Lenient Learning in a Multiplayer Stag Hunt

Daan Bloembergen Steven de Jong Karl Tuyls

Department of Knowledge Engineering
Maastricht University, P.O. Box 616, 6200 MD Maastricht

Abstract

This paper describes the learning dynamics of individual learners in a multiplayer Stag Hunt game, focussing primarily on the difference between lenient and non-lenient learning. We find that, as in 2-player games, leniency significantly promotes cooperative outcomes in 3-player games, as the basins of attraction of (partially) cooperative equilibria grow under this learning scheme. Moreover, we observe significant differences between purely selection-based models, as often encountered in related analytical research, and models that include mutation. Therefore, purely selection-based analysis might not always accurately predict the behavior of practical learning algorithms, which often include mutation.

1 Introduction

In this paper, we investigate the learning dynamics of individual learners in a generalized Stag Hunt game with more than two players, as proposed by Pacheco et al. [5]. We are mainly interested in the dynamics exhibited by a lenient learner in this game, as leniency has shown to be advantageous in a 2-player setting [2].

While extensive research has been performed in the domain of 2-player games [10], multiplayer games, which are much closer to real-world interactions, have started to receive attention only recently [3]. Rather than focusing on the N-player Prisoners' Dilemma (NPD), as is often done in literature, we follow Pacheco et al. [5] in their argument that the N-player Stag Hunt (NSH) is a more interesting and more appropriate game. In the Prisoners' Dilemma, any approach that establishes cooperative outcomes is in a way flawed, as cooperation is dominated by defection. Only in the iterated game may cooperation be a viable alternative, e.g., against a tit-for-tat strategy [1]. The Stag Hunt allows for far more interesting dynamics, as even in the one-shot game, there are two strong Nash equilibria. The cooperative equilibrium is payoff-dominant, while the defective equilibrium minimizes risk [8].

In addition to following the proposed generalization from 2- to multiplayer games, we generalize by considering a broader range of strategy-update approaches. Pacheco et al. [5] use a purely selection-based analytical approach, inspired by statistical physics. In addition, we also consider the possibility of updating strategies through individual (reinforcement) learning. We compare the learning dynamics of a purely selection-based learning algorithm, Finite Action-set Learning Automata (FALA) [9], with those of two learning algorithms that allow for mutation, namely lenient and regular Frequency-Adjusted Q-learning (FAQ-learning) [4]. Lenient FAQ-learning [2] is based on the concept of leniency [6], i.e., ignoring low rewards that may be due to initial mis-coordination. Especially in the initial phase of the learning process, actions that would have been optimal if agents coordinated correctly, may receive a suboptimal payoff because other agents are still learning to coordinate as well. This may drive agents to a suboptimal outcome. Leniency has been shown to alleviate this issue [7].

The paper is structured as follows. In the following section, we will present the multiplayer Stag Hunt game proposed by Pacheco et al. [5]. We also present an analysis of this game in terms of critical parameter settings and equilibria. In Section 3, we discuss Lenient FAQ-learning. Section 4 describes our experimental setup, i.e., the replicator equations corresponding to the learning algorithms (FALA, FAQ, Lenient FAQ), as well as results, i.e., how the different learning algorithms influence the basins of attraction of the multiplayer Stag Hunt equilibria. We conclude in Section 5.

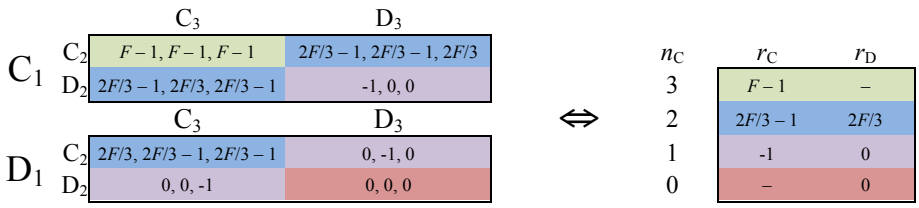

Figure 1: The full payoff table of a three-player Stag Hunt game (left) may be represented more compactly (right), due to the symmetric nature of the Nash equilibria.

2 Multiplayer Stag Hunt

The Stag Hunt [8] is a well known coordination game, in which two players go out on a hunt together. If they cooperate, they have a high chance of capturing a stag, constituting a high reward. On their own, the players can only hope to capture a hare, yielding a lower payoff. Should one player try to cooperate, while the other chooses to hunt alone (defects), the cooperator will fail and get nothing, whereas the defector can still get a hare. In this game, defection is a safe strategy as reasonable payoff is guaranteed independent of the other player's actions. Cooperation poses the risk of being left with nothing, but is more rewarding if the other player cooperates as well. This simple scenario is highly interesting as it maps perfectly to various other scenarios of human interaction involving social contract [8].

The *n*-player Stag Hunt

A straightforward generalization to an n-player Stag Hunt (NSH) has been defined by Pacheco et al. [5]. Suppose there are n players involved in the game. Cooperating incurs a cost c, defecting is free of charge. There is a threshold $m \leq n$ that defines the minimum number of cooperators needed to produce a public good. Above this threshold, the value of this public good depends linearly on the number of cooperators, n_C. The value is defined as $n_C \cdot F \cdot c$, where F is a multiplication factor. As a result, the payoff for a defector is given by $\prod_D = (n_C F c/n)\theta(n_C - m)$, where $\theta(x)$ is the Heaviside step function satisfying $\theta(x < 0) = 0$ and $\theta(x \geq 0) = 1$. The payoff for cooperators is given by $\prod_C = \prod_D - c$. For $m = 0$, the game is an n-player Prisoners' Dilemma, or discretized Public Goods Game.

Representing and analyzing the game

Analyses of normal-form games are generally limited to two players, since adding a third player requires a three-dimensional payoff table. Bukowski et al. [3] elaborately analyze and classify three-player normal-form games. Our analysis is aimed to be more intuitive, as we only focus on a specific type of games. An example of a payoff table for the 3-player Stag Hunt, flattened to two dimensions, is provided in Figure 1 (left). Given the definition of a Nash Equilibrium (i.e., no player can gain from unilaterally changing their strategy) and the fact that we look at a symmetric game (the players share a common payoff table), we may represent the NSH in a more compact payoff table, as shown for an example game in Figure 1. Players' strategy changes correspond to diagonal movements in this table ($\searrow$ or $\nwarrow$). The example shows a Stag Hunt with $n = 3$, $m = 2$, $c = 1$. For instance, $n_C = 3$ is a Nash Equilibrium if no player has the incentive to defect (move $\searrow$); for this to happen $F - 1 > 2F/3$ must be the case. We note that possible mixed equilibria are not visualized in either the full or the compact table.

The NSH can be shown to have *two* critical settings for the parameter F [5]. We demonstrate this for the example game in Figure 2. In this example game ($n = 3$, $m = 2$, $c = 1$), the two critical values of F are 3 and 1.5. Regardless of F, the fully defective joint strategy $n_C = 0$ is always a strong Nash Equilibrium. For $F > 3$, there is no incentive to deviate from $n_C = 3$ (a single deviating cooperator would obtain $8/3 < 3$), so $n_C = 3$ is a strong equilibrium. For $1.5 < F < 3$, we find a strong equilibrium for two cooperators and one defector ($n_C = 2$). A deviating cooperator obtains $0 < 1/3$, whereas a deviating defector obtains $1 < 4/3$. For $F < 1.5$, only the fully defective equilibrium remains.

More generally, there are two interesting regions of the compact payoff table, as visualized in Figure 3. For n players and a minimal coalition size of m, the interesting regions are around $n_C = n$ and $n_C = m$. As can be seen in the leftmost table, player switches from the fully cooperative equilibrium if $(n-1)F/n > F - c$, or $F < cn$. However, this also means that $(n-2)F/n > (n-1)F/n - c$, so $n_C = n - 1$ is not an equilibrium either, et cetera. Thus, $n_C = n$ is the only possible equilibrium in the leftmost table, i.e. for $F \geq cn$. In the rightmost table, we show that the situation changes below $n_C = m$, which may be an equilibrium. This minimal cooperative equilibrium is destroyed if $mF/n - c < 0$, i.e. if $F < cn/m$.

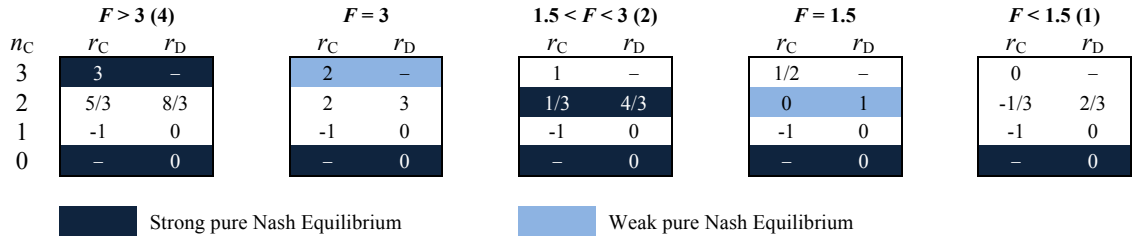

Figure 2: Different equilibria result from the parameter F ($n = 3$, $m = 2$, $c = 1$).

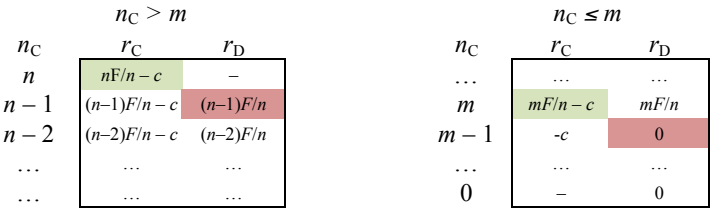

Figure 3: Interesting regions of the generalized compact payoff table.

As a result, the two critical parameter settings are $F = cn$ and $F = cn/m$. For $F > cn$, we expect full cooperation. For $cn > F > cn/m$, we expect m cooperators and $n - m$ defectors. For $F < cn/m$, we expect full defection. There are two exceptions: if $m = 1$, these two settings are identical, i.e. $F = cn$; if $m = n$, the only critical setting is $F = c$ since $n_C \not> m$.

3 Lenient Frequency Adjusted Q-learning

Many multi-agent learning scenarios take the form of coordination games. When multiple independent agents learn together in such an environment, it can often happen that they converge to suboptimal solutions. Initial mis-coordination on a globally optimal solution may result in decreased payoffs, and as a result the learner's preference for the corresponding action may also decrease. In the end, this can drive the agents away from the global optimum, resulting in suboptimal behavior. This effect can be reduced by introducing *leniency*, i.e., by ignoring initial mis-coordination. It has been shown that leniency can greatly improve the accuracy of an agent's projection of the search space in the beginning of the learning process [6]. It thereby overcomes the problem that initial mis-coordination might lead to suboptimal solutions in the long run.

Based on this notion of leniency, Panait et al. [7] presented an evolutionary model of lenient reinforcement learning and demonstrated the advantages of this model over traditional, non-lenient learners. In this model, leniency towards others is achieved by having a learning agent collect κ rewards (or payoffs) for a single action before it updates the value of this action based on the highest of those κ rewards. This results in a fixed degree of leniency, expressed by the value of κ. In previous work, we proposed a reinforcement learning algorithm called Lenient Frequency Adjusted Q-learning (Lenient FAQ) that implements the evolutionary model exactly, thereby inheriting its advantages [2].

The Lenient FAQ algorithm is a lenient version of Frequency Adjusted Q-learning (FAQ), a recent improvement over traditional Q-learning that is less sensitive to different initializations and therefore more robust [4]. The learning update rule of single-state FAQ-learning is given by

$$Q_i(t+1) \leftarrow Q_i(t) + min\left(\frac{\beta}{x_i}, 1\right) \cdot \alpha \left[r_i(t+1) - Q_i(t)\right] \tag{1}$$

where $Q_i(t)$ is the estimated value of action i at time t, x_i is the probability of selecting action i according to the learner's policy, β controls the frequency adjustment, α is a step size parameter, and $r_i(t)$ is the reward received for taking action i at time t. FAQ-learning only updates the value of the action that was previously selected; the value of all other actions remains unchanged. Based on the action-value function Q, a new policy can derived, e.g. by using the Boltzman exploration mechanism

$$x_i = \frac{e^{Q_i \cdot \tau^{-1}}}{\sum_j e^{Q_j \cdot \tau^{-1}}}. \tag{2}$$

FALA	$u_i = \sum_j \sum_k A_{i,j,k} y_j z_k$
	$\frac{dx_i}{dt} = \alpha x_i (u_i - x^T u)$
FAQ	$u_i = \sum_j \sum_k A_{i,j,k} y_j z_k$
	$\frac{dx_i}{dt} = \frac{\alpha x_i}{\tau}(u_i - x^T u) + x_i \alpha \sum_j x_j ln(\frac{x_j}{x_i})$
Lenient FAQ	$u_i = \sum_j \sum_k A_{ijk} y_j z_k \frac{\left[\left(\sum_{m,n:A_{imn} \leq A_{ijk}} y_m z_n\right)^\kappa - \left(\sum_{m,n:A_{imn} < A_{ijk}} y_m z_n\right)^\kappa\right]}{\sum_{m,n:A_{imn}=A_{ijk}} y_m z_n}$
	$\frac{dx_i}{dt} = \frac{\alpha x_i}{\tau}(u_i - x^T u) + x_i \alpha \sum_j x_j ln(\frac{x_j}{x_i})$

Table 1: Overview of the evolutionary dynamics of FALA [11], FAQ [11, 4] and Lenient FAQ [7], for three players with strategies x, y and z, and three-dimensional payoff matrix A.

This mechanism uses a temperature parameter τ to control the balance between exploration and exploitation. A high temperature drives the mechanism towards exploration by leveling the action selection probabilities, whereas a low temperature promotes exploitation by favoring actions with a high Q-value.

Lenient FAQ is based on the same mechanism as the lenient evolutionary model: κ rewards are collected for a particular action before the Q-value of that action is updated based on the highest of those rewards. The Q-value update itself it equal to that of standard FAQ, given in Equation 1.

It has been shown both theoretically [7] and empirically [2] that leniency is an advantageous strategy in 2-player coordination games. Increasing the degree of leniency makes it possible to guarantee an arbitrary high certainty of converging to the global optimum of the game.[1] In this paper, we investigate lenient learning in a coordination game with more than 2 players.

4 Experiments and Results

In this section we describe the experiments performed for the analysis of the game, together with the results. We compare lenient and non-lenient FAQ-learning, as well as the Finite Action-set Learning Automata (FALA) algorithm with a Linear Reward-Inaction (L_{R-I}) learning scheme [9]. The latter algorithm is chosen because it only includes selection, and no mutation. The algorithms are compared using various parameter settings for the 3-player Stag Hunt game defined in Section 2; of main importance for the analysis is the basin of attraction for the different pure strategy Nash equilibria.

Basin of Attraction

A basin of attraction for a certain equilibrium is defined as the region of the policy space for which learning will eventually converge to that equilibrium. In order to calculate a basin, we iterate over the evolutionary model (replicator equations) of the learning algorithm at hand. Starting from 1,000,000 uniformly-spaced points in the policy space, we evaluate to which equilibrium the dynamics converge. This way, we are able to calculate the region of the policy space that constitutes the basin of attraction for each equilibrium.

The evolutionary models are given in Table 1. In the replicator equations, x is the player's current strategy, u_i is the expected payoff for playing action i against two other players with strategies y and z, and therefore $x^T u$ is the average expected payoff. The probability of selecting action i, denoted as x_i, increases whenever the expected payoff of action i is larger than the average expected payoff, and decreases when the expected payoff is lower. For Lenient FAQ, the expected lenient payoff of an action is calculated by weighting each possible payoff by the probability that it is the maximum of κ random trials, as expressed by the extra summation terms in the equation for u_i [7].

As can be seen, FALA only includes selection, whereas FAQ and Lenient FAQ also include a mutation term. The influence of this term depends on the temperature τ, a high temperature leads to more mutation, whereas a low temperature prefers selection. For these experiments, we set the temperature to a low value (0.01) to enable convergence to strong equilibria. Furthermore, for FALA we set $\alpha = 0.001$, and for FAQ and Lenient FAQ we set $\alpha = 0.00001$, $\beta = 0.01$ and $\kappa \in \{3, 5\}$. The step size α is deliberately kept low to

[1] We note that lenient learning may not be a best response against itself; tailored counter-strategies that exploit leniency may exist in some classes of games.

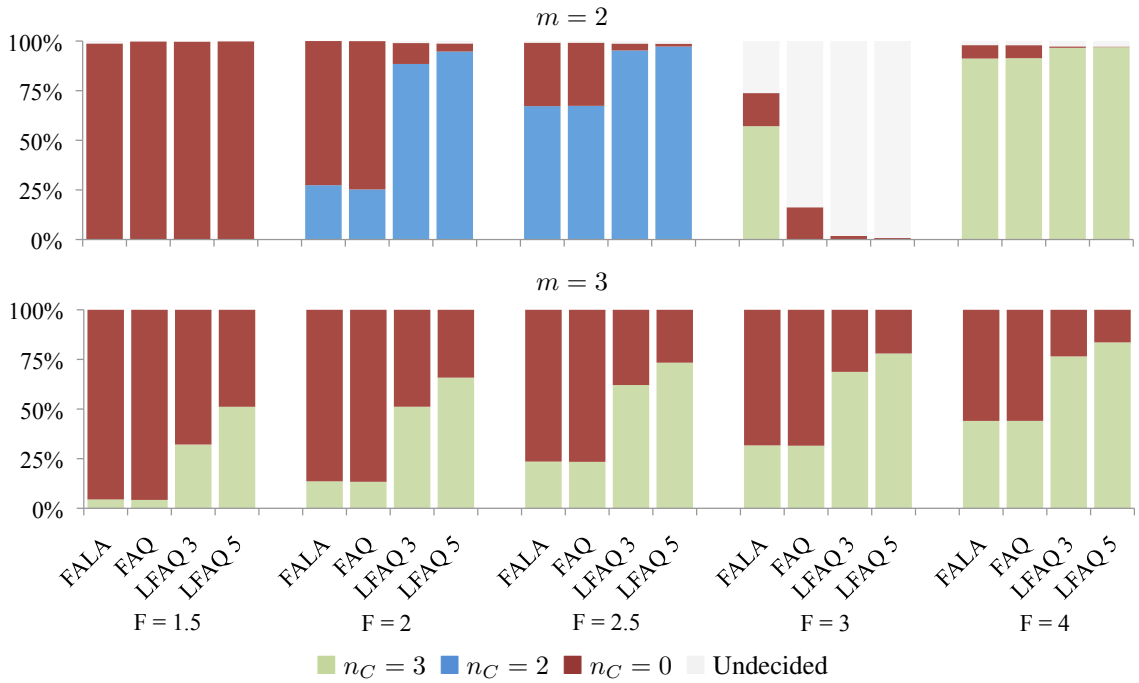

Figure 4: This chart shows the basin of attraction, in percentage of the policy space, for each of the pure Nash equilibria, for $n = 3$, $c = 1$, and varying m and F.

ensure that the dynamics of the actual learning algorithms closely resemble those of the evolutionary model. The other given settings are known to work well from experience [2].

Results for $n = 3$, $c = 1$, and varying m and F are visualized in Figure 4. For $m = 1$ the game is fully defective for $F < n$ and fully cooperative for $F > n$. This is not depicted in the figure.

For $m = 2$ (top), the two critical values are $F = 1.5$ and $F = 3$, as is clearly visible in the chart. For $F \leq 1.5$, the game is an NPD and therefore fully defective. For $1.5 < F < 3$ we observe the expected equilibria ($n_C = 2$ and $n_C = 0$). FALA and FAQ perform similarly, whereas Lenient FAQ increases the basin of attraction for the minimally cooperative equilibrium. For $F = 3$, we clearly observe a turning point where defect is still a strong NE, but the weak cooperative NE has a larger basin of attraction for FALA. FAQ and Lenient FAQ include mutation, and therefore do not fully converge. Instead, they end up in between the fully and partially cooperative equilibria, depending on the degree of leniency. This special scenario is analyzed in more detail below. For $F > 3$, the game is a NSH with a larger preference for the fully cooperative equilibrium. Here again Lenient FAQ has a larger basin for the fully cooperative equilibrium than the non-lenient learners.

For $m = 3$ (bottom), the critical value lies at $F = 1$ (see also Figure 3). The game is an NPD for $F \leq 1$, and a NSH for $F > 1$. We visualize only settings for F that yield an NSH as any NPD leads to a fully defective game for all algorithms. For $m = 3$ and $F > 1$ it is clear that FALA and FAQ again perform similarly, i.e., mutation does not play a big role as long as the temperature is kept low. In contrast, Lenient FAQ has a larger basin for the fully cooperative equilibrium, increasing with the degree of leniency.

(Non-)Convergence to weak Nash equilibria

As mentioned previously, the scenario where $m = 2$ and $F = 3$ proves interesting, as it is a switching point between two different classes of the game. In this scenario, $n_C = 0$ is a strong NE, and $n_C = 3$ is a weak NE. This means that, at the equilibrium $n_C = 3$, no player has an incentive to switch, but at the same time no player has an incentive to stay: both actions result in the same payoff. In order to analyze this scenario in more detail, we study the evolutionary dynamics as well as actual learning traces in the three-dimensional policy space, shown in Figure 5.

The learning traces are calculated by running the actual learning algorithms, starting at 27 uniformly distributed points in the policy space, and playing the iterated NSH. The simulations are run for 200,000 iterations to ensure approximate convergence (if it occurs), and results are averaged over 10 experiments to produce smooth learning traces. As there are only two actions in this game, $x_1 = 1 - x_2$ and therefore it is sufficient to only look at the probability of selecting the first action, x_1, without loss of information. Plotting

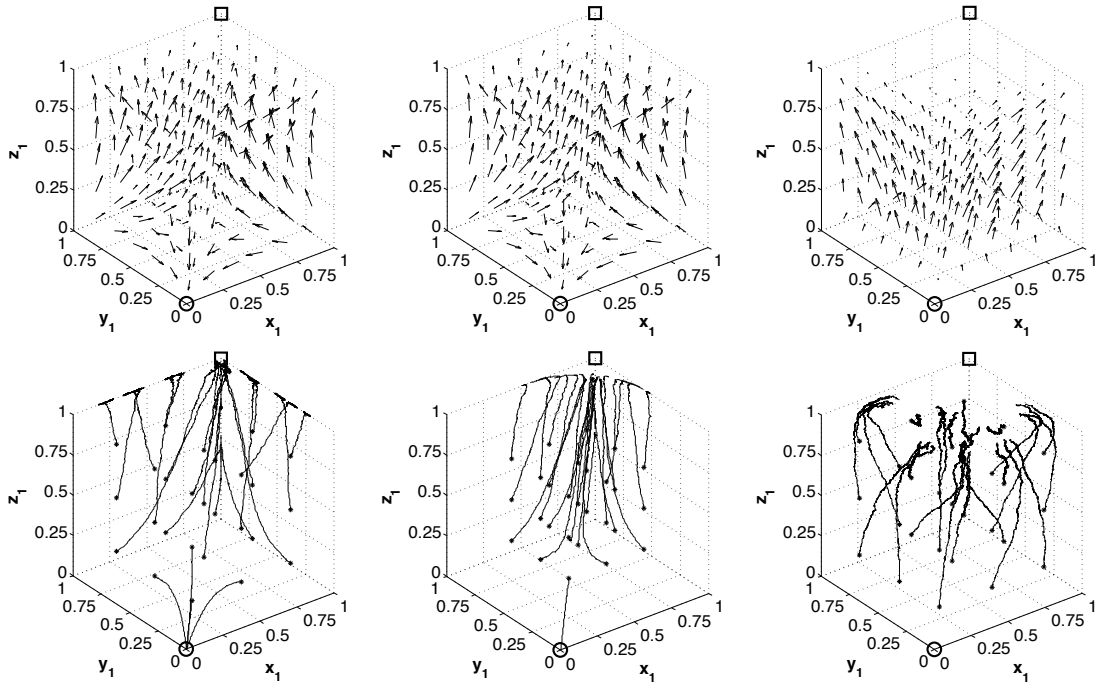

Figure 5: Evolutionary dynamics (top) and actual learning traces (bottom) of three different learning algorithms for $m = 2$ and $F = 3$. This game has two pure Nash equilibria, $\bigcirc$ ($n_C = 0$, strong NE) and $\square$ ($n_C = 3$, weak NE).

x_1 against y_1 and z_1 provides insight in to the learning dynamics of the different algorithms.

Figure 5 clearly illustrates the differences in behavior of the learners. We observe a notable distinction between the purely selection-based learning algorithm, FALA, and the selection-mutation algorithms FAQ and Lenient FAQ. Whereas FALA still converges to the weak equilibrium, FAQ and most notably Lenient FAQ are driven away from the equilibrium, even with a low temperature. With a higher temperature, the effect would be even more clear. More research in this area is currently being performed.

5 Conclusion

The multiplayer Stag Hunt game (NSH) [5] is an interesting and well-defined game for those researchers interested in scaling their analyses and approaches to game-theoretic interactions with more than two players. The NSH exhibits two interesting critical parameter settings for which equilibria shift. This is in contrast to (1) the multiplayer Prisoners' Dilemma, which always has one Pareto-dominated defective Nash equilibrium, regardless of chosen parameters and group sizes, and (2) the 2-player Stag Hunt, which always has two strong pure Nash equilibria. We provide an intuitive analysis to demonstrate these critical parameter settings.

Given the complexity of the game, we are interested in the learning dynamics of three different reinforcement learning algorithms: Finite Action-set Learning Automata, and lenient and regular Frequency Adjusted Q-learning. By applying both the evolutionary models of the learning algorithms as well as the algorithms themselves in simulation we find that the two non-lenient learners, FALA and FAQ, perform similarly in general but deviate considerably with respect to *weak* Nash equilibria. The mutation term of FAQ drives the algorithm away from weak equilibria whereas a purely selection-based learner such as FALA may still converge to such equilibria.

Moreover, we show that leniency offers the same advantages in 3-player coordination games as in 2-player games. Cooperative or partially cooperative equilibria have a larger basin of attraction under lenient learning than under regular learning, and the cooperative basin increases with increasing degree of leniency.

References

[1] R. Axelrod. *The Evolution of Cooperation.* Basic Books, New York., 1984.

[2] D. Bloembergen, M. Kaisers, and K. Tuyls. Empirical and theoretical support for leniency in cooperative games. In *Proc. of 10th Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, 2011.

[3] M. Bukowski and J. Miekisz. Evolutionary and asymptotic stability in symmetric multi-player games. *International Journal of Game Theory*, 33:41–54, 2004.

[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] J. M. Pacheco, F. C. Santos, Max O. Souza, and Brian Skyrms. Evolutionary dynamics of collective action in n-person stag hunt dilemmas. In *Proceedings of the Royal Society B: Biological Sciences*, volume 276, pages 315–321, 2009.

[6] L. Panait, K. Sullivan, and S. Luke. Lenience towards teammates helps in cooperative multiagent learning. In Nakashima, Wellman, Weiss, and Stone, editors, *Proc. of 5th Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, 2006.

[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] B. Skyrms. The stag hunt. *Proceedings and Addresses of the American Philosophical Association*, 75(2):31–41, 2001.

[9] 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.

[10] K. Tuyls and A. Nowe. Evolutionary Game Theory and Multi-Agent Reinforcement Learning. *The Knowledge Engineering Review*, 20:63–90, 2005.

[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.