Instituto de Sistemas e Robótica

Pólo de Lisboa

Convergence of independent adaptive learners[1]

Francisco S. Melo
Manuel C. Lopes

May 2007
RT-603-07

ISR Torre Norte
Av. Rovisco Pais, 1
1049-001 Lisboa
PORTUGAL

[1]This work was partially supported by Programa Operacional Sociedade do Conhecimento (POS_C) that includes FEDER funds. The first author acknowledges the PhD grant SFRH/BD/3074/2000.

Convergence of independent adaptive learners

Francisco S. Melo **Manuel C. Lopes**
Institute for Systems and Robotics
Instituto Superior Técnico
Av. Rovisco Pais, 1
1049-001 Lisboa,
PORTUGAL
{fmelo,macl}@isr.ist.utl.pt

Abstract

In this paper we analyze the convergence of independent adaptive learners in repeated games. We show that, in this class of games, independent adaptive learners converge to pure Nash equilibria, if they exist. We discuss the relation between our result and convergence results of adaptive play [22]. The importance of our result stems from the fact that, unlike adaptive play, no communication/action observability is assumed. We also relate this result to recent results on the convergence of weakened ficticious play processes for independent learners [11, 19]. Finally we present some experimental results to illustrate the main ideas of the paper.

1 Introduction

Game theory provides a mathematical framework to model situations in which several decision-makers interact. These interactions can occur at different levels, ranging from "games" in the common everyday sense to more complex interactions such as those taking place in economical or biological systems [15]. Situations where coordination/competition among several agents occur are naturally captured using game theoretic models; concepts such as Nash equilibrium define stable behavioral conventions from which no agent can profitably deviate. Such equilibria provide multi-agent counterparts to the concept of optimal behavior-rule in single-agent systems.

Game theory is traditionally used in economics, where it provides powerful models to describe interactions of economical agents. Recent years have witnessed an increasing interest from the computer science and robotic communities in applying game theoretic models to multi-agent systems [4, 5, 21]. For example, the interaction of a group of robots moving in a common environment can be naturally captured using a game theoretic model and their observed behavior suitably interpreted using game theoretic concepts.

When addressing game theory from a learning perspective, Boutilier [1] distinguishes two fundamental classes of learning agents: independent learners (IL) and joint-action learners (JAL). The former have no knowledge on other agents, interacting with the environment as if no other decision-makers existed. The latter, on the contrary, are aware of the existence of other agents and are capable of perceiving (*a posteriori*) their actions and rewards.

Learning algorithms considering JALs are easily implementable from standard single-agent reinforcement learning algorithms [6, 12, 13]. Action observability allows a learning agent to build statistics on the other agents' behavior-rules and act in a best-response sense. This is the underlying principle of standard methods such as fictitious play [2] or adaptive play [22].

However, in many practical applications it is not reasonable to assume the observability of other agents' actions. Most agents interact with their surroundings by relying on sensory information and action recognition is often far from trivial. With no knowledge on the other agents' actions and

payoffs, the problem becomes more difficult. Tan [18] and Claus and Boutilier [3] some empirical evidence is gathered that describes the convergence properties of reinforcement learning methods in multi-agent settings. In these works, the experimental performance of ILs is compared with that of JALs (using fictitious play). Wang and de Silva [21] and Crites and Barto [4] a similar comparison is performed for specific problems.

Lauer and Riedmiller [10] study independent learners in deterministic settings. They provide a learning algorithm that relies on strict assumptions on the other agents' behavior. In particular, the authors discuss the use of optimistic and pessimistic assumptions on the other agents' behavior and show their algorithm to converge in behavior to an optimal decision-rule. Kapetanakis and Kudenko [7, 8] an improvement is proposed that deals with more non-deterministic settings. Recent results have established the convergence of a variation of fictitious play for independent learners [11], first introduced by Van der Genugten [19].

In this paper, we propose *independent adaptive learning*, a variation of adaptive play for independent learners. This algorithm has an obvious advantage over the original adaptive learning algorithm [22], since it does not require each player to be able to observe the plays by the other agents. Furthermore, no *a priori* knowledge of the payoff function is required. We show that, in this class of games, independent adaptive learners converge to pure Nash equilibria, if they exist. We also present some experimental validation of our results.

2 Background

In this section we introduce some background material that will be used throughout the paper.

2.1 Strategic and repeated games

A strategic game is a possible model of interaction between decision-makers. Formally can be described as a tuple $\big(N, (\mathcal{A}_k), (r_k)\big)$, where N is the number of players, $\mathcal{A} = \times_{k=1}^{N} \mathcal{A}_k$ is the set of possible joint actions and r is a *reward function* or *payoff function*. For $k = 1, \ldots, N$, $\mathcal{A}_k$ represents the set of individual actions available to player k. The payoff function $r_k : \mathcal{A} \to \mathbb{R}$ is used to define a preference relation $\succsim_k$ on the set $\mathcal{A}$—the preference relation of player k. The payoff functions in strategic games can be represented by matrices, and such games are also known as *matrix games*.

We represent an element $a \in \mathcal{A}$ as a N-tuple $a = (a_1, \ldots, a_N)$ and refer it as a *joint action*. The tuple $a_{-k} = (a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_N)$ is a *reduced joint action*, and we write $a = (a_{-k}, a_k)$ to denote that the individual action of player k in the joint action a is a_k.

Notice that a strategic game is a *one-shot* game, in that *each play of the game is independent of any previous plays*. In particular, it is not possible to have memory effects in the players: even if the game is played repeatedly, at each play no player has any knowledge of previous plays of the game. If memory of past plays is possible, we refer to such a game as a *repeated game*. In a repeated game, N players repeatedly engage in a strategic game defined as usual as a tuple $\big(N, (\mathcal{A}_k), (r_k)\big)$. The repeated interaction allows the players to maintain, for example, statistics describing the strategies of the other players and use these statistics to play accordingly.

A strategic game is *zero-sum* or *strictly competitive* if it has 2 players and $r_1 = -r_2$, and *general-sum* otherwise. A general sum game is *fully cooperative* if $r_1 = \ldots = r_N$.

2.2 Nash equilibria

In strategic games (and other classes of games) there is an implicit assumption of rationality on the players. This means that each player k chooses from all its individual actions the best action according to the preference relation arising from r^k. However, the best action will often depend on the other player's choice of actions. This leads to the following definition.

A *Nash equilibrium* of a strategic game $\big(N, (\mathcal{A}_k), (r_k)\big)$ is an action profile $a^* \in \mathcal{A}$ such that, for every player $k = 1, \ldots, N$, $r_k(a^*) \geq r_k(a^*_{-k}, a_k)$, for all $a_k \in \mathcal{A}_k$. A Nash equilibrium can be interpreted as an action profile capturing a *steady-state play* in the game: if a^* is a Nash

equilibrium, no player benefits from individually deviating its play from a^*. We emphasize that not every strategic game has a Nash equilibrium.

So far, we have seen that payoff functions define a preference relation over the set $\mathcal{A} = \times_{k=1}^{N} \mathcal{A}_k$. However, it is often the case that the players choose their actions in a non-deterministic way. If this is the case, each payoff function r_k also translates the preferences of player k over possible *lotteries* over the actions in $\mathcal{A}$. A *strategy* for player k is a probability distribution over the set $\mathcal{A}_k$. A strategy σ_k assigns a probability $\sigma_k(a_k)$ to each action $a_k \in \mathcal{A}_k$. We say that player k follows strategy σ_k when playing the game $\big(N, (\mathcal{A}_k), (r_k)\big)$ if it chooses each action $a_k \in \mathcal{A}_k$ with probability $\sigma_k(a_k)$. If a strategy σ_k assigns probability 1 to some action $a_k \in \mathcal{A}_k$, then σ_k is a *pure strategy*. Otherwise, it is called a *mixed strategy*. We define the concepts of *joint strategy* and *reduced joint strategy* in a similar manner as defined for actions. The *support* of a strategy σ_k is the set of all actions $a_k \in \mathcal{A}_k$ such that $\sigma_k(a_k) > 0$.

A *mixed strategy Nash equilibrium* of a strategic game $\big(N, (\mathcal{A}_k), (r_k)\big)$ is a strategy profile σ^* such that, for every player $k = 1, \ldots, N$,

$$\mathbb{E}_{\sigma^*}\left[R_k\right] \geq \mathbb{E}_{(\sigma^*_{-k},\, \sigma_k)}\left[R_k\right]$$

for all strategies σ^k, where R_k is the random variable denoting the outcome of the game for player k. The next theorem was established by John F. Nash in 1950 [15].

Theorem 2.1. *Every strategic game $\big(N, (\mathcal{A}_k), (r_k)\big)$ with finite $\mathcal{A}$ has a mixed strategy Nash equilibrium.*

2.3 Fictitious play

Fictitious play is an iterative procedure originally proposed by Brown [2] to determine the solution for a strictly competitive game. This procedure was shown to converge in this class of games by Robinson [16] and later extended to other classes of games by several authors [9, 11, 14, 19, see, for example]).

In its original formulation, two players repeatedly engage in a strictly competitive game. Each player maintains an estimate of the other player's strategy as follows: let $N_t(a)$ denote the number of times that the individual action a was played up to (and including) the t^{th}play. At play t, player k estimates the other player's strategy to be

$$\hat{\sigma}_{-k}(a_{-k}) = \frac{N_t(a_{-k})}{t},$$

for each $a_{-k} \in \mathcal{A}_{-k}$. The expected payoff associated with each individual action of player k is then

$$EP(a_k) = \sum_{a_{-k} \in \mathcal{A}_{-k}} r_k(a_{-k}, a_k)\hat{\sigma}_{-k}(a_{-k}).$$

Player k can now choose its action from the set of best responses,

$$BR = \left\{ a_k \in \mathcal{A}_k \mid a_k = \arg\max_{u_k \in \mathcal{A}_k} EP(u_k) \right\}.$$

Robinson [16] showed that this methodology yields two sequences $\{\hat{\sigma}_1\}_t$ and $\{\hat{\sigma}_2\}_t$ converging respectively to σ_1^* and σ_2^* such that (σ_1^*, σ_2^*) is a Nash equilibrium for the game $\big(\{1,2\}, (\mathcal{A}_k), (r_k)\big)$.

In general, it is not possible to ensure that fictitious play converges in all games. However, for particular classes of games (see the references above), it is possible to establish the convergence of fictitious play, and this methodology can be used by a set of agents to learn a Nash equilibrium.

2.4 Adaptive play

Adaptive play was first proposed by Young [22] as an alternative method to fictitious play. The basic underlying idea is similar to fictitious play, but the actual method works differently from fictitious

play. For games which are *weakly acyclic*, adaptive play converges w.p.1 to a pure strategy Nash equilibrium, both in *beliefs* and in *behavior*.[1]

Let h be a vector of length m. We refer to any set of K samples randomly drawn from h without replacement as a *K-sample* and denote it generically by $K(h)$, where K and m are any two integers such that $1 \leq K \leq m$

Let $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ be a repeated game played at discrete instants of time $t = 1, 2, \ldots$. At each play, each player $k = 1, \ldots, N$ chooses an action $a_k(t) \in \mathcal{A}_k$ as described below, and the action profile $a(t) = (a_1(t), \ldots, a_N(t))$ is referred to as the *play at time t*. The history of plays up to time t is a vector $(a(1), \ldots, a(t))$.

Let K and m be as described above. At each time instant $t = 1, 2, \ldots$, each player $k = 1, \ldots, N$ chooses its action $a_k(t)$ as follows. For $t \leq m$, $a_k(t)$ is chosen randomly from $\mathcal{A}_k$; for $t \geq m + 1$, player k inspects K plays drawn without replacement from the most recent m plays. We denote by H_t the m most recent plays at time t. Let $N_K(a_{-k})$ be the number of times that the reduced action a_{-k} appears in the K-sample $K(H_t)$. Player k then uses $K(H_t)$ and determines the expected payoff $EP(a_k)$ for each $a_k \in \mathcal{A}_k$ as

$$EP(a_k) = \sum_{a_{-k} \in \mathcal{A}_{-k}} r_k(a^{-k}, a_k) \frac{N_K(a_{-k})}{K}$$

It then randomly chooses its action from the set of best responses,

$$BR = \left\{ a_k \in \mathcal{A}_k \mid a_k = \arg\max_{u_k \in \mathcal{A}_k} EP(u_k) \right\}.$$

Notice that this procedure is similar to fictitious play in that it chooses the best response action to the estimated reduced strategy $\hat{\sigma}_{-k}$. The only difference lies in the fact that adaptive play uses *incomplete history sampling*, while fictitious play uses the complete history.

Young [22] established the convergence of adaptive play for repeated games that are *weakly acyclic*. To properly introduce such result, let $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ be a strategic game with finite action-space $\mathcal{A} = \times_{k=1}^{N} \mathcal{A}_k$. The *best response graph* for Γ is a directed graph $\mathcal{G} = (V, E)$, where each vertex corresponds to a joint action (*i.e.*, $V = \mathcal{A}$) and any two actions $a, b \in \mathcal{A}$, are connected by a directed edge $(a, b) \in E$ if and only if $a \neq b$ and there is exactly one player k for which b_k is a best-response to the pure strategy a_{-k} and $a_{-k} = b_{-k}$. A strategic game $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ is then *weakly acyclic* if, given any vertex a in its best response graph, there is a directed path to a vertex a^* from which there is no exiting edge (a sink).

It should be clear that a sink as described in the previous definition corresponds necessarily to a strict Nash equilibrium. Given a weakly acyclic strategic game $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$, we denote by $L(a)$ the shortest path from the vertex a to a strict Nash equilibrium in the best response graph of Γ and by $L(\Gamma) = \max_{a \in \mathcal{A}} L(a)$. We are now in position to state the following theorem from Young [22]:

Theorem 2.2. *Let $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ be a weakly acyclic strategic game. If*

$$K \leq \frac{m}{L(\Gamma) + 2},$$

then adaptive play converges w.p.1 to a strict Nash equilibrium.

3 Independent adaptive leaning

In this section we describe *independent adaptive learning*, a variation of adaptive learning relying on independent learners. This algorithm has an obvious advantage over the original adaptive learning algorithm [22], since it does not require each player to be able to observe the plays by the other agents. Furthermore, no *a priori* knowledge of the payoff function is required.

[1] For a discussion on the differences between convergence in beliefs and convergence in behavior, see the works by Littman [12], Young [22].

3.1 Independent adaptive learning process

Let $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ be a repeated game played at discrete instants of time $t = 1, 2, \ldots$. At each play, each player $k = 1, \ldots, N$ chooses an action $a_k(t) \in \mathcal{A}_k$ and receives a reward $r_k(t)$. We are interested in developing a learning algorithm for independent players, *i.e.*, players that are not able to observe the plays of the others. Therefore, we consider that all plays, rewards referred henceforth concern a particular player k in Γ, except if explicitly stated otherwise. We refer to the pair $(a(t), r(t))$ as the play (of player k) at time t. The history of plays up to time t is a set $\mathcal{H}_t = \{(a(1), r(1)), (a(2), r(2)), \ldots, (a(t), r(t))\}$.

Let K and m be two integers $1 \leq K \leq m$. At each time instant $t = 1, 2, \ldots$, the player chooses its action $a(t)$ as follows. For $t \leq m$, $a(t)$ is chosen randomly from the corresponding action set $\mathcal{A}_k$; for $t \geq m + 1$, the player inspects K plays drawn without replacement from its most recent m plays. Suppose, for definiteness, that the selected plays corresponded to times $t_1, \ldots, t_k$. The expected payoff associated with each action $u \in \mathcal{A}_k$ is

$$EP(u) = \frac{\sum_{i=1}^{K} r(t_i)\mathbb{I}_u(a(t_i))}{\sum_{i=1}^{K} \mathbb{I}_u(a(t_i))},$$

where $\mathbb{I}_u(\cdot)$ is the indicator function for action $u \in \mathcal{A}_k$. Given $EP(u)$ for all $u \in \mathcal{A}_k$, the player now randomly chooses its action from the set

$$BR = \left\{ a \in \mathcal{A}_k \mid a = \arg\max_{u \in \mathcal{A}_k} EP(u) \right\}.$$

If one particular action $u \in \mathcal{A}_k$ is never played in the selected K plays, then the expected payoff should be taken as any sufficiently large *negative number* (we henceforth take it to be $-\infty$).

3.2 Convergence of the independent adaptive learning process

In this section we establish the convergence of our method by casting it as a variation of adaptive play as described by Young [22].

The main differences between our algorithm and the standard adaptive play lie on the fact that we do not assume any *knowledge of the payoff function* or any *observability of the actions of the other players*. Instead, we rely on the sampling process to implicitly provide this information.

Before introducing our main result, we need the following definition, adapted from the work by Singh et al. [17].

Definition 3.1 (GLIE strategy)**.** *A strategy σ_i is* greedy in the limit with infinite exploration (GLIE) *if it verifies the following conditions:*

- *each action is visited infinitely often;*
- *in the limit, the policy is greedy with respect to some payoff function r w.p.1.*

A well-known example of GLIE policy is Boltzmann exploration:

$$\mathbb{P}\left[A_t = a \mid r\right] = \frac{e^{r(a)/T_t}}{\sum_{u \in \mathcal{A}} e^{r(u)/T_t}},$$

where T_t is a temperature parameter that decays at an adequate rate (see the work by Singh et al. [17] for further details).

Theorem 3.1. *Let $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ be a weakly acyclic N-player game. If*

$$K \leq \frac{m}{L(\Gamma) + 2},$$

then every independent adaptive learner following a GLIE policy will converge to a best response strategy to the other players' strategies with probability 1.

PROOF We start by considering a fixed exploration rate $\lambda > 0$. In this situation, the independent adaptive learning process described in Subsection 3.1 yields an irreducible and aperiodic finite-state Markov chain whose state-space consists on the set of all m-long sequences of joint actions. This means that the sequence of histories provided by independent adaptive learning converges at an exponential rate to a stationary distribution as described by Young [22].

It is important to remark that in the paper by Young [22], the author considers an experimentation probability parameter ε, which defines the probability of a given player making a "mistake".[2] In our algorithm, if a particular action $u \in \mathcal{A}_k$ is never played in the selected K plays, then the associated expected payoff is $-\infty$. This means that, in our algorithm, the "mistakes" can arise due to the exploration (with probability ε) or due to the subestimation of action-values. Obviously, this does not affect the convergence of the chain but only the limiting distribution.

Young [22] showed that in weakly acyclic games, if $K \leq \frac{m}{L(\Gamma)+2}$, then as the experimentation probability ε approaches to zero, the limiting distribution "narrows" around the Nash equilibria in the game. This implies the convergence of the joint strategy to one such equilibrium w.p.1. Therefore, the conclusions of our theorem follow from this result as long as we show that the probability of making "mistakes" in our algorithm goes to zero at a suitable rate.

To see this, two important observations are in order. First of all, infinite exploration ensures that the probability of all players converging to a strategy other than a Nash equilibrium is 0. On the other hand, our assumption of a GLIE policy guarantees that $\lambda \to 0$ as $t \to \infty$, while always ensuring sufficient exploration. This naturally implies that the probability of making exploration "mistakes" decreases to zero. Furthermore, it also implies that Nash equilibria will be sampled with increasing probability—as the exploration decreases, Nash equilibria will be played more frequently and consequently more frequently sampled, and consequently more frequently played, and so on. But this finally implies that, as $t \to \infty$, the probability of making "mistakes" due to sub-evaluation also decreases to zero.

Finally, the probability of making "mistakes" goes to zero at a slower rate than the GLIE policy becomes greedy which, by construction, is slower than the rate of convergence of the above Markov chain to stationarity. This allows us to apply the desired result from the paper by Young [22] and the proof is complete. □ □

4 Experimental results

In this section we present the results of our method for several simple games.

4.0.1 Prisoner's dilemma

The prisoner's dilemma is a well-known game from game theory whose payoff function is represented in Figure 1. In this game, two criminal prisoners are persuaded to confess/rat on the other by being offered immunity. If none of the prisoners confess, they will be sentenced for a minor felony. If one of the prisoners confesses and the other remains silent, the one that confesses will be released while the other will be serve the full sentence. If both prisoners confess, they will not serve the full sentence, but still remain in jail for a long time.

This game is very interesting from a game theoretic point-of-view. In fact, both players would be better off by remaining silent, since they would both serve a short sentence. However, each player profits by confessing, no matter what the other player does. Therefore, both players will confess and therefore serve a long sentence. The joint action (R, R) is, therefore, a Nash equilibrium. This is clear from the best-response graph, depicted in Figure 2, where it is also clear that the game is weakly acyclic.

We have applied our algorithm to the prisoner's dilemma. We ran 1000 independent Monte-Carlo runs. Each run consisted of 900 plays of the game, for each of which we stored the received payoff. We used Boltzmann exploration with decaying temperature factor to ensure sufficient exploration of all actions. The results are depicted in Figure 3. Figure 3.a) presents the evolution

[2] We consider "mistakes" in the sense of Young [22].

	S	R
S	5, 5	-10, 20
R	20, -10	-5, -5

Figure 1: Payoff for the prisoner's dilemma. Each prisoner may opt by remaining silent (S) or by ratting on the other prisoner (R)

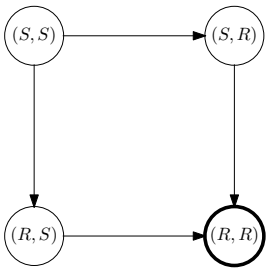
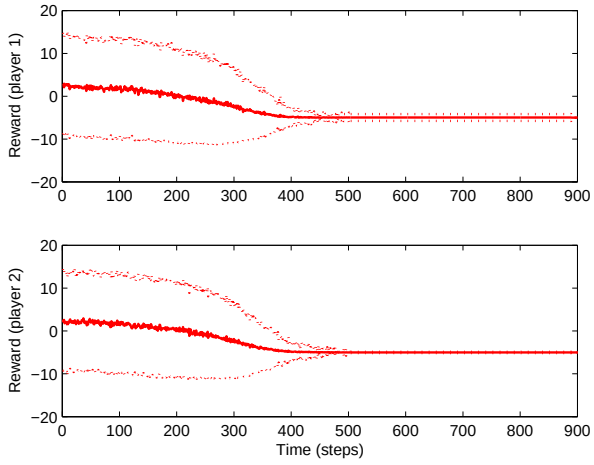
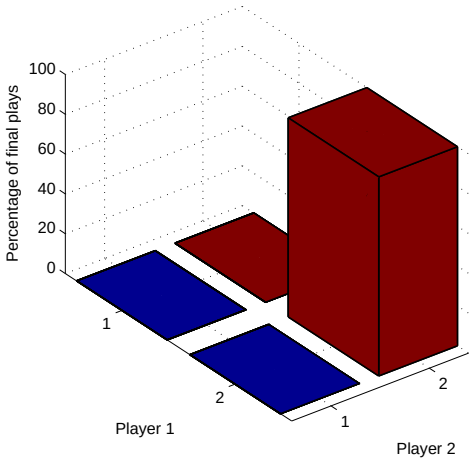


Figure 2: Best-response graph for the prisoner's dilemma.

of the received payoff, averaged over the 1000 runs (the dotted lines represent the standard deviation observed in the different runs). Figure 3.b) represents the percentage out of the 1000 runs that the algorithm converged to each joint strategy.



a) Evolution of expected payoff;



b) Limit strategies.

Figure 3: Learning performance in the prisoner's dilemma.

As mentioned, this game has a single Nash equilibrium, consisting of the pure strategy (R, R). To this joint strategy corresponds a payoff of $(-5, -5)$. By observing Figure 3.a) we can see that the average payoff received by each player converged to -5, indicating that the algorithm converged to the Nash equilibrium as expected. This is also clearly observable in Figure 3.b): the algorithm converged to the joint strategy (R, R) 100% of the 1000 runs.

4.0.2 Battle of Sexes

The battle of sexes, also known as the "Bach or Stravinsky" game is described by the payoff function in Figure 4. In this game, a couple must decide whether to go to a Bach concert or a Stravinsky concert. The man prefers the Bach concert over the Stravinsky concert, while the woman prefers Stravinsky over Bach. However, both prefer to attend a concert with company than alone.

Unlike the prisoner dilemma, this game has *two* pure Nash equilibria. This can be observed from the best-response graph in Figure 5, where it is also clear that the game is weakly acyclic. Notice that, in each Nash equilibrium, one of the players "submits" to the other player's will (hence the designation of "Battle of Sexes"), which is preferrable to attending the concert alone.

We applied our algorithm to this game. As in the prisoner's dilemma, we ran 1000 independent Monte-Carlo runs, each run consisting of 900 plays of the game. We again used Boltzmann exploration with decaying temperature factor. The results are depicted in Figure 6.

	B	S
B	20, 5	0, 0
S	0, 0	5, 20

Figure 4: Payoff for the battle of sexes.

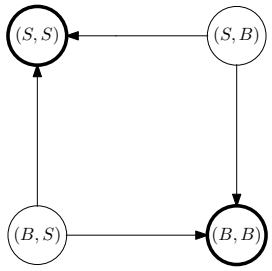
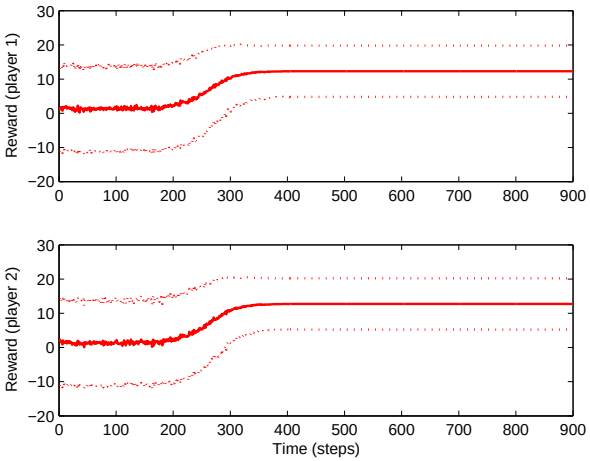
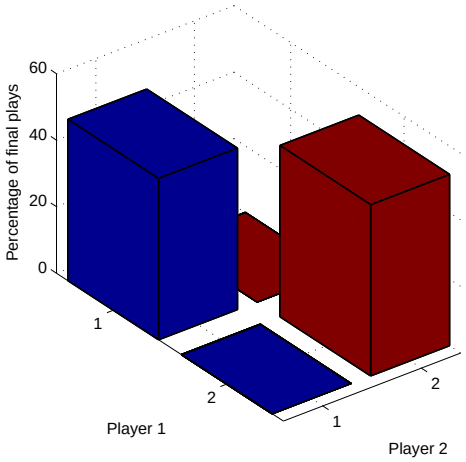


Figure 5: Best-response graph for the BoS.



a) Evolution of expected payoff;



b) Limit strategies.

Figure 6: Learning performance in the battle of sexes.

This game has two Nash equilibria, consisting of the pure strategies (B, B) and (S, S). Each such strategy rewards one of the players with a payoff of 20 and the other with a payoff of 5. Notice in Figure 6.a) that the average payoff received by each player converged to 12.5. This means that the algorithm converged approximately half of the times to each of the two Nash equilibria. This is confirmed by observing Figure 6.b), where the algorithm converged to each of the joint strategies (B, B) and (S, S) about 50% of the 1000 runs. This is an expected result: since the game is weakly acyclic, in each run the algorithm will converge to one of the two equilibria. However, there is no reason why one equilibrium is preferable to the other and, therefore, the algorithm will converge to each of the two equilibria with a 50% probability.

4.0.3 Diagonal game

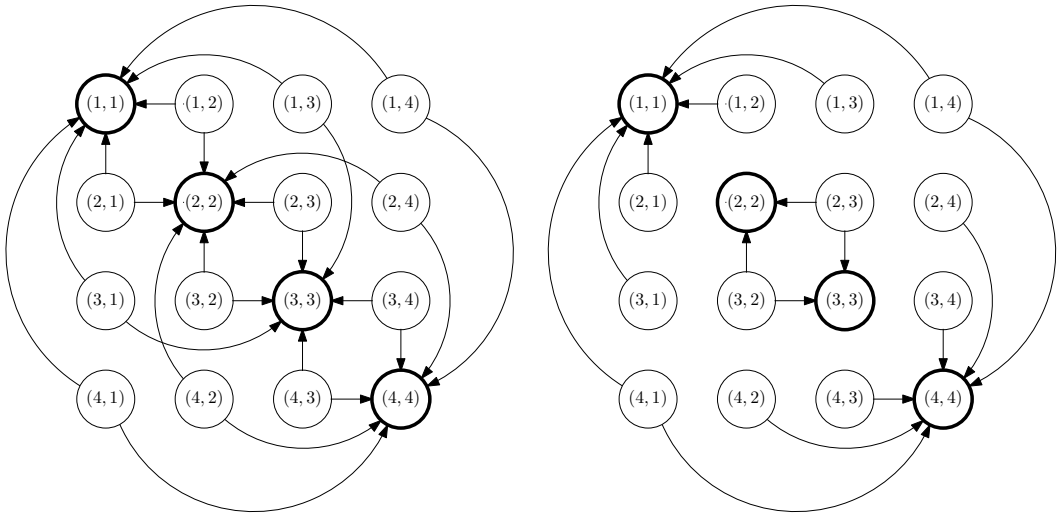
We next considered a 2-player, fully cooperative game described by the payoff function in Figure 7. We considered two values for the parameter ψ, namely $\psi = 0$ and $\psi = 0.1$. Notice that, in both situations, the diagonal elements corresponding to the joint actions $(1, 1)$, $(2, 2)$, $(3, 3)$ and $(4, 4)$ yield higher payoff than the remaining joint actions, as if rewarding the two players for "agreeing" upon their individual actions.

This game presents *four* pure Nash equilibria, corresponding to the diagonal elements in the payoff matrix (Figure 7). This motivates the naming of the game as the "diagonal game". The four Nash equilibria are evident from the best-response graphs in Figure 8 for the situations where $\psi = 0$ and $\psi = 0.1$. Notice, furthermore, that the game is weakly acyclic in both situations.

We applied our algorithm to both stances of the game, considering $\psi = 0$ and $\psi = 0.1$. We ran 1000 independent Monte-Carlo runs, each consisting of 900 plays of the game. The results are depicted in Figures 9 and 10.

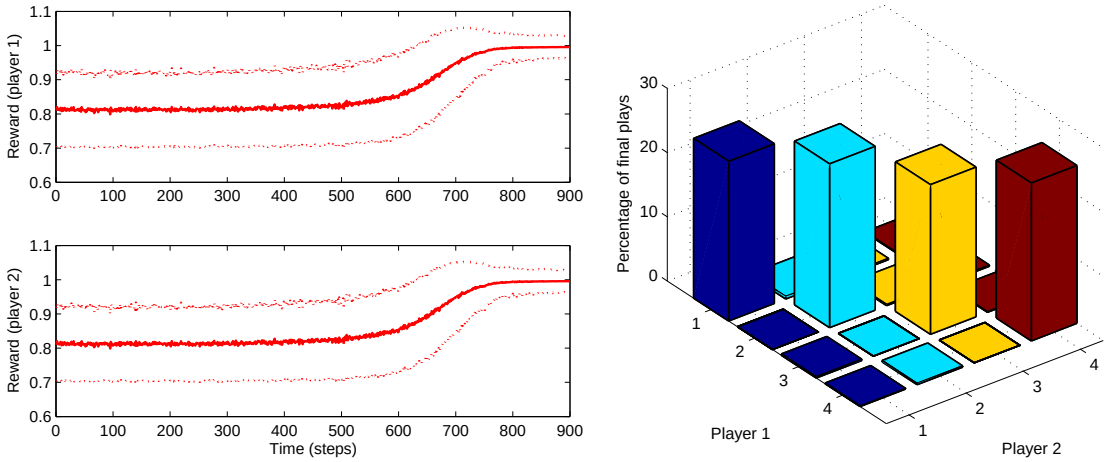
	1	*2*	*3*	*4*
1	1	0.75	0.75	0.75
2	0.75	$1-\psi$	0.75	0.75
3	0.75	0.75	$1-\psi$	0.75
4	0.75	0.75	0.75	1

Figure 7: Payoff for the fully cooperative, diagonal game.



a) $\psi = 0$; b) $\psi = 0.1$.

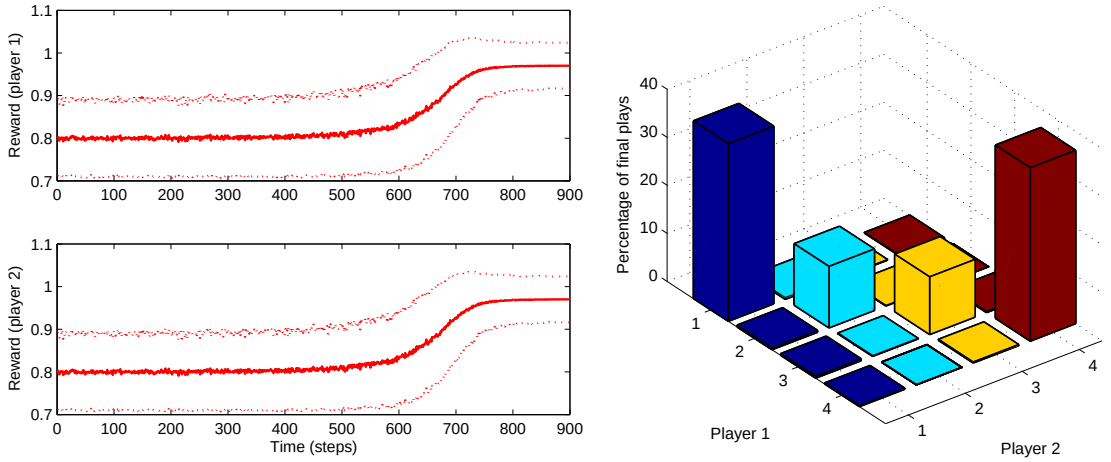
Figure 8: Best-response graphs for the diagonal game when $\psi = 0$ and $\psi = 0.1$.



a) Evolution of expected payoff; b) Limit strategies.

Figure 9: Learning performance in the diagonal game when $\psi = 0$.

We start by observing the results in Figure 9, concerning the situation where $\psi = 0$. In this situation, we have four Nash equilibria yielding a similar payoff of 1. This means that, as in the Battle of Sexes, the algorithm will expectedly converge to any of the four equilibria with



a) Evolution of expected payoff; b) Limit strategies.

Figure 10: Learning performance in the diagonal game when $\psi = 0.1$.

equal probability, since no equilibria is preferrable to the other. This is indeed the case, as seen in Figure 9: the average payoff to each player converges to 1 (Figure 9.a)) and the algorithm converged to each of the four equilibria about 25% of the times.

When considering the situation where $\psi = 0.1$ the situation is a little different. In this case, the four equilibria do not yield similar results and this will affect the convergence pattern of the algorithm. We start by noticing in Figure 10.a) that the expected payoff for both players converges to 0.975. This value has a precise interpretation that we provide next.

By close observation of the best-response graph in Figure 8.b) we notice, for example, that the equilibrium $(1, 1)$ can be reached from 7 different joint actions, namely, $(1, 1)$, $(1, 2)$, $(2, 1)$, $(1, 3)$, $(3, 1)$, $(1, 4)$ and $(4, 1)$. However, the joint actions $(1, 4)$ and $(4, 1)$ also lead to the equilibrium $(4, 4)$. This means that, out of the 16 possible joint actions, 5 lead to $(1, 1)$ and 2 other lead to $(1, 1)$ half of the times. This reasoning allows to conclude that we expect $(1, 1)$ to be the limit point of our algorithm $6/16 = 37.5\%$ of the times. The same reasoning can be applied to the equilibrium $(4, 4)$. As for the equilibria $(2, 2)$ and $(3, 3)$, the same reasoning leads to the conclusion that each of these equilibria will be reached $2/16 = 12.5\%$ of the times. These are, indeed, the results depicted in Figure 10.b) and further lead to the conclusion that the average expected payoff for each player is

$$r_{av} = 2 \times 0.375 \times 1 + 2 \times 0.125 \times 0.9 = 0.975.$$

4.0.4 Prejudice game

The prejudice game is a fully cooperative game described by the payoff function in Figure 7. This game models a situation in which two agents must (between the two) execute action 2. As long as action 2 is executed, the other agent can successfully execute action 1. However, if some of the players executes action 1 without the other executing action 2, they will both be penalized by receiving -1. The other situations are not important and are thus rewarded with a payoff of 0.

This game has multiple pure Nash equilibria, marked in bold in the best response graph (Figure 12). It is also a weakly acyclic game, so our algorithm can immediately be applied with guaranteed convergence. The results are depicted in Figure 13.

Conducting an analysis similar to the one in the previous game, we expect the algorithm to converge to the $(2, 2)$ equilibrium about 31.25% of the times. The equilibria $(1, 2)$ and $(2, 1)$ should be attained about 21.875% of the times and the remaining 4 equilibria about 6.25% of the times. By observing Figure 13.b) we can verify that this is indeed so. This leads to an average expected payoff for each player of

$$r_{av} = 0.3125 \times 1 + 2 \times 0.21875 \times 1 + 4 \times 0.0625 \times 0 = 0.75.$$

	1	*2*	*3*	*4*
1	-1	1	-1	-1
2	1	1	0	0
3	-1	0	0	0
4	-1	0	0	0

Figure 11: Payoff for the prejudice game.

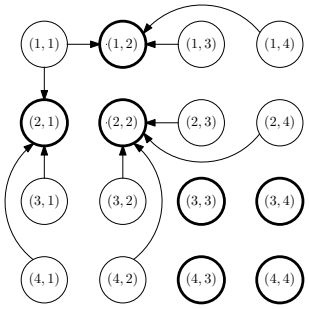
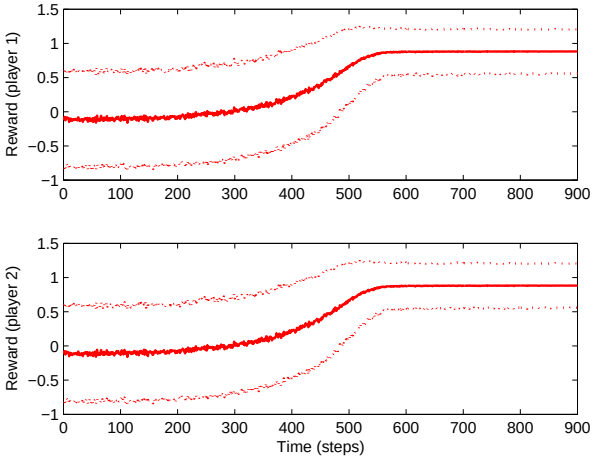
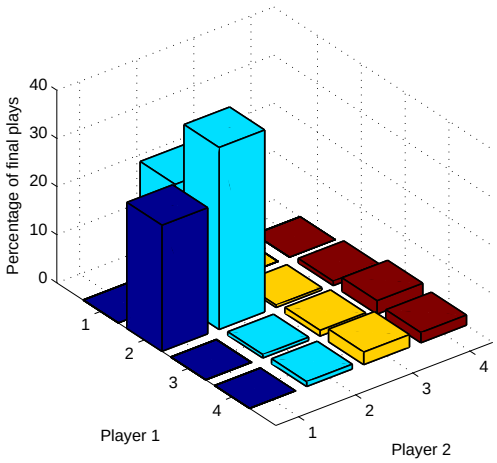


Figure 12: Best-response graph for the prejudice game.



a) Evolution of expected payoff;



b) Limit strategies.

Figure 13: Learning performance in the prejudice game.

We remark that the obtained value in Figure 13.a) is slightly superior, since in the reported results the algorithm converged to the sub-optimal equilibria with a frequency slightly below the expected value of 6.25%. This is justified by the random exploration: since the difference between the optimal and suboptimal equilibria is significant, the use of Boltzmann exploration "facilitates" the convergence to the optimal equilibria. We notice that this phenomenon is much less observable in the diagonal game with $\psi = 0.1$, since the difference between the optimal and the suboptimal equilibria is much less significant.

4.0.5 3-Player game

We now consider a fully cooperative 3-player game with multiple equilibria introduced by Wang and Sandholm [20]. In this game, 3 players have available 3 possible actions, α, β and γ. The players are rewarded maximum payoff if all 3 coordinate in the same individual action; they are rewarded a small payoff if all play different actions. Otherwise, they are penalized with a negative payoff.

The game has several Nash equilibria, marked in bold in the best-response graph in Figure 8. Clearly, the game is weakly acyclic.

We applied our algorithm to the game, running 1000 independent Monte-Carlo runs, each consisting of 900 plays of the game. The results are depicted in Figure 16.

Once again conducting an analysis similar to the one in the previous games, we expect the algorithm to converge to the optimal equilibria about 25.9% of the times and to the suboptimal equilibria about 3.7% of the times. As in the previous example, the use of Boltzmann exploration

	$\alpha\alpha$	$\alpha\beta$	$\alpha\gamma$	$\beta\alpha$	$\beta\beta$	$\beta\gamma$	$\gamma\alpha$	$\gamma\beta$	$\gamma\gamma$
α	10	-20	-20	-20	-20	5	-20	5	-20
β	-20	-20	5	-20	10	-20	5	-20	-20
γ	-20	5	-20	5	-20	-20	-20	-20	10

Figure 14: Payoff for the 3-player game by Wang and Sandholm [20].

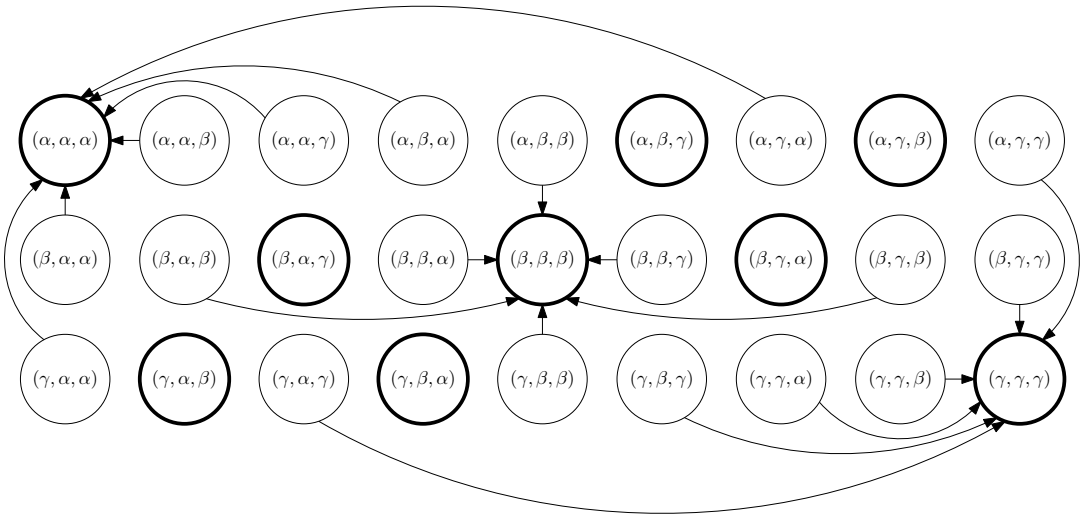
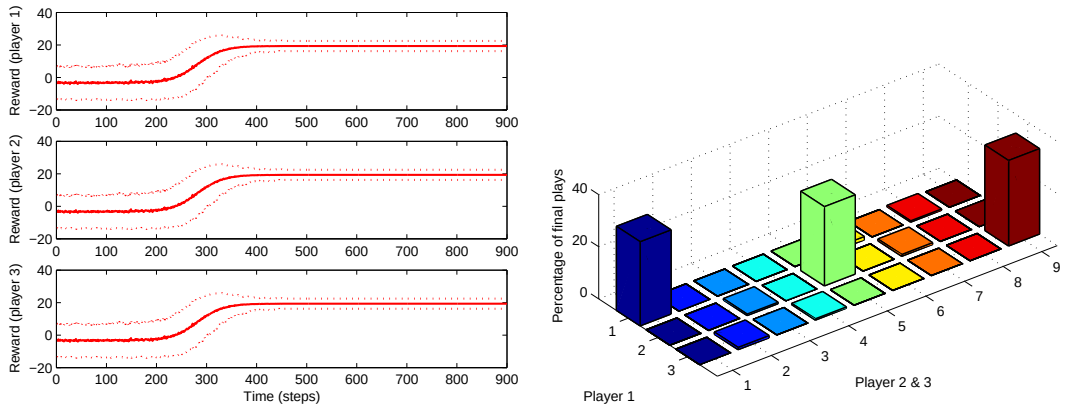


Figure 15: Best-response graph for the 3-player game by Wang and Sandholm [20].



a) Evolution of expected payoff; b) Limit strategies.

Figure 16: Learning performance in the 3-player game by Wang and Sandholm [20].

leads to a slight increase in the number of runs converging to the optimal equilibria and consequent decrease in the number of runs converging to the suboptimal equilibria (Figure 16.b)). This is also noticeable since the average payoff per player actually converges to 20 (Figure 16.a)), which indicates that each optimal equilibrium is actually reached about 1/3 of the times.

	1	*1*
1	5	0
2	0	20

Figure 17: Payoff for a zero-sum game.

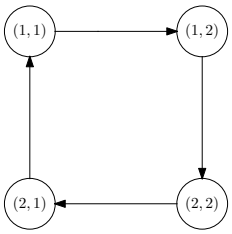
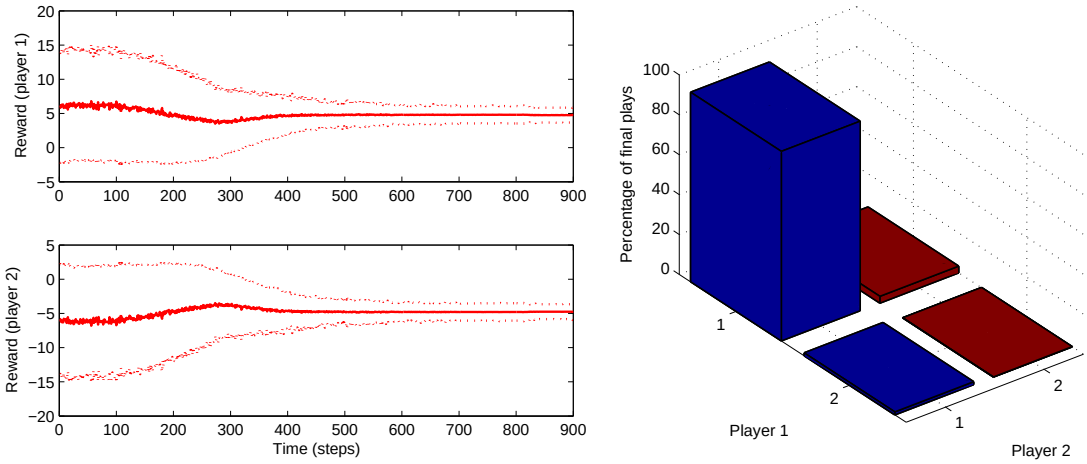


Figure 18: Best-response cyclic graph.

4.0.6 Cyclic game

Finally, we present a two-player, zero-sum game with no pure Nash equilibrium. The payoff function for the game is presented in Figure 17. Since this game has no pure Nash equilibrium, it cannot be weakly acyclic, as verified from the best-response graph in Figure 18. Therefore, it is not expected that our algorithm converges to an equilibrium, since the algorithm can only converge to pure strategies (and the equilibrium for this game is a mixed one).[3] We remark, however, that the Nash equilibrium for this game corresponds to an expected reward of 8 for player 1 and of -8 for player 2.

We applied our algorithm to the game, running 1000 independent Monte-Carlo runs, each consisting of 900 plays of the game. The results are depicted in Figure 16.



a) Evolution of expected payoff; b) Limit strategies.

Figure 19: Learning performance in the cyclic game.

Notice in Figure 19.a) that the average payoff received by player 1 converged to about 5 (and to -5 for player 2). This means that the algorithm converged to the pure strategy $(1, 1)$ as observed in Figure 19.b). Curiously, this is the pure strategy "closest" to the actual Nash equilibrium for the game.

5 Conclusions

In this work we generalized adaptive play [22] to situations where actions and payoffs are not observable. We showed that our algorithm converges with probability 1 to a (pure) Nash equilibrium if it exists. However, if no (pure) Nash equilibrium exists, and as seen in the example of the cyclic game, the algorithm may eventually converge to the pure strategy which is "closest" to

[3] The Nash equilibrium for this game consists on the mixed strategy that plays action 1 with a probability 0.8 and action 2 with probability 0.2.

a mixed strategy Nash equilibrium for the game. Our algorithm, independent adaptive learning, proceeds as in standard adaptive play by using incomplete sampling of finite length history of past actions/payoffs. To handle the lack of action observability, the algorithm requires infinite exploration to avoid getting "stuck" in non-equilibrium strategies. We provided a formal proof of convergence and some experimental results obtained with our algorithm in several games with different properties.

We are interested in extending the independent adaptive learning algorithm (or a variation thereof) to multi-state problems, such as Markov games. We are also interested in applying the algorithm to real world situations with a large number of agents with large action repertoires.

References

[1] C. Boutilier. Sequential optimality and coordination in multiagent systems. In *Proc. 16th Int. Joint Conf. Artificial Intelligence*, pages 478–485, 1999.

[2] George W. Brown. Some notes on computation of games solutions. Research Memoranda RM-125-PR, RAND Corporation, Santa Monica, California, 1949.

[3] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98)*, pages 746–752, 1998.

[4] Robert H. Crites and Andrew G. Barto. Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 33(2-3):235–262, 1998.

[5] Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. Game-theoretic control for robot teams. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, pages 1175–1181, 2004.

[6] Junling Hu and Michael P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*, pages 242–250, 1998.

[7] S. Kapetanakis and D. Kudenko. Reinforcement learning of coordination in cooperative multi-agent systems. In *Proc. 19th Nat. Conf. Artificial Intelligence*, pages 326–331, 2002.

[8] S. Kapetanakis and D. Kudenko. Improving on the reinforcement learning of coordination in cooperative multi-agent systems. In *Proc. 2nd Symp. Adaptive Agents and Multi-agent Systems*, pages 89–94, 2002.

[9] V. Krishna and T. Sjöström. On the convergence of fictitious play. Technical Report 1717, Institute of Economic Research, Harvard University, 1995.

[10] M. Lauer and M. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proc. 17th Int. Conf. Machine Learning*, pages 535–542, 2000.

[11] David S. Leslie and E. J.Collins. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, 2006.

[12] Michael L. Littman. Value-function reinforcement learning in Markov games. *Journal of Cognitive Systems Research*, 2(1):55–66, 2001.

[13] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In Ramon López de Mántaras and David Poole, editors, *Proceedings of the 11th International Conference on Machine Learning (ICML'94)*, pages 157–163, San Francisco, CA, 1994. Morgan Kaufmann Publishers.

[14] Don Monderer and Lloyd S. Shapley. Fictitious play property for games with identical interests. *Journal of Economic Theory*, 68:258–265, 1996.

[15] John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36:48–49, 1950.

[16] Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54:296–301, 1951.

[17] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvari. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3), 2000.

[18] M. Tan. *Multi-agent reinforcement learning: Independent vs. cooperative agents*, pages 487–494. 1997.

[19] Ben Van der Genugten. A weakened form of fictitious play in two-person zero-sum games. *International Game Theory Review*, 2(4):307–328, 2000.

[20] X. Wang and T. Sandholm. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In *Advances in Neural Information Processing Systems*, volume 15, pages 1571–1578. MIT Press, 2003.

[21] Ying Wang and Clarence W. de Silva. Multi-robot box-pushing: single-agent Q-learning vs. team Q-learning. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, pages 3694–3699, 2006.

[22] H. Peyton Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993.