A Game-Theoretic Model and Best-Response Learning Method for Ad Hoc Coordination in Multiagent Systems (Extended Abstract)

Stefano V. Albrecht
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, UK
s.v.albrecht@sms.ed.ac.uk

Subramanian Ramamoorthy
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, UK
s.ramamoorthy@ed.ac.uk

ABSTRACT

The *ad hoc coordination* problem is to design an ad hoc agent which is able to achieve optimal flexibility and efficiency in a multiagent system that admits no prior coordination between the ad hoc agent and the other agents. We conceptualise this problem formally as a *stochastic Bayesian game* in which the behaviour of a player is determined by its type. Based on this model, we derive a solution, called *Harsanyi-Bellman Ad Hoc Coordination* (HBA), which utilises a set of user-defined types to characterise players based on their observed behaviours. We evaluate HBA in the *level-based foraging* domain, showing that it outperforms several alternative algorithms using just a few user-defined types. We also report on a human-machine experiment in which the humans played *Prisoner's Dilemma* and *Rock-Paper-Scissors* against HBA and alternative algorithms. The results show that HBA achieved equal efficiency but a significantly higher welfare and winning rate.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]

Keywords

Ad Hoc Coordination; Stochastic Bayesian Games (SBG); Harsanyi-Bellman Ad Hoc Coordination (HBA)

1. INTRODUCTION

We are concerned with the *ad hoc coordination* problem, in which the goal is to design an *ad hoc agent* which is able to achieve optimal flexibility and efficiency in a multiagent system that admits no prior coordination between the ad hoc agent and the other agents. We are motivated by human-machine interaction problems such as robots used in nursing homes or rescue scenarios, and software agents used in online trading or video games. Here, the agent is expected to be able to quickly adapt to initially unknown behaviours, while at the same time produce consistently good results.

Game theorists have studied a related problem known as *incomplete information game*, in which the players have some private information which is relevant to their decision making. Harsanyi [5] introduced *Bayesian games* in which the private information of a player is represented by its *type*. While the concept of Bayesian games is useful to describe the ad hoc coordination problem, the learning processes and solutions studied therein are not directly applicable, since the focus has traditionally been on equilibrium attainment but not on efficiency. On the other hand, much work in intelligent agents focuses on efficiency, whilst often using some form of prior coordination (see [1] for a discussion). Thus, it is natural to ask if these fields can be combined in a useful way.

There have been several attempts to address incomplete information in multiagent systems, e.g. [3, 4, 6]. However, the assumptions made by the solutions proposed therein are relatively strong, which means that they only address certain aspects of the larger problem. For example, in [3, 4] it is assumed that all agents follow complex plans which specify roles and synchronised action sequences, and in [6] and affiliated works it is assumed that the behaviours of the other agents are a priori known and fixed, and that all agents have common payoffs. Furthermore, the problem descriptions in these works are of a procedural nature, associated with the specific tasks considered therein. That is, there is no formal model of the ad hoc coordination problem, general enough to accommodate a wide spectrum of problems.

In this work, we propose to model the problem as a *stochastic Bayesian game*, based on which we give formal definitions of flexibility, efficiency, and ad hoc coordination. We derive a best-response rule from this model, called *Harsanyi-Bellman Ad Hoc Coordination* (HBA), that utilises a set of user-defined types which, similar to [5], specify a player's payoffs and strategies. This solution is extended by mechanisms which allow it to register changed types and learn new types. We demonstrate the effectiveness of HBA in both simulated experiments and a human-machine experiment.

2. MODEL & SOLUTION

A *stochastic Bayesian game* (SBG) starts at time $t = 0$ in some initial state s^0. In state s^t, the player types $\theta_1^t, ..., \theta_n^t$ are sampled with probability $\Delta(t, (\theta_1^t, ..., \theta_n^t))$ (Δ is the *type distribution*) and each player i is informed about its own type θ_i^t. Based on the history $H^t = \langle s^0, a^0, s^1, a^1, ..., s^t \rangle$, each player i chooses an action a_i^t with probability $\pi_i(H^t, a_i^t, \theta_i^t)$ (π_i is player i's strategy). Given $a^t = (a_1^t, ..., a_n^t)$, the game transitions into state s^{t+1} with probability $T(s^t, a^t, s^{t+1})$ and every player i receives an individual payoff $u_i(s^t, a^t, \theta_i^t)$. This continues until the game reaches a terminal state.

⇒ A detailed account of this work can be found in [2].

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

Given a SBG Γ, we define the flexibility $F(\alpha|\Gamma,\mathbb{D})$ and the efficiency $E(\alpha|\Gamma,\mathbb{D})$ of ad hoc agent α (controlling player i) with respect to a set of type distributions $\mathbb{D}$ as

$$F(\alpha|\Gamma,\mathbb{D}) = \frac{1}{|\mathbb{D}|}\sum_{\Delta\in\mathbb{D}}\sum_{\rho\in\Phi}\Pr(\rho|\Gamma,\Delta)$$

$$E(\alpha|\Gamma,\mathbb{D}) = \frac{1}{|\mathbb{D}|}\sum_{\Delta\in\mathbb{D}}\sum_{\rho\in\Phi}\overline{\Pr}(\rho|\Gamma,\Delta)\frac{\left(\sum_{\tau=0}^{t_\rho-1}u_i(s_\rho^\tau,a_\rho^\tau,\alpha)\right)^{r_1}}{(t_\rho)^{r_2}}$$

where Φ is the set of all paths $\rho = \langle s_\rho^0,\theta_\rho^0,a_\rho^0,s_\rho^1,\theta_\rho^1,a_\rho^1,...,s_\rho^{t_\rho}\rangle$ such that $s_\rho^{t_\rho}$ is a terminal state in Γ, $\Pr(\rho|\Gamma,\Delta)$ is the probability of ρ in Γ with type distribution Δ, $\overline{\Pr}(\rho|\Gamma,\Delta)$ is the probability of ρ normalised over Φ, and $r_1, r_2 \geq 1$ specify the relative importance between payoff and time.

Based on the above definitions, the ad hoc coordination problem is to optimise the flexibility $F(\alpha|\Gamma,\mathbb{D})$ and efficiency $E(\alpha|\Gamma,\mathbb{D})$ of α in Γ with respect to a set of type distributions $\mathbb{D}$, subject to the constraint that α does not know the type spaces Θ_j in Γ (and, thus, the type distributions in Γ).

We derive a solution to this problem, called *Harsanyi-Bellman Ad Hoc Coordination* (HBA), by combining the concept of the Bayesian Nash equilibrium [5] with the Bellman optimality equation. Let Γ be an ad hoc coordination problem where ad hoc agent α controls player i and has access to user-defined type spaces $\Theta^*_{-i} = \times_{j\neq i}\Theta^*_j$. The best response rule HBA is defined as $a_i^t \sim \arg\max_{a_i} E^{a_i}_{s^t}(H^t)$ where

$$E^{a_i}_s(\hat{H}) = \sum_{\theta^*_{-i}\in\Theta^*_{-i}}\Pr(\theta^*_{-i}|H^t)\sum_{a_{-i}\in A_{-i}}Q^{a_i,-i}_s(\hat{H})\prod_{j\neq i}\pi_j(\hat{H},a_j,\theta^*_j)$$

is the expected long-term payoff for player i of taking action a_i in state s after history $\hat{H}$ ($a_{i,-i}$ denotes (a_i,a_{-i})), $\Pr(\theta^*_{-i}|H^t)$ is the *posterior* over opponent types θ^*_{-i} after history H^t, and $Q^a_s(\hat{H})$ is the expected long-term payoff for player i when joint action a is executed in state s after history $\hat{H}$.

Two properties of HBA are that it is optimal in self-play and that it achieves optimal efficiency against the class of *deterministic learners* [2]. For a more detailed description of our model and solution, including two extensions which enable HBA to recognise changed types (*TR-posteriors*) and learn new types (*conceptual types*), we refer to [2].

3. EXPERIMENTS

We tested various configurations of HBA and three alternative algorithms in the *level-based foraging* domain (Figure 1a), in which a group of agents (circles) attempts to collect foods (squares) in minimal time while also maximising their individual payoffs. Each agent and food has a random level (shown in centre), and a group of agents can collect a food only if the sum of their levels is at least as high as the food's level. All algorithms were implemented using the same reinforcement learning framework. The HBA configurations were "Cor" (always using correct types), "Gtw" (using TR-posteriors), and "Unl" (without TR-posteriors). Gtw and Unl used conceptual types (d1 and d2) to learn new types. The types of the other players were sampled from a large set of fixed and learning behaviours (of which HBA knew only a small fraction) and some players were changing their types periodically. Since optimal solutions were infeasible to compute, we measured the performance of humans using a graphical user interface. The results are shown in Figure 1b.

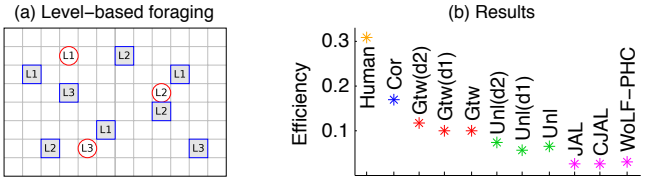

Figure 1: Simulated experiments. Results averaged over 1000 runs. Markers have same colour if difference insignificant (paired t-test, 5% sig. level).

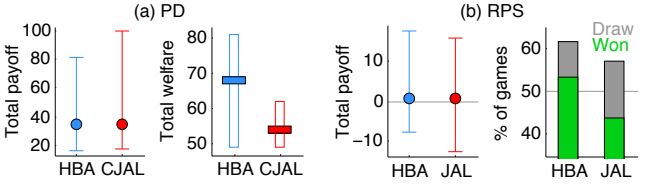

Figure 2: Human-machine experiment. Circles and whiskers show mean, min, and max. Welfare plot in (a) shows median and 25% / 75% percentiles.

We also conducted a human-machine experiment at the Royal Society Summer Science Exhibition 2012 in London.[1] Therein, the human participants played repeated *Prisoner's Dilemma* (PD) and *Rock-Paper-Scissors* (RPS) against HBA and alternative algorithms, where each game was played for 20 rounds. All algorithms were implemented using the same exact planning framework. We collected data from 427 participants, of which 186 played PD and 241 played RPS. The lowest and highest recorded ages were 9 and 72, respectively, with an average age of about 17. HBA used a small set of types (cf. Table 1 in [2]) and did not learn any new types. The results (Figure 2) show that HBA achieved equal efficiency but a significantly higher social welfare (sum of payoffs) and winning rate.

4. ACKNOWLEDGEMENTS

This work is partially supported by grants from the UK Engineering and Physical Sciences Research Council (EP/H012338/1), the European Commission (TOMSY Grant 270436, FP7-ICT-2009.2.1 Call 6) and a Royal Academy of Engineering Ingenious grant. S.A. is supported by the German National Academic Foundation.

5. REFERENCES

[1] S. Albrecht and S. Ramamoorthy. Comparative evaluation of MAL algorithms in a diverse set of ad hoc team problems. In *11th Autonomous Agents and Multiagent Systems*, 2012.

[2] S. Albrecht and S. Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. Technical report, School of Informatics, University of Edinburgh, 2012. http://wcms.inf.ed.ac.uk/ipab/autonomy/publications/SAlbrecht_tech_report.pdf.

[3] M. Bowling and P. McCracken. Coordination and adaptation in impromptu teams. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 53, 2005.

[4] M. Dias, T. Harris, B. Browning, E. Jones, B. Argall, M. Veloso, A. Stentz, and A. Rudnicky. Dynamically formed human-robot teams performing coordinated tasks. In *AAAI Spring Symposium*, 2006.

[5] J. Harsanyi. Games with incomplete information played by "Bayesian" players. Part I. The basic model. *Management Science*, 14(3):159–182, 1967.

[6] P. Stone and S. Kraus. To teach or not to teach? Decision making under uncertainty in ad hoc teams. In *9th Autonomous Agents and Multiagent Systems*, 2010.

[1] http://sse.royalsociety.org/2012/exhibits/robotic-soccer