A Game Theory Approach to Coordination in MAS

Matteo Golfarelli[1]

Abstract. In this paper, we study some aspects of the coordination problem faced by a population of *self-interest* interacting agents, from the perspective of 2-players repeated games. We propose an optimality criterion based on the *maximax* concept and we show the existence of an equilibrium point with respect to this criterion. Future work is aimed at using reinforcement learning techniques to evolve the population toward this point.

1 Introduction

The design of systems of multiple autonomous agents that interact in various ways (pursuing their own ends or compatible goals) has attracted a great deal of attention in AI [2]. Of special interest, due to their practical applicability to real situations, are systems in which *self-interested* agents interact within a common environment in order to reduce the cost of their own goal [4]. For example, consider some autonomous mobile robots, each with an assigned set of tasks to carry out. The self-interested nature of the robots will make them willing to coordinate their actions only if the resulting behavior will be useful to all of the participating agents. Many efforts have been made in DAI to study this kind of interaction. Of particular interest to us is the well-known Contract Net Protocol [3].

Within CNP coordination takes place by means of iterated negotiation cycles. CNP considers two kinds of players: the *announcer* and the *bidder*. At each cycle the announcer proposes a task to be exchanged; the bidder receiving the announcement formulates a bid consisting of a task that could be usefully exchanged. If the swap is individual rational for the announcer too (i.e. it reduces its execution cost) an award message will be sent to the bidder and the two tasks will be swapped.

In this paper, we study aspects of this coordination problem from the perspective of 2-player repeated games. Two agents repeatedly encounter a random situation, drawn from a finite set of possible states. Each situation requires coordinated action and is viewed as a single stage decision problem. This paper represents the first result of an ongoing project aimed at bridging the gap between non-zero sum games and multiagent systems.

2 A Game Theoretic Approach

From a game theory point of view the CNP framework can be modeled as follows: both the announcer and the bidder have a set of states, namely S_A, and S_B and a set of actions T_A, T_B. Agents repeatedly play *Two-Person Nonzero Sum Games* in which they are randomly placed in one of their states. The announcer in state $s_A \in S_A$ selects an action $a \in T_A$ and the bidder, seeing the announcer action a, selects its own action $b \in T_B$ to be performed according to its state $s_B \in S_B$ and the announcer action a. Both the players know the set of possible states for the opponent and the joint probability distribution over states $\Pr(s_A, s_B)$, but do not know the current state of the opponent. According to the selected actions they will receive respectively a reward $R_A(s_A, s_B, a, b)$ and $R_B(s_A, s_B, a, b)$ that is positive if the swap is completed and 0 otherwise. The game—called G—is entirely defined by the t-uple $(S_A, S_B, T_A, T_B, R_A, R_B)$ and the joint probability over states $\Pr(s_A, s_B)$.

The goal of the agents is to maximize the average reward along the infinite stages of the game:

$$G_A = \lim_{h\to\infty} E\left(\frac{1}{h}\sum_{t=0}^{h} R_A(s_A^t, s_B^t, a^t, b^t)\right) \tag{1a}$$

$$G_B = \lim_{h\to\infty} E\left(\frac{1}{h}\sum_{t=0}^{h} R_B(s_A^t, s_B^t, a^t, b^t)\right) \tag{1b}$$

which is achieved by maximizing the expected immediate reward at each time step, since the next state of the game does not depend on the actions performed by the agents[2].

3 Optimal policies

In order to maximize its reward each player must select a policy. An announcer (mixed) policy is defined by a mapping π_A from states to distributions of probability on actions: $\pi_A: S_A \to P(T_A)$. We denote with $\pi_A(s_A,a) \in [0,1]$ the probability that the announcer uses action a in state s_A. A bidder (mixed) policy is defined by a mapping π_B from states and opponent actions to distribution of probability on actions: $\pi_B: S_B \times T_A \to P(T_B)$. We denote with $\pi_B(s_B,a,b) \in [0,1]$ the probability that the bidder uses action b in state s_B against the announcer action a.

Formula (1) defines the agents objectives as maximizing the average reward that strictly depends on the opponent state and action. In the game theory literature, a usual solution is to evaluate each policy with respect to the opponent that makes it look the worst. This is the essence of the *minimax* policy, which is based on the assumption that the opponent will choose the strategy that minimizes the player's rewards. Although it is founded in the context of zero-sum games, this rationality does not fit our problem, since a gain from a player does not necessarily imply a loss for the other. In this context, it is more suited to consider that each agent will try to maximize its own rewards knowing that the opponent maximizes its reward too. This leads to a criterion of the kind *maximax*. In zero sum games, both criterions are equivalent since maximizing a player's reward is equivalent to minimizing its opponent's gain. In non-zero sum game, the use of the *maximax* criterion is natural, since it allows a cooperative behavior to be adopted.

An equilibrium point using the *maximax* criterion is attained if the policies π_A^* and π_B^* of agents A and B satisfy:

[1] DEIS – Università di Bologna, Viale Risorgimento 2, 40136 Bologna - ITALY. E-mail: mgolfarelli@deis.unibo.it

ECAI 98. *13th European Conference on Artificial Intelligence*
Edited by Henri Prade
Published in 1998 by John Wiley & Sons, Ltd.

[2] For this reason, we could equivalently use the expected infinite sum of discounted rewards.

$$\forall s_A \in S_A, \pi_A^*(s_A) =$$

$$\underset{p_A \in P(T_A)}{\operatorname{argmax}} \left[\sum_{a \in T_A} p_A(a) \sum_{s_B \in S_B} \Pr(s_B \mid s_A) \sum_{b \in T_B} \pi_B^*(s_B, a, b) \cdot R_A(s_A, s_B, a, b) \right], \quad (2a)$$

$$\forall s_B \in S_B,\ \forall a \in T_A, \pi_B^*(s_B, a) =$$

$$\underset{p_B \in P(T_B)}{\operatorname{argmax}} \left[\sum_{b \in T_B} p_B(b) \sum_{s_A \in T_A} \Pr(s_A \mid s_B, a, \pi_A^*) \cdot R_B(s_A, s_B, a, b) \right] \quad (2b)$$

where $\Pr(s_A \mid s_B, a, \pi_A^*)$ is the probability of agent A being in state s_A, knowing that π_A^* let him to choose action a and that agent B is in state s_B. This parameter may be calculated using Bayes' rule

$$\Pr(s_A \mid s_B, a, \pi_A^*) = \frac{\Pr(s_A \mid s_B) \cdot \pi_A^*(s_A, a)}{\sum_{s'_A \in S_A} p(s'_A \mid s_B) \cdot \pi_A^*(s'_A, a)} \quad (3)$$

starting from the joint probability of states. The existence of such a mixed equilibrium is demonstrated for every bi-matrix game [1]. However our game is not a bi-matrix game, except in the degenerated case where $|S_1| = |S_2| = 1$. Moreover, as our game is an alternating game, one could expect that the existence of a mixed equilibrium implies the existence of a pure equilibrium. But the presence of the Bayes rule (3) in the right hand side of (2b) prevents from such shortcut. Anyway, we can show that our game is completely equivalent to a larger bi-matrix game, where the available actions are the pure policies of our game.

In the game G, the agents first see their state and them randomly choose an action. Consider now the associated game where each agent first (randomly) chooses a deterministic (or pure) policy:

- $\mu_A: S_A \to T_A$ for player A,
- $\mu_B: S_B \times T_A \to T_B$ for player B,

then sees his state and apply the decision dictated by his policy in this state. This defines a $|T_A|^{|S_A|} \times |T_B|^{|S_B|.|T_A|}$ simultaneous bi-matrix game M where mixed policies are probability distributions over pure policies of G:

- $P_A \in P(\{\mu_A: S_A \to T_A\})$ for player A,
- $P_B \in P(\{\mu_B: S_B \times T_A \to T_B\})$ for player B.

As M is a bi-matrix game, it admits at least one mixed equilibrium.

The games G and M are related through the function Φ_A and Φ_B that maps their sets of mixed policies:

$$\forall s_A \in S_A, \forall a \in T_A: \left(\Phi_A(P_A)\right)(s_A, a) = \sum_{\substack{\mu_A : S_A \to T_A \\ \text{s.t. } \mu_A(s_A) = a}} P_A(\mu_A), \quad (4a)$$

$$\forall s_B \in S_B, \forall a \in T_A, \forall b \in T_B:$$

$$\left(\Phi_B(P_B)\right)(s_B, a, b) = \sum_{\substack{\mu_B : S_B \times T_A \to T_B \\ \text{s.t. } \mu_B(s_B, a) = b}} P_B(\mu_B), \quad (4b)$$

It can be showed [5] that Φ_A and Φ_B are mappings and that their inverse are, respectively,

$$\forall \mu_A: S_A \to T_A: \left(\Phi_A^{-1}(\pi_A)\right)(\mu_A) = \prod_{s_A \in S_A} \pi_A(s_A, \mu_A(s_A)) \quad (5a)$$

$$\forall \mu_B: S_B \times T_A \to T_B:$$

$$\left(\Phi_B^{-1}(\pi_B)\right)(\mu_B) = \prod_{s_B \in S_B} \prod_{a \in T_A} \pi_B(s_B, a, \mu_B(s_B, a)) \quad (5b)$$

Moreover, there is a complete equivalence of the two games, in terms of Nash equilibrium:

Theorem 1 (P_A, P_B) *is a Nash equilibrium of M if and only if* ($\Phi_A(P_A)$, $\Phi_B(P_B)$) *is a Nash equilibrium of G.*

This proves the existence of a mixed Nash equilibrium for the game G. Unfortunately, the uniqueness of such a point is not guaranteed and thus the agents are still faced with the problem of choosing a common policy. A normal suggestion for resolving the problem posed by the presence of several efficient outcomes is that the player selects a single outcome by bargaining with each other. Within the CNP, explicit communications are forbidden, but the agents can agree on a cooperative behavior "*expressing their preferences*" during the infinite stages of the game. [Boutilier and Claus] and [2 (to check)] study some mechanisms of this kind and provide hints to favor them in a population of adaptive agents.

4 Conclusion

This paper studies the multiagent interactions within the CNP from a game theory viewpoint. In particular, an optimality criterion based on the *maximax* concept is proposed and the conditions to obtain a cooperative behavior are drawn. Basically, we have established that some instances of the CNP can be modeled as a bi-matrix game whose dimension is exponential in the number of tasks to be completed and the number of states of the agents. Therefore, there is little hope to solve it exactly. However, recent research efforts have been made to solve this kind of problem adopting reinforcement learning techniques [2,6]. Our current research is aimed at addressing this issue within the CNP problem.

Acknowledgements: Thanks to T. Dean, L. Kaelbling and N. Meuleau for their comments and discussion of these ideas.

Reference

[1] R.J. Aumann and M. B. Maschler, *Repeated Games with Incomplete Information*, MIT Press, Cambridge, Massachusetts, 1995.

[2] C. Boutilier, Learning Conventions in Multiagent Stochastic Domains using Likelihood Estimates. In *Proc. 12th Conf. on Uncertainty in AI.* pp. 106-114, Portland, 1996.

[3] M. Golfarelli, D. Maio, S. Rizzi. Multiagent Path Planning Based on Task-Swap Negotiation. In *Proc. 16th UK Planning and Scheduling Workshop*, Durham, UK, 1997.

[4] M. Littman. Markov games as a framework for multiagent reinforcement learning. In *Proc. 11th Int. Conf. on Machine Learning*, pp. 157-163, San Francisco, CA, 1994.

[5] N. Meuleau and M. Golfarelli. *A model of partially observable state-game and its optimality*. In prep. 1998.

[6] T.W. Sandholm and R. H. Crites. Multiagent Reinforcement Learning in the Iterated Prisoner's Dilemma. *Biosystems* 37: 147-166. 1995