An Evolutionary Model of Multi-agent Learning with a Varying Exploration Rate

(Extended Abstract)

M. Kaisers, K. Tuyls
Eindhoven University of Tech.
Eindhoven, The Netherlands
{M.Kaisers, K.Tuyls}
@tue.nl

S. Parsons
Brooklyn College
Brooklyn, New York 11210
parsons
@sci.brooklyn.cuny.edu

F. Thuijsman
Maastricht University
Maastricht, The Netherlands
frank@micc.unimaas.nl

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*

General Terms

Algorithms, Theory, Auctions

Keywords

Multi-agent learning, Evolutionary game theory, Replicator dynamics, Q-learning, Auctions

1. INTRODUCTION

Multi-agent learning is a challenging problem and has recently attracted increased attention by the research community [4, 5]. It promises control over complex multi-agent systems such that agents enact a global desired behavior while operating on local knowledge.

This article contributes to the refinement of the theoretical framework for multi-agent learning, extending an evolutionary model for Q-learning to account for a time dependent temperature. This model of reinforcement learning with a varying exploration rate is tested in the domain of auctions. In particular, we consider the task of learning the probabilistic mix of three predefined trading strategies that yields the highest expected profit in a population of traders that choose between these strategies.

The advanced model shows that explorative learning behavior may overcome local optima and lead to a higher expected payoff than myopic, solely exploitive learning.

This abstract describes the current evolutionary model, derives its extension for a varying exploration rate and then summarizes experiments and results. It is concluded with a discussion of the contributions.

2. METHOD

This section explains the established evolutionary models of reinforcement learning and provides a complete model of Q-learning with a varying exploration rate.

Cite as: An Evolutionary Model of Multi-agent Learning with a Varying Exploration Rate, (Extended Abstract), M. Kaisers, K. Tuyls, S. Parsons, F. Thuijsman, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 1255–1256

The average learning behavior of Cross learning in the continuous time limit has been shown to converge to the replicator dynamics $\dot{x}_i = x_i \cdot (\pi_i(x) - x\pi(x))$, where x_i is the probability of action i to be played and $\pi_i(x)$ denotes the expected payoff of i against x [1].

Stateless Q-learning yields a temperature parameter that allows to balance exploration and exploitation. It proceeds in two steps: Action selection based on the temperature and applying a Q-value update to the played action. The Q-value update rule $Q_{t+1,i} = (1-\alpha)Q_{t,i} + \alpha r_t$ features a learning rate α and is applied to selected action i played in joint action s_t, after receiving the payoff or reward $r_t = u_i(s_t)$.

Given the Q-values, an action is selected based on the Boltzmann distribution, with probability x_i for action i:

$$x_i(Q, \tau_t) = \frac{e^{Q_i \tau_t^{-1}}}{\sum_j e^{Q_j \tau_t^{-1}}} \tag{1}$$

The learning dynamics from updating Q-values under fixed temperature have been proven to converge in their continuous limit to the evolutionary model derived in [6]:

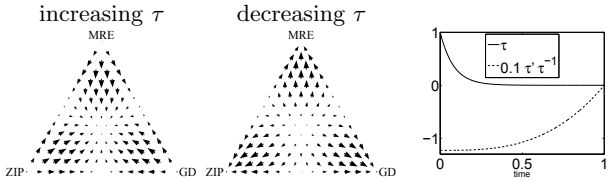
$$\dot{x_i} = \alpha x_i \left[\tau^{-1}\left(\pi_i(x) - x\pi(x)^T\right) + x \log x^T - \log x_i\right]$$

This model allows to study the average behavior of Q-learning with a fixed temperature. However, a change in temperature also alters the policy. Assuming fixed Q-values and a continuous temperature function τ_t, the change is:

$$\frac{d}{dt}x_i = \dot{\tau}_t \dot{x}_i = \frac{\dot{\tau}_t}{\tau_t} x_i \left(-\log\left(x_i\right) + x \log\left(x^T\right)\right) \tag{2}$$

The derived term consists of two factors. The first part $\frac{\dot{\tau}_t}{\tau_t}$ is dependent on t and determines the scale and sign of the resulting force. The remainder is solely dependent on x and can be computed for a grid of points in the simplex to give an intuition of the resulting force. Using $\forall t : \tau_t \geq 0$, the sign is determined by the derivative. As a consequence, the force field plot for the new term always has one of the two forms depicted in Figure 1, either for increasing or for decreasing the temperature. These plots feature arrows in the direction of x which are scaled proportionally to $|x|$. The examples in this article are based on three actions, for which the strategy space of x can be visualized in a simplex.

This figure also plots the selected temperature function over the interval $[0, 1]$ and shows how the term $\frac{\dot{\tau}_t}{\tau_t}$ ensures that the influence of temperature decrease on policy change vanishes over time.



(a) Force fields (b) Magnitude

Figure 1: The force field and its magnitude $\frac{\dot{\tau}_t}{\tau_t}$ that a change in temperature exhibits on the policy.

The complete evolutionary model for the average behavior of Q-learning with a varying exploration rate is given by the sum of the temperature change and learning forces:

$$\dot{x}_{t,i} = \underbrace{\alpha x_i \Big[\tau^{-1}\big(\pi_i(x) - x\pi(x)^T\big) + x\log x^T - \log x_i\Big]}_{\text{learning}} + \underbrace{\frac{\dot{\tau}_t}{\tau_t} x_i \Big(-\log x_i + x\log x^T\Big)}_{\text{temperature change}} \quad (3)$$

3. EXPERIMENTS

The derived model is tested on an example from the auction domain, which uses the same auction setup as Kaisers et al. [2] and applies the extended analysis. In sum, it is a simulated continuous double auction with three trading strategies described in [3]: Modified Roth-Erev (MRE), Zero Intelligence Plus (ZIP) and Gjerstad and Dickhaut (GD). A *heuristic payoff table* as proposed in [7] captures the average profit of each trading strategy for all possible mixtures of strategies in the competition of a finite number of traders.

Figure 2 shows the convergence of 200 example trajectories in the selection model and in the selection-mutation model with decreasing temperature. Each trajectory represents a learning process from a certain initial policy. The policy is not only subject to the forces of learning but rather to a linear combination of the forces of learning as in Figure 2 and temperature change as in Figure 1. It can be observed that the convergence behavior of the selection model which inherently features a fixed temperature is completely captured by the snapshot in the directional field plot. The selection-mutation model on the other hand features a continuously changing force field and cannot be captured by inspection of the directional and force field plots.

An analysis of 1000 trajectories with uniformly sampled initial policies showed the following convergence: In the selection model, 25.1% converged to the pure profile $(0, 0, 1)$ with payoff 78.51 and 74.9% converged to the mixed profile $(0.813, 0.187, 0)$ with payoff 97.27. This yields an expected payoff of 92.56 for the selection model. In contrast to this, 100% of the strategy space converge to $(0.811, 0.189, 0)$ with an expected payoff of 97.25 in the mutation model. The results imply that a population of agents that utilize the exploration scheme to overcome local optima may obtain a higher expected payoff than a population of myopic, absolutely rational learners.

4. CONCLUSIONS

The contributions of this article are two-fold and can be summarized as follows: On the one hand, the evolutionary model of Q-learning has been extended to account for a varying exploration rate. On the other hand, a case study in the domain of auctions has demonstrated that this model may deliver qualitatively different results, going beyond rational learners and considering a more complex model of learning, which may lead to global rather than local optima.

The authors wish to express their gratitude to Jinzhong Niu and Steve Phelps for their support. This research was partially sponsored by a TopTalent 2008 grant of the Netherlands Organisation for Scientific Research (NWO).

5. REFERENCES

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

[2] M. Kaisers, K. Tuyls, F. Thuijsman, and S. Parsons. Auction analysis by normal form game approximation. In *IEEE/WIC/ACM IAT*, 2008.

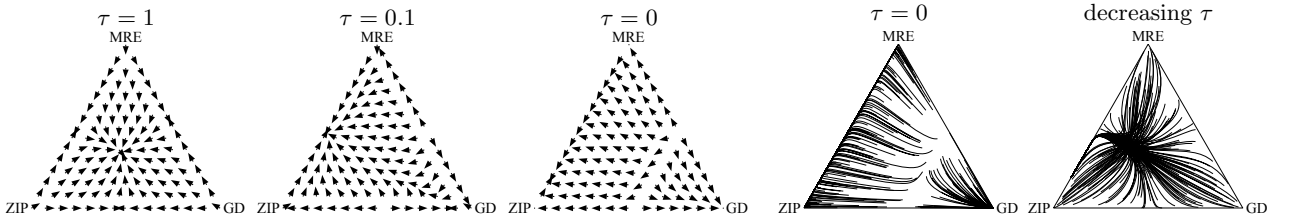
[3] S. Parsons, M. Marcinkiewicz, J. Niu, and S. Phelps. Everything you wanted to know about double auctions, but were afraid to (bid or) ask. Technical report, Brooklyn College, City University of New York, 2900 Bedford Avenue, Brooklyn, NY 11210, USA, 2006.

[4] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artif. Intell.*, 171(7):365–377, 2007.

[5] K. Tuyls and S. Parsons. What evolutionary game theory tells us about multiagent learning. *Artif. Intell.*, 171(7):406–416, 2007.

[6] K. Tuyls, P. 't Hoen, and B. Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12:115–153, 2005.

[7] W. E. Walsh, R. Das, G. Tesauro, and J. O. Kephart. Analyzing complex strategic interactions in multi-agent systems. In P. Gmytrasiewicz and S. Parsons, editors, *Proceedings of the Workshop on Game Theoretic and Decision Theoretic Agents*, pages 109–118, 2002.



(a) Directional field plot of learning dynamics (b) Trajectories with uniform starting positions

Figure 2: (a) Replicator dynamics for a set of fixed temperatures and (b) trajectories of the selection model for Cross learning compared to the new model for Q-learning with a varying exploration rate as in Equation (3)**.**