

Empirical and Theoretical Support for Lenient Learning

(Extended Abstract)

Daan Bloembergen, Michael Kaisers, Karl Tuyls
Maastricht University, P.O. Box 616, 6200MD, Maastricht, The Netherlands
{daan.bloembergen, michael.kaisers, k.tuyls}@maastrichtuniversity.nl

Categories and Subject Descriptors

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

General Terms

Algorithms, Theory

Keywords

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

1. RESEARCH SUMMARY

Recently, an evolutionary model of Lenient Q-learning (LQ) has been proposed, providing theoretical guarantees of convergence to the global optimum in cooperative multi-agent learning. However, experiments reveal discrepancies between the predicted dynamics of the evolutionary model and the actual learning behavior of the Lenient Q-learning algorithm, which undermines its theoretical foundation. Moreover it turns out that the predicted behavior of the model is more desirable than the observed behavior of the algorithm. We propose the variant Lenient Frequency Adjusted Q-learning (LFAQ) which inherits the theoretical guarantees and resolves this issue.

The advantages of LFAQ are demonstrated by comparing the evolutionary dynamics of lenient vs non-lenient Frequency Adjusted Q-learning. In addition, we analyze the behavior, convergence properties and performance of these two learning algorithms empirically. The algorithms are evaluated in the Battle of the Sexes (BoS) and the Stag Hunt (SH), while compensating for intrinsic learning speed differences. Significant deviations arise from the introduction of leniency, leading to profound performance gains in coordination games against both lenient and non-lenient learners.

1.1 Games and Learning

Reinforcement learning (RL) tries to maximize the numerical reward signal received from the environment as feedback on performed actions. This paper considers single-state RL. Each time step the agent performs an action i upon which it receives a reward $r_i \in [0, 1]$. Based on this reward the agent updates its policy which is defined as a probability distribution x over its actions, where x_i denotes the probability of selecting action i . The environment will be given by the following games, where the first player chooses

Cite as: Empirical and Theoretical Support for Lenient Learning (Extended Abstract), Daan Bloembergen, Michael Kaisers and Karl Tuyls, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. XXX-XXX.

Copyright © 2011, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

row i , and the second player chooses column j , and their payoff is given by the first and second entry of the matrix position (i, j) respectively.

$$\begin{array}{cc} & \begin{array}{cc} O & F \end{array} \\ \begin{array}{c} O \\ F \end{array} & \begin{pmatrix} 1, \frac{1}{2} & 0, 0 \\ 0, 0 & \frac{1}{2}, 1 \end{pmatrix} \end{array} \quad \begin{array}{cc} & \begin{array}{cc} S & H \end{array} \\ \begin{array}{c} S \\ H \end{array} & \begin{pmatrix} 1, 1 & 0, \frac{2}{3} \\ \frac{2}{3}, 0 & \frac{2}{3}, \frac{1}{3} \end{pmatrix} \end{array}$$

Battle of the Sexes **Stag Hunt**

A classical benchmark reinforcement learning algorithm is single-state Q-learning [6], which uses the action-value update

$$Q_i(t+1) \leftarrow Q_i(t) + \alpha[r_i(t+1) + \gamma \max_j Q_j(t) - Q_i(t)]$$

to refine its reward estimation Q for the taken action i at each time step t ; α controls the learning step size, and γ discounts future rewards. After each update of Q , the new policy is derived using the Boltzmann exploration mechanism that converts the action-value function Q to the probability distribution x :

$$x_i = \frac{e^{Q_i/\tau}}{\sum_j e^{Q_j/\tau}}$$

It has been shown that leniency, i.e., forgiving initial mis-coordination, can greatly improve the accuracy of an agent's reward estimation in the beginning of the learning process [4]. It thereby overcomes the problem that initial mis-coordination might lead to suboptimal solutions in the long run. Leniency towards others can be achieved by having the agent collect κ rewards for a single action before updating the value of this action based on the highest of those κ rewards [4].

The evolutionary model of LQ that delivers the theoretical guarantees is based on the evolutionary model of Q-learning, which was derived under the assumption that all actions are updated equally often [5]. However, the action-values in Q-learning are updated asynchronously: the value of an action is only updated when it is selected. Furthermore, the evolutionary model predicts more rational behavior than the Q-learning algorithm actually exhibits, and therefore [3] introduce the variation Frequency Adjusted Q-learning (FAQ) that simulates synchronous updates by weighting the action-value update inversely proportional to the action-selection probability:

$$Q_i(t+1) \leftarrow Q_i(t) + \frac{1}{x_i} \alpha \left[r(t+1) + \gamma \max_j Q_j(t) - Q_i(t) \right]$$

This paper proposes the Lenient Frequency Adjusted Q-learning (LFAQ) algorithm that combines the improvements of FAQ and Lenient Q-learning. The action-value update rule of LFAQ is equal to that of FAQ; the difference is that the lenient version collects κ rewards before updating its Q-values based on the highest of those rewards. An elaborate explanation of this algorithm can be found in [2].

2. EXPERIMENTS AND RESULTS

This section provides a validation of the proposed LFAQ algorithm, as well as an empirical comparison to non-lenient FAQ. A more elaborate evaluation of the performance of lenient vs. non-lenient learning algorithms can be found in [1].

Figure 1 presents an overview of the behavior of Lenient Q-learning and Lenient FAQ-learning in the Stag Hunt. The action-selection probability of both players' first action is plotted. The figure shows different initialization settings for the Q-values: pessimistic (left), neutral (center) and optimistic (right). The arrows represent the directional field plot of the lenient evolutionary model; the lines follow learning traces of the algorithm. These results show that the behavior of LQ deviates considerably from the evolutionary model, and depends on the initialization. LFAQ on the other hand is robust to different initialization values, and follows the evolutionary model precisely.

Lenient Q-learning

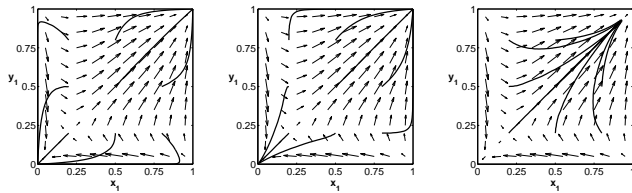


Figure 1: Validating LFAQ-learning.

Figure 2 shows the policy trajectories of FAQ, LFAQ, and a combination of both in Battle of the Sexes and the Stag Hunt. In BoS, LFAQ provides a clear advantage against non-lenient FAQ, indicated by a larger basin of attraction for its preferred equilibrium at $(0, 0)$. In SH, LFAQ outperforms FAQ also in self-play, with a larger basin of attraction for the global optimum at $(1, 1)$.

Finally, Figure 3 shows the average reward over time for FAQ (solid), LFAQ (dotted), FAQ mixed (dashed), and LFAQ mixed (dash-dot). Again, LFAQ has the advantage by achieving either a higher or similar average reward than FAQ.

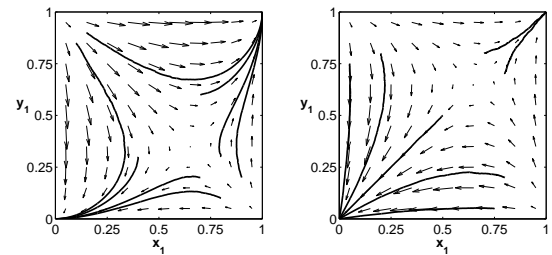
3. CONCLUSION

The proposed LFAQ algorithm combines insights from FAQ [3] and LQ [4] and inherits the theoretical advantages of both. Empirical comparisons confirm that the LFAQ algorithm is consistent with the evolutionary model derived by [4], whereas the LQ algorithm may deviate considerably. Furthermore, the behavior of LFAQ is independent of the initialization of the Q-values. In general, LFAQ performs at least as well as non-lenient learning in coordination games. As such, leniency is the preferable and safe choice in cooperative multi-agent learning.

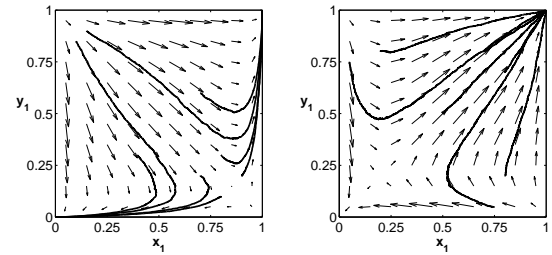
4. REFERENCES

[1] D. Bloembergen, M. Kaisers, and K. Tuyls. A comparative study of multi-agent reinforcement learning dynamics. In *Proc. of 22nd Belgium- Netherlands Conf. on Artif. Intel.*, 2010.

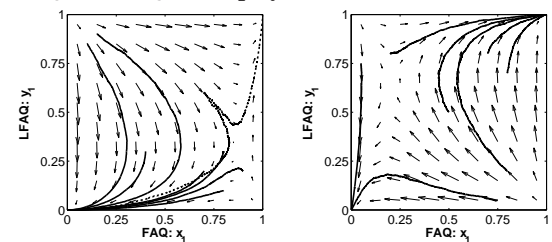
FAQ self play



LFAQ self play



FAQ vs. LFAQ mixed play



Battle of the Sexes

Stag Hunt

Figure 2: Comparing lenient and non-lenient FAQ.

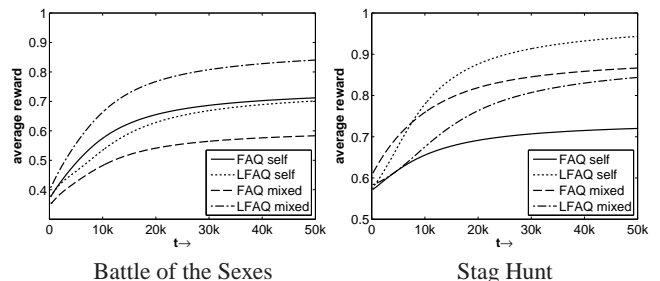
[2] D. Bloembergen, M. Kaisers, and K. Tuyls. Lenient frequency adjusted Q-learning. In *Proc. of 22nd Belgium-Netherlands Conf. on Artif. Intel.*, 2010.

[3] M. Kaisers and K. Tuyls. Frequency adjusted multi-agent Q-learning. In *Proc. of 9th Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 309–315, May, 10-14, 2010.

[4] L. Panait, K. Tuyls, and S. Luke. Theoretical advantages of lenient learners: An evolutionary game theoretic perspective. *Journal of Machine Learning Research*, 9:423–457, 2008.

[5] K. Tuyls, K. Verbeeck, and T. Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *Proc. of 2nd Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, 2003.

[6] C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.



Battle of the Sexes

Stag Hunt

Figure 3: Average reward plot.