamics for different p_a

In Figure 1 we show how the selection evolves along time, for relatively high p_a. Figure 1(a) is for $p_a = 10^{-1}$, i.e., each agent asks around with this rate. It is possible to see that although σ_a (`010`) is clearly selected more frequently, other joint actions also appear often, as, e.g. `012`. Interestingly, their counterparts `110` and `112`, which differ from `010` and `012` by $p = 2$ selecting S2 instead of G2, also appear relatively often. This indicates that agents in $p = 2$ try to adapt



(a) $p_a = 10^{-4}$ (b) $p_a = 10^{-4}$

Fig. 2. Evolution of dynamics for $p_a = 10^{-4}$

to the other two populations. In the end the performance is poor because this co-adaptation is disturbed by the high rate of changes by the social agents.

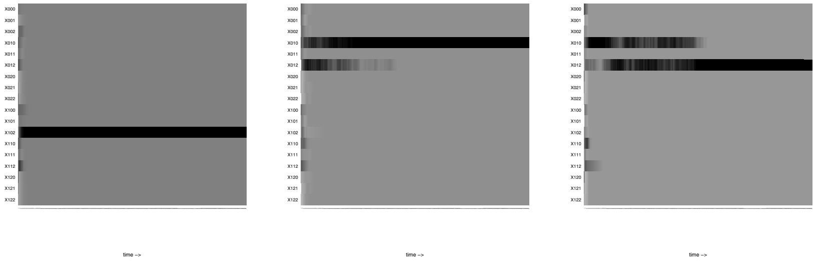
This overall picture improves a little bit with the reduction in p_a. When $p_a = 10^{-2}$ (not shown) and $p_a = 10^{-3}$, (Figure 1(b)) the convergence pattern is clearer but still it is not possible to affirm that one profile has established.

When we decrease the rate to $p_a = 10^{-4}$ (Figure 2), it is possible to observe that one of the two cases occur: either profile `102` (σ_c) establishes right in the beginning (Figure 2(a)), or there is a competition between `010` and `012`, with one or the other ending up establishing. For $p_a = 10^{-5}$ the pattern is pretty much the same as for $p_a = 10^{-4}$.

With further decrease in p_a, the time needed to either `010` or `012` establish decreases, if `102` has not already set. For instance, comparing Figure 3(b) to Figure 2(b), one sees that profile σ_a (`010`) established before in the former case.

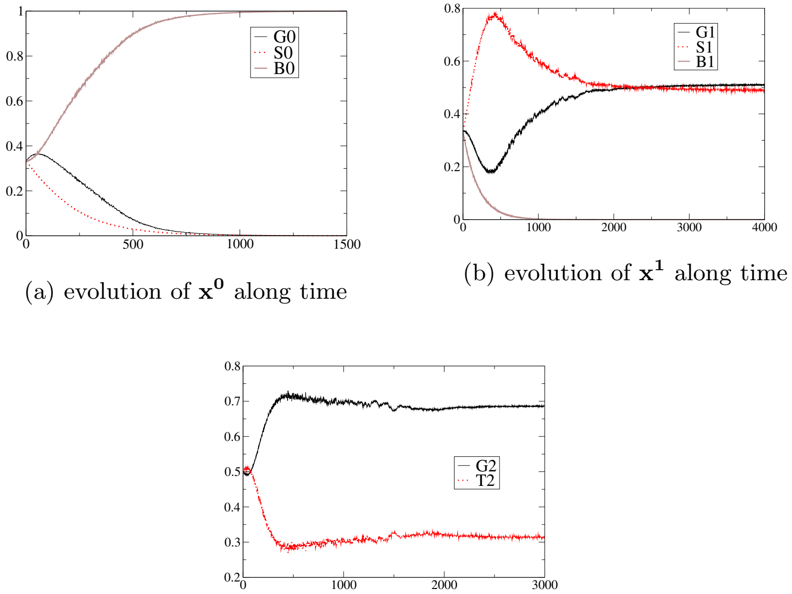
A remark is that the dominated strategy B1, in $p = 1$, is quickly discarded in all cases, even when $p_a = 10^{-1}$.

Regarding the individual 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 $\alpha = 0.5$. We show plots for ε starting at 1.0 with decay



(a) $p_a = 10^{-6}$ (b) $p_a = 10^{-6}$ (c) $p_a = 10^{-6}$

Fig. 3. Evolution of dynamics for $p_a = 10^{-6}$



(a) evolution of $\mathbf{x^0}$ along time

(b) evolution of $\mathbf{x^1}$ along time

(c) evolution of $\mathbf{x^2}$ along time

Fig. 4. Evolution of each $\mathbf{x^p}$ (y axis) along time (x axis) using QL

of 0.995 each time step, which, at the end of $\Delta = 5000$ turns ε of the order of 10^{-11}. Figure 4 depicts the evolution of each component of each vector $\mathbf{x^p}$ (i.e., state of each population in terms of strategies selected) along time. As seen in the three plots, B0 establishes for population $p = 0$, G1 and S1 are selected with near the same probabilities in $p = 1$, and G2 converges to nearly 0.7. This pattern is not a Nash equilibrium. However, the payoffs for the three population of agents are: ≈ 5 ($p = 0$), ≈ 5.4 ($p = 1$), and ≈ 7.3 ($p = 2$), which are very close to σ_d (see Table 3). Note that strategy B1, dominated, is quickly discarded (Figure 4(b)).

How agents have converged to this behavior is better explained by 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. Even a 3d plot, where one could see at least one variable per population, was rendered not informative. Thus we opted to show the behavior of selected variables in a pairwise fashion, namely B0 x G1, B0 x G2, G1 x G2 (Figure 5). It is possible to see that the fraction of agents using G2 (x_0^2), the black line, vertical component, has reached $\approx \frac{3}{4}$ (as in profile σ_d), but this was not stable and there was a shift to lower values, which has influenced also the fraction of agents using G1 (blue circles, vertical), where we also see a shift.

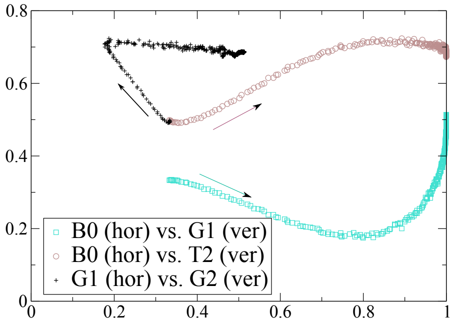


Fig. 5. Trajectories: B0 x G1, B0 x S2, G1 x G2

In short, some conclusions can be drawn from these simulations. First, simultaneous learning by the agents does not always lead to a Nash equilibrium when there are more than one of them, much less to the ESS computed for the corresponding RD of the static NFG. Whereas there is only one ESS among the three Nash equilibria in pure strategies (σ_c), depending on the p_a rate, and on the exploration rate ε, any of the three Nash equilibria (in pure strategies) may establish, or agents may be stuck at a sub-optimal state, even if, as in the case shown here, this state is very close to one of the Nash equilibria. This is in line with the result in [7], which prescribes gradual adjustments. The profile σ_c does establish fast (and does not shift) if it sets at all. When this is not the case, there is a competition between the other two. This competition is determined by agents in $p = 2$: from the payoff matrix (Table 2), one can see that only agents in $p = 2$ have different payoffs in profiles σ_a and σ_b. This leads to an important remark: agents in $p = 2$ have an asymmetric role in this game (not only due to the fact that they have less options), given that the establishment of profile σ_c is also related to a choice by agents in $p = 2$. Thus, in real-world traffic networks, such a study may prove key to determine which fraction of drivers to target when controlling traffic by means of route guidance.

4 Conclusion and Future Work

This paper contributes to the effort of analyzing route choices among population of agents. It shows that the use of models that assume stationarity may fail. An alternative approach is provided that helps the analysis of the dynamics of the RD of the static game, as well as the dynamics provided by two learning methods. In the case of the social learning, convergence is achieved depending on the rate of experimentation in the populations. Thus, anxious drivers may be stuck at sub-optimal equilibrium. For the individual Q-learning, results can be extrapolated to cases in which drivers tend to make too much experimentation (e.g., in response to broadcast of route information). In the case illustrated here, agents converge to a solution close to a Nash equilibrium, which is not an ESS. We remark that the use of a standard macroscopic modeling method (common practice in traffic engineering) would not have provided such insights.

Although the scenario used as illustration considers three populations only, each having a few actions, we claim that this is not an unrealistic simplification. In fact, in the majority of the situations a traffic engineer has to deal with, there is a small number of commodities (origin-destination pairs) that really matter, i.e., end up collecting (thus representing) several sources and sinks. Regarding the number of actions, it is equally the case that in the majority of the real-world cases drivers do not have more than a handful of options to go from A to B.

A future direction is to explicitly consider information broadcast to agents, in order to have a further way to model action selection.

Acknowledgement. The author is partially supported by CNPq. This project was funded by FAPERGS.

References

1. Wardrop, J.G.: Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers 1, 325–362 (1952)
2. Koutsoupias, E., Papadimitriou, C.: Worst-case equilibria. In: Meinel, C., Tison, S. (eds.) STACS 1999. LNCS, vol. 1563, pp. 404–413. Springer, Heidelberg (1999)
3. Papadimitriou, C.H.: Algorithms, games, and the internet. In: Proc. of the 33rd ACM Symp. on Theory of Computing (STOC), pp. 749–753. ACM Press (2001)
4. Roughgarden, T., Tardos, É.: How bad is selfish routing? J. ACM 49, 236–259 (2002)
5. Braess, D.: Über ein Paradoxon aus der Verkehrsplanung. Unternehmensforschung 12, 258 (1968)
6. Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. The MIT Press (1998)
7. Börgers, T., Sarin, R.: Learning through reinforcement and replicator dynamics. Journal of Economic Theory 77, 1–14 (1997)
8. Tuyls, K., Hoen, P.J., Vanschoenwinkel, B.: An evolutionary dynamical analysis of multi-agent learning in iterated games. Autonomous Agents and Multiagent Systems 12, 115–153 (2006)
9. Claus, C., Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 746–752 (1998)
10. Panait, L., Luke, S.: Cooperative multi-agent learning: The state of the art. Autonomous Agents and Multi-Agent Systems 11, 387–434 (2005)
11. Shoham, Y., Powers, R., Grenager, T.: If multi-agent learning is the answer, what is the question? Artificial Intelligence 171, 365–377 (2007)
12. Webb, J.N.: Game Theory – Decisions, Interaction and Evolution. Springer, London (2007)