FAQ-Learning in Matrix Games: Demonstrating Convergence near Nash Equilibria, and Bifurcation of Attractors in the Battle of Sexes

Michael Kaisers, Karl Tuyls
Maastricht University
P.O. Box 616
6200 MD Maastricht, The Netherlands

Abstract

This article studies Frequency Adjusted Q-learning (FAQ-learning), a variation of Q-learning that simulates simultaneous value function updates. The main contributions are empirical and theoretical support for the convergence of FAQ-learning to attractors near Nash equilibria in two-agent two-action matrix games. The games can be divided into three types: Matching pennies, Prisoners' Dilemma and Battle of Sexes. This article shows that the Matching pennies and Prisoners' Dilemma yield one attractor of the learning dynamics, while the Battle of Sexes exhibits a supercritical pitchfork bifurcation at a critical temperature of τ, where one attractor splits into two attractors and one repellent fixed point. Experiments illustrate that the distance between fixed points of the FAQ-learning dynamics and Nash equilibria tends to zero as the exploration parameter τ of FAQ-learning approaches zero.

Multi-agent learning has received increasing attention in the last years (Shoham, Powers, and Grenager 2007; Tuyls and Parsons 2007). It has ubiquitous applications in today's world, due to for instance infrastructures such as internet and mobile phone networks. The assumption of an automatic agent actually acting in isolation is usually at best a simplification. Since the multi-agent interactions are complex and difficult to predict, multi-agent learning gains popularity to find good policies to act in multi-agent systems, and a variety of approaches have been devised (Panait and Luke 2005).

Learning in multi-agent environments is significantly more complex than single-agent learning as the dynamics to learn are changed by the learning process of other agents. This makes predicting learning behavior of learning algorithms in multi-agent environments difficult. The research field is still rather young and despite raising interest in multi-agent learning, the theoretical framework is rather thin.

Recently, Evolutionary Game Theory (EGT) has been established as a tool to analyze independent reinforcement learning applied to multi-agent settings (Hofbauer and Sigmund 2002; Tuyls and Parsons 2007). EGT replaces the assumption of rationality from game theory by genetic operators such as natural selection and mutation, and studies the evolution of population models that are subdued to the genetic operators. A formal link of population models to reinforcement learning has been established (Börgers and Sarin 1997). Each population can be interpreted as a policy of one agent, and the genetic operators that induce change to the population correspond to the learning rule that updates the agent's policy. This allows to study the behavior and convergence properties of learning algorithms by analyzing the corresponding dynamical system from EGT. In addition, the stream of research exploiting the link between reinforcement learning and dynamical systems also encompasses some sources which are not subsumed by EGT, but do exploit the toolbox of dynamical systems in a similar way.

Seminal work has shown that Cross learning, a simple policy learner, becomes equivalent to the replicator dynamics when the learning rate is decreased to the infinitesimal limit (Börgers and Sarin 1997). The link between learning algorithms and dynamical systems in subsequent work is generally based on the limit of infinitesimal learning rates. This idea has been used in one of the first proofs of convergence for independent reinforcement learning in multi-agent settings. The proof shows that the average payoff of Infinitesimal Gradient Ascent, a policy gradient learning algorithm, converges to the Nash equilibrium payoff in two-agent two-action matrix games, even though actual policies may cycle (Singh, Kearns, and Mansour 2000). This result has been strengthened by introducing the Win-or-learn-fast (WoLF) learning speed modulation. The policies of Infinitesimal Gradient Ascent with WoLF learning rates are proven to converge to the Nash equilibrium policies in two-agent two-action games (Bowling and Veloso 2002). In contrast to other reinforcement learning algorithms like Q-learning, Infinitesimal Gradient Ascent assumes that the agents possess a lot of information about the payoff structure. In particular, agents are usually not able to compute the gradient of the reward function, which is necessary for this algorithm. Variations of Infinitesimal Gradient Ascent have been devised to tackle this issue, but these are beyond the scope of this article.

The basis of this article is the evolutionary game theoretic model of Q-learning (Tuyls, Verbeeck, and Lenaerts 2003), which assumes simultaneous action updates for the sake of the analysis. However, the analysis of the simplified evolutionary model reveals game theoretically more desirable be-

havior than actual Q-learning. As a response, the variation Frequency Adjusted Q-learning (FAQ-learning) has been proposed, which simulates simultaneous action updates and inherit the theoretical guarantees (Kaisers and Tuyls 2010). The evolutionary model used in the research presented below will therefore be referred to as FAQ-learning dynamics.

A notable extension to FAQ-learning is Lenient FAQ-learning, which combines the advantages of lenient learning in coordination games with FAQ-learning (Bloembergen, Kaisers, and Tuyls 2011). While the before mentioned articles about Q-learning assume Boltzmann action selection and primarily study learning trajectories in the policy space, there also exist a number of publications considering ϵ-greedy Q-learning, concentrating on the dynamics in the Q-value space (Gomes and Kowalczyk 2009; Wunder, Littman, and Babes 2010).

This article adds to the theoretical guarantees of FAQ-learning provided in (Kaisers and Tuyls 2010) by proving convergence to fixed points near Nash equilibria in two-agent two-action games, and by demonstrating a bifurcation in a specific class of games. For the analysis, games are divided into three classes, corresponding to Matching Pennies, Prisoners' Dilemma and Battle of Sexes type games. Theoretical examination reveals that Matching Pennies and Prisoners' Dilemma type games yield one attracting fixed point for FAQ-learning. The Battle of Sexes type games feature one attractor for high exploration (temperature), and a supercritical pitchfork bifurcation at a critical temperature, below which there are two attracting and one repelling fixed point. Fixed points in all games approach Nash equilibria as the temperature tends to zero.

The remainder of this article is structured as follows: The following section introduces concepts from reinforcement learning and (evolutionary) game theory. Subsequently, the learning dynamics of FAQ-learning in two-agent two-action matrix games are examined theoretically. The following section presents simulation experiments that illustrate the learning behavior and convergence near Nash equilibria in three representative games. Finally, conclusions summarize the main contributions of this article and give directions for future work.

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 FAQ-learning is provided.

Game Theory

Game theory models strategic interactions in the form of games. Each player has a set of actions, and a preference over the joint action space which is captured in the numerical payoff signal. For two-player games, the payoffs can be given in a bi-matrix (A, B), that gives the payoff for the row player in A, and the column player in B (see Figure 1).

$$\begin{pmatrix} A_{11}, B_{11} & A_{12}, B_{12} \\ A_{21}, B_{21} & A_{22}, B_{22} \end{pmatrix}$$

Figure 1: General payoff bi-matrix (A, B) for two-agent two-action games.

A core solution concept in game theory is the Nash equilibrium. Assume the row player plays his actions with probability $(x, 1-x)$, and the column player plays his actions with probability $(y, 1-y)$. The policy pair (x, y) is a Nash equilibrium if and only if no player can gain by unilaterally deviating from his policy. Let e_i denote the i^{th} unit vector. Formally, (x_e, y_e) comprises a Nash equilibrium iff $\forall i : x_e A y_e \geq e_i A y_e$ and $x_e B y_e \geq x_e B e_i$.

Q-learning

Q-learning was invented to maximize discounted payoffs in a multi-state environment by Watkins, see (Watkins and Dayan 1992). 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 (Bowling 2000). 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:

$$Q_i(t+1) \leftarrow Q_i(t) + \alpha \left(r_i(t) + \gamma \max_j Q_j(t) - Q_i(t) \right)$$

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 (Sutton and Barto 1998). 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:

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

The parameter τ lends its interpretation as temperature from the domain of thermodynamics. 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 (Kaisers et al. 2009). Within the scope of this article, the temperature is kept constant for analytical simplicity and coherence with the derivations in (Tuyls, 't Hoen, and Vanschoenwinkel 2005; Tuyls, Verbeeck, and Lenaerts 2003).

FAQ-learning is equivalent to Q-learning, except for the update rule. The magnitude of each learning step for action i is adjusted by the inverse of the action probability x_i (computed at time t according to Eq. 1). FAQ-learning simulates simultaneous action updates by increasing the learning steps of less frequently selected actions.

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

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 (Maynard-Smith 1982). 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 (Taylor and Jonker 1978). 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 crossover. 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 (Hirsch, Smale, and Devaney 2002; Hofbauer and Sigmund 2002).

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 (Börgers and Sarin 1997; Tuyls and Parsons 2007; Tuyls, 't Hoen, and Vanschoenwinkel 2005). 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 (Börgers and Sarin 1997):

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

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 $\sum_i \dot{x}_i = 0$. This article only discusses two-action games, 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 $(\dot{x}_1, \dot{y}_1)$. Using $h = (1, -1)$ and eliminating x_2 and y_2, the equations can be reduced to:

$$\dot{x}_1 = \alpha x_1(1 - x_1)\left[y_1 h A h^T + A_{12} - A_{22}\right]$$
$$\dot{y}_1 = \alpha y_1(1 - y_1)\left[x_1 h B h^T + B_{21} - B_{22}\right]$$

The behavior of *Cross learning*, a simple policy iterator, has been shown to converge to the replicator dynamics in the infinitesimal time limit (Börgers and Sarin 1997). Based on these insights, an analogical relation between Q-learning and an extension of the replicator dynamics has been derived in (Tuyls, Verbeeck, and Lenaerts 2003), which the following subsection elaborates.

FAQ-learning dynamics

In (Tuyls, Verbeeck, and Lenaerts 2003) the authors extended the work of Borgers et al. of (Börgers and Sarin 1997) to Q-learning. More precisely, they derived the dynamics of the Q-learning process under the simplifying assumption of simultaneous action updates. This yields the following system of differential equations, describing precisely the FAQ-learning dynamics for a two-player stateless matrix game (Kaisers and Tuyls 2010):

$$\dot{x}_i = x_i\alpha\left(\tau^{-1}\left[e_iAy - xAy\right] - \log x_i + \sum_k x_k \log x_k\right)$$
$$\dot{y}_j = y_j\alpha\left(\tau^{-1}\left[xBe_j - xBy\right] - \log y_j + \sum_l y_l \log y_l\right) \quad (2)$$

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 (Tuyls, 't Hoen, and Vanschoenwinkel 2005; Tuyls, Verbeeck, and Lenaerts 2003).

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 2, the corresponding payoff matrix can be found in Figure 4.

Theory

This section delineates the theoretical support for convergence of FAQ-learning. The dynamical system defined by $\dot{x}$ and $\dot{y}$ yields a number of fixed points, which may be attracting or repelling. Since learning trajectories converge to attractors, the local stability (attracting or repelling) is the main condition that is analyzed.

For notational convenience, we define auxiliary variables a, b and functions K_1, K_2 to simplify the FAQ dynamics $\dot{x}, \dot{y}$ from Equation 2, where we also drop the index for the remainder of this section. Let the row player play policy $(x, 1-x)$ against the column player with policy $(y, 1-y)$.

$$\begin{aligned}
a_1 &= A_{11} - A_{21}\\
a_2 &= A_{12} - A_{22}\\
b_1 &= B_{11} - B_{12}\\
b_2 &= B_{21} - B_{22}\\
h &= (1, -1)\\
hAh^T &= a_1 - a_2\\
hBh^T &= b_1 - b_2\\
K_1(x,y) &= \tau_1^{-1}\left[yhAh^T + a_2\right] - \log\frac{x}{1-x}\\
K_2(x,y) &= \tau_2^{-1}\left[xhBh^T + b_2\right] - \log\frac{y}{1-y}\\
\dot{x} &= \alpha x(1-x)K_1(x,y)\\
\dot{y} &= \alpha y(1-y)K_2(x,y)
\end{aligned}$$

At a fixed point $\dot{x} = \dot{y} = 0$. Since FAQ-learning with positive exploration parameter τ only covers the open set of policies (x, y) with $x, y \notin \{0, 1\}$, we know that $\alpha x(1-x) \neq 0$. Hence, $\dot{x} = \dot{y} = 0$ implies $K_1(x, y) = K_2(x, y) = 0$.

The local stability can be analyzed by checking the eigenvalues of the Jacobian matrix $J = \begin{bmatrix} \frac{\partial\dot{x}}{\partial x} & \frac{\partial\dot{x}}{\partial y} \\ \frac{\partial\dot{y}}{\partial x} & \frac{\partial\dot{y}}{\partial y} \end{bmatrix}$ at a fixed point (Hirsch, Smale, and Devaney 2002).

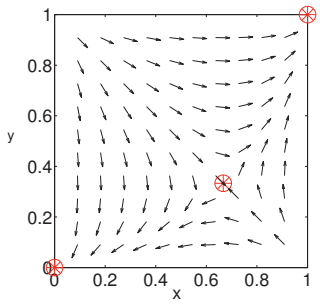


Figure 2: An example of a replicator dynamics plot, showing the dynamics of the Battle of Sexes game.

$$\begin{aligned}
\frac{\partial\dot{x}}{\partial x} &= \alpha\left[(1-2x)K_1(x,y) - 1\right]\\
\frac{\partial\dot{x}}{\partial y} &= \alpha x(1-x)\tau_1^{-1}hAh^T\\
\frac{\partial\dot{y}}{\partial x} &= \alpha y(1-y)\tau_2^{-1}hBh^T\\
\frac{\partial\dot{y}}{\partial y} &= \alpha\left[(1-2y)K_2(x,y) - 1\right]
\end{aligned}$$

Since we have just established that $K_1(x, y) = K_2(x, y) = 0$ at mixed fixed points, this can be plugged in to the Jacobian.

$$J(x,y) = \begin{bmatrix} -\alpha & \alpha x(1-x)\tau^{-1}hAh^T \\ \alpha y(1-y)\tau^{-1}hBh^T & -\alpha \end{bmatrix}$$

The eigenvalues can be computed using the quadratic formula.

$$\begin{aligned}
\lambda_{1/2} &= -\alpha \pm \frac{1}{2}\sqrt{4\frac{\partial\dot{x}}{\partial y}\frac{\partial\dot{y}}{\partial x} + (-\alpha - (-\alpha))^2}\\
&= -\alpha \pm \sqrt{\frac{\partial\dot{x}}{\partial y}\frac{\partial\dot{y}}{\partial x}}\\
&= -\alpha \pm \alpha\sqrt{x(1-x)y(1-y)\tau_1^{-1}hAh^T\tau_2^{-1}hBh^T}
\end{aligned}$$

Dynamical systems theory has established that the fixed point is locally attracting if $\forall\lambda : \mathsf{real}(\lambda) \leq 0$ and $\exists\lambda : \mathsf{real}(\lambda) < 0$ (Hirsch, Smale, and Devaney 2002). This leads to the following condition for stability, which will be denoted $C(x, y) \leq 1$:

$$\begin{aligned}
\alpha\left[-1 \pm \sqrt{x(1-x)y(1-y)\tau_1^{-1}\tau_2^{-1}hAh^ThBh^T}\right] &\leq 0\\
-1 \leq \sqrt{x(1-x)y(1-y)\tau_1^{-1}\tau_2^{-1}hAh^ThBh^T} &\leq 1\\
C(x,y) = x(1-x)y(1-y)\tau_1^{-1}\tau_2^{-1}hAh^ThBh^T &\leq 1
\end{aligned}$$

Since $x, (1-x), y, (1-y), \tau_1, \tau_2$ all are positive, this condition holds independent of x, y if $hAh^ThBh^T \leq 0$, leading to eigenvalues with $\mathsf{real}(\lambda) = -\alpha < 0$. In other words, games that satisfy $hAh^ThBh^T \leq 0$ have only attracting

fixed points. These games already cover all Matching Pennies type games and some Prisoners' Dilemma type games.

The following system of equations defines the stability boundary using two conditions for the fixed point, and one for local stability.

$$\tau_1 \log \frac{x}{1-x} - a_2 = yhAh^T$$
$$\tau_2 \log \frac{y}{1-y} - b_2 = xhBh^T$$
$$x(1-x)y(1-y)hAh^T hBh^T \leq \tau_1\tau_2$$

This set of equations can be solved numerically for any specific game to obtain fixed points and their stability property. The following general discussion will provide support for convergence in all three classes, especially discussing the characteristic number $hAh^T hBh^T$ associated with each game.

Class 1 Matching Pennies type games: I. $a_1a_2 < 0$, II. $b_1b_2 < 0$, and III. $a_1b_1 < 0$.

In order to link these conditions to the stability property, consider that $hAh^T hBh^T = a_1b_1 - a_1b_2 - a_2b_1 + a_2b_2$. Assumptions I and II imply $a_1a_2b_1b_2 > 0$, hence a_1b_2 and a_2b_1 are either both positive or both negative. Dividing out III one finds $a_2b_2 < 0$. Assume a_1b_2 is negative, then $a_1b_2a_1a_2 > 0$ leads to the contradiction $a_1^2 < 0$. Since all numbers in the matrix need to be real, we conclude $a_1b_2 > 0$ and $a_2b_1 > 0$. In sum, $hAh^T hBh^T < 0$, which leads to the eigenvalues λ of the Jacobian matrix to have $\mathsf{real}(\lambda) = -\alpha$ as explained above. The fixed point is necessarily attracting in matching pennies games, since $\forall\lambda, \mathsf{real}(\lambda) < 0$.

Class 2 Prisoners' dilemma type games: I. $a_1a_2 > 0$ and II. $b_1b_2 > 0$.

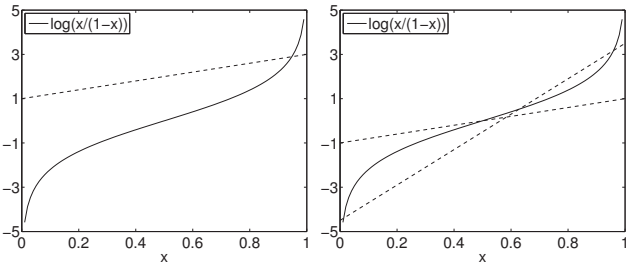
Games of this class can have both positive and negative characteristic numbers. Games with $hAh^T hBh^T < 0$ yield necessarily attracting fixed points for the same reason as in Class 1. However, a large number of games of this type have positive characteristic numbers, e.g., for symmetric games $hAh^T hA^T h^T \geq 0$. It remains to show that games with III. $(a_1 - a_2)(b_1 - b_2) \geq 0$ have attracting fixed points.

From I and II we know that $[ya_1 + (1-y)a_2] \neq 0$ and $[xb_1 + (1-x)b_2] \neq 0$. This implies that there is only one solution to $K_1(x,y) = K_2(x,y) = 0$:

$$y\frac{a_1}{\tau_1} + (1-y)\frac{a_2}{\tau_1} = \log\frac{x}{1-x}$$
$$x\frac{b_1}{\tau_2} + (1-x)\frac{b_2}{\tau_2} = \log\frac{y}{1-y}$$

Figure 3 plots an example of the first equation. The temperature τ determines the point of intersection between the two lines: If a_1 and a_2 are positive, then $x \to 1$ as $\tau \to 0$. If a_1 and a_2 are negative, then $x \to 0$ as $\tau \to 0$. Equivalent conditions hold for y in relation with b_1 and b_2.

It is trivial to check that the stability condition holds for sufficiently large temperatures. Since $x(1-x)$ goes to zero faster than τ_1, and similarly $y(1-y)$ goes to zero faster than τ_2, the stability condition $x(1-x)y(1-y)hAh^T hBh^T \leq \tau_1\tau_2$ holds for all temperatures $\tau > 0$.



Prisoners' Dilemma Battle of Sexes

Figure 3: The Prisoners' Dilemma features one fixed point, because there is exactly one intersection between the linear combination of $\frac{a_1}{\tau_1}$ and $\frac{a_2}{\tau_1}$ with the log function. The Battle of Sexes on the other hand features one or three fixed points, depending on the slope of the linear combination.

Class 3 Battle of Sexes type games: I. $a_1a_2 < 0$, II. $b_1b_2 < 0$, and III. $a_1b_1 > 0$.

The first two imply $a_1a_2b_1b_2 > 0$, hence a_1b_2 and a_2b_1 are either both positive or both negative. Dividing out the third assumption we find $a_2b_2 > 0$. Assume a_1b_2 is positive, then $a_1b_2a_1a_2 < 0$ leads to the contradiction $a_1^2 < 0$. Since all numbers in the matrix need to be real, we conclude $a_1b_2 < 0$ and $a_2b_1 < 0$. As a result, the characteristic number $(a_1 - a_2)(b_1 - b_2) = a_1b_1 - a_1b_2 - a_2b_1 + a_2b_2 > 0$.

From I and II we know that $[ya_1 + (1-y)a_2]$ and $[xb_1 + (1-x)b_2]$ both cross zero. Figure 3 illustrates the difference between the Prisoners' Dilemma and the Battle of Sexes. It shows the function $\mathsf{log}\frac{x}{1-x}$ and the linear interpolation between $\frac{a_1}{\tau_1}$ and $\frac{a_2}{\tau_1}$. Large values of τ lead to one intersection, while sufficiently small values of τ lead to three intersections and corresponding fixed points.

The stability condition $x(1-x)y(1-y)hAh^T hBh^T \leq \tau_1\tau_2$ is satisfied for large τ. At the critical temperature τ_{crit}, the stability condition holds with equality, leading to a supercritical pitchfork bifurcation of the fixed points in τ. Below the critical temperature, two fixed points approach pure Nash equilibria and are stable for the same reasons as the fixed point in the Prisoners' Dilemma. In addition, one fixed point remains mixed, and $x(1-x)$ as well as $y(1-y)$ is clearly bound away from zero. As a result, this fixed point is not stable below the critical temperature.

Experiments

This section illustrates the convergence behavior, and the effect of the exploration parameter τ on the distance of fixed points to Nash equilibria. Each class of two-agent two-action games is represented by one specific game. The payoff bimatrices (A, B) for Matching Pennies (Class 1), Prisoners' Dilemma (Class 2), and Battle of Sexes (Class 3) are given in Figure 4.

Let the row player play policy $(x, 1-x)$ against the column player with policy $(y, 1-y)$. The Nash equilibria of these games lie at $(\frac{1}{2}, \frac{1}{2})$ for the Matching Pennies, $(1,1)$ for the Prisoners' Dilemma, and at $(0,0), (1,1)$, and $(\frac{2}{3}, \frac{1}{3})$ for the Battle of Sexes.

$$\begin{array}{c} \\ H \\ T \end{array}\begin{array}{c} \begin{array}{cc} H & T \end{array} \\ \begin{pmatrix} 1,0 & 0,1 \\ 0,1 & 1,0 \end{pmatrix} \end{array} \quad \begin{array}{c} \\ C \\ D \end{array}\begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} 1,1 & 5,0 \\ 0,5 & 3,3 \end{pmatrix} \end{array} \quad \begin{array}{c} \\ O \\ F \end{array}\begin{array}{c} \begin{array}{cc} O & F \end{array} \\ \begin{pmatrix} 2,1 & 0,0 \\ 0,0 & 1,2 \end{pmatrix} \end{array}$$

Matching Pennies Prisoner's Dilemma Battle of the Sexes

Figure 4: Payoff bi-matrices (A, B) for three representative games.

Note, that the joint policy space is completely characterized by the pair (x, y) in the unit square. The dynamical system can be inspected by plotting the replicator dynamics $(\dot{x}, \dot{y})$ over the unit square, indicating the direction of the vector field by arrows. This allows to determine the course location of attractors by inspection. In addition, the fixed points have been computed, and are marked with circles.

Figure 5 shows the empirical analysis of FAQ-learning: Matching Pennies, Prisoners' Dilemma and Battle of Sexes. The top three rows show replicator dynamics (arrows) and the computed fixed points (circles) for different temperature parameters τ (first $\tau = \infty$, second $\tau = 0.72877$, third $\tau = 0$). The fixed points move between these discrete values for τ as indicated by the lines of the last row. For reference, all fixed points computed for the discrete values are also marked in the last row.

The dynamics of FAQ-learning are independent of the game when the exploration parameter τ tends to infinity. For finite temperatures, the three games exhibit very different behavior. However, In the limit of $\tau \to 0$, the fixed points coincide with the Nash equilibria of the game (see row 3).

The Matching Pennies yields one mixed equilibrium, which is also an attracting fixed point of the FAQ-learning dynamics for any positive τ. In the limit $\tau \to 0$, the fixed point's stability weakens to Lyapunov stability (points that start close will stay close to the fixed point, but not necessarily converge to it).

The Prisoners' Dilemma yields one pure equilibrium, and one mixed fixed point that is always attracting. The lower the temperature τ, the closer the fixed point moves towards the equilibrium. It is also stable in the limit $\tau \to 0$.

The Battle of Sexes yields three equilibria. However, for high values of τ, it only yields one attracting fixed point that moves from $(\frac{1}{2}, \frac{1}{2})$ toward the mixed equilibrium $(\frac{2}{3}, \frac{1}{3})$. This fixed point splits in a supercritical pitchfork bifurcation at the critical temperature $\tau_{crit} \approx 0.72877$ and at position $(x, y) \approx (0.5841, 0.4158)$. For low temperatures $\tau < \tau_{crit}$, this game yields three fixed points that move closer to the corresponding equilibria as τ is decreased. The two fixed points moving toward the pure equilibria $(0, 0)$ and $(1, 1)$ are attracting, and the third one moving toward $(\frac{2}{3}, \frac{1}{3})$ is repelling.

The relation between the exploration parameter τ of FAQ-learning and the distance between fixed points and Nash equilibria is closely examined in Figure 6. It shows that the distance is constant zero for Matching Pennies, and monotonically decreasing toward zero for the other two games. Notably, the two emerging fixed points in the Battle of Sexes result in the same distance plot, due to a certain symmetry

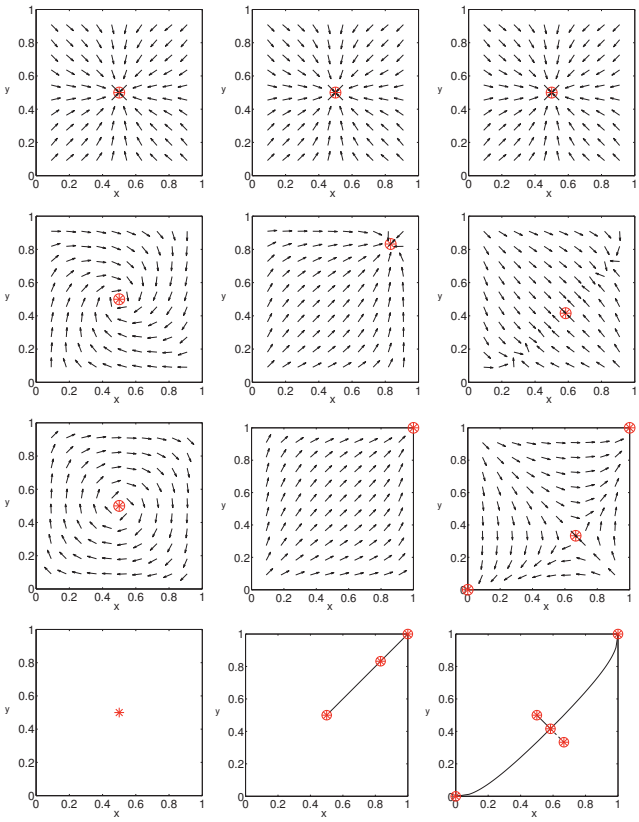


Figure 5: Replicator dynamics (arrows) and fixed points (dots) for the Prisoners' Dilemma, Matching Pennies and Battle of Sexes with temperatures $\tau \in \{\infty, 0.72877, 0\}$ from top to third row. Last row shows trajectories of fixed points as temperature is decreased, revealing the bifurcation of fixed points in the Battle of Sexes. All indicated fixed points are attracting, except for the mixed fixed point that tends to $(\frac{2}{3}, \frac{1}{3})$ after bifurcation (indicated with a dashed line).

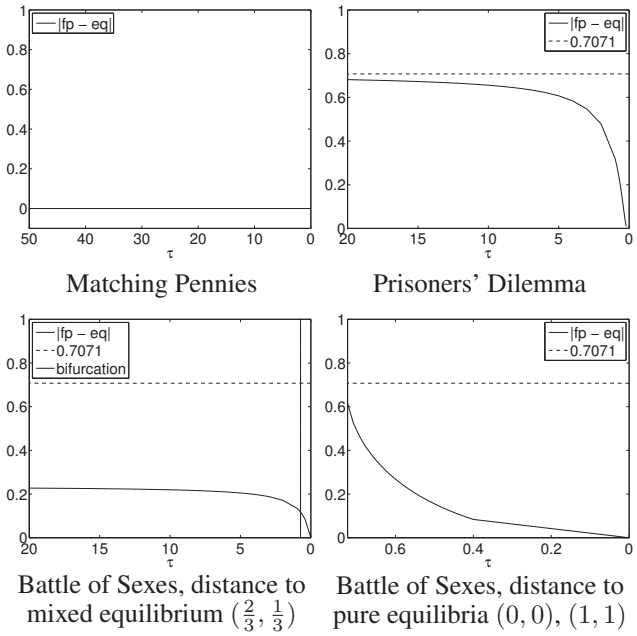
of the game.

In sum, FAQ-learning converges to fixed points in the three representative games Matching Pennies, Prisoners' Dilemma and Battle of Sexes. In addition, these fixed points can be moved arbitrarily close to the Nash equilibria of these games by choosing an exploration parameter τ close to zero.

Conclusions

The contributions of this article are two-fold: First, it is shown theoretically that fixed points of FAQ-learning are attracting in Matching Pennies and Prisoners' Dilemma type games, and that a supercritical pitchfork bifurcation occurs in Battle of Sexes type games. Second, representative example games of each category demonstrate that fixed points approach Nash equilibria, and illustrate the bifurcation of fixed points in the Battle of Sexes.

This article contributes to the framework of multi-agent



Matching Pennies | Prisoners' Dilemma

Battle of Sexes, distance to mixed equilibrium $(\frac{2}{3}, \frac{1}{3})$ | Battle of Sexes, distance to pure equilibria $(0, 0)$, $(1, 1)$

Figure 6: This figure shows the relation between exploration parameter τ and the distance between fixed points (fp) of FAQ-learning dynamics and Nash equilibria (eq). As $\tau \to 0$, the distance $|\text{fp} - \text{eq}| \to 0$ as well.

learning by deepening the understanding of convergence properties of independent reinforcement learning in multi-agent settings. The method has been shown to work in the policy space, and naturally extends to the Q-value space, which allows to generalize insights to standard Q-learning. By doing so, future work will strengthen the theoretical guarantees and their impact to a wide array of applications.

THIS RESEARCH WAS PARTIALLY SPONSORED BY A TOPTALENT2008 GRANT OF THE NETHERLANDS ORGANISATION FOR SCIENTIFIC RESEARCH (NWO).

References

Bloembergen, D.; Kaisers, M.; and Tuyls, K. 2011. Empirical and theoretical support for lenient learning. In *AAMAS '11: Proceedings of The 10th International Conference on Autonomous Agents and Multiagent Systems*.

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

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

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

Gomes, E., and Kowalczyk, R. 2009. Modelling the dynamics of multiagent q-learning with ϵ-greedy exploration (short paper). In Decker, Sichman, S., and Castelfranchi., eds., *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, 1181–1182.

Hirsch, M. W.; Smale, S.; and Devaney, R. 2002. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press.

Hofbauer, J., and Sigmund, K. 2002. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Kaisers, M., and Tuyls, K. 2010. Frequency adjusted multi-agent q-learning. In *Proc. of 9th Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, 309–315.

Kaisers, M.; Tuyls, K.; Parsons, S.; and Thuijsman, F. 2009. 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*, 1255–1256. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

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

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

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

Singh, S.; Kearns, M.; and Mansour, Y. 2000. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 541–548. Morgan Kaufman.

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

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

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

Tuyls, K.; 't Hoen, P. J.; and Vanschoenwinkel, B. 2005. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems* 12:115–153.

Tuyls, K.; Verbeeck, K.; and Lenaerts, T. 2003. 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*, 693–700. New York, NY, USA: ACM.

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

Wunder, M.; Littman, M.; and Babes, M. 2010. Classes of multiagent q-learning dynamics with ϵ-greedy exploration. In *Proceedings of the 27th International Conference on Machine Learning*.