# Convergence of Independent Adaptive Learners*

Francisco S. Melo and Manuel C. Lopes

Institute for Systems and Robotics,
Instituto Superior Técnico,
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 in self play, if they exist, and to a best response strategy against stationary opponents. We discuss the relation between our result and convergence results of adaptive play [1]. 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 [2,3]. Finally we present experimental results illustrating the main ideas of the paper.

## 1 Introduction

Game theory is traditionally used in economics, where it provides powerful models to describe interactions of economical agents. However, recent years have witnessed an increasing interest from the computer science and robotic communities in applying game theoretic models to multi-agent systems. 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 [4] distinguishes two fundamental classes of learning agents: *independent learners* (IL) and *joint-action learners* (JAL). The former have no knowledge on the other agents, interacting with the environment as if no other decision-makers existed. In particular, they are unable to observe the rewards and actions of the other agents. Joint action leaners, 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 [5]. 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 [6] or adaptive play [1]. Joint action observability is also commonly assumed in several domains studied in the economic literature (*e.g.*, auctions or

* Work partially supported by POS_C that includes FEDER funds. The first author acknowledges the PhD grant SFRH/BD/3074/2000.

J. Neves, M. Santos, and J. Machado (Eds.): EPIA 2007, LNAI 4874, pp. 555–567, 2007.

exchanges[1]) and several learning algorithms are available that make use of such assumption [7,8].

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. In [9,4] some empirical evidence is gathered that describes the convergence properties of reinforcement learning methods in multi-agent settings. In [10], the authors study independent learners in deterministic settings. Posterior works [11,12] address non-deterministic settings. Recent results have established the convergence of a variation of fictitious play for independent learners [3]. In a different approach, Verbeeck et al. [13] propose an independent learning algorithm for repeated games that converges to a *fair periodical policy* that periodically alternates between several Nash equilibria.

In this paper, we propose and analyze the performance of *independent adaptive learning*, a variation of adaptive play for independent learners. This algorithm has an obvious advantage over the original adaptive learning algorithm [1], 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. Our results show that a very simple learning approach, requiring no communication or knowledge on the other agents, is still able to exhibit a *convergent* and *rational* behavior, in the sense of [14]. This means that independent adaptive learning is able to attain a Nash equilibrium in self-play and converge to a best-response strategy against stationary opponents. We show that, in weakly acyclic repeated games, independent adaptive learners converge to pure Nash equilibria, if they exist. This convergence is attained in both *beliefs* and *behavior*. We experimentally validate our results in several simple games.

## 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 tuple $\big(N, (\mathcal{A}_k), (r_k)\big)$, where $N$ is the number of players, $\mathcal{A}_k$ is the set of *individual actions* of player $k$, $k = 1, \ldots, N$ and $\mathcal{A} = \times_{k=1}^{N} \mathcal{A}_k$ is the set of *joint actions* for the group. Each function $r^k : \mathcal{A} \to \mathbb{R}$ is a *reward function* or *payoff function*, defining a preference relation on the set $\mathcal{A}$.

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* or *action profile*. 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 $a_k$.

[1] Exchanges are also known as double actions.

In strategic games it is not possible to have memory effects in the players. 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

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 $a_k \in \mathcal{A}_k$. In a Nash equilibrium no player benefits from individually deviating its play from $a^*$. We emphasize that not every strategic game has a Nash equilibrium.

A *strategy* for player $k$ is a probability distribution over the set $\mathcal{A}_k$. A strategy $\sigma_k$ assigns a probability $\sigma_k(a_k)$ to each action $a_k \in \mathcal{A}_k$. We say that player $k$ follows strategy $\sigma_k$ 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 $\sigma_k$ assigns probability 1 to some action $a_k \in \mathcal{A}_k$, then $\sigma_k$ is a *pure strategy*. Otherwise, it is called a *mixed strategy*. We define the concepts of *joint strategy* or *strategy profile* and *reduced joint strategy* in a similar manner as defined for actions. The *support* of a strategy $\sigma_k$ 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 $\sigma^*$ such that, for any strategy $\sigma$ and for every player $k = 1, \ldots, N$,

$$\mathbb{E}_{\sigma^*}[R_k] \geq \mathbb{E}_{(\sigma^*_{-k},\, \sigma_k)}[R_k] \tag{1}$$

where $\mathbb{E}_{\sigma^*}[\cdot]$ is the expectation conditioned on the strategy $\sigma^*$ and $R_k$ is the random variable denoting the outcome of the game for player $k$. The Nash equilibrium is *strict* if (1) holds with a strict inequality. 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 [6] to determine the solution for a strictly competitive game. This procedure was shown to converge in this class of games in [15] and later extended to other classes of games by several authors (see, for example, [3]).

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 $t^{\text{th}}$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 [15] showed that this methodology yields two sequences $\{\hat{\sigma}_1\}_t$ and $\{\hat{\sigma}_2\}_t$ converging respectively to $\sigma_1^*$ and $\sigma_2^*$ such that $(\sigma_1^*, \sigma_2^*)$ is a Nash equilibrium for the game $\big(\{1,2\}, (\mathcal{A}_k), (r_k)\big)$.

### 2.4 Adaptive Play

Adaptive play was first proposed by Young [1] 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 with probability 1 (w.p.1) to a pure strategy Nash equilibrium, both in *beliefs* and in *behavior*.

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 two given integers 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 [1] 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 $\Gamma$ 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 *weakly acyclic* if, given any vertex in its best response graph there is a directed path to a vertex from which there is no exiting edge (a sink).

A sink as described in the previous definition corresponds necessarily to a strict Nash equilibrium. Given a weakly acyclic strategic game $\Gamma$, we denote by $L(a)$ the shortest path from the vertex $a$ to a strict Nash equilibrium in the best response graph of $\Gamma$ and by $L(\Gamma) = \max_{a \in \mathcal{A}} L(a)$. Young [1] showed that for any weakly acyclic strategic game, adaptive play converges w.p.1 to a strict Nash equilibrium as long as $K \leq \frac{m}{L(\Gamma)+2}$.

## 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 [1], 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.

### 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 and rewards referred henceforth concern a particular player $k$ in $\Gamma$, 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 the set $\{u\}$ with $u \in \mathcal{A}_k$.[2] 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 in [1].

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 definitions.

**Definition 1 (Greedy strategy).** *An individual strategy* $\sigma_k$ *is* greedy with respect to (w.r.t.) a payoff function $r$ *if it assigns probability* 1 *to the action* $a^* = \arg\max_{a \in \mathcal{A}^k} r(a)$.

**Definition 2 (GLIE strategy [16]).** *An individual strategy* $\sigma_k$ *is* greedy in the limit with infinite exploration (GLIE) *if (i) each action is visited infinitely often and (ii) 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:

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

where $T_t$ is a temperature parameter that decays at an adequate rate (see [16] for further details).

**Theorem 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.*

[2] The indicator function for a set $A$, $\mathbb{I}_A$, takes the value 1 when the argument is in $A$ and 0 otherwise.

*Proof.* To prove our result we make use of two results from [1]. In this paper, Young showed that in weakly acyclic games, if $K \leq \frac{m}{L(\Gamma)+2}$, then as the *experimentation probability* 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. The experimentation probability in [1] defines the probability of a player choosing non-greedy actions (*i.e.*, making a "mistake").

To prove our result, we make use of the results from [1] by first considering a fixed, positive exploration rate. The exploration rate in our algorithm plays the role of the "experimentation probability" in [1]. 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. The conclusions of our theorem now follow from the results in [1] as long as we show that the probability of making "mistakes" in our algorithm goes to zero at a suitable rate, *i.e.*, slower than the aforementioned Markov chain converges to stationarity.

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, "mistakes" can arise either due to the exploration or to the subestimation of action-values.

Two important observations are now 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 the probability of exploration goes to zero 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.

These two remarks lead to the conclusion that 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 [1] and the proof is complete. □

## 4 Experimental Results

In this section we present the results of our method for several simple games. In each game, we applied our algorithm by running 1000 independent Monte-Carlo trials, each trial consisting of 900 plays of the same game. We used Boltzmann exploration with decaying temperature factor to ensure sufficient exploration of

all actions. We present in Figures 3.a), 6.a), 9.a) and 12.a) the average evolution of the received payoff for each game (solid line) and the corresponding standard deviation (in dashed line) for each game. We also present in Figures 3.b), 6.b), 9.b) and 12.b) the percentage of trials that the algorithm converged to each joint strategy in each game.

**Prisoner's dilemma.** The prisoner's dilemma is a well-known game from game theory whose payoff function is represented in Fig. 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 confessesing is released while the other serves the full sentence. If both prisoners confess, they do not serve the full sentence, but still remain in jail for a long time.

| | $S$ | $R$ |
|---|---|---|
| $S$ | 5, 5 | -10, 20 |
| $R$ | 20, -10 | -5, -5 |

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

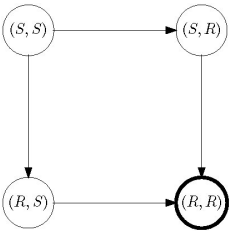
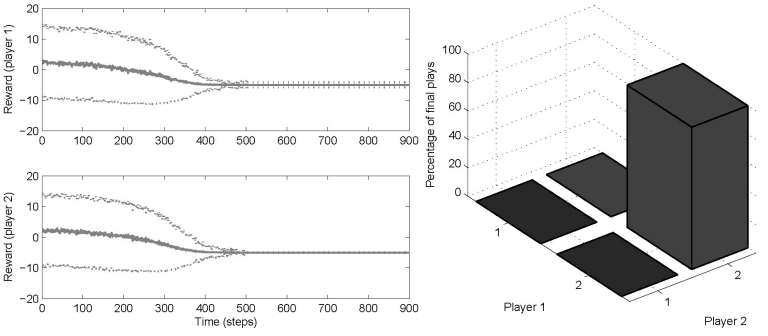



**Fig. 2.** Best-response graph for the prisoner's dilemma

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 Fig. 2, where it is also clear that the game is weakly acyclic.

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 Fig. 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 Fig. 3.b): the algorithm converged to the joint strategy $(R, R)$ 100% of the 1000 runs.

**Diagonal game.** We next considered a 2-player, fully cooperative game described by the payoff function in Fig. 4. Notice that 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.




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

**Fig. 3.** Learning performance in the prisoner's dilemma

| | *1* | *2* | *3* | *4* |
|---|---|---|---|---|
| *1* | 1 | 0.75 | 0.75 | 0.75 |
| *2* | 0.75 | 0.9 | 0.75 | 0.75 |
| *3* | 0.75 | 0.75 | 0.9 | 0.75 |
| *4* | 0.75 | 0.75 | 0.75 | 1 |

**Fig. 4.** Payoff for the fully cooperative, diagonal game

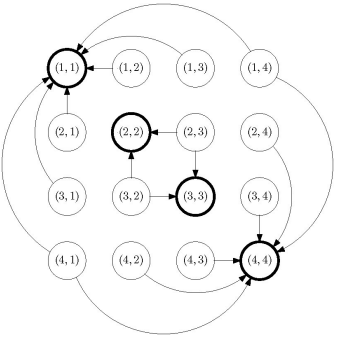


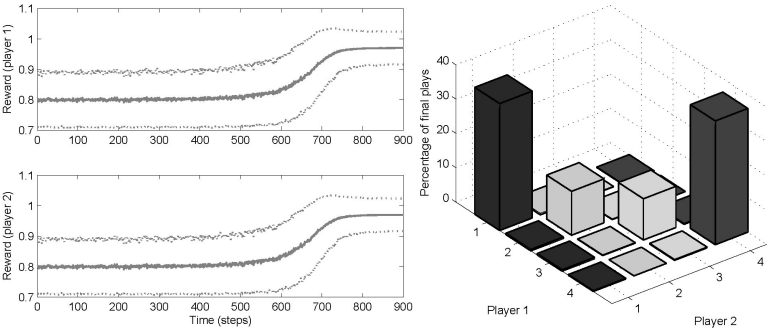

**Fig. 5.** Best-response graphs for the diagonal game

This game presents *four* pure Nash equilibria, corresponding to the diagonal elements in the payoff matrix (Fig. 4). This motivates the naming of the game as the "diagonal game". The four Nash equilibria are evident from the best-response graph in Fig. 5. Notice, furthermore, that the game is weakly acyclic.

We applied our algorithm to both stances of the game and depicted the results in Fig. 6.

Notice that the four equilibria do not yield similar payoffs and this will affect the convergence pattern of the algorithm. We start by noticing in Fig. 6.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 Fig. 5.b) we notice, for example, that the equilibrium $(1, 1)$ can be reached from 7 different joint actions. 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 $(3, 3)$,




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

**Fig. 6.** Learning performance in the diagonal game when $\psi = 0.1$

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 Fig. 6.b) and further lead to the conclusion that the average expected payoff for each player is $r_{av} = 2 \times 0.375 \times 1 + 2 \times 0.125 \times 0.9 = 0.975$.

**3-Player game.** We now consider a fully cooperative 3-player game with multiple equilibria introduced in [17]. In this game, 3 players have available 3 possible actions, $\alpha$, $\beta$ and $\gamma$. 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.

| | $\alpha\alpha$ | $\alpha\beta$ | $\alpha\gamma$ | $\beta\alpha$ | $\beta\beta$ | $\beta\gamma$ | $\gamma\alpha$ | $\gamma\beta$ | $\gamma\gamma$ |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 10 | -20 | -20 | -20 | -20 | 5 | -20 | 5 | -20 |
| $\beta$ | -20 | -20 | 5 | -20 | 10 | -20 | 5 | -20 | -20 |
| $\gamma$ | -20 | 5 | -20 | 5 | -20 | -20 | -20 | -20 | 10 |

**Fig. 7.** Payoff for the 3-player game from [17]

The game has several Nash equilibria, marked in bold in the best-response graph in Fig. 5. Clearly, the game is weakly acyclic.

We applied our algorithm to the game. The results are depicted in Fig. 9.

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. The use of Boltzmann exploration 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 (Fig. 9.b)). This is also noticeable since the average payoff per player actually converges to 20 (Fig. 9.a)), which indicates that each optimal equilibrium is actually reached about 1/3 of the times.

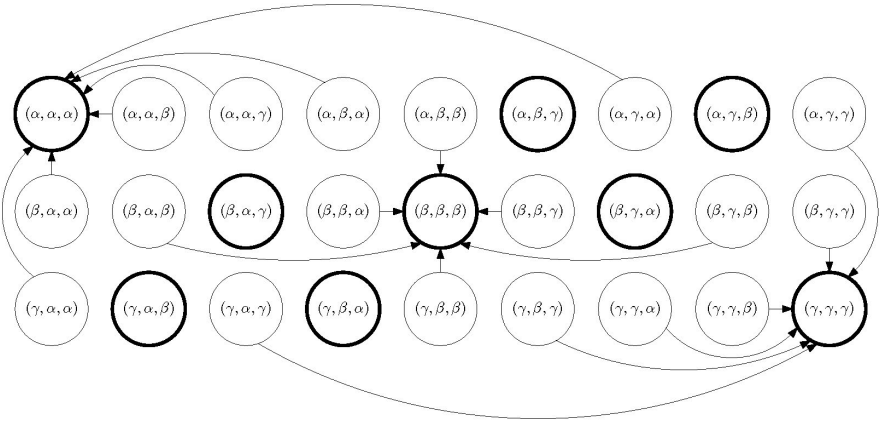
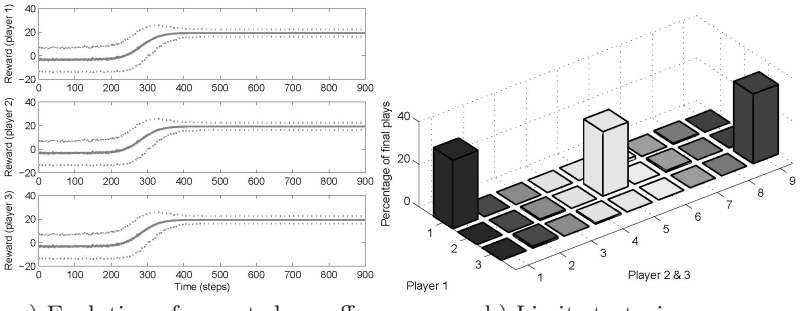



**Fig. 8.** Best-response graph for the 3-player game from [17]




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

**Fig. 9.** Learning performance in the 3-player game from [17]

**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 Fig. 10. Since this game has no pure Nash equilibrium, it cannot be weakly acyclic, as verified from the best-response graph in Fig. 11. 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 Fig. 9.

Notice in Fig. 12.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

[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.

| | *1* | *1* |
|---|---|---|
| *1* | 5 | 0 |
| *2* | 0 | 20 |

**Fig. 10.** Payoff for a zero-sum game

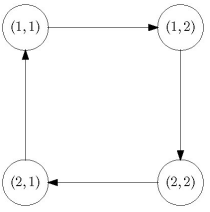
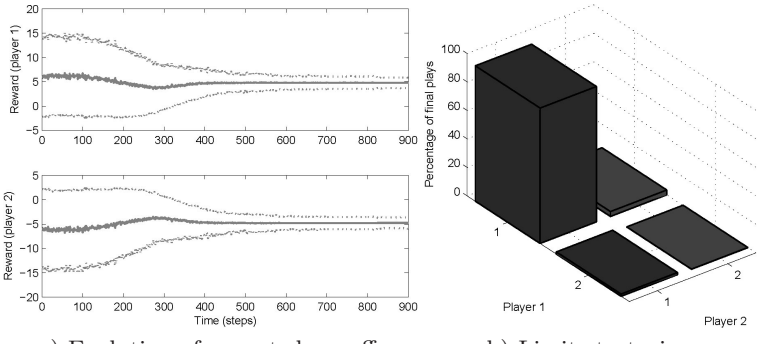



**Fig. 11.** Best-response cyclic graph




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

**Fig. 12.** Learning performance in the cyclic game

to the pure strategy $(1, 1)$ as observed in Fig. 12.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 [1] 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 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. Further experimental results can be found in [18].

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. Young, H.P.: The evolution of conventions. Econometrica 61(1), 57–84 (1993)
2. Van der Genugten, B.: A weakened form of fictitious play in two-person zero-sum games. International Game Theory Review 2(4), 307–328 (2000)
3. Leslie, D.S., Collins, E.J.: Generalised weakened fictitious play. Games and Economic Behavior 56(2), 285–298 (2006)
4. Claus, C., Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems. In: AAAI, pp. 746–752 (1998)
5. Littman, M.L.: Value-function reinforcement learning in Markov games. Journal of Cognitive Systems Research 2(1), 55–66 (2001)
6. Brown, G.W.: Some notes on computation of games solutions. Research Memoranda RM-125-PR, RAND Corporation, Santa Monica, California (1949)
7. He, M., Leung, H.F., Jennings, N.R.: A fuzzy logic based bidding strategy for autonomous agents in continuous double auctions. IEEE Trans. Knowledge and Data Engineering 15(6), 1345–1363 (2002)
8. Bagnall, A.J., Toft, I.E.: Zero intelligence plus and Gjerstad-Dickhaut agents for sealed bid auctions. In: Kudenko, D., Kazakov, D., Alonso, E. (eds.) Adaptive Agents and Multi-Agent Systems II. LNCS (LNAI), vol. 3394, Springer, Heidelberg (2005)
9. Tan, M.: Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Readings in Agents, pp. 487–494 (1997)
10. Lauer, M., Riedmiller, M.: An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In: ICML, pp. 535–542 (2000)
11. Kapetanakis, S., Kudenko, D.: Improving on the reinforcement learning of coordination in cooperative multi-agent systems. In: Symp. AAMAS, pp. 89–94 (2002)
12. Lauer, M., Riedmiller, M.: Reinforcement learning for stochastic cooperative multi-agent-systems. In: AAMAS, pp. 1516–1517 (2004)
13. Verbeeck, K., Nowé, A., Parent, J., Tuyls, K.: Exploring selfish reinforcement learning in repeated games with stochastic rewards. JAAMAS 14, 239–269 (2006)
14. Bowling, M., Veloso, M.: Rational and convergent learning in stochastic games. In: IJCAI 2001. Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 1021–1026 (2001)
15. Robinson, J.: An iterative method of solving a game. Annals of Mathematics 54, 296–301 (1951)
16. Singh, S., Jaakkola, T., Littman, M., Szepesvari, C.: Convergence results for single-step on-policy reinforcement-learning algorithms. Machine Learning 38(3) (2000)
17. Wang, X., Sandholm, T.: Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In: NIPS, vol. 15, pp. 1571–1578 (2003)
18. Melo, F., Lopes, M.: Convergence of independent adaptive learners. Technical Report RT-603-07, Institute for Systems and Robotics (2007)