Modelling the Dynamics of Multiagent Q-learning with ϵ-greedy Exploration

(Extended Abstract)

Eduardo Rodrigues Gomes
Swinburne University of Technology
Hawthorn, 3122, Victoria, Australia
egomes@groupwise.swin.edu.au

Ryszard Kowalczyk
Swinburne University of Technology
Hawthorn, 3122, Victoria, Australia
rkowalczyk@groupwise.swin.edu.au

ABSTRACT

We present a framework to model the dynamics of Multiagent Q-learning with ϵ-greedy exploration. The applicability of the framework is tested through experiments in typical games selected from the literature.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence

Keywords

Multiagent Learning, Reinforcement Learning, Q-learning, ϵ-greedy

1. INTRODUCTION

The development of mechanisms to understand and model the expected behaviour of multiagent learners is becoming increasingly important as the area rapidly find application in a variety of domains [3, 1]. In this paper we present a framework to model the behaviour of Q-learning agents using the ϵ-greedy exploration mechanism. For this, we analyse a continuous-time version of the Q-learning update rule and study how the presence of other agents and the ϵ-greedy mechanism affect it. We then model the problem as a system of difference equations which is used to theoretically analyse the expected behaviour of the agents. The applicability of the framework is tested through experiments in typical games selected from the literature.

2. MULTIAGENT Q-LEARNING

Multiagent Q-learning is a natural extension of single-agent Q-learning to multiagent scenarios. In this approach, the agents are equipped with a standard Q-learning algorithm each and learn independently without considering the presence of each other in the environment. The rewards and the state transitions, however, depend on the joint actions of all agents. The task of a Q-learning agent is to learn a mapping from environment states to actions so as to maximize a numerical reward signal [2]. In each step, the agent receives a signal from the environment indicating its state $s \in S$ and chooses an action $a \in A$. Once the action is performed, it changes the state of the environment, generating a reinforcement signal $r \in R$ that is then used to evaluate the quality of the decision by updating the corresponding $Q(s,a)$ values using the equation $Q(s,a) = Q(s,a) + \alpha(r(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a))$, where $0 < \alpha < 1$ is the learning rate and $0 < \gamma < 1$ is the discount rate. By updating $Q(s,a)$, the agent eventually makes it converge to the $Q^*(s,a)$, the optimal Q-values. The optimal policy is then followed by selecting the actions where the Q^*-values are maximum.

An important component of Q-learning is the action selection mechanism, which is responsible for selecting the actions that the agent will perform during the learning process. Its purpose is to harmonize the trade-off between exploitation and exploration such that the agent can reinforce the evaluation of those actions it already knows to be good but also explore new actions. One of the popular mechanisms for action selection is the ϵ-greedy. This mechanism selects a random action with probability ϵ and the best action, i.e. the one that has the highest Q-value at the moment, with probability 1-ϵ. As such, it can be seen as defining a probability vector over the action set of the agent for each state.

3. A MODEL OF MULTIAGENT Q-LEARNING

We now present our model for multiagent Q-learning with ϵ-greedy exploration. For simplicity of explanation, we consider scenarios composed of 2 agents with 2 actions each and a single state. The reward functions of the agents in this case can be described using 2×2 matrices and the Q-learning update rule can be simplified to $Q_{a_i} = Q_{a_i} + \alpha(r_{a_i} - Q_{a_i})$, where Q_{a_i} is the Q-value for action i of agent a.

We start by rewriting the update rule for the first agent as $Q_{a_i}(k+1) - Q_{a_i}(k) = \alpha(r_{a_i}(k+1) - Q_{a_i}(k))$. This difference equation describes the absolute growth in Q_{a_i} between times k and $k+1$. To obtain its continuous time version, consider $\Delta t \in [0,1]$ to be a small amount of time and $Q_{a_i}(k+\Delta t) - Q_{a_i}(k) \approx \Delta t \times \alpha(r_{a_i}(k+\Delta t) - Q_{a_i}(k))$ to be the approximate growth in Q_{a_i} during Δt. Dividing both sides of the equation by Δt, $\frac{Q_{a_i}(k+\Delta t) - Q_{a_i}(k)}{\Delta t} \approx \alpha(r_{a_i}(k+\Delta t) - Q_{a_i}(k))$, and taking the limit for $\Delta t \to 0$, $\lim_{\Delta t \to 0} \frac{Q_{a_i}(k+\Delta t) - Q_{a_i}(k)}{\Delta t} \approx \alpha(r_{a_i}(k) - Q_{a_i}(k))$, we obtain $\frac{dQ_{a_i}(k)}{dt} \approx \alpha(r_{a_i}(k) - Q_{a_i}(k))$, which is an approximation for the continuous time version of the difference equation.

Cite as: Modelling the Dynamics of Multiagent Q-learning with Є-greedy Exploration, (Extended Abstract), Eduardo Rodrigues Gomes, Ryszard Kowalczyk, *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. 1181–1182

The general solution of the differential equation above can be found by integration: $Q_{a_i}(k) = Ce^{-\alpha t} + r_{a_i}$, where C is the constant of integration. As e^{-x} is a monotonic function and $\lim_{x\to\infty} e^{-x} = 0$, it is easy to observe that the limit of this equation when $t \to \infty$ is r_{a_i}.

If we consider that only the first agent is learning and that the second is using a pure strategy, and assuming that the rewards are noise-free, playing a particular action will always generate the same reward for the first agent. In this case, Q_{a_i} will monotonically increase or decrease towards r_{a_i}, for any initial value of Q_{a_i}. More specifically, the function is monotonically increasing if $Q_{a_i}(0) < r_{a_i}$ and monotonically decreasing if $Q_{a_i}(0) > r_{a_i}$.

If the second agent is using a mixed strategy and the game is played repeatedly, then r_{a_i} can be replaced by $E[r_{a_i}] = \sum_j a_{ij} y_j$, which is the expected payoff of the first agent given the mixed strategy $\mathbf{y}$ of the second. The, the differential equation becomes $\frac{dQ_{a_i}(k)}{dt} \approx \alpha(E[r_{a_i}(k)] - Q_{a_i}(k))$ and its solution is $Q_{a_i}(k) = Ce^{-\alpha t} + E[r_{a_i}]$.

Thus, if the adversary is not learning, Q_{a_i} will move in expectation towards $E[r_{a_i}]$ in a monotonic fashion. With a learning adversary, however, the situation is more complex. A learning adversary can change its probability vector, which affects the expected rewards. Changes in the expected reward will modify the direction field associated with the differential equation and, consequently, the equilibrium points of it. Changes will also modify the limit and, possibly, the direction of Q_{a_i}. Hence, it is important to identify when they will occur.

The ϵ-greedy mechanism updates the probability vector whenever a new action becomes the one with the highest Q-value. Thus, we need to identify the intersection points in the functions of the adversary. It follows that the overall behaviour of the agent depends on these intersection points as they define which values Q_{a_i} will converge to.

During the learning process, the actions have different probabilities of being played meaning that the Q-values are updated at different *speeds*. To simulate this behaviour, we define the growth in the Q-values as directly proportional to the probabilities. The equation becomes $\frac{dQ_{a_i}(k)}{dt} \approx x_i(k)\alpha(E[r_{a_i}(k)] - Q_{a_i}(k))$, where $x_i(k)$ is the probability of playing action i at time k.

Thus, based on the observations above, the expected behaviour for the Q-values can be modelled by the system of equations:

$$Q_{a_i}(k+1) = Q_{a_i}(k) + x_i(k)\alpha(\sum_j a_{ij} y_j(k) - Q_{a_i}(k))$$

$$Q_{b_i}(k+1) = Q_{b_i}(k) + y_i(k)\alpha(\sum_j b_{ij} x_j(k) - Q_{b_i}(k))$$

$$x_i(k) = \begin{cases} (1-\epsilon) + (\epsilon/n), & \text{if } Q_{a_i}(k) \text{ is currently the highest} \\ \epsilon/n, & \text{otherwise} \end{cases}$$

$$y_i(k) = \begin{cases} (1-\epsilon) + (\epsilon/n), & \text{if } Q_{b_i}(k) \text{ is currently the highest} \\ \epsilon/n, & \text{otherwise} \end{cases}$$

where A and B are the payoff matrices, $\mathbf{x}$ and $\mathbf{y}$ are the probability vectors, and Q_a and Q_b are the vectors of Q-values for the first and second agents respectively. Having the model for the Q-values, we can derive the expected behaviour of the agents by tracking the actions with highest Q-value over the learning process of each agent.

4. EXAMPLES

Figure 1 plots the comparison between the theoretical curves of Q-values obtained by the model with the curves found in experiments with the Prisoner Dillema (PD) and the Battle of the Sexes (BS). The experimental curves show the median Q-values over 5000 learning experiments. For the PD the initial Q-values are set to $Q_a = [0, 1]$ and $Q_b = [1, 0]$, and the learning parameters set to $\alpha = 0.1$ and $\epsilon = 0.4$. For the BS the initial configurations are: $Q_a = [2, 1]$, $Q_b = [2, 4]$, $\alpha = 0.1$ and $\epsilon = 0.1$. The payoff tables for the PD and the BS are respectively:

$$\text{A} = \begin{bmatrix} 1 & 5 \\ 0 & 3 \end{bmatrix} \ \text{B} = \begin{bmatrix} 1 & 0 \\ 5 & 3 \end{bmatrix} \qquad \text{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \ \text{B} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

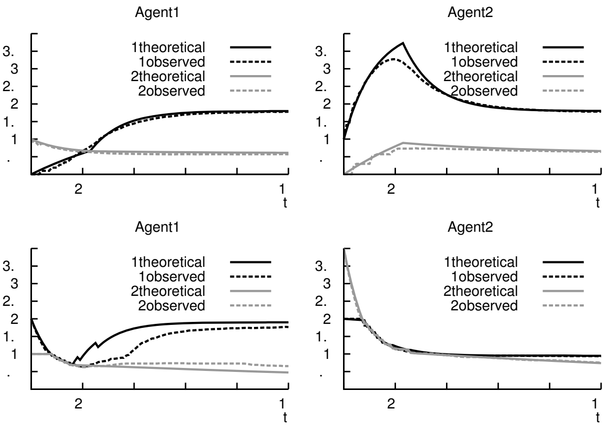

Figure 1: Graphs for the Prisoners Dilemma (upper) and the Battle of the Sexes (lower)

5. CONCLUSIONS

In this paper we have presented a framework to model the behaviour of Q-learning agents using the ϵ-greedy exploration mechanism. For this, we analysed a continuous-time version of the Q-learning update rule and studied how the presence of other agents and the ϵ-greedy mechanism affect it. We then modelled the problem as a system of difference equations which was used to calculate the expected evolution of the Q-values and, consequently, the expected behaviour of the agents. The application of the model in the typical games selected from the literature has shown its feasibility. The model was able to capture all the major trends found in the experiments.

As far as we are aware, none of the existing models for multiagent learning dynamics explores the specific case of Multiagent Q-learning with ϵ-greedy exploration. The next step is to extend the approach to multi-state scenarios.

6. REFERENCES

[1] K. Tuyls et al. A Selection-mutation Model for Q-learning in Multi-agent Systems. In *Proceedings of the AAMAS'03*, pages 693–700. ACM, 2003.

[2] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[3] J. M. Vidal and E. H. Durfee. Predicting the Expected Behavior of Agents that Learn about Agents: the CLRI framework. *AAMAS*, 6(1):77–107, Jan. 2003.