A Comparative Study of Multi-agent Reinforcement Learning Dynamics

Daan Bloembergen Michael Kaisers Karl Tuyls

Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Abstract

Multi-agent learning plays an increasingly important role in solving complex dynamic problems in today's society. Recently, an evolutionary game theoretic approach to multi-agent reinforcement learning has been proposed as a first step towards a more general theoretical framework. This article contributes to a better understanding of multi-agent reinforcement learning dynamics in both homogeneous and heterogeneous games using an evolutionary game theory perspective. Simulation experiments are performed in the domain of 2×2 normal form games with the learning algorithms Lenient and non-lenient Frequency Adjusted Q-learning, Finite Action-set Learning Automata and Polynomial Weights Regret Minimization. The results show that evolutionary game theory provides an efficient way to predict the behavior, convergence properties and performance of reinforcement learners. In general, leniency is found to be the preferable choice in cooperative games. Furthermore, the non-lenient learning algorithms do not show significant differences when their intrinsic learning speed is compensated for.

1 Introduction

Recent years have seen an increasing interest in multi-agent learning within the field of Artificial Intelligence (AI) [8]. The dynamic and complex nature of many multi-agent environments makes reinforcement learning (RL) a preferred learning technique in such cases. Single-agent RL has already been studied in much detail and acquired a strong theoretical foundation [9]. This allowed for the construction of proofs of convergence for several RL methods, e.g., Q-learning [12]. However, multi-agent RL still lacks such a general theoretical framework, despite some specific theoretical proofs of convergence (e.g., [2]). Recently, an evolutionary game theoretic approach to reinforcement learning has been taken up that might fill this gap [11].

This article combines the result of simulation experiments with insights from evolutionary game theory, in order to provide a thorough analysis of the qualitative as well as quantitative aspects of reinforcement learning. Such an analysis provides a means to link certain behavioral properties of a learning method to its performance. Furthermore, different RL methods are analyzed both in homogeneous and in heterogeneous games, thereby being able to compare results from both scenarios. Recently, [5] have studied several RL methods in a similar setting. This article contributes to these results by emphasizing the difference between traditional RL methods (such as studied by [5]) and lenient methods, i.e., methods that forgive miscoordination in the initial learning phase. The latter have been shown theoretically to improve convergence in coordination games [7]. This article extends on and confirms these results empirically.

The remainder of this article is structured as follows. Section 2 provides an overview of reinforcement learning and its link to evolutionary game theory. The experimental setup is described in section 3. Sections 4 and 5 present the results of the homogeneous and heterogeneous experiments. Finally, section 6 summarizes and concludes this article.

2 Background

This section presents the necessary background for the experiments performed in this article. We limit the theoretical discussion to a brief overview of reinforcement learning and its relation to evolutionary game theory. For a more extensive introduction to these fields, the reader is referred to [9] and [3], respectively.

2.1 Reinforcement learning

In reinforcement learning (RL), an agent has to learn by trial and error interaction with the environment. It performs action i with probability x_i and receives reward r_i as a feedback that indicates the desirability of the resulting environment state. Two types of RL methods are generally distinguished: value-based methods, which estimate the expected discounted future reward Q_i, from which then a policy x is derived; and policy-based methods, which learn directly in the policy space. This article studies the following four RL methods.

Frequency Adjusted Q-learning (FAQ) [4] is a variation of the value-based Q-learning method, that modulates the learning step size to be inversely proportional to the action selection probability. This modulation leads to more rational behavior that is less susceptible to initial over-estimation of the action values. The update rule for FAQ learning is $Q_i(t+1) \leftarrow Q_i(t) + \min\left(\frac{\beta}{x_i}, 1\right) \alpha \left[r(t+1) + \gamma \max_j Q_j(t) - Q_i(t)\right]$, where α and β are learning step size parameters, and γ is the discount factor. The Boltzmann action-selection mechanism is used with a temperature τ: $x_i = \frac{e^{Q_i \cdot \tau^{-1}}}{\sum_j e^{Q_j \cdot \tau^{-1}}}$.

Lenient FAQ-learning (LFAQ) is a variation of FAQ learning. Leniency has been shown to improve convergence to the optimal solution in coordination games [7]. Leniency is introduced by having the FAQ method collect κ rewards for an action, before updating this action's Q-value based on the highest perceived reward.

Finite Action-set Learning Automata (FALA) [10] is a policy-based learning method. This article considers the Linear Reward-Inaction update scheme. It updates its action selection probability based on a fraction α of the reward received. The probability is increased for the selected action, and decreased for all other actions. The update rules for FALA are $x_i(t+1) \leftarrow x_i(t) + \alpha r_i(t+1)(1 - x_i(t))$ if i is the action taken at time t, and $x_j(t+1) \leftarrow x_j(t) - \alpha r_i(t+1) x_j(t)$ for all actions $j \neq i$.

Regret Minimization (RM) [5] is another policy-based learning method. It updates its policy based on the loss (regret) incurred for playing that policy, with respect to some other policy. This article studies the Polynomial Weights method, that calculates the loss with respect to the optimal policy in hindsight. Again, a learning step size parameter α controls the update process. The method updates the weight of the actions by $w_i(t+1) \leftarrow w_i(t)\left(1 - \alpha l_i(t+1)\right)$. Normalization of these weights gives the action selection probabilities.

2.2 The evolutionary game theoretic approach

Evolutionary game theory (EGT) adopts the idea of evolution from biology to describe how agents optimize their behavior without complete information [6]. It therefore provides a solid basis to study the decision making process of boundedly rational agents in an uncertain environment. EGT utilizes concepts such as evolutionarily stable strategies (ESS) and replicator dynamics (RD) to describe behavior and convergence of a population of agents playing a certain game. Supposing that each player is represented by a population consisting of pure strategies, the fact that a player plays action A with probability p can be translated as a fraction p of the population playing pure strategy A. In a two-player game, the coupled replicator equations that describe the change in the frequency distribution over the pure strategies are given by

$$\frac{dx_i}{dt} = x_i[(Ay)_i - x^T Ay] \tag{1}$$

$$\frac{dy_i}{dt} = y_i[(Bx)_i - y^T Bx] \tag{2}$$

where x (y) is the frequency distribution for player 1 (2), and A (B) represents its individual payoff matrix.

Recently, a formal relation between evolutionary game theory and reinforcement learning has been established, by showing that a specific RL method, Cross learning, converges in the continuous time limit to the replicator dynamics [1]. Based on this result, several authors have derived evolutionary models for different RL methods. Table 1 presents the evolutionary models of FAQ [11], LFAQ [7], FALA [11] and RM [5].

3 Experimental setup

Table 1 shows the four games that are used for the learning experiments. All are 2 × 2 normal form games, meaning that they are two-player games in which each player has to choose between two actions. Three

Table 1: Overview of the evolutionary dynamics of the studied learning methods. Only the dynamics of the first player are given; the dynamics of the second player can be found by substituting B for A, swapping x and y, and swapping the matrix indexes in the u_i rule of LFAQ.

Method	Evolutionary model
FAQ	$\frac{dx_i}{dt} = \frac{\alpha x_i}{\tau}[(Ay)_i - x^T Ay] + x_i \alpha \sum_j x_j ln(\frac{x_j}{x_i})$
LFAQ	$u_i = \sum_j \frac{A_{ij} y_j \left[\left(\sum_{k:A_{ik} \leq A_{ij}} y_k\right)^\kappa - \left(\sum_{k:A_{ik} < A_{ij}} y_k\right)^\kappa\right]}{\sum_{k:A_{ik} = A_{ij}} y_k}$
	$\frac{dx_i}{dt} = \frac{\alpha x_i}{\tau}(u_i - x^T u) + x_i \alpha \sum_j x_j ln(\frac{x_j}{x_i})$
FALA	$\frac{dx_i}{dt} = \alpha x_i[(Ay)_i - x^T Ay]$
RM	$\frac{dx_i}{dt} = \frac{\lambda x_i[(Ay)_i - x^T Ay]}{1 - \lambda[\max_k (Ay)_k - x^T Ay]}$

distinct classes of 2×2 normal form games can be identified [3]. The first class consists of games with one pure Nash equilibrium, such as the Prisoner's Dilemma. The second class of games has two pure and one mixed Nash equilibrium. The Battle of the Sexes, Stag Hunt and Coordination game belong to this class. Finally, the third class of games has only one mixed Nash equilibrium; an example is the Matching Pennies game. Results from the latter are left out as these do not contribute additional insights for the discussion.

$$\begin{array}{c|cc} & C & D \\ \hline C & \frac{3}{5},\frac{3}{5} & 0,1 \\ D & 1,0 & \frac{1}{5},\frac{1}{5} \end{array} \qquad \begin{array}{c|cc} & O & F \\ \hline O & 1,\frac{1}{2} & 0,0 \\ F & 0,0 & \frac{1}{2},1 \end{array} \qquad \begin{array}{c|cc} & S & H \\ \hline S & 1,1 & 0,\frac{2}{3} \\ H & \frac{2}{3},0 & \frac{2}{3},\frac{2}{3} \end{array} \qquad \begin{array}{c|cc} & O & F \\ \hline O & 1,1 & 0,0 \\ F & 0,0 & \frac{1}{2},\frac{1}{2} \end{array}$$

Prisoner's Dilemma | Battle of the Sexes | Stag Hunt | Coordination Game

Figure 1: Normalized payoff matrices for the four different games.

3.1 Convergence speed

When two different learners oppose each other in a game, they may not learn equally fast. This can lead to artifacts caused by the mere difference in learning speed rather than a true differentiation in qualitative learning dynamics. Knowing the relation between the learners' step sizes and their convergence speed makes it possible to select the step sizes in such a way that the different methods learn equally fast in self play, which ensures a fair competition in mixed play.

Table 2 shows the average number of iterations needed to converge for the different learning methods, in the Prisoners' Dilemma using the step size $\alpha = 0.001$. These averages are calculated by running 250 simulations with starting points uniformly distributed over the policy space, and measuring the number of iterations needed for the learner to ϵ-converge ($\epsilon = 0.001$). The other parameters are set to $\beta = \tau = 0.01$ and $\gamma = 0$ for FAQ and LFAQ, and $\kappa = 5$ for LFAQ. These results also provide a means to level the convergence speed of the different algorithms. For example, to level the convergence speed of FAQ and

Table 2: Mean convergence time of the different learners in the Prisoner's Dilemma (rounded averages over 250 simulations with uniformly distributed starting points), learning speed ratio ρ and convergence time given modulated learning rate $\alpha\rho$.

PD	$\alpha = 0.001$	ρ	$\alpha = 0.001\rho$
FAQ	46388	1.00	46388
LFAQ	270003	5.82	46625
RM	38413	0.83	46655
FALA	38116	0.82	46121

Table 3: The learning speed ratio ρ for different games and learners, with respect to FAQ, as calculated from 250 simulations with uniformly distributed starting points for each pair of learners.

	PD	SH	BoS	CG
FAQ	1.00	1.00	1.00	1.00
LFAQ	5.82	7.81	7.20	5.41
RM	0.83	0.97	0.91	0.95
FALA	0.82	0.89	0.90	0.95

RM, the ratio ρ between their respective average numbers of iterations k can be calculated as

$$\rho_{RM,FAQ} = \frac{k_{RM}}{k_{FAQ}}.$$

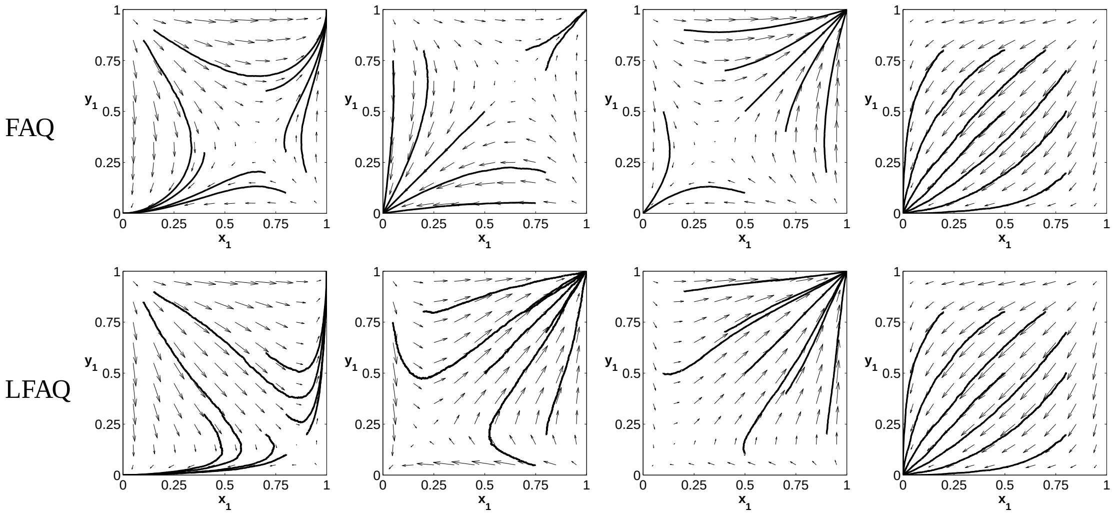
This ratio might differ depending on the game, since the learners' behavior also depends on the game, especially when multiple equilibria are present. The resulting values of ρ are indicated in Table 2, and additional experiments with the modulated learning rate $\alpha\rho$ show that it compensates the intrinsic learning speed differences. The same procedure can be applied to the other games, using the corresponding values for ρ, with similar results, summarized in Table 3. The values found in this table are used throughout the remainder of the experiments.

4 Self play

Self play is the standard form of learning experiments, in which each competing player implements the same learning method. All experiments use the same parameter settings: step size $\alpha = 0.001$ times the ratio given in Table 3; for (L)FAQ, $\beta = 0.01$, $\tau = 0.01$ and $\gamma = 0$. This section describes the behavior of the learning methods in self-play. Convergence properties and performance are evaluated in Section 5 in comparison to mixed play results.

The behavior of a learner over time can be visualized using a trajectory plot or by plotting the directional field of the corresponding replicator dynamics. Here, a combination of both is used to show how the individual learning trajectories relate to their evolutionary prediction. All trajectory plots show the average trajectory over 10 runs of 50,000 iterations each (100,000 for the Prisoner's Dilemma).

Figure 2 shows the difference in behavior of FAQ and LFAQ in the four games studied. FAQ is taken as a representative example of the non-lenient learners, whose behavior is very similar. The results show that both algorithms indeed behave as predicted by their evolutionary model. In the Prisoner's Dilemma, there is not much variation in the trajectories or predictions for the two different learners. As can be seen, all trajectories converge to the game's Nash equilibrium (D,D), which in the plot lies at (0,0). The directional field shows that indeed all possible initial policies will eventually converge to this equilibrium, for both types of learners. In the three games belonging to the second category, with multiple equilibria, a clear distinction can be seen between the three non-lenient learners, FAQ, FALA and RM, and the lenient learner LFAQ.



Battle of the Sexes Stag Hunt Coordination Game Prisoners' Dilemma

Figure 2: Policy trajectories of FAQ and LFAQ in three different games. In these games, the trajectories of FALA and RM are similar to FAQ.

In the Battle of the Sexes, the trajectories converge to the same equilibria for both types of learners, but do so in different ways. The lenient learner has a different mixed equilibrium much closer to (O,F) which indicates that each player sticks to its preferred action much longer by ignoring lower payoffs. The higher the degree of leniency, the closer the mixed equilibrium gets to (O,F).

The Stag Hunt, and to a lesser extend also the Coordination Game, clearly shows the advantage of leniency in cooperative environments. In these games, both player prefer the same equilibrium, in both cases (1,1) in the plot. In the Stag Hunt game, non lenient learners prefer the safer risk dominant equilibrium (0,0), where both hunt for Hare. The lenient learner in this case forgives mistakes made by the other player, and can therefore reach the Pareto optimal equilibrium (S,S) in most cases. In the Coordination game this effect is less strong, since there is no risk dominant equilibrium. However, again the lenient learner is able to reach the Pareto optimal equilibrium more often. This effect is described in more detail in the next section.

5 Mixed play

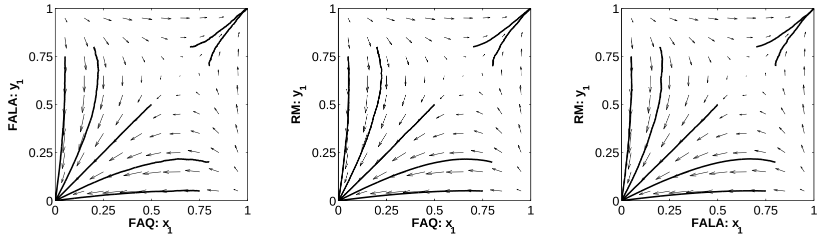
In the mixed play experiments, games are played by heterogeneous pairs of players, meaning that both players implement different learning methods. The results of these experiments are compared with those of the self play experiments in the previous section. This provides insight into how the behavior of a learner depends on the behavior of its opponent. Moreover, the results indicate how well the learners do against different opponents by looking at their performance.

Again, all experiments use the same parameter settings. All methods use step size $\alpha = 0.001$ times the ratio given in Table 3. For (L)FAQ, $\beta = 0.01$, $\tau = 0.01$ and $\gamma = 0$. This section is divided in three parts, describing the behavior, convergence properties, and performance of the learning methods respectively.

5.1 Behavior

The behavior of the learners is again analyzed by running simulations with several different starting points, and plotting the resulting trajectories together with the directional field of the mixed replicator dynamics. For each starting point, 10 simulations are run and the resulting trajectories are averaged. Each simulation consists of 50,000 iterations (100,000 in the Prisoner's Dilemma).

Again, the non-lenient learners behave very similar to each other, and most deviations occur only when LFAQ is involved. For example, Figure 3 shows the behavior of combinations of the three non-lenient learners in the Stag Hunt game. The behavior of these combinations of learners does not show any significant deviation. Moreover, the behavior is very similar to the self play behavior of these learners, see for example the self play of FAQ in Figure 2.



FAQ - FALA

FAQ - RM

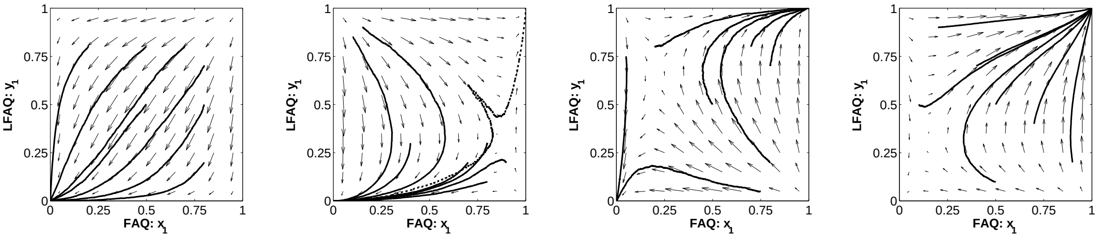
FALA - RM

Figure 3: Policy trajectories of combinations of non-lenient learners in the Stag Hunt game.

When LFAQ is involved, on the other hand, the resulting behavior tends to differ from any of the learners' self play behavior. Figure 4 shows the behavior of a combination of FAQ and LFAQ in four different games. In these games, both FALA and RM in combination with LFAQ show very similar results. The depicted trajectories clearly deviate from the self play behavior of any of the learners. It is interesting to note that the evolutionary prediction is still correct, which shows that also in mixed play the replicator dynamics provide a very useful analytical tool.

5.2 Convergence properties

The effect of the behavioral changes described in the previous section can best be analyzed by looking at the changing basins of attraction of the different games. This analysis is limited to the three games with multiple equilibria, since the Prisoner's Dilemma has only one basin of attraction that either fills the whole policy space, or is empty.



Prisoner's Dilemma Battle of the Sexes Stag Hunt Coordination Game

Figure 4: Policy trajectories of mixed play between FAQ and LFAQ in four different games. The results of the Prisoners' Dilemma resemble self-play behavior. LFAQ has a larger basin of its preferred equilibrium in the Battle of Sexes, and it supports convergence to Pareto optimal equilibria in common interest games. The dotted trajectory line in Battle of the Sexes indicates that not all simulations converged to the same equilibrium for this starting point.

As noted before, the most interesting changes occur when LFAQ is involved, otherwise results resemble self play. Since RM, FALA and FAQ showed almost identical behavior, FAQ is used as a representative for these three non-lenient learners. Figure 5 shows the basins of attraction in the three games for FAQ and LFAQ, both in self play and when the two play against each other. The basins are calculated by iterating the replicator equations over a 100×100 grid, and represented as a solid line indicating their border. The directional field of the replicator dynamics is also shown to provide a clear overview of the convergence properties.

The figures show several interesting properties of mixed play learning. In the Stag Hunt game, for example, both learners have almost opposite basins of attraction in self play, with FAQ converging to (H,H) and LFAQ to (S,S) in the largest part of the policy space. When these two learners play against each other, the resulting basins of attraction appear to be a mix between those two opposites. A similar effect can be seen in the Coordination Game, although in this case the difference is much smaller as the original basins of attraction are more similar.

Also interesting to note is that in the Battle of the Sexes, FAQ and LFAQ show similar convergence properties in self play, but LFAQ profits in the mixed scenario: a larger part of the policy space converges to (0,0), which corresponds to the preferred equilibrium (F,F) of LFAQ, being player 2. When the learners switch sides, again LFAQ 'wins' and FAQ 'loses' in a larger part of the policy space. The results for all combinations of learners are summarized in Table 4.

Table 4: Percentage of the policy space belonging to the basin of attraction of the various equilibria, for different combinations of learners. Pareto optimal equilibria are indicated with $*$.

	SH		BoS		CG	
	(H,H)	(S,S)*	(F,F)	(O,O)	(F,F)	(O,O)*
FAQ self play	74.3	25.7	49.5	49.5	25.7	73.9
FAQ - LFAQ	37.3	62.7	68.3	31.7	16.7	83.3
LFAQ self play	19.0	80.9	49.5	49.5	10.8	89.2

5.3 Performance

The performance of the learners is analyzed by looking at the average reward earned during game play. The average reward of learning method A against learning method B is calculated as the average over 1,000 simulations where A is player 1 and B is player 2, and another 1,000 simulations where B is player 1 and A is player 2. The starting points of the simulations are uniformly distributed over the policy space. These results are compared to the self play results of lenient and non-lenient learners.

Figure 6 shows the average reward over time for FAQ and LFAQ in self play and mixed play. In the Prisoner's Dilemma not much variation is seen in the results, indicating that both learners do equally well in this game. In the Stag Hunt game, LFAQ performs better than FAQ in self play, but it does worse in mixed play. This can be explained by the fact that FAQ still prefers to play action H in the beginning, which leads to a lower payoff for LFAQ when playing S. In the Battle of the Sexes, LFAQ clearly gains from mixed play, whereas FAQ looses. Finally, in the Coordination Game the mixed result lies between the two learners'

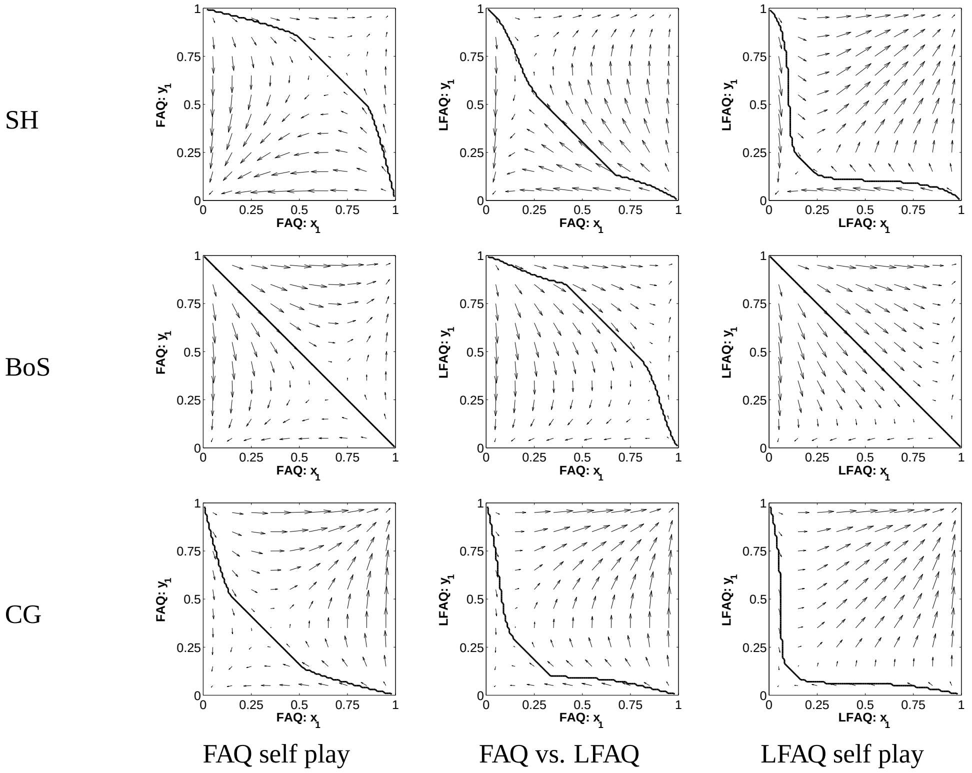
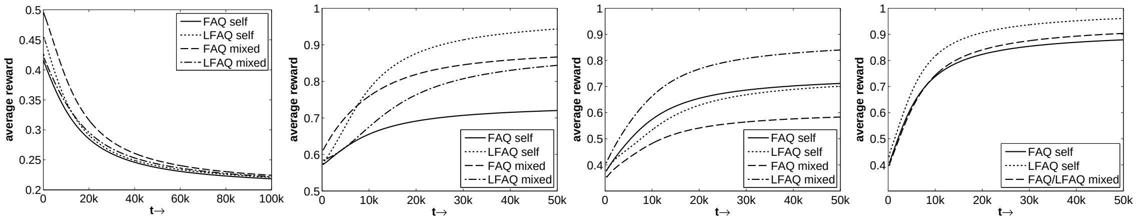


Figure 5: Overview of the basins of attraction of FAQ, LFAQ, and the combination between both learners as representative examples of the difference between non-lenient and lenient learners.

results in self play. In this game, both players always receive the same payoff and therefore their cumulative reward in mixed play is exactly equal.



Prisoner's Dilemma Stag Hunt Battle of the Sexes Coordination Game

Figure 6: Average reward over time for FAQ (solid), LFAQ (dotted), FAQ mixed (dashed), and LFAQ mixed (dash-dot).

It is also possible to calculate the expected average reward of the learners given in Table 5 using the basins of attraction calculated in Table 4 and the games' payoff matrices. These evolutionary expectations are in line with the simulation-based findings presented in Figure 6, which shows that the replicator dynamics are not only useful in describing the behavior and convergence of the learners, but can also accurately predict their performance.

6 Discussion and Conclusions

A method has been designed to level the convergence speed of different learners in self play by calculating a modulation factor for the learning step size. This is required to ensure a fair competition in the case of mixed play, and to rule out artifacts based on quantitative rather than qualitative differences between the learners. The self play experiments, described in Section 4, show that all learners behave as predicted by their evolutionary models. There are notable differences between the behavior of LFAQ and the other, non-

Table 5: Expected average reward for FAQ and LFAQ in self play and mixed play, based on the games' basins of attraction and payoff matrices. These results show that leniency is a weakly dominant choice in cooperative games, as it achieves at least as high reward against any opponent.

	SH		BoS		CG	
	Player 1	Player 2	Player 1	Player 2	Player 1	Player 2
FAQ self play	0.75	0.75	0.74	0.74	0.87	0.87
FAQ - LFAQ	0.88	0.88	0.66	0.84	0.92	0.92
LFAQ self play	0.94	0.94	0.74	0.74	0.95	0.95

lenient learners. In common interest cooperative games such as the Stag Hunt, LFAQ converges to the Pareto dominant equilibrium more often than the other learners, thereby achieving a higher average reward.

Similar effects are seen in the mixed play experiments of Section 5. Again, there is a difference between the non-lenient learners, and the lenient learner LFAQ. Most notably, in the Battle of the Sexes LFAQ is able to push the learning process towards its preferred equilibrium more often than the non-lenient learners, leading to a higher average reward in mixed play for LFAQ and a lower reward for its opponent. In the cooperative games with common interest, LFAQ 'teaches' its opponent to converge to the Pareto optimal equilibrium more often, which leads to a higher payoff for the other player and a lower payoff for LFAQ itself. In general, LFAQ performs at least as well against a specific opponent as the other investigated learners do. As such, it is the preferable and safe choice for cooperative games.

Furthermore, it has been shown that the replicator dynamics can efficiently describe the behavior and convergence properties of the learners both in self play and in mixed play. Moreover, the replicator dynamics can be used to predict the performance of the learners in specific games, using the game's basins of attraction and payoff matrix to compute the expected average reward of the learners.

References

[1] T. Börgers and R. Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77:1–14, 1997.

[2] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

[3] H. Gintis. *Game Theory Evolving*. University Press, Princeton, NJ, 2nd edition, 2009.

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

[5] T. Klos, G.J. Van Ahee, and K. Tuyls. Evolutionary dynamics of regret minimization. Technical report, 2010.

[6] J. Maynard Smith and G. R. Price. The logic of animal conflict. *Nature*, 246(2):15–18, 1973.

[7] 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.

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

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

[10] M.A.L. Thathachar and P.S. Sastry. Varieties of learning automata: An overview. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 32(6):711–722, 2002.

[11] K. Tuyls, P.J. 't Hoen, and B. Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Journal of Autonomous Agents and Multi-Agent Systems*, 12(1):115–153, 2006.

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