Emerging coordination in infinite team Markov games

Francisco S. Melo*
Institute for Systems and Robotics
Instituto Superior Técnico
1049-001 Lisboa, Portugal
fmelo@isr.ist.utl.pt

M. Isabel Ribeiro
Institute for Systems and Robotics
Instituto Superior Técnico
1049-001 Lisboa, Portugal
mir@isr.ist.utl.pt

ABSTRACT

In this paper we address the problem of coordination in multi-agent sequential decision problems with infinite state-spaces. We adopt a game theoretic formalism to describe the interaction of the multiple decision-makers and propose the novel *approximate biased adaptive play* algorithm. This algorithm is an extension of biased adaptive play to team Markov games defined over infinite state-spaces. We establish our method to coordinate with probability 1 in the optimal strategy and discuss how this methodology can be combined with approximate learning architectures. We conclude with two simple examples of application of our algorithm.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems, Coherence and coordination*

General Terms

Algorithms, Theory

Keywords

Team Markov games, coordination, biased adaptive play

1. INTRODUCTION

Research on cooperative multi-agent systems (MAS) typically focuses on three fundamental issues [4]: the *task* to accomplish, the *mechanism of cooperation* and the *performance* of the MAS. In this paper we adopt the model of *team Markov games* or *fully cooperative Markov games* to describe the interaction of multiple decision-makers that must cooperatively complete a pre-specified task. The use of this interaction model settles two of the fundamental issues referred above, by considering a reward structure that, simultaneously, *defines the task* and is used to *evaluate the performance* of the team.

The class of problems considered herein describe multi-agent sequential decision tasks in which all decision-makers must commit upon a common joint behavior. We assume that no explicit communication takes place. Instead, consensus in this common joint strategy must *emerge* from the mutual interaction among the different agents and with the environment.[1] Therefore, with respect to the third of the above issues, we feature cooperation as *coordination*: the multiple decision-makers must *coordinate* their individual decisions to yield an optimal joint behavior.

The paper contributes an extension of *biased adaptive play* [19] to Markov games with infinite state-spaces (henceforth referred as infinite Markov games). We identify the conditions under which our method, dubbed as *approximate biased adaptive play* (ABAP), is guaranteed to coordinate in all but a negligible part of the state-space. Like the widely known *fictitious play* process [3], ABAP relies on approximate statistics describing the strategies of each player to achieve coordination.

Coordination in multi-agent sequential decision making problems has been a widely covered topic of research [2, 5, 7, 9, 10, 19]. However, few works address the problem of coordination in *infinite domains*. In [8], coordination graphs are used to achieve coordination in infinite multi-agent problems. The coordination mechanism uses structured communication and a variable elimination procedure to achieve coordination. In [9], coordination graphs are also used to achieve coordination in continuous domains, this time with no communication assumed.

The ABAP method proposed in this paper differs from the previously referred methods in several aspects. First of all, ABAP assumes that no communication takes place. Furthermore, no assumption is made regarding previous knowledge or the coordination algorithm of the other decision-makers. In particular, we do not assume that all decision-makers follow the same decision-making or coordination algorithm. This is an important advantage of ABAP: in the presence of a heterogeneous group of decision-makers, ABAP is still able coordinate to the best decision-rule possible if, for some reason, the other decision-makers act sub-optimally.

The paper is organized as follows. We start by describing the original biased adaptive play (BAP) algorithm as proposed in [19]. We proceed by describing the framework of team Markov games used throughout the paper. We then present our main contribution: we describe the ABAP algo-

*From January 2008, Francisco S. Melo is with the School of Computer Science, Carnegie Mellon University.

Cite as: Emerging coordination in infinite team Markov games, F. Melo and I. Ribeiro, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, MüllerandParsons(eds.),May,12-16.,2008,Estoril,Portugal, pp.355-362

Copyright © 2008, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

[1]The consideration of no explicit communication can be supported by several arguments (bandwidth constraints, cost of communication, possible added complexity to the problem, etc.). We do not pursue such argument here and refer to several works that discuss these issues in greater detail [6, 18].

rithm and establish its convergence properties. Finally, we conclude the paper with a simple illustrative example and discuss some issues to be addressed in future work.

2. BIASED ADAPTIVE PLAY

We start by introducing some terminology and notation.

2.1 Strategic games

A N-player *strategic game* is 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 all *joint actions* and r_k is a function assigning a utility or payoff $r_k(a)$ to player k, when the joint action is $a \in \mathcal{A}$.[2] A *joint action* $a \in \mathcal{A}$ is a tuple $a = (a_1, \ldots, a_N)$ and we denote by a_{-k} a *reduced action*, obtained by removing the individual action a_k from a.

An individual *strategy* σ_k is a probability distribution over $\mathcal{A}_k$ and defines the probability of player k playing each action $a_k \in \mathcal{A}_k$ in the game. A strategy σ_k is a *pure strategy* if $\sigma_k(a_k) = 1$ for some action $a_k \in \mathcal{A}_k$ and *a mixed strategy* otherwise. A *joint strategy* is a vector $\sigma = (\sigma_1, \ldots, \sigma_N)$ of individual strategies and $\sigma(a)$ represents the probability of the joint action a being played when all agents follow the joint strategy σ. We refer to σ_{-k} as a *reduced joint strategy* or simply as *reduced strategy*, obtained from σ by removing the individual strategy of player k.

The individual strategy σ_k^* of player k is a *best response* to a reduced strategy σ_{-k} if player k cannot improve its expected reward using any other individual strategy σ_k, *i.e.*, if

$$\mathbb{E}_{(\sigma_k^*, \sigma_{-k})}\left[r_k(a)\right] \geq \mathbb{E}_{(\sigma_k, \sigma_{-k})}\left[r_k(a)\right]. \qquad (1)$$

A *Nash equilibrium* is a joint strategy $\sigma^* = (\sigma_1^*, \ldots, \sigma_N^*)$ in which each individual strategy σ_k^* is a best response to the reduced strategy σ_{-k}^*. Every finite strategic game has at least one Nash equilibrium [13]. A Nash equilibrium σ^* is *strict* if the inequality in (1) is strict for every $\sigma_k^* \in \sigma^*$.

A game in which $r_1(a) = \ldots = r_N(a)$ for all $a \in \mathcal{A}$ is *fully cooperative*. In this class of games there is always (at least) one *Pareto optimal* pure Nash equilibrium that yields maximum payoff for all players. In this paper, we consider only fully cooperative games.

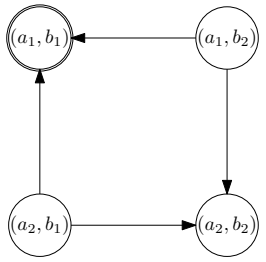
2.2 Biased adaptive play

When considering finite, fully cooperative games, *fictitious play* [3] is known to *converge in beliefs* to a Nash equilibrium [12]. If agents follow fictitious play, each agent maintains an estimate on the strategy of the other agents. Convergence in beliefs means that as $t \to \infty$, the estimates of all agents will converge to a Nash equilibrium. However, such guarantees do not extend in behavior, *i.e.*, it is not guaranteed that the policies of the fictitious play agents will converge to a Nash equilibrium [20]. And, if there are multiple such equilibria with different values, even if convergence is attained there are no guarantees that the limit equilibrium is the one with highest value.

Consider, for example, the 2-player, fully cooperative game in Fig. 1. The boldface entry represents the Pareto optimal equilibrium, (a_1, b_1). However, the action (a_2, b_2) is also a Nash equilibrium, and there are no guarantees that fictitious play will not converge to such equilibrium. The problem is even more evident if $r(a_2, b_2) = 10$. In this case, both equilibria are equally "desirable" and it may happen that one agent chooses the equilibrium (a_1, b_1) while the other chooses the equilibrium (a_2, b_2), which leads to the non-equilibrium joint action (a_1, b_2). This problem is known as an *equilibrium selection problem* in the game theory literature, or as a *coordination problem* in the multi-agent systems literature [1]. As seen in [19], BAP effectively settles the equilibrium selection problem and ensures that all players coordinate in a Pareto optimal Nash equilibrium.

	b_1	b_2
a_1	**10**	-20
a_2	-20	5

a) Payoff matrix



b) Best-response graph

Figure 1: Simple two-agent, two-action team strategic game. The Pareto optimal equilibrium is marked in bold in the payoff matrix and with a double line in the best-response graph.

We now briefly describe the BAP mechanism for repeated games, as introduced in [19].[3] Let $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ be a repeated 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 $V = \mathcal{A}$ and, given any two vertices $a, b \in V$, $(a, b) \in E$ if and only if $a \neq b$ and there is exactly one agent k for which b_k is a best response to a_{-k} and $a_{-k} = b_{-k}$. A best response graph is build by considering all joint actions in $\mathcal{A}$ as vertices and setting a directed edge from a joint action a to a joint action b if the two actions are composed of the same individual actions for all agents except one. For that single agent k, b_k is a best response to the reduced action a_{-k}. The best-reponse graph for the 2-agent, 2-action example above is depicted in Fig. 1.b).

We need one additional concept. Let $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ be a matrix game and $D \subset \mathcal{A}$ a set containing some of the Nash equilibria in Γ (and no other joint actions).

Definition 1. *A strategic game $\Gamma = \big(N, (\mathcal{A}_k), (r_k)\big)$ is* 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. It is* weakly acyclic with respect to (w.r.t.) the bias set D *if, given any vertex a in the best response graph of Γ, there is a directed path to either a Nash equilibrium in D or a strict Nash equilibrium.*

Now, considering a fully cooperative repeated game $\Gamma = \big(N, (\mathcal{A}_k), r\big)$, we construct an auxiliary *virtual game* $VG = \big(N, (\mathcal{A}_k), r_V\big)$, where $r_V(a) = 1$ if a is an optimal equilibrium for Γ and $r_V(a) = 0$ otherwise. By setting $D = \{a \in \mathcal{A} \mid r_V(a) = 1\}$, the game VG is weakly acyclic w.r.t. the set D. By construction, all Nash equilibria in VG correspond to Pareto optimal equilibria in Γ.

[2] We use the notation $\times_{k=1,\ldots,N} X_k$ to represent the cartesian product of N sets $X_k, k = 1, \ldots, N$.

[3] A repeated game is a Markov game with a single state. Notice that, unlike strategic games which are *one-shot games*, repeated games are played repeatedly.

Let K and m be two integers such that $1 \leq K \leq m$ and let $H(t)$ be a vector with the last m joint plays at the t^{th} play of the game. We refer to any set of K samples randomly drawn from $H(t)$ without replacement as a *K-sample* and denote it as $\mathcal{K}(H(t))$. A player k following BAP draws a K-sample $\mathcal{K}(H(t))$ from the history of the m most recent plays and checks if

1. There is a joint action $a^* \in D$ such that, for all the actions $a \in \mathcal{K}(H(t))$, $a_{-k} = a^*_{-k}$;
2. There is at least one action $a^* \in D \cap \mathcal{K}(H(t))$.

If these two conditions are verified, player k is "lead to believe" that the remaining players have coordinated in the action a^*_{-k} in D. Therefore, if conditions 1 and 2 are met, player k chooses its best response a^*_k from the joint action

$$a^* = \arg\max_{a \in H(t)} \{\tau \mid A(\tau) = a \mid A(\tau) \in K(H(t)) \text{ and } a \in D\},$$

where $A(t)$ denotes the joint action at time t. If either 1 or 2 (or both) does not hold, then player k uses the K-sample to estimate the strategies of the other players and chooses its action as a best response to this estimate. It has been shown that BAP ensures coordination with probability 1 (w.p.1) as $t \to \infty$ as long as $m \geq K(N+2)$ (see Theorems 1 and 3 and Lemma 4 in [19]).

3. MARKOV MODELS

In this section we introduce some important concepts regarding Markov chains, processes and games. These concepts will later be used in establishing our main result.

3.1 Markov chains and processes

A *time-homogeneous Markov chain* is a discrete-time stochastic process $\{X(t)\}$ defined by a pair $(\mathcal{X}, \mathsf{P})$, where $\mathcal{X}$ is the state-space and P is a transition probability kernel defining the transition probabilities

$$\mathsf{P}(x, U) = \mathbb{P}\left[X(t) \in U \mid X(t-1) = x\right],$$

which are independent of the particular time instant t considered. Given an arbitrary measurable set $U \subset \mathcal{X}$, the *first return time to* U, τ_U, is defined as

$$\tau_U = \min_{t \in \mathcal{T}} \{X(t) \in U, \quad t \geq 1\}.$$

A Markov chain is *ψ-irreducible* if, for any $x \in \mathcal{X}$,

$$\mathbb{P}\left[\tau_U < \infty \mid X(0) = x\right] > 0 \tag{2}$$

for any measurable set $U \subset \mathcal{X}$ such that $\psi(U) > 0$ and ψ is maximal in the sense that if ν is some other measure verifying (2), then $\nu \ll \psi$.

If η_U is the number of visits to a measurable set $U \subset \mathcal{X}$ in an infinite trajectory of the chain, the set U is said to be *Harris recurrent* if, for any $x \in \mathcal{X}$,

$$\mathbb{P}\left[\eta_U = \infty \mid X(0) = x\right] = 1.$$

A ψ-irreducible Markov chain is *Harris recurrent* if all measurable sets $U \subset \mathcal{X}$ such that $\psi(U) > 0$ are Harris recurrent.

Let now $\{X(t)\}$ be a $\mathcal{X}$-valued *controlled* Markov chain. The transition probabilities for the chain are now given by the action-dependent kernel

$$\mathsf{P}^a(x, U) = \mathbb{P}\left[X(t+1) \in U \mid X(t) = x, A(t) = a\right],$$

for any measurable set $U \subset \mathcal{X}$. The $\mathcal{A}$-valued process $\{A(t)\}$ represents the control process: $A(t)$ is the control action at time instant t and $\mathcal{A}$ is the finite set of possible actions. A decision-maker must determine the control process $\{A(t)\}$ so as to maximize the functional

$$V(\{A(t)\}, x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(X(t), A(t)) \mid X(0) = x\right], \tag{3}$$

where $0 \leq \gamma < 1$ is a discount-factor and $R(x, a)$ represents a random "reward" received for taking action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$. We assume that there is a bounded, deterministic function $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \longrightarrow \mathbb{R}$ assigning a reward $r(x, a, y)$ every time a transition from x to y occurs after taking the joint action a and such that

$$\mathbb{E}\left[R(x, a)\right] = \int_{\mathcal{X}} r(x, a, y) \mathsf{P}^a(x, dy).$$

The tuple $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$ thus defined is a *Markov decision process* (MDP).

Given an MDP $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, the *optimal value function* V^* is defined for each state $x \in \mathcal{X}$ as

$$V^*(x) = \max_{\{A(t)\}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^t R(X(t), A(t)) \mid X(0) = x\right] \tag{4}$$

and verifies

$$V^*(x) = \max_{a \in \mathcal{A}} \int_{\mathcal{X}} \left[r(x, a, y) + \gamma V^*(y)\right] \mathsf{P}^a(x, dy), \tag{5}$$

which is a form of the Bellman optimality equation. The optimal Q-values $Q^*(x, a)$ are defined for each state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ as

$$Q^*(x, a) = \int_{\mathcal{X}} \left[r(x, a, y) + \gamma V^*(y)\right] \mathsf{P}^a(x, dy). \tag{6}$$

If $V^*(x)$ "measures" the total discounted reward obtained during an expectedly optimal trajectory starting at state x, $Q^*(x, a)$ measures the total discounted reward obtained during an expectedly optimal trajectory starting at state x when the first action is a.

3.2 Team Markov games

Markov games [14] can be interpreted as generalizations of MDPs to multiple decision-makers. Therefore, a Markov game is a tuple $\left(N, \mathcal{X}, (\mathcal{A}_k), \mathsf{P}, (r_k), \gamma\right)$, where N is the number of players, $\mathcal{X}$ is the state-space, $\mathcal{A} = \times_{k=1}^{N} \mathcal{A}_k$ is the set of joint actions, P is the controlled transition kernel and r_k is the reward function for player k.

In this paper, we are interested in *fully cooperative* Markov games, also known as *team Markov games*.[4] In team Markov games all players share the same goal, which is to maximize the total expected reward over all joint control sequences $\{A(t)\}$. This total expected reward is defined as in (3), where now $R(x, a)$ is the random reward received by *all* players for taking the joint action a in state x. It is immediate to define the *optimal value function* V^* for a team Markov game as in (4), where now $A(t)$ stands for the joint action at time t. This optimal value function also verifies (5) and we can define the optimal Q-function, Q^*, as in (6).

It is also straightforward to extend the concepts of individual strategy, joint strategy and reduced strategy from

[4]In the literature, team Markov games are sometimes referred as multi-agent MDPs (MMDPs) [2].

strategic games to team Markov games. For example, an individual strategy for player k is a *state and time-dependent* probability distribution $\sigma_k(t)$ over the set $\mathcal{A}_k$. The corresponding control sequence $\{A_k(t)\}$ should verify

$$\mathbb{P}\left[A_k(t) = a_k \mid X(t) = x\right] = \sigma_k(t; x, a).$$

We write $V^{\sigma(t)}(x)$ instead of $V(\{A(t)\}, x)$ whenever the control sequence $\{A(t)\}$ is generated by the joint strategy $\sigma(t)$, and refer to $V^{\sigma(t)}$ as the *value function* associated with strategy $\sigma(t)$. A joint or individual strategy that does not depend on t is said to be *stationary*.

Two important remarks are now in order. First of all, if the definition of V^* and the existence of an optimal joint control strategy follow immediately from the corresponding results for MDPs, the fact that the decision process in team Markov games is distributed implies that coordination must be addressed explicitly [2].

On the other hand, we note that the function Q^* defines at each state $x \in \mathcal{X}$ a strategic game $\Gamma_x = \left(N, (\mathcal{A}_k), Q^*(x, \cdot)\right)$ that we refer as a *stage game*. If the players coordinate in an optimal Nash equilibrium in each stage game Γ_x, they coordinate in a Pareto optimal Nash equilibrium for the team Markov game [1]. Each stage-game is *fully cooperative* and *weakly acyclic* and we can thus apply BAP to each such game. As seen in Section 2, BAP converges to a Pareto optimal Nash equilibrium as $t \to \infty$. Therefore, all players will coordinate in a Pareto optimal Nash equilibrium for the team Markov game as long *as every state $x \in \mathcal{X}$ is visited infinitely often*. In [19] this idea is used to ensure coordination in team Markov games with finite state-space. However, if $\mathcal{X}$ is not finite, this is generally not possible, as we discuss in the continuation.

4. APPROXIMATE BAP

As seen in the previous section, if BAP is to be applied to a team Markov game, coordination at each stage game requires that the corresponding state be visited a sufficient number of times. Recall that BAP uses incomplete samples from the history of past plays to estimate the average strategies of the players in the game, providing a method to choose upon a best response to such strategy, as long as the game is known. The successive visits to each state provide each player with a sample of the other players' strategies in the particular state considered. Hence the need of "infinite" visits to every state in order to ensure convergence [19].

Formally, the condition of infinite visits amounts to requiring the underlying Markov chain to be irreducible (every state is "visitable") and recurrent (each "visitable" state is visited infinitely often). In the infinite state-space case, these conditions translate in ψ-irreducibility (all but a negligible part of the state-space is "visitable") and Harris recurrence (every "visitable" region is visited infinitely often). We discuss these requirements further ahead.

In adapting BAP to cope with infinite state-spaces, coordination at each state should rely not only in past visits to that particular state but should also use the information provided by plays in *several nearby states*. The intuition behind this idea can be easily clarified. Each agent k can no longer use the past history at a particular state x to infer the other agents' strategy in that state, since there is the possibility that it was never visited before. Instead, agent k will assume that *the policies of the other agents do not change significatively in the states sufficiently close to x*, this clearly depending on the continuity of Q^* in x. If this assumption holds, agent k can use the past history at nearby states to estimate the strategy of the other agents *at state x*. To implement this idea, we rely on the distance between two states x and y in $\mathcal{X}$ as an indication on the "closeness" of the states. As will soon become apparent, the use of such approximation mechanism suitably adapts BAP to team Markov games with infinite state-spaces while ensuring coordination in all but a negligible part of the state-space.

Let $\Gamma = \left(N, \mathcal{X}, (\mathcal{A}_k), \mathsf{P}, r, \gamma\right)$ be a team Markov game with compact state-space $\mathcal{X} \subset \mathbb{R}^p$ and finite joint action-space $\mathcal{A}$. Let Q^* be the optimal Q-function for Γ and define, for each $x \in \mathcal{X}$, the team matrix game $\Gamma_x^* = \left(N, (\mathcal{A}_k), Q^*(x, \cdot)\right)$. To introduce and analyze ABAP, we resort to an auxiliary process $\{Y(t)\}$ evolving in $\mathcal{X}$. We assume this process $\{Y(t)\}$ to be a ψ-irreducible and Harris recurrent Markov chain, with a irreducibility measure ψ that is absolutely continuous w.r.t. the Lebesgue measure in $\mathbb{R}^p$.

At each time instant t, N agents engage in the repeated game $\Gamma_{Y(t)}^*$ where $Y(t)$ is the state of the auxiliary process $\{Y(t)\}$ at time t. The sole purpose of the agents is to coordinate in a Pareto optimal equilibrium strategy in each state-game Γ_x^*; the agents have no knowledge otherwise on the Markov game Γ or on the auxiliary process $\{Y(t)\}$, and consider the payoffs $Q^*(x, \cdot)$ at different state-games Γ_x^* to be independent. This technical artifice allows us to discard the effect of the joint actions of the agents on the state evolution of the Markov game. The agents merely visit the states in $\mathcal{X}$ along the trajectories of $\{Y(t)\}$ and coordinate in each visited stage-game Γ_x^*.[5]

Consider the past history up to time t,

$$\mathcal{H}(t) = \{y(0), a(0), y(1), a(1), \ldots, y(t-1), a(t-1)\},$$

where the sequence $\{y(t)\}$ is a sample trajectory of the process $\{Y(t)\}$ and each joint action $a(\tau)$ corresponds to that chosen by the agents in game $\Gamma_{y(\tau)}^*$. At each time instant t, each agent determines the distance between the current state $Y(t)$ and each state $y(\tau)$ occurring in $\mathcal{H}(t)$, given by $\|Y(t) - y(\tau)\|$ for some norm $\|\cdot\|$ in $\mathbb{R}^p$. It then chooses m occurrences from this history so as to minimize the corresponding distance. The sample set thus obtained, denoted as $S_m(Y(t), \mathcal{H}(t))$, contains the m elements in $\mathcal{H}(t)$ closer to $Y(t)$, *i.e.*, those minimizing

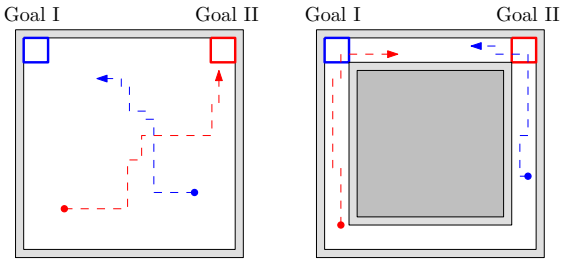
$$\sum_{i=1}^{m} \|Y(t) - y(t_i)\|. \tag{7}$$

We remark that a particular state $x \in \mathcal{X}$ may occur in $S_m(Y(t), \mathcal{H}(t))$ more than once. On the other hand, if two occurrences $y(t_i)$ and $y(t_j)$ verify

$$\|Y(t) - y(t_i)\| = \|Y(t) - y(t_j)\|$$

and only one such occurrence must be chosen, then the most recent one should be picked (*e.g.*, if $t_j > t_i$ above, then $y(t_j)$ would be chosen). We also notice that, due to the ψ-irreducibility and Harris recurrence of the Markov chain, given any state $x \in \mathcal{X}$ and a corresponding neighborhood U with positive ψ-measure, there is a time $T_0(x, U)$ such that, w.p.1, $S_m(x, \mathcal{H}(t)) \subset U$ for $t > T_0$. Roughly speaking

[5]We note that, as discussed in Section 6, this assumption has little impact on the validity of our result and merely aims at simplifying the exposition and the proof.



a) Example 1: Single-room b) Example 2: Corridor

Figure 2: Example of two continuous indoor environments.

this means that, as $t \to \infty$, the elements in $S_m(Y(t), \mathcal{H}(t))$ will all lie in a neighborhood of $Y(t)$ on which the optimal policies are "similar". Therefore, once the set $S_m(Y(t), \mathcal{H}(t))$ is determined, ABAP proceeds as standard BAP by drawing a K-sample from the m plays in $S_m(Y(t), \mathcal{H}(t))$.

The following result establishes the convergence of ABAP.

THEOREM 2. *Let $\{Y(t)\}$ be a Markov chain evolving on $\mathcal{X}$ as described above. In particular, assume that the chain is ψ-irreducible and Harris recurrent, with irreducibility measure ψ absolutely continuous w.r.t. the Lebesgue measure. Suppose that N agents following ABAP engage at each time step in the coordination games described above. Suppose that Q^* is continuous in $\mathcal{X}$ in all but a ψ-null set of states. Then all agents coordinate in a Pareto optimal equilibrium strategy w.p.1 in ψ-almost every state in $\mathcal{X}$, as long as the conditions for convergence of standard BAP are met.*

PROOF. See the Appendix. □

5. ILLUSTRATIVE EXAMPLES

We now analyze two applications of ABAP in simple multi-robot navigation tasks. Consider the two indoor environments depicted in Fig. 2.

Two mobile robots (I and II) must navigate to the corresponding goal regions, signaled with the bold, colored lines. Both environments are 1×1 squares, and the state of each robot at each time instant is a pair $(\mathbf{x}, \mathbf{y})$ of coordinates.[6] The coordinates of the corners in the goal regions are $(1, 1)$ and $(0, 1)$, respectively, and the corresponding goal regions are 0.1×0.1 squares, as depicted in Fig. 2. We denote the goal region for robot k by G_k and by G the cartesian product of G_{I} and G_{II}. In their trajectories, the robots learn must not to crash into each other by avoiding to lie in the same 0.1×0.1 area simultaneously (see Fig. 3 for an illustration). We denote the state of robot k at time t by $X_k(t)$, $k = \mathrm{I}, \mathrm{II}$. The state of the robot group is a pair $X(t) = (X_{\mathrm{I}}(t), X_{\mathrm{II}}(t))$ and can take any value in $([0; 1] \times [0; 1]) \times ([0; 1] \times [0; 1])$.

Each robot has 4 actions available, namely N, S, E and W. Each individual action moves the robot 0.3 in the corresponding direction (with some zero-mean Gaussian noise) within the limits of the depicted walls. We consider the movements of the robots to be independent of each other.

[6]We use boldface symbols $\mathbf{x}$ and $\mathbf{y}$ to denote the physical coordinates of one robot to distinguish these from the symbols x and y used to denote generic elements of the state-space $\mathcal{X}$.

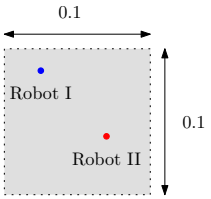


Figure 3: Situation of possible crash.

Both navigation problems can be modeled by team Markov games $\left(N, \mathcal{X}, (\mathcal{A}_k), \mathsf{P}, r, \gamma\right)$ where

- $N = 2$;
- $\mathcal{X} = ([0; 1] \times [0; 1]) \times ([0; 1] \times [0; 1])$;
- $\mathcal{A}_k = \{N, S, E, W\}$ for $k = \mathrm{I}, \mathrm{II}$;
- For each problem, the transition probabilities are defined by a kernel P given by

$$\mathsf{P}^a(x, U) = \mathsf{P}_{\mathrm{I}}^{a_{\mathrm{I}}}(x_{\mathrm{I}}, U_{\mathrm{I}})\mathsf{P}_{\mathrm{II}}^{a_{\mathrm{II}}}(x_{\mathrm{II}}, U_{\mathrm{II}})$$

  where the kernels P_{I} and P_{II} define the single-robot transition probabilities according to the description above and $U = U_{\mathrm{I}} \times U_{\mathrm{II}}$;
- The reward function r is defined as

$$r(x, a, y) = \begin{cases} 20 & \text{if } y \in G; \\ -10 & \text{if } \|y_{\mathrm{I}} - y_{\mathrm{II}}\|_\infty < 0.1; \\ 0 & \text{otherwise;} \end{cases}$$

- We consider $\gamma = 0.95$.

To test ABAP, we first computed the optimal Q-function for both problems, using a random exploration strategy for 10^5 time steps. To this purpose, we used Q-learning with soft-state aggregation [15], where we considered a partition of the joint state-space into 81 non-uniform aggregated states. This preliminary step is necessary as ABAP coordinates given the optimal Q-function. We then allowed the agents to "learn to coordinate" using ABAP for 10^3 time steps. Notice that, because of the finite learning time, we need to store a finite history of length 10^3.[7]

The total reward obtained during learning is depicted in Fig. 4. It is worth remarking that the slope of the learning curves, corresponding to the total reward obtained during learning, provide a rough indication of the performance of the robots. For purposes of comparison, we also present the total reward obtained with an uncoordinated group of robots interacting for the same period of time. It is clear that, after an initial period where the ABAP robots "evaluate" the strategies of the other robots, the group apparently converges to a coordinated joint strategy (the slope of the curve becomes positive). We also note that this actually happens after not so many iterations. It is nevertheless important to remark that when the slope of the learning curve becomes positive the robots need not have coordinated in *every state*, but only on those around the followed joint trajectory to the goal. However, the difference between the

[7]Since these problems have a total of only 16 joint actions, storing the complete history of 1000 plays requires less than 2 Kb of memory.

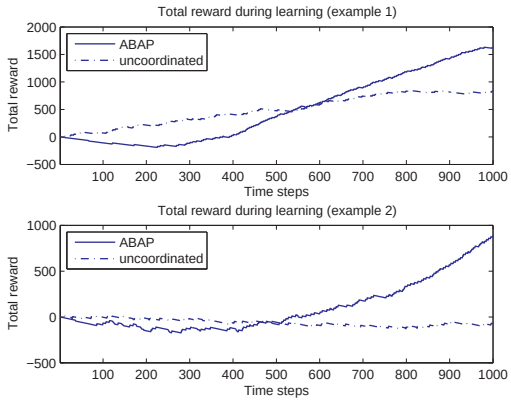


Figure 4: Cumulative reward during the 10^3-time-units learning period for both problems. Example 1 refers to single-room scenario and Example 2 refers to the corridor scenario.

Table 1: Comparative results of ABAP vs. no coordination in both environments. The reported results were obtained after the learning period was complete. We present the average total discounted reward obtained over $2,000$ Monte-Carlo runs. Once again, Example 1 refers to single-room scenario and Example 2 refers to the corridor scenario.

	Method	Total Disc. Reward
Example 1	Uncoordinated	6.882
	ABAP	44.5076
Example 2	Uncoordinated	2.016
	ABAP	55.286

coordinated and the uncoordinated groups is evident: even if initially the ABAP group performs worse than the uncoordinated group (since all robots initially "experiment" to better explore the space of possible joint strategies), after nearly 500 time-steps the performance of the former has already surpassed that of the latter.

To further understand the difference in performance between the coordinated and the uncoordinated robots, we tested the learnt strategies in the corresponding environments. We ran the learnt policies for each environment during 50 time units and determined the total discounted reward obtained in each case. Table 1 represents the final results obtained and Fig. 5 depicts the corresponding temporal evolution. We ran $2,000$ independent Monte-Carlo trials and present the average total discounted reward obtained in both scenarios. For the purpose of comparison, we also present the results obtained with no coordination mechanism.

Note that the values presented in Table 1 correspond to value attained by each group at the end of the Monte-Carlo trials as depicted in Fig. 5. Also, the curves in Fig. 5 are much smoother than those in Fig. 4 because of the Monte-Carlo averaging.

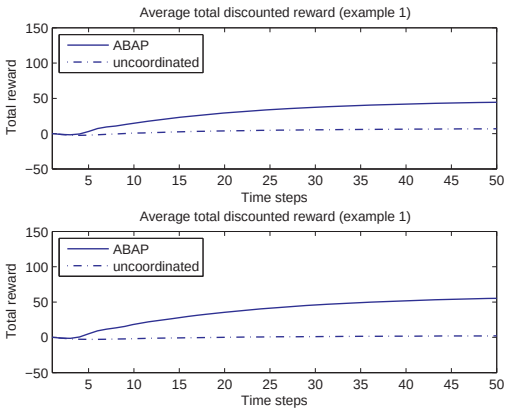


Figure 5: Temporal evolution of the total discounted reward obtained using the learnt joint strategy along the 50-time-units trials, averaged over the $2,000$ Monte Carlo trials. As before, Example 1 refers to single-room scenario and Example 2 refers to the corridor scenario.

It is worth observing that the slope of the uncoordinated team in Fig. 5 is similar to the one observed in Fig. 4. This is expected since there is no adaptation of the joint strategy on the uncoordinated team. On the other hand, the ABAP team exhibits an exponential-like curve, this clearly due to the effect of the discount factor. It is also worth mentioning that, unlike the uncoordinated team, the ABAP team performs better in the corridor scenario than on the single-room scenario. The observed difference is due to two main factors: the size of the environment (that influences the performance of both teams) and the how critical coordination is (that greatly influences the performance of the uncoordinated team in Example 2 and thus leads to the observed difference).

6. DISCUSSION

We now discuss several important issues referred along the text and postponed to these concluding remarks.

"On-strategy" coordination: When describing the ABAP algorithm, considered an auxiliary process $\{Y(t)\}$ that allowed to separated the control of the dynamics of the game and the problem of coordination. The coordination mechanism thus obtained was "off-strategy", in that the actions of the players did not affect the dynamics of the underlying Markov chain. As argued before, the purpose of this mathematical device was to aleviate our analysis from concerns on the underlying behavior of the Markov chain.

In a team Markov game, the coordination mechanism will always be "on-strategy": the actions of the players *will* influence the evolution of the underlying chain. Nevertheless, the requirements of ψ-irreducibility and Harris recurrence in Theorem 2 must still hold to ensure that the conclusions of the latter theorem to hold. Therefore, the strategies used by the players prior to coordination should be crafted so as to ensure these properties.

We remark, however, that the requirements of most approximate learning methods in terms of the underlying Markov chain are usually much stronger than ψ-irreducibility

or Harris recurrence. Moreover, the use of *GLIE strategies* (greedy in the limit with infinite exploration) easily settles this need, as it ensures sufficient exploration of the state-space to guarantee both ψ-irreducibility and Harris recurrence of the underlying chain. A GLIE strategy converges to the optimal (coordinated) strategy as $t \to \infty$ and guarantees that all "significant" parts of the state-space are visited infinitely often.[8] Theorem 3 of [19] describes how GLIE policies can be combined with standard BAP. The extension of this result to ABAP is immediate.

Storage of infinite histories: The ABAP algorithm, as formulated, stores the complete history $\mathcal{H}(t)$ of the process. Since we are considering the algorithm to eventually run along an infinite trajectory, storing the complete history would be infeasible. However, in any practical implementation such requirement can easily be alleviated without any loss in performance. This was particularly evident from the results portrayed in the previous section.

In fact, at each state $X(t)$ ABAP uses K samples drawn from the m points closest to $X(t)$. This means that, in practice, implementation of ABAP can rely on a fixed-size history, chosen sufficiently large to properly sample the state-space in a representative way. The exact length of the history to be chosen will depend on the *irreducibility measure* associated with the sampled chain and with the *support* of the optimal Q-function for the game. For example, if ABAP is combined with approximate learning algorithms (see ahead), the history to be maintained can be crafted from the function approximation architecture.

In any case, there are numerous applications in which the agents are only allowed to "learn" for a finite period of time, as seen in the examples in Section 5. After this finite learning period, the agents all stick the learnt strategy, and no further storing of past plays is necessary.

Absolute continuity of ψ with respect to μ^{Leb}: Another requirement in Theorem 2 is related with the absolute continuity of the irreducibility measure ψ with respect to the Lebesgue measure μ^{Leb}.

This requirement can also be alleviated, although requiring a more evolved proof. The central idea is as follows: if ψ is not absolutely continuous w.r.t. μ^{Leb}, there must be at least one probability atom $\boldsymbol{\alpha} \subset \mathcal{X}$. Each such atom is visited infinitely often (due to the Harris recurrence of the chain) and the argument proceeds by reducing the coordination problem to each such atom. A complete, formal proof of this result will be provided in a longer version of the paper.

Combination of ABAP with approximate learning algorithms: In all developments considered in this paper we implicitly admitted that the optimal Q-function was known. In this situation, coordinating in the Markov game amounts to coordinating in each of the corresponding stage games. However, since ABAP addresses coordination in infinite Markov games, it must often rely on some approximation of Q^*, since only in very particular problems can this function be represented exactly in a computer. Therefore, it may prove useful to construct the sets $S_m(x, \mathcal{H}(t))$ taking into account the approximation considered.

[8]An example of a GLIE strategy is Boltzmann exploration with decreasing temperature factor. More details on GLIE strategies can be found in [16].

For example, suppose that Q^* is represented as a linear combination of a set of basis functions $\phi_1, \ldots, \phi_M$. Then $S_m(X(t), \mathcal{H}(t))$ could be chosen to minimize

$$\sum_i \|\phi(X(t)) - \phi(x(t_i))\|$$

instead of (7).

Another important aspect is concerned with *simultaneous learning and coordination*. In many situations of interest, decision-makers must learn or approximate the function Q^* and coordinate at the same time. When dealing with infinite state-spaces, approximate learning methods must be used, such as interpolation based Q-learning [17] or Q-learning with soft state-aggregation [15]. Convergence of these methods usually requires the underlying Markov chain be *geometrically ergodic*, which is a stronger condition than ψ-irreducibility or Harris recurrence, as the latter two are implied by the former [11].

Extending such convergence results to team Markov games will require similar conditions, which are compatible with the conditions of Theorem 2. Therefore, combination of ABAP with any of the mentioned learning algorithms is far from complicated and can be attained by mimicking the procedure in [19] with due modifications.

Acknowledgments

This work was partially supported by the Portuguese Fundação para a Ciência e a Tecnologia under the Carnegie Mellon-Portugal Program and the Information and Communications Technologies Institute (ICTI) (www.icti.cmu.edu) and also under Programa Operacional Sociedade do Conhecimento (POS_C) that includes FEDER funds. The views and conclusions contained in this document are those of the authors only.

7. REFERENCES

[1] C. Boutilier. Planning, learning and coordination in multiagent decision processes. In *Proc. 6th Conf. Theoretical Aspects of Rationality and Knowledge*, pp. 195–210, 1996.

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

[3] G. Brown. Some notes on computation of games solutions. Research Memoranda RM-125-PR, RAND Corporation, Santa Monica, CA, 1949.

[4] Y. Cao, A. Fukunaga, and A. Kahng. Cooperative mobile robotics: Antecedents and directions. *Autonomous Robots*, 4(1):1–23, 1997.

[5] G. Chalkiadakis and C. Boutilier. Coordination in multiagent reinforcement learning: A Bayesian approach. In *Proc. 2nd Int. Joint Conf. Autonomous Agents and Multiagent Systems*, pp. 709–716, 2003.

[6] E. Durfee, V. Lesser, and D. Corkill. Coherent cooperation among communicating problem solvers. *IEEE Trans. Computers*, 36(11):1275–1291, 1987.

[7] F. Fischer, M. Rovatsos, and G. Weiss. Hierarchical reinforcement learning in communication-mediated multiagent coordination. In *Proc. 3rd Int. Joint Conf. Autonomous Agents and Multiagent Systems*, pp. 1334–1335, 2004.

[8] C. Guestrin, M. Lagoudakis, and R. Parr. Coordinated reinforcement learning. In *Proc. 19th Int. Conf. Machine Learning*, pp. 227–234, 2002.
[9] J. Kok, M. Spaan, and N. Vlassis. An approach to noncommunicative multiagent coordination in continuous domains. In *Proc. 12th Belgian-Dutch Conf. Machine Learning*, pp. 46–52, 2002.
[10] M. Lauer and M. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multiagent systems. In *Proc. 17th Int. Conf. Machine Learning*, pp. 535–542, 2000.
[11] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.
[12] D. Monderer and L. Shapley. Fictitious play property for games with identical interests. *Journal of Economic Theory*, 68:258–265, 1996.
[13] J. Nash. Equilibrium points in n-person games. *Proc. National Academy of Sciences*, 36:48–49, 1950.
[14] L. Shapley. Stochastig games. *Proc. National Academy of Sciences*, 39:1095–1100, 1953.
[15] S. Singh, T. Jaakkola, and M. Jordan. Reinforcement learning with soft state aggregation. In *Adv. Neural Information Proc. Systems 7*, pp. 361–368, 1994.
[16] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvari. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–310, 2000.
[17] C. Szepesvári and W. Smart. Interpolation-based Q-learning. In *Proc. 21st Int. Conf. Machine Learning*, pp. 100–107, 2004.
[18] J. Tsitsiklis and M. Athans. On the complexity of decentralized decision making and detection problems. *IEEE Trans. Automatic Control*, 30(5):440–446, 1985.
[19] X. Wang and T. Sandholm. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In *Adv. Neural Information Proc. Systems 15*, pp. 1571–1578, 2003.
[20] H. Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993.

APPENDIX

A. PROOF OF THEOREM 2

The proof of this theorem will require several intermediate results before being properly established.

We first assume Q^* to be continuous. This means that the function $V^a(x) = Q^*(x, a)$ is continuous for each $a \in \mathcal{A}$. Take an arbitrary point $x \in \mathcal{X}$ and an arbitrary action $a_0 \in \mathcal{A}$. Then, one of two statements below holds:

1. $Q^*(x, a_0) < \max_{a\in\mathcal{A}} Q^*(x, a)$. If this is the case, due to the continuity of Q^* in x, the inequality above holds for some neighborhood U of x. In other words, there is a neighborhood U of x such that
$$Q^*(y, a_0) < \max_{a\in\mathcal{A}} Q^*(y, a), \forall y \in U.$$
This has an interesting implication: for every point $x \in \mathcal{X}$ there is a neighborhood U such that
$$\mathbf{opt}(y) \subset \mathbf{opt}(x), \tag{8}$$
for all $y \in U$, where $\mathbf{opt}(x)$ is the set of optimal joint actions at state x.
2. $Q^*(x, a_0) = \max_{a\in\mathcal{A}} Q^*(x, a)$. If this is the case, two possible situations can occur:
   (a) There is a neighborhood U of x such that $a_0 \in \mathbf{opt}(y)$ for all $y \in U$;
   (b) Given any neighborhood U of x there is a point $y \in U$ such that $a_0 \notin \mathbf{opt}(y)$;

Denote by $D(a_0)$ the set of points $x \in \mathcal{X}$ verifying 2b and define the sets $D = \bigcup_{a\in\mathcal{A}} D(a)$ and $C = \mathcal{X} - D$. We now show that

LEMMA 3. *Given the sets C and D above, $D = \partial C$.*

PROOF. We prove the lemma by establishing that $\partial C \subset D$ and that $D \subset \partial C$.

From 1 and 2a, we see that a point $x \in C$ has a neighborhood U such that $U \cap D = \emptyset$. Then $C = \mathbf{int}(C)$ and since $D = \mathcal{X} - C$, $\partial C \subset D$. Since C and D are complementary and $C = \mathbf{int}(C)$, the conclusion of the lemma follows. □

Since $D = \partial C$, it is immediate that D is closed and therefore measurable. In turn, C must be open (in the subspace topology) and also measurable. We now proceed with the following result.

LEMMA 4. *The set D defined above verifies $\mu^{\text{Leb}}(D) = 0$.*

PROOF. Recall that the function Q^* is continuous in x. Therefore, the function $V^*(x) = \max_{a\in\mathcal{A}} Q^*(x, a)$ is also continuous. We define a new function $G^a(x) = V^a(x) - V^*(x)$. Clearly, G^a is continuous and $G^a(x) \leq 0$ for all $x \in \mathcal{X}$. We will show the set
$$\Omega_{G^a} = \{x \in \mathcal{X} \mid G^a(x) < 0\}$$
to be a p-dimensional topological manifold. Clearly, such set is a subset of $\mathbb{R}^p$ and, hence, Hausdorff and second countable (in the subspace topology). On the other hand, any point $x \in \Omega_{G^a}$ has a neighborhood $U \subset \Omega_{G^a}$, due to the continuity of G^a. This neighborhood is a neighborhood in $\mathbb{R}^p$ and therefore Ω_{G^a} is locally Euclidean and a topological manifold of dimension p. Its boundary is a manifold of dimension $p-1$ and its Lebesgue measure is therefore zero.

By construction, we have that
$$\partial C \subset \bigcup_{a\in\mathcal{A}} \partial\Omega_{G^a},$$
and the conclusion follows. □

We remark that, for each point $x \in C$, there is a neighborhood U such that $\mathbf{opt}(x) = \mathbf{opt}(y)$, for all $y \in U$. This can be seen by noticing that a point in C either verifies Condition 1 or Condition 2a for every action $a \in \mathcal{A}$. Therefore, given one such point x and corresponding neighborhood U, it is immediate that the virtual game obtained by setting to 1 all optimal actions and to 0 all non-optimal actions is the same in every point in U. This implies that, if $\psi(U) > 0$, there is a time T_0 such that, w.p.1, $S_m(x, H_t) \subset U$ for $t > T_0$ and ABAP reduces to BAP around x. Since, for all $t > T_0$ all K-samples are drawn from $S_m(x, H_t)$, convergence of standard BAP ensures that, for all points in C, ABAP coordinates in an optimal Nash equilibrium w.p.1. Therefore, since ψ is absolutely continuous w.r.t. μ^{Leb}, Lemma 4 suffices to conclude that convergence to an optimal policy in all but a ψ-null set of points. Now if Q^* is continuous in all but a -null set of points, the previous proof holds for every point x in which Q^* is continuous (with some care when defining the p-dimensional manifolds Ω_{G^a}), and the proof is complete. □