Reducing the Complexity of Multiagent Reinforcement Learning

Andriy Burkov
DAMAS Laboratory
Laval University
G1K 7P4, Quebec, Canada
burkov@damas.ift.ulaval.ca

Brahim Chaib-draa
DAMAS Laboratory
Laval University
G1K 7P4, Quebec, Canada
chaib@damas.ift.ulaval.ca

ABSTRACT

It is known that the complexity of the reinforcement learning algorithms, such as Q-learning, may be exponential in the number of environment's states. It was shown, however, that the learning complexity for the goal-directed problems may be substantially reduced by initializing the Q-values with a "good" approximative function. In the multiagent case, there exists such a good approximation for a big class of problems, namely, for goal-directed stochastic games. These games, for example, can reflect coordination and common interest problems of cooperative robotics. The approximative function for these games is nothing but the relaxed, single-agent, problem solution, which can easily be found by each agent individually. In this article, we show that (1) an optimal single-agent solution is a "good" approximation for the goal-directed stochastic games with action-penalty representation and (b) the complexity is reduced when the learning is initialized with this approximative function, as compared to the uninformed case.

General Terms

Algorithms, Performance, Experimentation

Keywords

Multiagent learning, Stochastic games, Q-learning, Initialization

1. INTRODUCTION

One of the major challenges of learning in multiagent systems (MAS) is the state space dimensionality. This is due to the fact that the cardinality of the state space of MAS grows exponentially with the number of agents, because in such environment each state is composed of the individual "positions" of all agents acting in that environment.

In this paper, we address the problem of multiagent learning complexity reduction in a specific context, namely, in goal-directed stochastic games with action-penalty representation. In such context, all agents have their respective goals and the rewards of making an action are negative in all states except the goal state. Our main contribution consists in using the results of single-agent planning in a relaxed, single-agent, problem as a heuristic function to initialize the agents' multiagent Q-values in all unknown states of the initial multiagent problem. The idea is to focus the learning process on a relatively small relevant region of the entire state space and, by so doing, to reduce the calculation time required to learn a multiagent solution.

In the next section, we give a short description of our framework and, more precisely, the assumptions we made about the structure of the environment and the agents' initial knowledge.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
AAMAS'07 May 14–18 2007, Honolulu, Hawai'i, USA.
Copyright 2007 IFAAMAS .

2. NOTATION AND CONCEPTS

Stochastic games (SGs) combine MDPs and matrix games to formalize temporally extended agent interaction. An SG is a tuple $(n, \mathbf{S}, \mathcal{A}^{1\ldots n}, T, R^{1\ldots n})$, where n is the number of agents, $\mathbf{S}$ is the set of states $\mathbf{s} \in \mathbf{S}$ now represented as vectors, $\mathcal{A}^j$ is the set of actions $a^j \in \mathcal{A}^j$ available to agent j, $\mathbf{A}$ is the joint action space $\mathcal{A}^1 \times \ldots \times \mathcal{A}^n$, T is the transition function: $\mathbf{S} \times \mathbf{A} \times \mathbf{S} \mapsto [0, 1]$, R^j is the reward function for agent j: $\mathbf{S} \times \mathbf{A} \mapsto \mathbb{R}$ and $\mathbf{s}_0 \in \mathbf{S}$ is the initial state.

The goal of each agent in an SG is to maximize its expected utility in each state of the game. In the stochastic game framework, the "expected utility" is a combination of two expectations in the sense that the agents in an SG aim to maximize their expected utilities over other players' joint strategy in each matrix game (state), and their temporal utility over all future matrix games. Formally, for an agent j, the discounted utility U^j of a state $\mathbf{s}$ is defined as follows:

$$U^j(\Pi(\mathbf{s})) = u^j(\Pi(\mathbf{s})) + \gamma \sum_{\mathbf{s}' \in \mathbf{S}} T(\mathbf{s}, \Pi(\mathbf{s}), \mathbf{s}') U^j(\Pi(\mathbf{s}')) \quad (1)$$

where u^j is the "immediate" expected utility of a matrix game $\mathbf{s}_t$ for the agent j, Π is the policy of joint strategies of players, which defines a strategy profile $\Pi(\mathbf{s})$ for each state $\mathbf{s} \in \mathbf{S}$. A solution in SGs is called "equilibrium". Equilibrium is a joint strategy, where no agent is interested to unilaterally change its own strategy.

The algorithm we use as a basis for our approach is called Adaptive Play Q-learning (APQ) [1]. This algorithm is based on Q-learning combined with the Adaptive Play [3] to calculate for each player a policy of best response to the other players' strategies.

978-81-904262-7-5 (RPS) ©2007 IFAAMAS

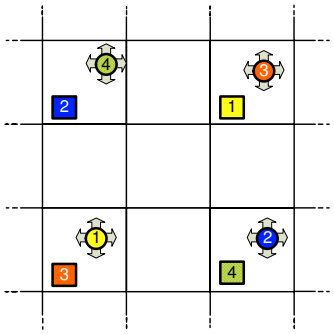

Figure 1: A fragment of the four-robot grid world environment containing the start and goal positions of agents.

2.1 Adaptive Play Q-learning

Formally, each player j playing the Adaptive Play saves in memory a history $H_t^j = \{\mathbf{a}_{t-p}^{-j}, \ldots, \mathbf{a}_t^{-j}\}$ of the last p joint actions played by the other players. To select a strategy to play at time $t+1$ each player randomly and irrevocably samples from H_t^j a set of examples of length l, $\hat{H}_t^j = \{\mathbf{a}_{k_1}^{-j}, \ldots, \mathbf{a}_{k_l}^{-j}\}$, and calculates the empiric distribution $\hat{\Pi}^{-j}$ as an approximation of the real reduced profile of strategies played by the other players, using the following rule: $\hat{\Pi}_{\mathbf{a}^{-j}}^{-j} = \frac{C(\mathbf{a}^{-j}, \hat{H}_t^j)}{l}$, where $C(\mathbf{a}^{-j}, \hat{H}_t^j)$ is the number of times that the joint action $\mathbf{a}^{-j}$ was played by the other players according to the set $\hat{H}_t^j$. Given the probability distribution over the other players' actions, $\hat{\Pi}^{-j}$, the player j plays its best response, $BR^j(\hat{\Pi}^{-j})$, to this distribution with some exploration. If there are several equivalent best responses, the player j randomly chooses one of them. Young [3] proved the convergence of the Adaptive Play to an equilibrium when played in self-play for a big class of games such as the coordination and common interest games. APQ is an extension of Adaptive Play to the SG context. In APQ, the following modified Q-learning update rule is used:

$$\begin{aligned}\hat{Q}^j(\mathbf{s},\mathbf{a}) \quad \leftarrow \quad & (1-\alpha)\hat{Q}^j(\mathbf{s},\mathbf{a}) + \alpha[R^j(\mathbf{s},\mathbf{a}) \\ & + \gamma \max_{a^j \in \pi^j(\mathbf{s}')} U^j(\hat{\Pi}(\mathbf{s}') \cup \{\pi^j(\mathbf{s}')\})]\end{aligned}$$

In the above update rule, j is an agent, $\mathbf{a}$ is a joint action played by the agents in state $\mathbf{s} \in \mathbf{S}$, $\hat{Q}^j(\mathbf{s},\mathbf{a})$ is the current value for player j of playing the joint action $\mathbf{a}$ in state $\mathbf{s}$, $R^j(\mathbf{s},\mathbf{a})$ is the immediate reward the player j receives if the joint action $\mathbf{a}$ is played in the state $\mathbf{s}$ and $\pi^j(\mathbf{s}')$ are all possible *pure* strategies that are available for player j.

2.2 Q-values Initialization

In our approach, we made several assumptions about the model of the environment. First, SGs, where agents are intended to act, are assumed to be goal directed with action-penalty representation. The second assumption is that the multiagent environment applies additional restrictions on the reward and transition functions of the underlying MDP. I.e., the multiagent penalties for all state-action pairs may be only higher than the corresponding single-agent values and the multiagent transitions in the direction of optimal single-agent actions may be only more uncertain. More formally, we assume that,

$$R^j(\mathbf{s},\mathbf{a}) \leq R^j(s^j,a^j)\ \forall \mathbf{s} = \langle s^j, \mathbf{s}^{-j}\rangle,\ \mathbf{a} = \langle a^j, \mathbf{a}^{-j}\rangle$$

where $\mathbf{s}$ is a multiagent state, $\mathbf{a}$ is a joint action, s^j and a^j correspond to j's position in $\mathbf{s}$ and action in $\mathbf{a}$, $R^j(\mathbf{s},\mathbf{a})$ is the reward of j when $\mathbf{a}$ is played in $\mathbf{s}$ and $R^j(s^j,a^j)$ is the corresponding single-agent reward.

In turn, given the same rewards in multiagent and single-agent cases, the multiagent transition function relates to the single-agent one by affecting the utilities as follows:

$$U^j(\Pi(\mathbf{s})) \leq U^j(\overset{\star}{\pi}{}^j(s^j)) \qquad \forall \Pi, \forall \mathbf{s}$$

where $\mathbf{s} = \langle s^j, \mathbf{s}^{-j}\rangle$, $U^j(\Pi(\mathbf{s}))$ is defined by equation (1) and $U^j(\overset{\star}{\pi}{}^j(s^j))$ is the utility of the corresponding single-agent state of j in $\mathbf{s}$.

As is easy to see, in that case, a single-agent problem (MDP) is an appropriate *relaxation* of the multiagent problem (SG). We use an optimal solution of this relaxation to initialize the Q-values of APQ. MDP may be solved with a variety of techniques (value iteration, reinforcement learning, heuristic search, etc). In our approach, we suppose that all agents are able to calculate an optimal single-agent policy before starting the learning in multiagent context.

In order to ensure the tractability of the Q-learning algorithm, the Q-values of all state-action pairs must be initialized with some monotonic and admissible function [2]. The monotonicity property of Q-values corresponds to the consistence of the heuristic function in the heuristic search terminology and means that the triangle inequality holds. In turn, admissibility means that for all state-action pairs $-\hat{Q}^j(\mathbf{s},\mathbf{a})$ never overestimates $-\overset{\star}{Q}{}^j(\mathbf{s},\mathbf{a})$.

According to our approach, multiagent Q-values are initialized using precalculated single-agent state utilities and transition function as follows:

$$\hat{Q}^j(\mathbf{s}, \langle a^j, \mathbf{a}^{-j}\rangle) \leftarrow \overset{\star}{Q}{}^j(s^j,a^j)\ \ \forall \mathbf{a}^{-j} \qquad (2)$$

where $\mathbf{s}$ is a multiagent state, s^j is the j's component of the vector $\mathbf{s}$ (in other words, s^j is the agent j's state in the corresponding single-agent world) and $\overset{\star}{Q}{}^j(s^j,a^j)$ is an optimal single-agent Q-value that is calculated from the single-agent solution and the model as follows:

$$\overset{\star}{Q}{}^j(s^j,a^j) = R(s^j,a^j) + \gamma \sum_{s'^j} T^j_{s,a,s'} U(\overset{\star}{\pi}{}^j(s'^j))$$

where $T^j_{s,a,s'}$ denotes $T(s^j,a^j,s'^j)$, the single-agent transition function, and $U(\overset{\star}{\pi}{}^j(s^j))$ is the utility of the single-agent state s^j according to the optimal policy $\overset{\star}{\pi}{}^j$.

THEOREM 1. *If in a goal directed stochastic game with action-penalty representation, Q-values $\hat{Q}(\mathbf{s},\mathbf{a})$ are initialized using the utilities of the corresponding single-agent state-action pairs according to equation (2), then these Q-values are admissible and monotonic.*

The proof of the above Theorem results from the following two Lemmas.

LEMMA 1. *If in a goal directed stochastic game with action-penalty representation, Q-values of agent j, $\hat{Q}^j(\mathbf{s},\mathbf{a})$, are initialized according to equation (2), then these Q-values are monotonic.*

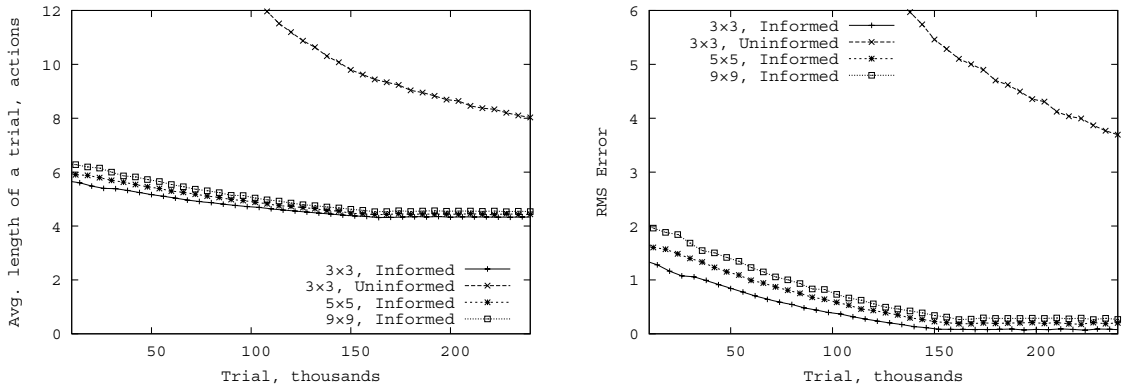

Figure 2: The dynamics of the learning process.

LEMMA 2. *If in a goal directed stochastic game with action-penalty representation, Q-values of agent j, $\hat{Q}^j(\mathbf{s}, \mathbf{a})$, are initialized according to equation (2), then these Q-values are admissible.*

Thus, the initialization we proposed is admissible and monotonic, therefore the learning complexity of the initialized algorithm is expected to be reduced as compared to the zero-initialized case. In the next section, we provide the results of the experiments produced on several examples of the four-robot grid world problem, which justify this hypothesis.

3. EXPERIMENTS

A four-agent grid world may be depicted as presented in Figure 1. For each robot, j, the set of actions is $\mathcal{A}^j = \{N, S, W, E\}$. These actions have stochastic effect. If an action is successful, robot changes its position on the grid to the intended cell, otherwise its position remains unchanged. Each action has a negative reward associated with it. In our example, we set the reward of -0.04 for any action in any cell except the goal cells, where the rewards for all actions are 0. In the case of robot collision, no transition is made and all affected robots obtain the reward of -0.1. Thus, robots are interested in attainment of their respective goals by making a minimal number of actions and avoiding collisions. Hence, this is a goal directed coordination stochastic game with action-penalty representation and, therefore, in self-play, APQ initialized with monotonic and admissible Q-values is expected converge to an equilibrium in self-play.

We tested our algorithm in a zero-initialized (called "uninformed") case and in a case (called "informed"), when Q-values were initialized using single-agent solution, calculated assuming that the agent is alone in the environment (no collisions are possible). The grids we considered were 3×3, 5×5 and 9×9 cells. The dynamics of the learning process is presented in Figure 2. The left diagram reflects the average number of actions made by the agents before they reached their goals with respect to the number of learning trials. The right diagram represents the evolution of the root-mean-square (RMS) error. In the uninformed case, the results are presented for the 3×3 grid only, since the uninformed algorithm became intractable on our machine starting from the 5×5 cells. This was due the fact that the uninformed agents explored almost entire state space, while the informed algorithm visited a very small relevant subset of the state space. As one can see, while the heuristically initialized Q-learning converged to an optimal solution after merely $250,000$ trials in all grids, the uninformed algorithm was still far from the convergence even in a very small grid. Notice that the convergence of the informed learning depends weakly on the dimensions of the grid, while in the uninformed case the size of the grid is critical for the tractability of the algorithm. This means that the heuristic function permits "focusing" the learning on a small subset of the total state space, which is the main reason for such considerable learning complexity reduction.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we showed that the initialization of multiagent Q-values using a precalculated single-agent solution of the relaxed problem permits reducing the complexity of the learning process. We also showed that such initialization is admissible and monotonic for the problems that can be modeled as a goal-directed stochastic game with action-penalty representation. By producing a set of empirical tests on a multiagent coordination problem, we showed that the uninformed multiagent learning algorithm quickly becomes intractable, while the informed, heuristically initialized, algorithm remains tractable with the growth of the state space, while being weakly sensible to that growth due to the strict focusing on a small subset of relevant states. In our future work, we intend to extend the applicability of our approach to the general form stochastic games.

5. REFERENCES

[1] O. Gies and B. Chaib-draa. Apprentissage de la coordination multiagent : une méthode basée sur le Q-learning par jeu adaptatif. *Revue d'Intelligence Artificielle*, 20(2-3):385–412, 2006.

[2] S. Koenig and R. G. Simmons. The effect of representation and knowledge on goal-directed exploration with reinforcement-learning algorithms. *Machine Learning*, 22:227–250, 1996.

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