Reinforcement Learning to Play An Optimal Nash Equilibrium in Team Markov Games

Xiaofeng Wang
Electrical and Computer Engineering Department
Carnegie Mellon University
Pittsburgh, PA 15213
xiaofeng@andrew.cmu.edu

Tuomas Sandholm
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
sandholm@cs.cmu.edu

Abstract

Multiagent learning is a key problem in game theory and AI. It involves two interrelated learning problems: identifying the game and learning to play. These two problems prevail even in team games where the agents' interests do not conflict. Even team games can have multiple Nash equilibria, only some of which are optimal. We present *optimal adaptive learning* (OAL), the first algorithm that converges to an optimal Nash equilibrium for any team Markov game. We provide a convergence proof, and show that the algorithm's parameters are easy to set so that the convergence conditions are met. Our experiments show that existing algorithms do not converge in many of these problems while OAL does. We also demonstrate the importance of the fundamental ideas behind OAL: incomplete history sampling and biased action selection.

1 Introduction

Multiagent learning is a key problem in game theory and AI. For a decade, computer scientists have worked on extending reinforcement learning (RL) [7] to multiagent settings [11, 15, 5, 17]. In game-theoretic terms, multiagent RL is the problem of learning to play Nash equilibria in Markov games (aka. stochastic games) [16] when the game structure is unknown. This has been studied under various types of Markov games such as zero-sum Markov games [9], general-sum Markov games [6, 8] and team Markov games [2].

Multiagent RL in Markov games involves two interrelated learning problems: identifying the game and learning to play. These two problems prevail even in team Markov games where the agents have common interests. Even team Markov games can have multiple Nash equilibria, only some of which are optimal (that is, maximize sum of the agents' discounted payoffs). The agents should learn (enough of) the game structure, and to *coordinate* their play on an optimal Nash equilibrium.

Hu and Wellman presented an RL algorithm for Markov games where the agents learn to play Nash equilibrium if there is exactly one Nash equilibrium [6]. Littman introduced an RL algorithm where the agents learn the structure of any Markov game, but the agents might not learn to play (coordinate) [8]. Claus and Boutilier designed an RL algorithm for team Markov games. The agents learn to play a Nash equilibrium, but sometimes a suboptimal one [2]. Learning to play an optimal Nash equilibrium in team Markov games has been posed as one of the important open problems in multiagent reinforcement learning [9]. Equilibrium selection is an infamous problem in game theory also.

We present *optimal adaptive learning* (OAL), the first algorithm that converges to an optimal Nash equilibrium for any team Markov game (Section 3). It works for any number of agents. We prove its convergence, and show that OAL's parameters are easy to set so that the convergence conditions are met (Section 4). Our experiments show that existing algorithms do not converge in many of these problems while OAL does (Section 5). We also demonstrate the importance of the fundamental ideas behind OAL: incomplete history sampling and biased action selection.

2 The setting

2.1 MDPs and reinforcement learning (RL)

In a Markov decision problem, there is one agent in the environment. A fully observable Markov decision problem (MDP) is a tuple $\langle S, A, R, T\rangle$ where S is a finite state space; A is the space of actions the agent can take; $R : S \times A \rightarrow \Re$ is a payoff function ($R(s,a)$ is the expected payoff for taking action a in state s); and $T : S \times A \times S \rightarrow [0,1]$ is a transition function ($T(s,a,s')$ is the probability of ending in state s', given that action a is taken in state s). An agent's deterministic *policy* (aka. *strategy*) is a mapping from states to actions. We denote by $\pi(s)$ the action that policy π prescribes in state s. The objective is to find a policy π that maximizes $\sum_{t=0}^{\infty} \gamma^t E(r_t|\pi)$, where r_t is the payoff at time t, and $\gamma \in [0,1)$ is a discount factor. There exists a deterministic optimal policy π^* [12]. The Q-function for this policy, Q^*, is defined by the set of equations $Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \max_{a' \in A} Q^*(s',a')$. At any state s, the optimal policy chooses $\arg\max_a Q^*(s,a)$ [10].

Reinforcement learning (RL) [7] can be viewed as a sampling method for estimating Q^* when the payoff function R and/or transition function T are unknown. $Q^*(s,a)$ can be approximated by a function $Q_t(s,a)$ calculated from the agent's experience up to time t. The model-based approach uses samples to generate models of R and T, and then iteratively computes $Q_t(s,a) = R_t(s,a) + \gamma \sum_{s' \in S} T_t(s,a,s') \max_{a' \in A} Q_{t-1}(s',a')$.

Based on Q_t, a *learning policy* assigns probabilities to actions at each state. If the learning policy has the "*Greedy in the Limit with Infinite Exploration*" (GLIE) property, then Q_t will converge to Q^* (with either a model-based or model-free approach [7]) and the agent will converge in behavior to an optimal policy [14]. Using GLIE, every state-action pair is visited infinitely often, and in the limit the action selection is greedy with respect to the Q-function w.p.1. One common GLIE policy is Boltzmann exploration [14].

2.2 Multiagent reinforcement learning in team Markov games when the game is unknown

A natural extension of an MDP to multiagent environments is a *Markov game* (aka. *stochastic game*) [16]. In this paper we focus on *team Markov games*, that are Markov games where each agent receives the same payoff. In other words, there are no conflicts between the agents, but learning the game structure and learning to coordinate are nevertheless highly nontrivial.

Definition 1 *A team Markov game (aka identical-interest stochastic game) Γ is a tuple $\langle \alpha, S, A, R, T \rangle$, where α is a set of n agents; S is a finite state space; $A = \times A_{i=1\ldots n}$ is a joint action space of n agents; $R : S \times A \to \Re$ is the common payoff function; and $T : S \times A \times S \to [0,1]$ is a transition function.*

The objective of the n agents is to find a deterministic *joint policy (aka. joint strategy aka. strategy profile)* $\pi = \{\pi_{i=1\ldots n}\}$ (mapping from states to joint actions) so as to maximize the expected sum of their common discounted payoffs. The Q-function, $Q^\pi(s,a)$, is the expected sum of discounted payoffs given that the agents play joint action a in state s and follow policy π thereafter. The optimal Q-function $Q^*(s,a)$ is the Q-function for (each) optimal policy π^*. So, Q^* captures the game structure. The agents generally do not know Q^* in advance. Sometimes, they know neither the payoff structure nor the transition probabilities.

A joint policy $\{\pi_{i=1\ldots n}\}$ is a *Nash equilibrium* [1] if each individual policy is a best response to the others. That is, for all $i \in \alpha$, $s \in S$ and any individual policy π'_i, $Q^*(s, \{\pi_i(s)\} \cup \pi_{-i}(s)) \geq Q^*(s, \{\pi'_i(s)\} \cup \pi_{-i}(s))$, where π_{-i} is the joint policy of all agents except agent i. A Nash equilibrium is *strict* if the inequality above is strict. An *optimal Nash equilibrium* π^* is a Nash equilibrium that gives the agents the maximal expected sum of discounted payoffs. In team games, each optimal Nash equilibrium is an optimal joint policy (and there are no other optimal joint policies).

A joint action a is optimal in state s if $Q^*(s,a) \geq Q^*(s,a')$ for all $a' \in A$. If we treat $Q^*(s,a)$ as the payoff of joint action a in state s, we obtain a team game in matrix form. We call such a game a *state game* for state s. An optimal joint action in s is an optimal Nash equilibrium of that state game. Thus, the task of optimal coordination in a team Markov game boils down to having all the agents play an optimal Nash equilibrium in each state game [1].

However, a *coordination problem* arises if there are multiple optimal Nash equilibria. The 3-player game in Table 1 has three optimal Nash equilibria: $\{a1, b1, c1\}$, $\{a2, b2, c2\}$ and $\{a3, b3, c3\}$. To act optimally, the agents have to coordinate which equilibrium they will play. This problem is called *equilibrium selection*. A learning approach widely advocated to tackle this problem is *fictitious play* [3]: each agent i assumes that the other agents use a stationary (but unknown) joint policy, and i models that policy based on the frequency of past observations. Each agent i uses its best-response strategy to this model. Such *myopic adjustment* converges to a Nash equilibrium in many games including all team games. However, it does not generally converge to an *optimal* Nash equilibrium.

	$b1c1$	$b1c2$	$b1c3$	$b2c1$	$b2c2$	$b2c3$	$b3c1$	$b3c2$	$b3c3$
$a1$	10	-20	-20	-20	-20	5	-20	5	-20
$a2$	-20	-20	5	-20	10	-20	5	-20	-20
$a3$	-20	5	-20	5	-20	-20	-20	-20	10

Table 1: *A three-player coordination game*

3 Optimal adaptive learning (OAL) algorithm

Sometimes, agents may learn to play a game by coordinating their actions in a different game. For example, in robosoccer, adding artificial payoffs may help guide the robots' learning. Likewise, we construct a *virtual game* (VG) for each state s of a team Markov game to eliminate all the strict suboptimal Nash equilibria in that state. Let $VG^*(s,a)$ be the payoff that the agents

[1] Throughout the paper, every Nash equilibrium that we discuss is also a subgame perfect Nash equilibrium. (This refinement of Nash equilibrium was first introduced in [13] for different games).

receive from the VG in state s for a joint action a. We let $VG^*(s,a) = 1$ if $a = \arg\max_{a' \in A} Q^*(s,a')$ and $VG^*(s,a) = 0$ otherwise. For example, the VG for the game in Table 1 gives payoff 1 for each optimal Nash equilibrium ($\{a1, b1, c1\}$,$\{a2, b2, c2\}$, and $\{a3, b3, c3\}$), and payoff 0 to every other joint action.

Virtual games simplify the coordination task for the learning agents. For example for the game in Table 1, the VG is *weakly acyclic*.

Definition 2 (Weakly acyclic game [18]) *Let G be an n-player game in matrix form. The best-response graph of G takes each joint action $a \in A$ as a vertex and connects two vertices a and a' with a directed edge $a \rightarrow a'$ if and only if 1) $a \neq a'$; 2) there exists exactly one agent i such that a'_i is a best response to a_{-i} and $a'_{-i} = a_{-i}$. We say the game G is weakly acyclic if, from any initial vertex a, there exists a directed path to some vertex a^* from which there is no outgoing edge.*

To tackle the equilibrium selection problem for weakly acyclic games, Young [18] proposed a special fictitious play algorithm called *adaptive play* (AP), which works as follows. Let $a^t \in A$ be a joint action played at time t over an n-player game in matrix form. Fix integers k and m such that $1 \leq k \leq m$. When $t \leq m$, each agent i randomly chooses its actions. Starting from $t = m+1$, each agent looks back at the m most recent plays $h_t = (a^{t-m}, a^{t-m+1}, \ldots, a^{t-1})$ and randomly (without replacement) selects k samples from h_t. Let $K_{t+1}(a_{-i})$ be the number of times that a reduced joint action $a_{-i} \in A_{-i}$ (a joint action without agent i's individual action) appears in the k samples at $t+1$. Let $u_i(a)$ be agent i's payoff given that joint action a has been played. Agent i calculates its expected payoff w.r.t its individual action a_i as $EP(a_i) = \sum_{a_{-i} \in A_{-i}} u_i(\{a'_i\} \cup a_{-i}) \frac{K_{t+1}(a_{-i})}{k}$, and then randomly chooses an action from a set of best responses: $BR^t_i = \{a_i \mid a_i = \arg\max_{a'_i \in A_i} EP(a'_i)\}$.

Young showed that AP in a weakly acyclic game converges to a strict deterministic Nash equilibrium w.p.1. Thus, AP on the VG for the game in Table 1 leads to an equilibrium with payoff 1. In that example, any such equilibrium is an optimal Nash equilibrium for the original game.

Unfortunately, the above treatment cannot be trivially extended to all VGs because not all VGs are weakly acyclic. In fact, the VG in Table 2 has no strict Nash equilibria. As we show in our experiments, AP will not converge to any (deterministic) joint policy, and furthermore, it does not even learn to avoid the 0 payoff. In summary, the applicability of AP is limited.

	$b1$	$b2$
$a1$	1	1
$a2$	1	0

Table 2: *A virtual game with no strict Nash equilibria.*

In order to address more general settings, we now modify the notion of weakly acyclic game and adaptive play to accommodate weak optimal Nash equilibria.

Definition 3 (Weakly acyclic game w.r.t a biased set (WAGB))*: Let D be a biased set containing some of the Nash equilibria of a game G (and no other joint policies). Game G is a WAGB if, from any initial vertex a, there exists a directed path to either a Nash equilibrium inside D or a strict Nash equilibrium.*

We can convert any VG to a WAGB by setting the biased set D to include all joint policies that give payoff 1 (and no other joint policies). To solve such a game, we introduce a new learning algorithm for equilibrium selection. It enables each agent to deterministically select a best-response action once any Nash equilibrium in the biased set is attained (even if there exist several best responses when the Nash equilibrium is not strict). This is different from AP where players randomize their action selection when there are multiple best-response actions. We call our approach *biased adaptive play* (BAP).

BAP works as follows. Let D be the biased set composed of some Nash equilibria of a game in matrix form. Let SP_t be the set of k samples drawn at time t, without replacement, from among the most recent m joint actions. If

1. there exists a joint action $a' \in D$ such that for all $a \in SP_i$, $a_{-i} \subset a$ and $a_{-i} \subset a'$, and
2. there exists at least one joint action a such that $a \in SP_i$ and $a \in D$,

then agent i chooses its best-response action a_i such that $a_i \in a^{t'}$ and $t' = \max\{T \mid a^T \in SP_t \wedge a^T \in D\}$. That is, a_i is contained in the most recent play of a Nash equilibrium inside D. On the other hand, if the two conditions above are not met, then agent i chooses its best-response action in the same way as AP. As we will show, BAP (even with GLIE exploration) on a WAGB converges w.p.1 to either a Nash equilibrium in D or a strict Nash equilibrium.

So far we tackled learning of coordination in team Markov games where the game structure is known. Our real interests are in learning when the game is unknown. In multiagent reinforcement learning, Q^* is asymptotically approximated with Q_t. Let $VG_t(s,a)$ be the virtual game w.r.t $Q_t(s,a)$. Our question is how to construct VG_t so as to assure $VG_t \rightarrow VG^*$ w.p.1. Our method of achieving this makes use of the notion of ε-optimality (although our OAL algorithm, presented later, will find an actual optimum).

Definition 4 *Let ε be a positive constant. A joint action a is ε-optimal at state s and time t if $Q_t(s,a) + \varepsilon \geq \max_{a'} Q_t(s,a')$ for all $a' \in A$. We denote the set of ε-optimal joint actions at state s and time t as $A^{\varepsilon_t}(s)$.*

The idea is to use a decreasing ε-bound ε_t to estimate $A^{\varepsilon_t}(s)$ at state s and time t. All the joint actions belonging to the set are treated as optimal Nash equilibria in the virtual game VG_t which give agents payoff 1. If ε_t converges to zero at a rate slower than Q_t converges to Q^*, then $VG_t \to VG^*$ w.p.1. We make ε_t proportional to a function $B(N_t) \in [0,1]$ which decreases slowly and monotonically to zero with N_t, where N_t is the smallest number of times that any state-action pair has been sampled so far. Now, we are ready to present the entire *optimal adaptive learning* (OAL) algorithm. As we will present thereafter, we craft $B(N_t)$ carefully using an understanding of the convergence rate of a model-based RL algorithm that is used to learn the game structure.

Optimal adaptive learning algorithm (for agent i)

1. **Initialization** $t = 0$. For all $s \in S$ and $a \in A$ do $n_t(s,a) = 1$, $T_t(s,a,s') = \frac{1}{|s|}$ and $R_t(s,a) = 0$. $\varepsilon_t = C$; $A^{\varepsilon_t}(s) = A$. $D = A$.

2. **Learning of coordination policy**
If $t \leq m$, randomly select an action, otherwise do
   (a) Update the virtual game VG_t at state s: $VG_t(s,a) = 1$ if $a \in A^{\varepsilon_t}(s)$ and $VG_t(s,a) = 0$ otherwise. Set $D = \{a \mid VG_t(s,a) = 1\}$.
   (b) According to GLIE exploration, with an exploitation probability p do
      i. Randomly select (without replacement) k records from m recent observations of others' joint actions played at state s.
      ii. Calculate expected payoff of individual action a_i over the virtual game $VG_t(s,a)$ at current state s as follows: $EP(s,a_i) = \sum_{a_{-i} \in A_{-i}} \frac{K_t(s,a_{-i})}{k} VG_t(s, \{a_i\} \cup a_{-i})$. Construct the best-response set at state s and time t: $BR_i^t(s) = \{a_i | a_i = \arg\max_{a_i' \in A_i} EP(s,a_i')\}$.
      iii. If conditions 1) and 2) of BAP are met, choose a best-response action with respect to the biased set D. Otherwise, randomly select a best-response action from $BR_i^t(s)$.
   Otherwise, randomly select an action to explore.

3. **Off-policy learning of game structure** Q^*
   (a) Observe state transition $s \to s'$ and payoff r_t under the joint action a. Do
      i. $n_t(s,a) \leftarrow n_t(s,a) + 1$.
      ii. $R_t(s,a) \leftarrow R_t(s,a) + \frac{1}{n_t(s,a)}(r_t - R_t(s,a))$.
      iii. $T_t(s,a,s') \leftarrow$
      $T_t(s,a,s') + \frac{1}{n_t(s,a)}(1 - T_t(s,a,s'))$.
      iv. For all $s" \in S$ and $s" \neq s'$ do $T_t(s,a,s") \leftarrow (1 - \frac{1}{n_t(s,a)}) T_t(s,a,s")$.
   (b) $Q_{t+1}(s,a) \leftarrow$
   $R_t(s,a) + \gamma \sum_{s \in S} T_t(s,a,s') \max_{a' \in A} Q_t(s',a')$.
   (c) $t \leftarrow t+1$. $N_t \leftarrow \min_{s \in S, a \in A} n_t(s,a)$.
   (d) If $\varepsilon_t > CB(N_t)$
      i. $\varepsilon_t \leftarrow CB(N_t)$.
      ii. Update $Q_t(s,a)$ for all (s,a) using (b).
      iii. $A^{\varepsilon_t}(s) \leftarrow$
      $\{a | Q_t(s,a) + \varepsilon_t \geq \max_{a' \in A} Q_t(s,a')\}$.

Here, $n_t(s,a)$ is the number of times a joint action a has been played in state s (by time t). C is a positive constant (any value will work). $K_t(s,a_{-i})$ is the number of times that a joint action $a_{-i} \in A_{-i}$ appears in agent i's k samples (at time t) from the most recent m joint actions taken at state s.

4 Proof of convergence of OAL

In this section, we prove that OAL converges to an optimal Nash equilibrium. Throughout, we make the common RL assumptions: payoffs are bounded, and the number of states and actions is finite. The proof is organized as follows. In Section 4.1 we show that OAL learns optimal coordination if the game is known. Specifically, we show that BAP against a WAGB with known game structure converges to a Nash equilibrium under GLIE exploration. Then in Section 4.2 we show that OAL will learn the game structure. Specifically, any virtual game can be converted to a WAGB which will be learned surely. Finally, these two tracks merge in Section 4.3 which shows that OAL will learn the game structure and optimal coordination.

4.1 Learning to coordinate in a known game

In this section, we first model our biased adaptive play (BAP) algorithm as a stationary Markov chain. In the second half of this section we then model BAP with GLIE exploration as a nonstationary Markov chain.

4.1.1 BAP as a stationary Markov chain

Consider BAP with randomly selected initial m plays. We take the initial *history* $h_m = (a^1, a^2, \ldots, a^m)$ as the initial state of the Markov chain. The definition of the other states is inductive: A successor of state h is any state h' obtained by deleting the left-most element of h and appending a new right-most element. The only exception is that all the states $h = (a, a, ..., a)$ with a being either a member of the biased set D or a strict Nash equilibrium are grouped into a unique terminal state h_T. Any state directing to $h \in h_T$ is treated as directly connected to h_T.

Let P be the state transition matrix of the above Markov chain. Let h' be a successor of h, and let $a = \{a_1, a_2, \ldots, a_n\}$ (n players) be the new element that was appended to the right of h to get h'. Let $p_{h,h'} > 0$ be the transition probability from h to h'. Now, $p_{h,h'} > 0$ if and only if for each agent i, there exists a sample of size k in h to which a_i is i's best response according to the action-selection rule of BAP. Because agent i chooses such a sample with a probability independent of time t, the Markov chain is stationary. Finally, due to our clustering of multiple states into a terminal state h_T, for any state h connected to h_T, we have $p_{h,h_T} = \sum_{h' \in h_T} p_{h,h'}$.

In the above model, once the system reaches the terminal state, each agent's best response is to repeat its most recent action. This is straightforward if in the actual terminal state $h = (a, a, ..., a)$ (which is one of the states that were clustered to form the terminal state), a is a strict Nash equilibrium. On the other hand, if a is only a weak Nash equilibrium (in this case, $a \in D$), BAP biases each agent to choose its most recent action because conditions 1) and 2) of BAP are satisfied. Therefore, the terminal state h_T is an absorbing state of the finite Markov chain.[2]

Theorem 1 *Let G be a weakly acyclic game w.r.t. a biased set D. Let $L(a)$ be the length of the shortest directed path in the best-response graph of G from a joint action a to either an absorbing vertex or a vertex in D, and let $L_G = \max_a L(a)$. If $m \geq k(L_G+2)$, then, w.p.1, biased adaptive play in G converges to either a strict Nash equilibrium or a Nash equilibrium in D.*

Proof. The proof is similar to the proof of Theorem 1 in [18] except for the treatment of the terminal state h_T. Because the WAGB G has a biased set D, BAP (when biased to the same set D) in this game has exactly one absorbing state (h_T) in the Markov chain model. We show that this Markov chain has a unique stationary distribution in which h_T occurs w.p.1.

The basic idea is to show that there exists a state-independent and time-independent pair (M, p), where M is a positive integer and p is a positive probability, such that starting from any state h, the probability that the agents attain h_T within M periods is at least p. This implies that the probability of not reaching h_T in λM steps drops exponentially to zero with λ.

Starting from a random state $h_{t+1} = (a^{t-m+1}, \ldots, a^t)$ at time $t+1$ ($t \geq m$), there exists a positive and time-independent probability that all agents keep on sampling the same k plays $\xi_0 = (a^{t-k+1}, \ldots, a^t)$ in h_{t+1} for k consecutive times and choose the same best responses from their best-response sets. This results in a new state with identical last k elements: $\xi_1 = (a, \ldots, a)$. If $a \in D$, conditions (1) and (2) of BAP are satisfied. Hence, each agent sticks to its individual action contained in a w.r.t ξ_1. If a is a strict Nash equilibrium, the agents adhere to their unique best responses w.r.t ξ_1. Since there is a positive probability that all agents sample ξ_1 in the next $m-k$ periods, in the above two cases the learning process will converge to h_T with that probability. This reasoning is based on $m \geq 2k$ (agents' memory m should be large enough to contain both ξ_0 and ξ_1) and takes m periods.

If a is neither a strict Nash equilibrium nor a member of set D, after the first k periods, there is a positive and time-independent probability that in the next k periods, exactly one agent i samples ξ_1 and takes the same best response a_i to ξ_1 which makes $a' = a_{-i} \cup \{a_i\}$ the second vertex on the shortest directed best-response path starting from a and leading to either a strict Nash equilibrium or a Nash equilibrium in D, and other agents sample ξ_0 and take their individual best response contained in a_{-i}. This brings in a new consecutive k plays of a' which is denoted as ξ_2. If, again, a' belongs to set D or is a strict Nash equilibrium, the logic in the last paragraph applies and the terminal state h_T is attained with a positive probability. This needs $m \geq 3k$ and happens in $m+k$ periods. Otherwise, with the same reasoning, there is a positive probability that agents further follow the shortest directed best-response path starting from a (with length $L(a)$). At the nth step along the path, agents need to sample $\xi_0, \ldots, \xi_n$ so as to produce ξ_{n+1}. Thus, agents' memory m must be no smaller than $(n+2)k$ and if $L(a) = n+1$, agents totally take $M = m + nk$ steps to get into the absorbing state. With this induction, we have that agents have positive, time-independent probability to complete this best-response path in $m + L(a)k$ periods given the memory sizes $m \geq (L(a)+2)k$. $L_G = \max_a L(a)$ is the upper bound of $L(a)$. Denote the positive probability of starting from state h and getting to h_T within $M = m + L_G k$ as p_h. Thus, the time-independent and state-independent probability that agents reach h_T within M periods is at least $p = \min_h p_h > 0$. □

Theorem 1 says that the stationary Markov chain for BAP in a WAGB (given $m \geq k(L_G+2)$) has a unique stationary distribution in which only the terminal state has probability 1 to appear.

4.1.2 BAP with GLIE exploration as a nonstationary Markov chain

In the next section, we will discuss the problem of identifying the game. To accomplish that, the learners need to use exploration. In this section we show that such exploration does not hurt the other learning task, namely that of learning to coordinate. We show this by first modeling BAP with GLIE exploration as a *non-stationary Markov chain.*

With GLIE exploration, at every time step t, each *joint action* occurs with positive probability. This means that the system transitions from the state it is in to any of the successor states with positive probability. In GLIE exploration, the agents' action-selection be-

[2] Akin to how we modeled BAP as a stationary Markov chain above, Young modeled adaptive play (AP) as a stationary Markov chain [18]. There are two differences. First, unlike AP's, BAP's action selection is biased. Second, in Young's model, it is possible to have several absorbing states while in our model, at most one absorbing state exists (for any team game, our model has exactly one absorbing state). This is because we cluster all the absorbing states into one. This allows us to prove our main convergence theorem.

comes increasingly greedy over time. In the limit, with probability one, the transition probabilities converge to those of BAP with no exploration. Therefore, we can model the learning process with a sequence of transition matrices $\{P_t\}_{t=1}^{\infty}$ such that $\lim_{t\to\infty} P_t = P$, where P is the transition matrix of the stationary Markov chain describing BAP without exploration.

Our objective here is to show that on a WAGB, BAP with GLIE exploration will converge to the ("clustered") terminal state. To do that, we need the following lemma [4].

Lemma 2 [3] *Let P be the finite transition matrix of a stationary Markov chain with a unique stationary distribution g. Let $\{P_t\}_{t=1}^{\infty}$ be a sequence of finite transition matrices. Let f be a probability vector and denote $f^{(l,j)} = f \cdot P_l \cdots P_{l+j}$. If $\lim_{t\to\infty} P_t = P$, then $\lim_{j\to\infty} f^{(l,j)} = g$ for all $l > 0$.*

Using this lemma, we can prove the following theorem.

Theorem 3 (BAP with GLIE) *On a WAGB G, w.p.1, BAP with GLIE exploration (and $m \geq k(L_G + 2)$) converges to either a strict Nash equilibrium or a Nash equilibrium in D.*

Proof. Theorem 1 shows that BAP against a WAGB given $m \geq k(L_G + 2)$ converges to the unique terminal state almost surely. Thus, the stationary Markov chain has a unique limit distribution that only the terminal state h_T has probability 1 to appear. Guaranteed by the properties of the GLIE learning policy, the nonstationary Markov chain of BAP with GLIE has a transition matrix sequence converging to the transition matrix of the above stationary chain w.p.1. Thus, according to Lemma 2, starting from any state, BAP with GLIE against a WAGB G given $m \geq k(L_G + 2)$ converges almost surely to the unique stationary distribution of the stationary chain which is to stay at the terminal state forever. $\square$

4.2 Learning the virtual game

So far, we have shown that if the game structure is known in a WAGB, then BAP will converge to the terminal state. To prove optimal convergence of the OAL algorithm, we need to further demonstrate that 1) every virtual game is a WAGB, and 2) in OAL, the "temporary" virtual game VG_t will converge to the "correct" virtual game VG^* w.p.1. The first of these two issues is handled by the following lemma:

Lemma 4 *The virtual game VG of any n-player team state game is a weakly acyclic game w.r.t a biased set that contains all the optimal Nash equilibria, and no other joint actions. (By the definition of a virtual game, there are no strict Nash equilibria other than optimal ones.) The length of the shortest best-response path $L_{VG} \leq n$.*

Lemma 4 implies that BAP in a known virtual game with GLIE exploration will converge to an optimal Nash equilibrium. This is because (by Theorem 3) BAP in a WAGB will converge to either a Nash equilibrium in a biased set D or a strict Nash equilibrium, and (by Lemma 4) any virtual game is a WAGB with all such Nash equilibria being optimal.

Proof. Consider a team state game G with the optimal joint action set (also the biased set) D. Let $a^t = \{a_1, \ldots, a_n\}$ be any joint action outside D. Without loss of generality, we can *re-arrange* the individual actions inside a^t such that: 1) there exists an integer $I \in [1, n]$ such that for some $a' = \{a'_1, \ldots, a'_n\}$ in D, $a_i = a'_i$ when $i < I$; 2) no $a' \in D$ satisfying 1) also contains any a_i with $i \geq I$. According to the formation of virtual game, a'_I is agent I's one of the best responses to a^t. Thus, $a^{t+1} = a^t_{-I} \cup \{a'_I\}$ is one of the successors of a^t in the best-response graph of G. If $I = n$, $a^{t+1} \in D$. Otherwise, we have that a'_{I+1} is agent $I+1$'s one of the best response to a^{t+1} and thus $a^{t+2} = a^{t+1}_{-(I+1)} \cup \{a'_{I+1}\}$ is the next vertex after a^{t+1}. Following this line of reasoning, we construct a directed best-response path from a^t to a'. Since $1 \leq I \leq n$, The length of this path $L(a^t)$ is at most n. Thus, we get that the VG is a weakly acyclic game w.r.t set D with $L_{VG} \leq n$. $\square$

The following two lemmas are the last link of our proof chain. They show that OAL will cause agents to obtain the correct virtual game almost surely.

Lemma 5 *In any team Markov game, (part 3 of) OAL assures that as $t \to \infty$, $\max_{s\in S, a\in A}(|Q_t(s,a) - Q^*(s,a)|) \leq C\sqrt{\frac{\log\log N_t}{N_t}}$ for some constant $C > 0$ w.p.1.*

Proof. In the OAL algorithm, the model-based update rule is:

$$Q_{t+1}(s,a) = R_t(s,a) + \gamma \sum_{s'\in S} T_t(s,a,s') \max_{a'\in A} Q_t(s',a') \quad (1)$$

where, T_t and R_t are the estimates of the environmental model (transition function T and payoff function R) according to the agent's experience up to time t. Define $\hat{Q}_t$ function as: $\hat{Q}_{t+1}(s,a) = R_t(s,a) +$

[3] This lemma is easily obtained by using Theorem V.4.5, Theorem V.4.4 of [4] and their proofs. Theorem V.4.5 says that if the stationary chain is ergodic (i.e., with a unique stationary distribution) and P_t converges to P, the nonstationary Markov chain is "strongly ergodic" (that is, it has a unique stationary distribution) and the conditions of Theorem V4.4 are met. The proof of Theorem V4.4 shows that for all $l > 0$, $\prod_{t=l}^{\infty} P_t$ converges to a constant matrix with each row equal to g (it is the unique *probability left eigenvector* of P given the chain is ergodic.). This suffices to show that starting from any distribution, the nonstationary Markov chain will converge to the unique stationary distribution of the stationary chain.

$\gamma \sum_{s' \in S} T_t(s,a,s') \max_{a' \in A} Q^*(s',a')$. According to the *triangle inequality*, we have

$$|Q_{t+1}(s,a) - Q^*(s,a)| \le |Q_{t+1}(s,a) - \hat{Q}_{t+1}(s,a)| + |\hat{Q}_{t+1}(s,a) - Q^*(s,a)| \tag{2}$$

There are two components on the right-hand side of (2), let us study the second component first.

$$\begin{aligned}|\hat{Q}_{t+1}(s,a) - Q^*(s,a)| &\le |R_t(s,a) - R(s,a)| \\ &+\gamma \sum_{s' \in S} |T_t(s,a,s') - T(s,a,s')|| \max_{a' \in A} Q^*(s',a')| \\ &\le |R_t(s,a) - R(s,a)| \\ &+\gamma \max_{s' \in S} |T_t(s,a,s') - T(s,a,s')||S|M\end{aligned} \tag{3}$$

where, $|S|$ is the number of states and M is the upper bound of $|Q^*(s,a)|$.

According to the OAL algorithm, $R_t(s,a)$ and $T_t(s,a,s')$ are sample means of the random variables with the means $R(s,a)$ and $T(s,a,s')$. Thus, *the law of large number* says that they will converge to $R(s,a)$ and $T(s,a,s')$ w.p.1. In addition, the convergence rate of (3) will follow *the law of the iterated logarithm.* That is, it converges to 0 at a rate asymptotically no less than $O(\sqrt{\frac{\log\log N_t}{N_t}})$ w.p.1, where N_t is the minimal number of samplings on all state-action pairs of the MDP up to time t.

We can study the other component on the right-hand side of (2) as follows:

$$\begin{aligned}&|Q_{t+1}(s,a) - \hat{Q}_{t+1}(s,a)| \\ &\le \gamma \sum_{s' \in S} |T_t(s,a,s')|| \max_{a' \in A} Q_t(s',a') - \max_{a' \in A} Q^*(s',a')| \\ &\le \gamma \max_{s' \in S, a' \in A} |Q_t(s',a') - Q^*(s',a')|\end{aligned} \tag{4}$$

Denote *maximum norm* as $\|.\|$. Let $\Delta_t = \|\hat{Q}_t - Q^*\|$, $t_{j=1 \dots m-1}$ be the largest time before t_{j+1} when the state-action pair (s,a) with $\|Q_{t_{j+1}} - Q^*\|$ was updated with (1), $t_1 = 0$ and $t_m = t$. We can further decompose (4) by using the triangle inequality as:

$$\begin{aligned}&|Q_{t+1}(s,a) - \hat{Q}_{t+1}(s,a)| \le \gamma^m \|Q_0 - \hat{Q}_0\| \\ &+\gamma^m \Delta_{t_1} + \ldots + \gamma^{m-k+1} \Delta_{t_k} + \ldots + \gamma \Delta_{t_m}\end{aligned} \tag{5}$$

According to the OAL algorithm, all state-action pairs will be updated with (1) once the minimal number of samplings N_t increases. Therefore, $m \ge N_t$. That is, each state-action pair has been updated at least N_t times before (including) time t. Denote t_k as the moment when the minimal number of samplings attained $\lfloor N_t/2 \rfloor$ (clearly $k \ge \lfloor N_t/2 \rfloor$ and $m - k \ge \lfloor N_t/2 \rfloor$). We have

$$\begin{aligned}&|Q_{t+1}(s,a) - \hat{Q}_{t+1}(s,a)| \le \gamma^{N_t} \|Q_0 - \hat{Q}_0\| \\ &+\frac{\gamma^{\lfloor N_t/2 \rfloor + 1}}{1-\gamma} \max_{i=1 \dots k} \Delta_{t_i} + \frac{\gamma}{1-\gamma} \max_{j=k+1 \dots m} \Delta_{t_j}\end{aligned} \tag{6}$$

Consider the last component on the right-hand side of (6). Following a similar discussion as on (3), the convergence of Δ_{t_j} follows the *law of the iterated logarithm* with the number of samplings at least $\lfloor N_t/2 \rfloor$. Thus, its convergence rate will be asymptotically no less than $O(\sqrt{\frac{\log\log \lfloor N_t/2 \rfloor}{\lfloor N_t/2 \rfloor}})$ w.p.1. This converges slower than the other two components on the right-hand side of (6) when $N_t \to \infty$. Therefore, we get the results of the lemma. □

Using Lemma 5, the following lemma is easy to prove.

Lemma 6 *Consider any team Markov game. Let θ_t be the event that for all $t' > t$, $VG_{t'} = VG^*$ in the OAL algorithm in a given state. If 1) $B(N_t)$ decreases monotonically to zero ($\lim_{t\to\infty} B(N_t) = 0$), and 2) $\lim_{t\to\infty} \frac{\sqrt{\log\log N_t/N_t}}{B(N_t)} = 0$, then $\lim_{t\to\infty} Pr\{\theta_t\} = 1$.*

Proof. Let θ'_t be an event that $\max_a |Q_{t'}(s,a) - Q^*(s,a)| < \frac{C}{2} B(N_t)$ in state s for all $t' > t$, where $C > 0$ is the initial value of the ε-bound. With 2), we know that $\lim_{t\to\infty} \frac{C' \sqrt{\frac{\log\log N_t}{N_t}}}{CB(N_t)} = 0$ for any positive constant C'. Therefore, according to Lemma 5, for any $\varepsilon > 0$, there is a time $T(\varepsilon)$ such that if $t > T(\varepsilon)$, $Pr\{\theta'_t\} > 1 - \varepsilon$. We denote the set of all optimal joint actions in s by $A^*(s)$. Let $a \in A^*(s) = \{\hat{a} | \hat{a} = \arg\max_{a' \in A} Q^*(s,a')\}$ and $b = \arg\max_{(a' \in A) \wedge (a' \notin A^*(s))} Q^*(s,a')$. Let $\Delta = |Q^*(s,a) - Q^*(s,b)|$. With 1), we know $B(N_t)$ converging to 0. Therefore, there is a time T' such that $CB(N_t) < \Delta/2$ for all $t > T'$. Let $\hat{T} = \max(T(\varepsilon), T')$. Therefore, if $t > \hat{T}$, with a probability larger than $1 - \varepsilon$, we have: starting from t, for any joint actions $b \notin A^*(s)$, $Q_t(s,b) + CB(N_t) < Q^*(s,b) + \Delta/4 + \Delta/2 \le \max_a Q^*(s,a) - \Delta/4 < \max_a Q_t(s,a)$. And for any action $a \in A^*(s)$, $Q_t(s,a) + CB(N_t) > Q^*(s,a) + \frac{C}{2} B(N_t) > \max_a Q_t(s,a)$. Intuitively, this means that from t on, no suboptimal joint actions are in $A^{\varepsilon_t}(s)$ and no optimal joint actions are out of $A^{\varepsilon_t}(s)$. This implies $VG_t = VG^*$ for any time after t. Therefore, we have $Pr\{\theta_t\} > 1 - \varepsilon$ for all $t > \hat{T}$. □

Lemma 6 states that if the criterion for including a joint action among the ε-optimal joint actions in OAL is not made strict too quickly (quicker than the iterated logarithm), then agents will identify all optimal joint actions with probability one. In this case, they set up the correct virtual game. It is easy to make OAL satisfy this condition. For example, any function $B(N_t) = N_t^{\eta - \frac{1}{2}}$ satisfies the condition if $0 < \eta < \frac{1}{2}$.

4.3 Main convergence theorem: Learning the game structure and coordination simultaneously

Now we are ready to prove that OAL converges to an optimal Nash equilibrium in any team Markov game, even when the game structure is unknown. The idea is to show that the OAL agents learn the game structure (VGs) and the optimal coordination policy (over these VGs). OAL tackles these two learning problems simultaneously—specifically, it interleaves BAP (with GLIE exploration) with learning of game structure. However, the convergence proof does not make use of this fact. Instead, the proof proceeds by showing that the VGs are learned first, and coordination second (the learning algorithm does not even itself know when the switch occurs, but it does occur w.p.1).[4]

Theorem 7 (Optimal convergence) *In any team Markov game among n agents, if (1) $m \geq k(n+2)$, and (2) $B(N_t)$ satisfies Lemma 6, then the OAL algorithm converges to an optimal Nash equilibrium w.p.1.*

Proof. According to [1], a team Markov game can be decomposed into a sequence of state games. The optimal equilibria of these state games form the optimal policy π^* for the game. By the definition of GLIE exploration, each state in the finite state space will be visited infinitely often w.p.1. Thus, it is sufficient to only prove that the OAL algorithm will converge to the optimal policy over individual state games w.p.1.

Let θ_t be the event that $VG_{t'} = VG^*$ at that state for all $t' > t$. Let ε_1 be any positive constant. If Condition (2) of the theorem is satisfied, by Lemma 6 there exists a time $T(\varepsilon_1)$ such that $Pr\{\theta_t\} > 1 - \varepsilon_1$ if $t > T(\varepsilon_1)$.

If θ_t occurs and Condition (1) of the theorem is satisfied, by Theorem 3, OAL will converge to either a strict Nash equilibrium or a Nash equilibrium in the biased set w.p.1. Furthermore, by Lemma 4, we know that the biased set contains all of the optimal Nash equilibria (and nothing else), and there are no strict Nash equilibria outside the biased set. Therefore, if θ_t occurs, then OAL converges to an optimal Nash equilibrium w.p.1.[5] Let ε_2 be any positive constant, and let ξ_τ be the event that the agents play an optimal joint action at a given state for all $t' \geq \tau$. With this notation, we can reword the previous sentence: there exists a time $\Upsilon(\varepsilon_2, t)$ such that if $\tau > \Upsilon(\varepsilon_2, t)$, then $Pr\{\xi_\tau | \theta_t\} > 1 - \varepsilon_2$.

Put together, there exists a time $\Psi(\varepsilon_1, \varepsilon_2)$ such that if $\tau > \Psi(\varepsilon_1, \varepsilon_2)$, then $Pr\{\xi_\tau\} > Pr\{\xi_\tau | \theta_t\} Pr\{\theta_t\} > (1 - \varepsilon_1)(1 - \varepsilon_2) > 1 - \varepsilon_1 - \varepsilon_2$. Because ε_1 and ε_2 are only used in the proof (they are not parameters of the OAL algorithm), we can choose them to be arbitrarily small. Therefore, OAL converges to an optimal Nash equilibrium w.p.1. □

It is easy to set the parameters of OAL so that the convergence conditions are met. The parameters m and k should be set so that $m \geq k(n+2)$. Any positive value of C will guarantee convergence, but different values may lead to different convergence speed. $B(N_t)$ can be, for example, $N_t^{\eta - \frac{1}{2}}$, for any value η, $0 < \eta < \frac{1}{2}$. Another example that works is $B(N_t) = \sqrt{\frac{\log N_t}{N_t}}$.

5 Experimental results

We present experiments to demonstrate the performance of OAL in practice. We use *expected value Boltzmann exploration* as the exploration policy. That is, each agent i chooses its action a_i with probability $\frac{e^{EP(a_i)/Te}}{\sum_{a_i'} e^{EP(a_i')/Te}}$, where $EP(a_i)$ is the *expected payoff* as defined in the OAL algorithm (Section 3) and Te is a *temperature* parameter which decreases slowly with the number of rounds that the game is played. This is a GLIE exploration policy. For both OAL and JAL, the temperature parameter was initialized to $Te_0 = 100$ and was decreased over time using the following formula: $Te_t = .99\ Te_{t-1}$. For OAL, we used $C = 3.5$ and $B(N_t) = N_t^{-0.4}$.

5.1 OAL vs. JAL

The first experiment compared our OAL algorithm to Claus and Boutilier's joint action learner (JAL) algorithm [2]. The game is the 3-player game of Table 1, where the *players do not know the game structure.*

OAL converged to an optimal Nash equilibrium on each of the 1,000 runs. JAL converged to a suboptimal Nash equilibrium 424 times (and to an optimal Nash equilibrium in the rest).

To also compare the algorithms in the face of noise, we conducted an experiment where the payoff of each joint action was uniformly distributed between 0 and the value in the game matrix. Again, OAL converged to an optimal Nash equilibrium on each of the 1,000 runs. JAL converged to a suboptimal Nash equilibrium 437 times (and to an optimal Nash equilibrium in the rest).

5.2 Impact of m and k on convergence speed

The next experiment studied the impact of the OAL parameters m and k on convergence speed in the same game with the same noise model (Figure 1). Again, the agents did not know the game structure in advance.

With $m = 5$ and $k = 1$, the algorithm converges very slowly. This is because the agents' action selection is subject to noise (agents will deliberately explore suboptimal joint actions), and a small k gives agents high probability to sample such error-containing memory—and thus to make bad decisions. In the experiment, even if there exists only one exploration action in an

[4] One could try to identify the game first, and once it is learned, one could try to use some (other) learning technique to learn coordination. However, that approach suffers from not knowing when the game has been identified. Therefore, one would not know when one can switch to the second learning task.

[5] Theorem 3 requires $m \geq k(L_{VG} + 2)$. If Condition (1) of our main theorem is satisfied ($m \geq k(n + 2)$), then by Lemma 4, we do have $m \geq k(L_{VG} + 2)$.

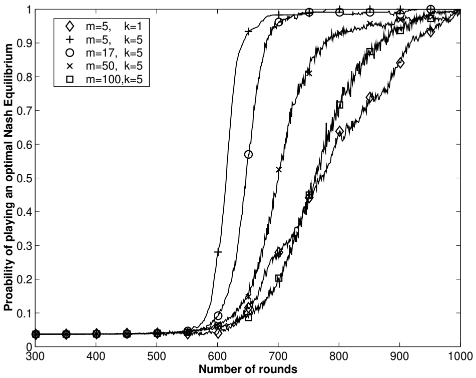

Figure 1: *Impact of m and k on convergence speed. The results are averaged over 300 runs.*

agent's memory, the agent has a 1/5 chance of choosing it and making a decision solely based on it.

For a fixed k, convergence becomes slower if m increases. With a large m, an agent keeps some out-of-date information which increases the agent's chance of making a decision based on such old memory. Therefore, a small m should be preferred. As can be seen in the figure, fastest convergence was achieved with $m = k = 5$. Interestingly, this parameter combination violates condition (2) of our main convergence theorem (Theorem 7). This suggests that the condition is not *necessary*—at least not in all VGs.

5.3 Necessity of *incomplete* sampling

However, the next analysis and experiment show that we cannot drop condition (2) from the convergence theorem in general. They highlight the importance of *incomplete* memory sampling, a key idea within OAL.

We first show analytically that without exploration, OAL with complete sampling does not converge in general—even if the game structure is known. In the

	$b1$	$b2$
$a1$	1	0
$a2$	0	1

Table 3: *A game where BAP with complete memory sampling does not converge.*

game in Table 3, if $m = k = 1$ (that is, BAP uses complete sampling) and agents start BAP with any suboptimal joint action such as $\{a1, b2\}$, they will never coordinate to an optimal joint action. In fact, we uncovered a series of 2-player team games that have this property: The players a and b have w actions each such that all the joint actions $\{ai, b(w-i+1)\}$ $(i = 1 \ldots w)$ are suboptimal (give payoff 0) while others are optimal (give payoff 1). Over such a game, players with memory $m = k = w - 1$ and an initial history consisting of $w - 1$ different suboptimal joint actions will be trapped into a loop composed of all suboptimal joint actions. We will call such games M_w.

Next, we show experimentally that incomplete sampling is necessary in BAP even with exploration. Figure 2 shows an experiment in game M_3 between BAP players that know the game structure. In that game, two players have 3 actions each ($ai = 1, 2, 3$ and $bi = 1, 2, 3$) and their joint actions $\{a1, b3\}$, $\{a2, b2\}$ and $\{a3, b1\}$ give payoff 0 while other joint actions give payoff 1. With $m = k = 2$ (that is, complete sampling of memory), we conducted 1000 trials. In 32 runs, the agents coordinate their actions to a loop composed of the three suboptimal joint actions. Figure 2 illustrates the frequency of these suboptimal joint actions (always averaged over the past 10 rounds).

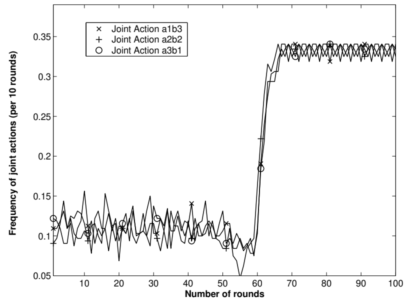

Figure 2: *Nonconvergence of OAL with complete sampling. The results are averaged over the 32 trials.*

When we satisfied condition (2) by setting $m = 5$ and $k = 2$, the agents found an optimal joint action in each of the 1000 trials. (They also converged on each of the 1,000 trials even if condition (2) was slightly violated: $k = 2$ and $5 > m > 2$. Therefore, further research is required to investigate whether condition (2) is necessary in general.)

5.4 Necessity of biased action selection

The last experiment demonstrates the necessity of biased action selection, another key idea behind OAL. The experiment is on the game shown in Table 2. Our experiments shows that without biased action-selection, AP with GLIE does not converge even in such a simple game (whose structure is known).

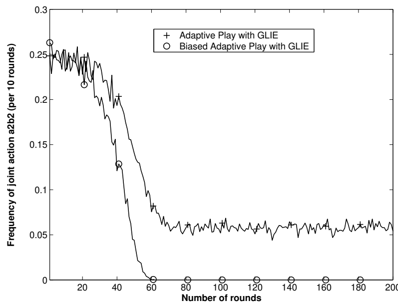

Figure 3: *Frequency of suboptimal joint actions. The results are averaged over 200 trials.*

6 Conclusions and future research

Multiagent RL involves two interrelated learning problems: identifying the game and learning to play. These two problems prevail even in team games where the agents' interests do not conflict. Even team games can have multiple Nash equilibria, only some of which are optimal. We presented *optimal adaptive learning* (OAL), the first algorithm that converges to an optimal Nash equilibrium for any team Markov game.

In OAL, agents shape the individual state games of a team Markov game into a sequence of *virtual games* and then coordinate their actions on these VGs using *biased adaptive play*. We proved that such play converges to an optimal Nash equilibrium for any team Markov game. Simultaneously, (with an appropriately decreasing ε-bound), the agents learn the structure of these virtual games almost surely. Therefore, the RL algorithm (OAL) that combines these two techniques is assured to converge to an optimal joint policy.

OAL uses a model-based approach to learn the game structure. We plan to consider model-free approaches such as Q-learning as well. Another research thrust is to extend the algorithm to general-sum Markov games.

References

[1] C.Boutilier. Planning, learning and coordination in multi-agent decision processes. In *TARK*, 1996.

[2] C.Claus and C.Boutilier. The dynamics of reinforcement learning in cooperative multi-agent systems. In *AAAI*, 1998.

[3] D.Fudenberg and D.K.Levine. *The theory of learning in games*. MIT Press, 1998.

[4] D.L.Isaacson and R.W.Madsen. *Markov chain: theory and applications*. John Wiley and Sons, Inc, 1976.

[5] G.Weiβ. Learning to coordinate actions in multi-agent systems. In *IJCAI*, 1993.

[6] J.Hu and W.P.Wellman. Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, 1998.

[7] L.Kaelbling, M.L.Littman, and A.W.Moore. Reinforcement learning: A survey. *JAIR*, 1996.

[8] M.Littman. Friend-or-Foe Q-learning in general sum game. In *ICML*, 2001.

[9] M.L.Littman. Value-function reinforcement learning in markov games. *J. of Cognitive System Research*, 2:55–66, 2000.

[10] M.L.Purterman. *Markov decision processes-discrete stochastic dynamic programming*. John Wiley, 1994.

[11] M.Tan. Multi-agent reinforcement learning: independent vs. cooperative agents. In *ICML*, 1993.

[12] R.A.Howard. *Dynamic programming and Markov processes*. MIT Press, 1960.

[13] R. Selten. Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft*, 12:301–324, 1965.

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

[15] S.Sen, M.Sekaran, and J. Hale. Learning to coordinate without sharing information. In *AAAI*, 1994.

[16] F. Thusijsman. *Optimality and equilibrium in stochastic games*. Centrum voor Wiskunde en Informatica, 1992.

[17] T.Sandholm and R.Crites. Learning in the iterated prisoner's dilemma. *Biosystems*, 37:147–166, 1995.

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