Frequency Adjusted Multi-agent Q-learning

Michael Kaisers and Karl Tuyls
Maastricht University
Maastricht, The Netherlands
{michael.kaisers, k.tuyls} @maastrichtuniversity.nl

ABSTRACT

Multi-agent learning is a crucial method to control or find solutions for systems, in which more than one entity needs to be adaptive. In today's interconnected world, such systems are ubiquitous in many domains, including auctions in economics, swarm robotics in computer science, and politics in social sciences. Multi-agent learning is inherently more complex than single-agent learning and has a relatively thin theoretical framework supporting it. Recently, multi-agent learning dynamics have been linked to evolutionary game theory, allowing the interpretation of learning as an evolution of competing policies in the mind of the learning agents. The dynamical system from evolutionary game theory that has been linked to Q-learning predicts the expected behavior of the learning agents. Closer analysis however allows for two interesting observations: the predicted behavior is not always the same as the actual behavior, and in case of deviation, the predicted behavior is more desirable. This discrepancy is elucidated in this article, and based on these new insights Frequency Adjusted Q- (FAQ-) learning is proposed. This variation of Q-learning perfectly adheres to the predictions of the evolutionary model for an arbitrarily large part of the policy space. In addition to the theoretical discussion, experiments in the three classes of two-agent two-action games illustrate the superiority of FAQ-learning.

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

1. INTRODUCTION

Today's world shows numerous examples of interconnected systems, ranging from the ubiquitous internet to high tech multi-robot applications. The assumption of a system being actually isolated from any other actor can rarely be upheld. Hence, problems from a variety of domains are naturally modeled as multi-agent systems to account for their inherent structure and complexity. Their complexity however makes those systems hard to understand and even harder to predict. Multi-agent learning has been acknowledged to be a valuable tool to control or find solutions to these systems [15, 17]. Significant progress has been facilitated in various applications, ranging from auctions and swarm robotics to predicting political decisions [6, 11, 14, 16].

Learning in multi-agent environments is significantly more complex than single-agent learning, as the optimal behavior to learn depends on other agents' policies. These policies are in turn changed according to the other agents' learning strategies, which makes the first agents learning goal a moving target. All agents face this same situation, while chasing their own dynamic learning goal they indirectly influence and move the learning goals of other agents. This makes predicting the behavior of learning algorithms in multi-agent systems difficult. In such non-stationary environments, the Markov property does not hold, which makes all proofs of convergence to optimal policies from single-agent learning that are based on that assumption inapplicable. This limits the theoretical backbone available for multi-agent learning. Furthermore, the agents may not only be situated in a non-stationary environment but may also need to deal with incomplete information and communication limits.

More recently, evolutionary game theory has been linked to reinforcement learning and provides useful insights into the learning dynamics [3, 7, 21, 22]. In particular, this link has provided insights into the dynamics and convergence properties of current state of the art multi-agent reinforcement learning algorithms such as Q-learning. It allows to study the resilience of equilibria, visualize the basins of attraction and fine tune parameters. Experiments comparing Q-learning to its evolutionary model reveal two interesting facts: one, the learning traces partly deviate significantly from the predicted dynamics, and two, in case of deviation the prediction is more desirable than the actual learning behavior. This paper is the first to examine this issue in depth for multi-agent Q-learning, giving a detailed elaboration of the causes for the occasional mismatch. Subsequently, Frequency Adjusted Q- (FAQ-) learning is proposed as a variation of Q-learning that complies with the prediction of the evolutionary model derived in [21, 22]. Although FAQ-learning can be applied to multi-state problems as introduced, this article evaluates it in single-state games for

Cite as: Frequency Adjusted Multi-agent Q-learning, M. Kaisers and K. Tuyls, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 309-315

the sake of clarity and coherence with related work [2, 3, 22]. Furthermore, the selected games suffice to show the improved behavior and the applicability of the evolutionary game theoretic framework.

In essence, Q-learning fails to comply with its prediction because actions are updated at different frequencies. The newly proposed variation compensates the difference in frequencies by modulating the learning step size for each action separately. Thereby, initialization dependencies are removed and convergence progresses through more rational policy trajectories, i.e., in expectation never moving away from dominant actions. It has been shown that modulating the learning rate can improve learning performance, e.g., Bowling et al. have modulated the learning rate anti-proportional to the success of the current strategy [5]. The here presented approach is different in that it considers the learning rate of each action separately, compensating the fact that an action which is selected more often receives more updates and thereby has its estimation updated more quickly.

The remainder of this article is structured as follows: Section 2 introduces basic concepts from reinforcement learning and evolutionary game theory. Using these preliminaries, Section 3 discusses the exact relation between Q-learning and its evolutionary model, which leads to the derivation of FAQ-learning. An empirical evaluation of the anomalies and their alleviation in FAQ-learning are presented in Section 4. Finally, Section 5 concludes the article with a discussion of the new algorithm and its position in the evolutionary game theoretic framework.

2. BACKGROUND

This section introduces the main concepts from reinforcement learning and evolutionary game theory that this article is based on. In particular, Q-learning and its relation to evolutionary game theory are discussed. The general concept of replicator dynamics is explained and the specific replicator dynamics model that has been linked to Q-learning is provided. The latter are analyzed in-depth in Section 3.

2.1 Q-learning

Q-learning was invented to maximize discounted payoffs in a multi-state environment [23]. It was originally studied in single-agent learning, where the learning process is markovian from the agent's point of view, i.e., the policy change only depends on the current and known states of the world. This article discusses single-state multi-agent Q-learning, which has an established but imperfect relation to evolutionary game theory. In multi-agent learning, the environment is not markovian from an agents point of view, as the optimal policy to learn changes due to the adaptation of other agents. Consequently, proofs from single-agent learning may not hold or may require stronger assumptions [4]. The discussion of single-state games is the first step to establish a new framework for the analysis of multi-agent learning.

By definition, the Q-learner repeatedly interacts with its environment, performing action i at time t, and receiving reward $r_i(t)$ in return. It maintains an estimation $Q_i(t)$ of the expected discounted reward for each action i. This estimation is iteratively updated according to the following equation, known as the Q-learning update rule, where α denotes the learning rate and γ is the discount factor:

$$Q_i(t+1) \leftarrow Q_i(t) + \alpha \left(r_i(t) + \gamma \operatorname*{argmax}_j Q_j(t) - Q_i(t) \right)$$

Let k be the number of actions, and let x_i denote the probability of selecting action i, such that $\sum_{i=1}^{k} x_i = 1$. Furthermore, let $x(Q) = (x_1, \ldots, x_k)$ be a function that associates any set of Q-values with a policy. The most prominent examples of such policy generation schemes are the ϵ-greedy and the Boltzmann exploration scheme [18]. This article exclusively discusses Q-learning with the Boltzmann exploration scheme. Boltzmann exploration is defined by the following function, mapping Q-values to policies, and balancing exploration and exploitation with a temperature parameter τ:

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

The parameter τ lends its interpretation as temperature from the domain of physics. High temperatures lead to stochasticity and random exploration, selecting all actions almost equally likely regardless of their Q-values. In contrast to this, low temperatures lead to high exploitation of the Q-values, selecting the action with the highest Q-value with probability close to one. Intermediate values prefer actions proportionally to their relative competitiveness. In many applications, the temperature parameter is decreased over time, allowing initially high exploration and eventual exploitation of the knowledge encoded in the Q-values. An examination of the Q-learning dynamics under time dependent temperatures is given in [12]. Within the scope of this article, the temperature is kept constant for analytical simplicity and coherence with the derivations in [21, 22].

2.2 Evolutionary game theory

Evolutionary game theory takes a rather descriptive perspective, replacing hyper-rationality from classical game theory by the concept of natural selection from biology [13]. It studies the population development of individuals belonging to one of several species. The two central concepts of evolutionary game theory are the replicator dynamics and evolutionary stable strategies [19]. The replicator dynamics presented in the next subsection describe the evolutionary change in the population. They are a set of differential equations that are derived from biological operators such as selection, mutation and cross-over. The evolutionary stable strategies describe the possible asymptotic behavior of the population. However, their examination is beyond the scope of this article. For a detailed discussion, we refer the interested reader to [9, 10].

2.3 Replicator dynamics

The replicator dynamics from evolutionary game theory formally define the population change over time. A population comprises a set of individuals, where the species that an individual can belong to relate to pure actions available to a learner. The utility function $r_i(t)$ that assigns a reward to the performed action can be interpreted as the Darwinian fitness of each species i. The distribution of the individuals on the different strategies can be described by a probability vector that is equivalent to a policy for one player, i.e., there is one population in every agent's mind. The evolutionary pressure by natural selection can be modeled by the replicator equations. They assume this population to evolve such that successful strategies with higher payoffs than average grow while less successful ones decay. These dynamics are formally connected to reinforcement learning [3, 20, 21].

Let the policy of a player be represented by the probability vector $x = (x_1, \ldots, x_k)$, where x_i indicates the probability to play action i, or the fraction of the population that belongs to species i. The dot notation will be used to denote differentiation over time, i.e. $\dot{x}_i = \frac{dx_i}{dt}$. The replicator dynamics that relate to the learning process of *Cross Learning*, a simple learning automaton, are given by the following set of differential equations [3]:

$$\dot{x}_i = x_i \left[E\left[r_i(t)\right] - \sum_j^k x_j E\left[r_j(t)\right] \right]$$

This is a *one-population* model. In order to describe a *two-population* model relating to two-agent matrix games played by Cross learners, let e_i denote the i^{th} unit vector, and x and y be policy vectors for a two-player matrix game, where the utility functions are given by $\forall t : E\left[r_i(t)\right] = e_i A y$ and $\forall t : E\left[r_j(t)\right] = x B e_j$ for player one and two respectively. The corresponding replicator dynamics are given by the following set of differential equations:

$$\dot{x}_i = x_i \left[e_i A y - x A y\right]$$
$$\dot{y}_j = y_j \left[x B e_j - x B y\right]$$

The change in the fraction playing action i is proportional to the difference between the expected payoffs $e_i A y$ and $x B e_i$ of action i against the mixing opponent, and the expected payoff $x A y$ and $x B y$ of the mixed strategies x and y against each other. Hence, above average actions get stronger while below average actions decay. The replicator dynamics maintain the probability distribution, thus $\sum_i \dot{x}_i = 0$. The examples used in this article are constraint to two actions, which implies $\dot{x}_1 = -\dot{x}_2$ and $\dot{y}_1 = -\dot{y}_2$. The policy space is completely described by the unit square (x_1, y_1), in which the replicator dynamics can be plotted as arrows in the direction of $(\dot{x}_1, \dot{y}_1)$.

The behavior of *Cross learning*, a simple policy iterator, has been shown to converge to the replicator dynamics in the infinitesimal time limit [3]. Based on these insights, an analogical relation between Q-learning and an extension of the replicator dynamics has been derived in [22], which the following subsection elaborates.

2.4 Q-learning dynamics

In [22] the authors extended the work of Borgers et al. of [3] to Q-learning. More precisely, they derived the dynamics of the Q-learning process, which yielded the following system of differential equations, describing the learning dynamics for a two-player stateless matrix game:

$$\begin{aligned} \dot{x}_i &= x_i \alpha \left(\tau^{-1} \left[e_i A y - x A y\right] - \log x_i + \sum_k x_k \log x_k \right) \\ \dot{y}_j &= y_j \alpha \left(\tau^{-1} \left[x B e_j - x B y\right] - \log y_j + \sum_l y_l \log y_l \right) \end{aligned} \quad (2)$$

with x, y the policies, α the learning rate, τ temperature parameter, A, B the payoff matrices, and e_i the i^{th} unit vector. The striking part of this result was that the equations contain a selection part equal to replicator dynamics, and a mutation part. For an elaborate discussion in terms of selection and mutation operators we refer to [21, 22].

With this model, it now became possible to get insight into the learning process, its traces, basins of attraction, and stability of equilibria, by just examining the coupled system of replicator equations and plotting its force and directional fields. An example plot of the dynamics of the game Battle of Sexes is given in Figure 1, the corresponding payoff table can be found in Figure 3.

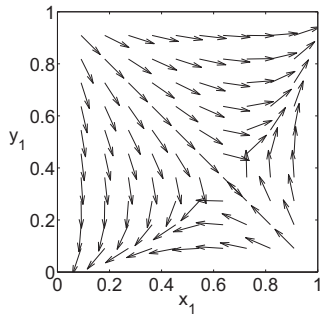

Figure 1: An example of a replicator dynamics plot, showing the dynamics of the Battle of Sexes game.

Borgers et al. observed that the actual learning traces of Cross learning may deviate from the predicted behavior [3]. Similarly, we observed that the behavior of the Q-learning process does not always match the derived Q-learning dynamics. While the correspondence between algorithm and model improves under smaller learning rates in Cross learning, these deviations are systematic and non-negligible for Q-learning. The next section analyzes and elucidates why this is the case.

3. ANALYSIS

The evolutionary model defined by Equation 2 should predict the learning behavior of Q-learning accurately. However, significant deviations of the learning trajectories from the model can be observed. Figure 2 shows an example of such deviations in the Prisoners' Dilemma game, for which the payoff table is given in Figure 3. The remainder of this section analyzes why these anomalies occur and how they can be accounted for.

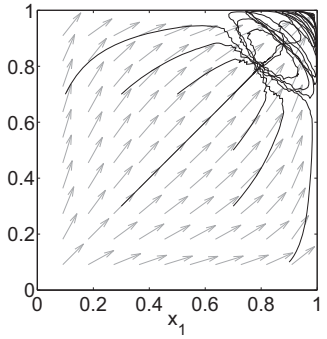

Figure 2: An example of the discrepancy between Q-learning and its evolutionary prediction in the Prisoner's Dilemma game. The arrows indicate the expected policy change derived from the extended replicator dynamics, and the solid lines show Q-learning trajectories, obtained from running the algorithm with a small learning rate.

3.1 Deriving the cause of discrepancies

When action a is selected, the Q-values are changed according to $\Delta Q_i(t) = Q_i(t+1) - Q_i(t)$:

$$\Delta Q_i(t) = \begin{cases} \alpha \left(r_i(t) + \gamma \operatorname{argmax}_j Q_j(t) - Q_i(t) \right) & \text{if i=a} \\ 0 & \text{otherwise} \end{cases}$$

The policy x determines the frequency with which each Q-value is updated and influences the expected Q-value change. The expected reward $E[r_i(t)]$ also depends on the environment and the other agents. The resulting expected Q-value change incurred by the Q-learning update rule is given by:

$$E[\Delta Q_i(t)] = x_i \cdot \alpha \underbrace{\left(E[r_i(t)] + \gamma \operatorname{argmax}_j Q_j(t) - Q_i(t) \right)}_{T}$$

The authors of [2, 8] independently arrived at the same expected change of Q-values. However, these sources explicitly consider ϵ-greedy exploration, which does not allow to describe the process as policy iteration.

Next, the continuous time limit of the Q-value change will be derived, using the term T for notational convenience. Taking the continuous time limit of a learning algorithm is inspired by [3], which describes a policy learner with infinitesimal time steps and shows that the process of multi-agent *Cross-learning* converges to the replicator dynamics in the continuous time limit. In the learning algorithm, updates proceed in discreet iterations of $\Delta t = 1$.

$$E[Q_i(t+1) - Q_i(t)] = 1 \cdot x_i \cdot \alpha T$$

The continuous time limit can be constructed by changing the basis for time from 1 to δ and then taking the limit of δ to zero. Let the learning rate α_δ in the new frame of reference be decomposed into $\alpha_\delta = \delta\alpha$, i.e., when time δ passes, only a proportional fraction of the update is incurred.

$$E[Q_i(t+\delta) - Q_i(t)] = \delta \cdot x_i \cdot \alpha T$$

Which in the continuous time limit $\delta \to 0$, again using the dot notation for differentiation to time, becomes:

$$E\left[\dot{Q}_i\right] = x_i \cdot \alpha \Big(E[r_i(t)] + \gamma \operatorname{argmax}_j Q_j(t) - Q_i(t) \Big)$$

Due to the fact that in the infinitesimal time limit an infinite number of updates is perceived, the expected change equals the actual change in that limit, i.e., $E\left[\dot{Q}_i\right] = \dot{Q}_i$. Formally, $\forall \epsilon > 0$, however small, $\exists \delta > 0 : \epsilon = k\delta$, with $k \to \infty$ and $E\left[\frac{dQ_i}{\epsilon}\right] = E\left[\frac{dQ_i}{k\delta}\right] = \frac{1}{k} E\left[\frac{dQ_i}{\delta}\right]$. According to the law of large numbers, the mean approaches the expected value for large sample sizes k. Hence, $\frac{1}{k} E\left[\frac{dQ_i}{\delta}\right] = \frac{dQ_i}{k\delta} = \frac{dQ_i}{\epsilon}$.

$$\dot{Q}_i = x_i \cdot \alpha \Big(E[r_i(t)] + \gamma \operatorname{argmax}_j Q_j(t) - Q_i(t) \Big)$$

In the following, the link of the continuous time version of Q-learning to an extension of the replicator dynamics is discussed.

A relation of Q-learning to evolutionary game theory was derived in [21]. However, the derivation starts from the following assumption:

$$\dot{Q}_i = \alpha \left(E[r_i(t)] + \gamma \operatorname{argmax}_j Q_j(t) - Q_i(t) \right)$$

This differs from the actual $\dot{Q}_i$ by a factor of x_i. This discrepancy between the model and the update rule explains observed anomalies and initialization dependencies. This discrepancy can be resolved in two ways: one, deriving the actual replicator dynamics of Q-learning, incorporating the factor x_i, or two, adapting the Q-learning update rule to fit the model.

The derived evolutionary game theory model of [22] predicts *more rational* policy trajectories than Q-learning actually exhibits. For example, if two actions are over-estimated by the current Q-values and the dominant action receives more updates due to being selected more often, the dominant action will loose it's over-estimation more quickly and Q-learning may policy-wise move away from this dominant action. Such behavior is undesirable because the problem of over- and under-estimation is prevalent in the application of the algorithm. The Q-values need some initialization which must not be based on knowledge of the rewards. This leads to initial errors in the estimate. In practice, this is overcome by sufficient initial exploration, but the amount of exploration that suffices may differ from case to case and if underestimated, i.e., if exploration is decreased prematurely, the same problems of wrong estimates re-occur. Another drawback of moving away from dominant actions is the decrease of expected reward for a period of time, which may in some applications be worse than an almost monotonically ascending expected reward with a slightly lower accumulated payoff. This may be the case when dependent processes rely on the profit that is generated from this game, e.g. humans commonly prefer monotonically increasing income over temporarily decreasing income, even if the cumulated reward is lower [1]. In addition, the actual dynamics of Q-learning are not independent of the Q-values and can therefore not be sufficiently described in the policy space.

As a consequence, rather than deriving the actual replicator dynamics for Q-learning, this article presents an alternative update rule for Q-learning, i.e., Frequency Adjusted Q- (FAQ-) learning that perfectly fits the model for an arbitrarily large subspace of the policy space. In particular, the update rule is adapted to compensate the frequency term x_i in the expected Q-value change.

3.2 Frequency Adjusted Q-learning

This subsection introduces Frequency Adjusted Q- (FAQ-) learning, which is derived to inherit the more desirable game theoretical behavior of the evolutionary game theory model that was invented to describe Q-learning. *FAQ-learning* is equivalent to Q-learning, except for the update rule for which it uses the following adapted version:

$$Q_i(t+1) \leftarrow Q_i(t) + \frac{1}{x_i} \alpha \left(r_i(t) + \gamma \operatorname{argmax}_j Q_j(t) - Q_i(t) \right)$$

Using the same reasoning as in the previous section, the continuous time limit of this process converges to the following equation:

$$\dot{Q}_i = \alpha \Big(E[r_i(t)] + \gamma \operatorname{argmax}_j Q_j(t) - Q_i(t) \Big)$$

This means that FAQ-learning matches the assumption made in [21] precicely, while regular Q-learning differs by a factor of x_i. The experiments in Section 4 show how that translates to anomalies and differences between Q-learning and its prediction, while an exact match of FAQ-learning and the

evolutionary game theoretical model can be observed. However, this update rule is only valid in the infinitesimal limit of α, otherwise $\frac{\alpha}{x_i}$ may become larger than 1. This would allow the Q-values to escape the convex hull of experienced rewards. That in turn breaks the learning algorithm. In fact, a maximal learning step should be very small to yield reasonable convergence behavior, i.e., $\frac{\alpha}{x_i} << 1$. Consequently, this idealized version of FAQ-learning cannot be applied numerically. We propose the following generalized model of FAQ-learning with a new model parameter $\beta \in [0, 1)$:

$$Q_i(t+1) \leftarrow Q_i(t) + \min\left(\frac{\beta}{x_i}, 1\right) \cdot \alpha\left(r_i(t) + \gamma \operatorname*{argmax}_j Q_j(t) - Q_i(t)\right)$$

Let us inspect the properties of this update rule, considering that the behavior changes at $\frac{\beta}{x_i} = 1$, which is at $x_i = \beta$. For notational convenience, the time dependency is dropped from $x_i(t)$, $Q_i(t)$, and $r_i(t)$ in the following equation:

$$x_i \geq \beta : E\left[\Delta Q_i\right] = \frac{\beta}{x_i}\alpha\left(E\left[r_i\right] + \gamma \operatorname*{argmax}_j Q_j - Q_i\right)$$
$$x_i < \beta : E\left[\Delta Q_i\right] = \alpha\left(E\left[r_i\right] + \gamma \operatorname*{argmax}_j Q_j - Q_i\right)$$

If $\beta = 1$, this model degenerates to regular Q-learning, therefore this value is excluded from the allowed range of β. If $0 \leq \beta < 1$, the limit of $\alpha \to 0$ makes this model equivalent to idealized FAQ-learning with a learning rate of $\alpha\beta$, i.e., the behavior converges to the replicator dynamics derived in [21]. Numerical simulation needs to choose finitely small α. In that case, the dynamics for $x_i \geq \beta$ are equivalent to idealized FAQ-learning with learning rate $\alpha\beta$, while the dynamics for $x_i < \beta$ equal those of regular Q-learning with learning rate α. Hence, the maximal learning step is defined by α and needs to be reasonably small, while the size of the subspace that behaves like idealized FAQ-learning is controlled by β. For both parameters, smaller values are more desirable regarding the path of convergence, but lead to an increase in the required number of iterations. By choosing β arbitrarily small, the learner can be made to behave according to the evolutionary model for an arbitrarily large part of the policy space.

The examples given in this article will empirically evaluate FAQ-learning with $\beta = \alpha$ to obtain a smooth convergence to the true Q-values, while maintaining the correct update behavior for a large part of the policy space.

4. EXPERIMENTS AND RESULTS

This section compares Q-learning and FAQ-learning trajectories to their evolutionary game theoretic predictions. For the sake of clarity, the empirical evaluation is restricted to two-player two-action normal form games. This type of games can be characterized by a payoff bi-matrix (A, B), where for any joint action (i, j) the payoff to player one and two are given by A_{ij} and B_{ij} respectively. Figure 3 gives the payoff matrices of three representative examples of this class of games, corresponding to the selection in [21, 22]: the Prisoners' Dilemma (PD), the Battle of Sexes (BoS), and Matching Pennies (MP). They represent the classes of games with one pure Nash Equilibrium (PD), with one mixed and

	D	C
D	$1, 1$	$5, 0$
C	$0, 5$	$3, 3$

	B	S
B	$2, 1$	$0, 0$
S	$0, 0$	$1, 2$

	H	T
H	$1, -1$	$-1, \ 1$
T	$-1, \ 1$	$1, -1$

Figure 3: Reward matrices for Prisoners' Dilemma (left, *Defect* or *Cooperate*), Battle of Sexes (right, *Bach* or *Stravinski*) and Matching Pennies (bottom, *Head* or *Tail*).

two pure Nash Equilibria (BoS), and with one mixed Nash Equilibrium (MP).

For Boltzmann action selection, policies do not uniquely identify the Q-values they are generated from. Translation of all Q-values by an equal amount does not alter the policy, which is solely dependent on the difference between the Q-values. For example, the Q-value pair $(0, 1)$ generates the same policy as $(1, 2)$. The replicator dynamics describe the policy change depending on the policy, while the learning update rule incurs a policy change dependent on the policy and the Q-values. In order to compare Q-learning and FAQ-learning to the evolutionary prediction, learning trajectories showing the update rule's effect are given for several translations of initial Q-values. In particular, the initial Q-values are centered around the minimum, mean or maximum possible Q-value, given the game's reward space. The two spaces are related according to the following equation for the minimum, and similarly for the other values:

$$Q_{min} = \sum_{i=0}^{\infty} \gamma^i r_{min} = \frac{1}{1-\gamma} r_{min}$$

This relates to $\{0, 2\frac{1}{2}, 5\}$ for the Prisoners' Dilemma, $\{0, 1, 2\}$ for Battle of Sexes and $\{-1, 0, 1\}$ for Matching Pennies if $\gamma = 0$, and the tenfold if $\gamma = 0.9$.

Figure 4 shows trajectories obtained from running the learners with $\gamma = 0.9$, $\alpha = 10^{-6}$ for Q-learning, and $\alpha = \beta = 10^{-3}$ for FAQ-learning, with a fixed temperature $\tau = 0.1$. The trajectories yield 200 thousand iterations in all but the right two cases of the top row, which show 500 thousand iterations.

While regular Q-learning shows significantly different learning behavior depending on the initialization, FAQ-learning merely increases the noise for higher values in the initialization. The noise is caused by larger learning steps, as the Q-value change includes a term $-\alpha Q_i(t)$, which is clearly proportional to the magnitude of the Q-values. Nonetheless, the expected direction of change remains unaffected in FAQ-learning.

In comparison to the evolutionary prediction, the FAQ-learning trajectories always follow the predicted expected change, while Q-learning trajectories deviate from it depending on the initialization. The behavior of Q-learning and FAQ-learning are most similar to each other for the mean reward initialization. However, tweaking the initialization does not remove but only alleviates the deviations, and knowing the exact reward space violates the assumption of many applications. In addition, the Prisoners' Dilemma shows qualitatively significant differences even for the mean initialization.

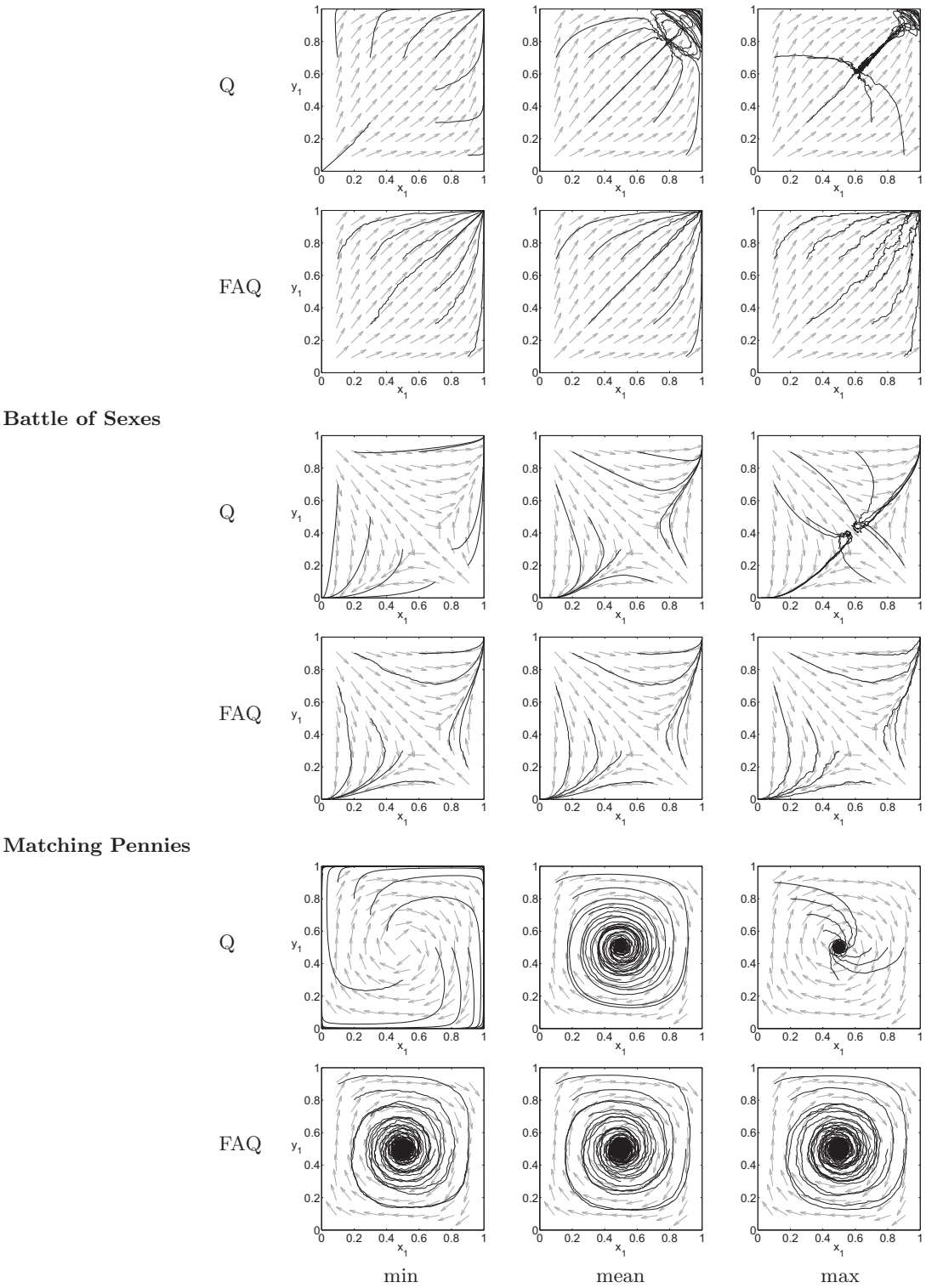

Figure 4: Comparison of Q-learning to FAQ-learning with various Q-value initializations in the Prisoners' Dilemma, the Battle of Sexes and Matching Pennies. The Q-values are initialized centered at the minimum (left), mean (center) and maximum (right) possible Q-value given the reward space of the game.

5. DISCUSSION AND CONCLUSIONS

The results have shown empirical confirmation of the match between trajectories of the newly proposed FAQ-learning algorithm and its evolutionary prediction. These results have been found to be qualitatively insensitive to the values of γ and α, as long as α is reasonably small. Given the Q-value space and a specific temperature τ, the most extreme policy can be computed using the policy generating function given in Equation 1. Hence, a temperature τ can be selected such that $x_i \geq \beta$ is guaranteed in FAQ-learning, and the algorithm behaves according to ideal FAQ-learning. Using analogous derivations as in Section 3.2, Frequency Adjusted Sarsa (FAS) can be shown to behave equivalently in single-state environments. The contributions of this article can be summarized as follows: The deviation of Q-learning from its evolutionary model has been analyzed and explained. Based on the new insights, FAQ-learning was devised and it is shown to comply with the evolutionary prediction for an arbitrarily large part of the policy space.

Further experiments are required to verify the performance gain in multi-state domains and real applications; the relation between the learning speed $\frac{\alpha\beta}{x_i}$ in FAQ- and α in Q-learning is critical for the speed and quality of convergence that is achieved and needs further investigation. Future work will combine these insights with progress on the methodology of analyzing multi-agent learning from the perspective of evolutionary game theory. Eventually, this will contribute to a solidified theoretical framework for the understanding and prediction of multi-agent learning dynamics.

This research was partially sponsored by a TopTalent2008 grant of the Netherlands Organisation for Scientific Research (NWO).

6. REFERENCES

[1] Dan Ariely. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. HarperCollins, February 2008.

[2] Monica Babes, Michael Wunder, and Michael Littman. Q-learning in two-player two-action games. In *Autonomous Learning Agents Workshop at the 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, May 2009.

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

[4] Michael Bowling. Convergence problems of general-sum multiagent reinforcement learning. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pages 89–94. Morgan Kaufmann, 2000.

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

[6] Bruce Bueno de Mesquita. Game theory, political economy, and the evolving study of war and peace. *American Political Science Review*, 100(4):637–642, November 2006.

[7] C. M. Gintis. *Game Theory Evolving*. University Press, Princeton, June 2000.

[8] Eduardo Gomes and Ryszard Kowalczyk. Modelling the dynamics of multiagent q-learning with ϵ-greedy exploration (short paper). In Sierra Decker, Sichman and Castelfranchi, editors, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 1181–1182, Budapest, Hungary, May 10–15, 2009.

[9] Morris W. Hirsch, Stephen Smale, and Robert Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 2002.

[10] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 2002.

[11] Shlomit Hon-Snir, Dov Monderer, and Aner Sela. A learning approach to auctions. *Journal of Economic Theory*, 82:65–88, November 1998.

[12] Michael Kaisers, Karl Tuyls, Simon Parsons, and Frank Thuijsman. An evolutionary model of multi-agent learning with a varying exploration rate. In *AAMAS '09: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 1255–1256, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[13] J. Maynard-Smith. *Evolution and the Theory of Games*. Cambridge University Press, December 1982.

[14] Shervin Nouyan, Roderich Groß, Michael Bonani, Francesco Mondada, and Marco Dorigo. Teamwork in self-organized robot colonies. *Transactions on Evolutionary Computation*, 13(4):695–711, August 2009.

[15] L. Panait and S. Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, November 2005.

[16] S. Phelps, M. Marcinkiewicz, and S. Parsons. A novel method for automatic strategy acquisition in n-player non-zero-sum games. In *AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 705–712, Hakodate, Japan, 2006. ACM.

[17] 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.

[18] R. Sutton and A. Barto. *Reinforcement Learning: An introduction*. MA: MIT Press, Cambridge, 1998.

[19] P. D. Taylor and L. Jonker. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40:145–156, 1978.

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

[21] K. Tuyls, P. J. '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.

[22] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 693–700, New York, NY, USA, 2003. ACM.

[23] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.