Adaptive Agents on Evolving Networks

Extended PhD Thesis Abstract

Ardeshir Kianercy
University of Southern California

In this work we study the learning dynamics for agents playing games on networks. We propose a model of network formation in repeated games where players strategically adopt actions and connections simultaneously using a reinforcement learning scheme which is called Boltzmann-Q-learning. This adaptation scheme in the continuous time limit has a proven relation to the evolutionary game theory through replicator dynamics.

We assume that the agents adapt to their environment through a simple reinforcement mechanism. Among different reinforcement schemes, here we focus on (stateless) Q-learning. Within this scheme, the agents' strategies are parameterized through so called Q–functions that characterize relative utility of a particular strategy. After each round of game, the Q functions are updated according to the following rule:

$$Q^i_{xy}(t+1) = Q^i_{xy}(t) + \alpha[R^i_{x,y} - Q^i_{xy}(t)] \tag{1}$$

where $R^i_{x,y}$ is the expected reward of agent x for playing action i with agent y, and α is a parameter that determines the learning rate (which can be set to $\alpha = 1$ without a loss of generality). . Here we focus on Boltzmann action selection mechanism, where the probability x_i of selecting the action i is given by

$$p^i_{xy} = \frac{e^{\beta Q^i_{xy}}}{\sum_{\tilde{y},j} e^{\beta Q^j_{x\tilde{y}}}} \tag{2}$$

where the *temperature* $T > 0$ controls exploration/exploitation tradeoff: for $T \to 0$ the agent always acts greedily and chooses the strategy corresponding to the maximum Q–value (exploitation), whereas for $T \to \infty$ the agents' strategy choices are completely random (exploration).

In the continuos time limit, one obtains the following dynamics describing the evolution of agent x probability choosing action i and play with agent y, with respect to time:

$$\frac{\dot{p}^i_{xy}}{p^i_{xy}} = \sum_j A^{ij}_{xy} p^j_{yx} - \sum_{i,j,\tilde{y}} A^{ij}_{x\tilde{y}} p^i_{x\tilde{y}} p^j_{\tilde{y}x} + T \sum_{\tilde{y},j} p^j_{x\tilde{y}} \ln \frac{p^j_{x\tilde{y}}}{p^i_{xy}} \tag{3}$$

First, we consider the dynamics of Q–learning in two–player two–action games with Boltzmann exploration mechanism. For any non–zero exploration rate the dynamics is *dissipative*, which guarantees that agent strategies converge to rest points that are generally different from the game's Nash Equlibria (NE). We provide a comprehensive characterization of the the rest point structure for different games, and examine the sensitivity of this structure with respect to the noise due to exploration. Our results indicate that for a class of games with multiple NE the asymptotic behavior of learning dynamics can undergo drastic changes at a critical exploration rate.

We demonstrated that, depending on the game, the rest point structure of the learning dynamics is different. Namely, for games with a single NE (either pure or mixed) there is a single globally stable rest point for any positive exploration rate. Furthermore, we examined the impact of exploration (noise) on the asymptotic behavior, and showed that in games with multiple NE the rest point structure undergoes a bifurcation so that above a critical exploration rate only one globally stable solution persists.

We suggest that the latter observation can be useful for validating various hypotheses about possible learning mechanisms in experiments. Indeed, most empirical studies so far has been limited to games with a single equilibrium, such as matching pennies, where the dynamics is rather insensitive to the exploration rate. We believe that for different games (such as coordination or chicken game), the fine–grained nature of the rest point structure, and specifically, its sensitivity to the exploration rate, can provide much richer information about learning mechanisms employed by the agents.

In the next step we investigate the learning dynamics of agents on network. We now make the assumption that the agents' strategies can be factorized as follows:

$$p^i_{xy} = c_{xy} p^i_x \ , \ \sum\nolimits_y c_{xy} = 1, \ \sum\nolimits_i p^i_x = 1. \tag{4}$$

Here c_{xy} is the probability that the agent x will initiate a game with the agent y, whereas p^i_x is the probability that he will choose action i. Thus, the assumption behind this factorization is that the probability that the agent will perform action i does not depend on whom the game is played against. Substituting 4 in 3 yields

$$\dot{c}_{xy} p^i_x + c_{xy} \dot{p}^i_x = c_{xy} p^i_x \Bigg[\sum_j a^{ij}_{xy} c_{yx} p^j_y - \sum_{i,y,j} a^{ij}_{x,y} c_{xy} c_{yx} p^i_x p^j_y - T \left[\ln c_{xy} + \ln p^i_x - \sum_y c_{xy} \ln c_{xy} - \sum_j p^j_x \ln p^j_x \right] \Bigg] \tag{5}$$

Next, we take a summation of both sides in Equation 5, once over y and then over i, and make use of the normalization conditions in

Eq. 4 to obtain the following system:

$$\frac{\dot{p}_x^i}{p_x^i} = \sum_{\tilde{y},j} A_{x\tilde{y}}^{ij} c_{x\tilde{y}} c_{\tilde{y}x} p_{\tilde{y}}^j - \sum_{i,j,\tilde{y}} A_{x\tilde{y}}^{ij} c_{x\tilde{y}} c_{\tilde{y}x} p_x^i p_{\tilde{y}}^j + T \sum_j p_x^j \ln(p_x^j/p_x^i) \quad (6)$$

$$\frac{\dot{c}_{xy}}{c_{xy}} = c_{yx} \sum_{i,j} A_{xy}^{ij} p_x^i p_y^j - \sum_{i,j,\tilde{y}} A_{x\tilde{y}}^{ij} c_{x\tilde{y}} c_{\tilde{y}x} p_x^i p_{\tilde{y}}^j + T \sum_{\tilde{y}} c_{x\tilde{y}} \ln(c_{x\tilde{y}}/c_{xy}) \quad (7)$$

Equations 6 and 7 are the replicator equations that describe the collective and mutual evolution of the agents' strategies and the network structure, by taking into account explicit coupling between the strategies and link weights.

We now consider general two actions games in the case when there is no exploration, $T = 0$. We assume that the reward matrix is the same for all pairs (x, y), $A_{xy} = A$,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad (8)$$

Let p_α, $\alpha \in \{x, y, \dots, \}$, denote the probability for agent α to play action 1. For agent x the expected reward of choosing action one is denoted by r_x^1 and the expected reward of choosing playmate y by r_{xy} and the average expected reward by R_x

$$r_x^1 = \sum_{\tilde{y}} ((a_{11} - a_{12}) p_{\tilde{y}} + a_{12}) c_{x\tilde{y}} c_{\tilde{y}x} \quad (9)$$

$$r_{xy} = c_{yx}(a p_x p_y + b p_x + d p_y + a_{22}) \quad (10)$$

$$R_x = \sum_{\tilde{y}} (a p_x p_{\tilde{y}} + b p_x + d p_{\tilde{y}} + a_{22}) c_{x\tilde{y}} c_{\tilde{y}x} \quad (11)$$

In the last equation we have defined the following parameters,

$$a = a_{11} - a_{21} - a_{12} + a_{22} \quad (12)$$

$$b = a_{12} - a_{22} \quad (13)$$

$$d = a_{21} - a_{22} \quad (14)$$

For $T = 0$, the learning dynamics attains the following form:

$$\frac{\dot{p}_x}{p_x} = r_x^1 - R_x \quad (15)$$

$$\frac{\dot{c}_{xy}}{c_{xy}} = r_{xy} - R_x \quad (16)$$

These equations have a simple intuitive meaning. Indeed, Equation 15 asserts that the probability for agent x to play action 1 increases at a rate that is equal to the expected payoff for playing action 1 relative to the overall payoff R_x. Similarly, Equation 16 reads that the probability for agent x to play with agent y increases at a rate equal to the expected payoff for playing with y relative to the payoff averaged over all the other agents (and strategies).

We should note that generally, the replicator dynamics (and Nash equilibria) in matrix games are invariant with respect to adding any column vector to the payoff matrix. However, this invariance does not hold in the present networked game. The reason for this is the following: if an agent does not have any incoming links (i.e., no other agent plays with him/her), then he always gets a zero reward. Thus, the zero reward of an isolated agent serves as a reference point. This poses a certain problem. For instance, consider the game of Prisoner's Dilemma where the payoff for mutual defection is P: In general, the outcome of the game should not depend on P as long as the structural properties of the payoff matrix is the same.

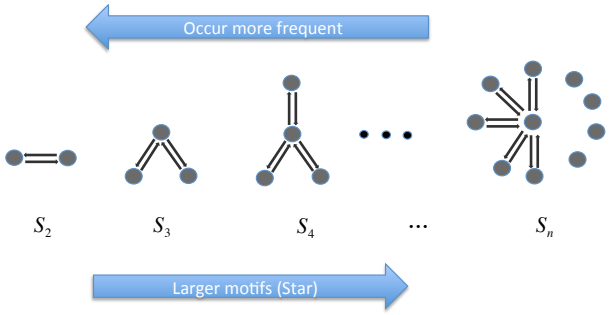

Figure 1: The graphical illustration of motifs structure.

However, in our case the situation is different. Indeed, if $P < 0$, an agent might decide to avoid the game by isolating himself (i.e., linking to agents that do not reciprocate), whereas for $P > 0$ the agent might be better of participating in a game.

To resolve this issue, we assume that every time a partner of agent x refuses to play, x receives a negative payoff $-c_p < 0$, which can be viewed as a *cost of isolation*. It can be shown that the introduction of this cost merely means adding a constant to the reward matrix in the replicator learning dynamics (see Section.**??**). The adjusted reward matrix elements a_{ij} are:

$$a_{ij} = b_{ij} + c_p \quad (17)$$

where B is the game reward matrix and similar for all agents.

We then proceed to a comprehensive analysis for 3-player two-action games which is the minimum system size where structural dynamics is important. In particular, we provide a complete characterization of Nash equilibria in such games, and examine the local stability of the rest-points of the learning dynamics. Our results indicate that at zero temperature the dynamics reaches different configurations based on the cost of isolation. By tuning the cost, the stability of the fixed points changes abruptly at critical values of *cost of isolation*. In the stable configurations agents must choose actions deterministically. In other words, although Coordination and Chicken games allow mixed-strategy Nash equilibria, these equilibria are not stable, and cannot be achieved dynamically via learning.

In addition to the three agent system, we also examined the behavior of the co-evolving system for larger number of agents. Except several specific cases, obtaining analytical results in this case is extremely difficult, and one has to generally resort to numerical integrations and/or simulations.

We observed in those numerical simulations that there is a reciprocating connection between couple of players with only one central player. In other words the network structure of the learning dynamic consists of $star$ motifs. The star graph S_n , is a tree on n nodes with one node having vertex degree $n - 1$ and the other $n - 1$ has vertex degree 1 as it is shown in Figure 1.

As for future work, we intend to go beyond the three–agent systems at $T = 0$ and examine larger systems for any $T > 0$. We also plan to examine large systems asymptotic behavior for full spectrum of exploration rate T and isolation cost c_p values.