Valuing Search and Communication in Partially-Observable Coordination Problems (Extended Abstract)

Simon A. Williamson, Archie C. Chapman and Nicholas R. Jennings
Electronics & Computer Science
University of Southampton
Southampton, SO17 1BJ, UK.
{acc,saw1,nrj}@ecs.soton.ac.uk

ABSTRACT

In this paper we extend the class of *Bayesian coordination games* to include explicit observation and communication. This general class of problems includes the canonical multi–door multi–agent Tiger problem. We argue that this class of games is appropriate for situations where the agents observation, communication and payoff–earning actions are limited by some common resource, without introducing arbitrary penalties for communicating (unlike most existing approaches).

Categories and Subject Descriptors

I.2.11 [Computing Methodologies]: Distributed Artificial Intelligence—*Multiagent systems*

General Terms

Economics, Theory

Keywords

Coordination games, agent communication

1. INTRODUCTION

Information is central to coordination in partially observable multi–agent systems. That is, by observing the world and communicating their beliefs, agents can better coordinate their actions to achieve their goals. However, in many real–world problems, information gathering and communication are not free. Rather, they are subject to restrictions on availability, bandwidth or timing, or must be carried out at the expense of some other action. Consequently, agents must reason over the costs and benefits of an observation and communication policy, and its effects on other agents' decision (i.e. the policy's stability) before utilising it. We argue that these activities should be treated like any other action: they consume scarce resources that then cannot be used to perform other actions, but may confer some benefit to the agent. As such, the costs and benefits of gathering information and communicating can be better expressed as the value of their effects on the agents' chances of achieving underlying goals rather than using arbitrary costs.

In this paper, we use the representation of communication from decentralised POMDPs (e.g. [3], [2]) to develop a framework for valuing observations and communication in a new class of games — iterated Bayesian Coordination games with explicit observing and communicating actions. Here, agents can request observations or communicate to refine their view of the world before they commit to any action *within the game*. As a result, agents can reason about coordinating by individually identifying the state or by using communication to share each others' beliefs about the world. However, depending on the relative costs of communication and observing, each policy will be appropriate in different contexts. This reasoning is not possible in traditional Bayesian games because they ignore costly communication and information gathering: Furthermore, traditionally, the Bayesian games considered are not large enough to warrant an information gathering strategy, and communication and information gathering are not considered. In contrast, our model captures a broad class of problems, including a multi–door multi–agent extension of the Tiger problem [1].

Cite as: Valuing Search and Communication in Partially-Observable Coordination Problems (Extended Abstract), Williamson, Chapman & Jennings, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1461-1462

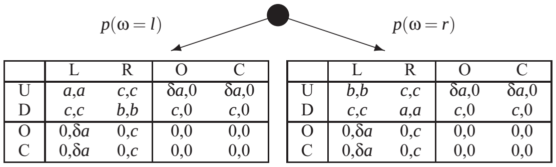
2. BAYESIAN COORDINATION GAMES

A noncooperative game consists of a set of agents $N = 1,\dots,n$, and for each agent $i \in N$, a set of *strategies* $S_i = \{1,\dots,m_i\}$, and a *utility function* $u_i : S \rightarrow \mathbb{R}$, where $S = \cup_{i=1}^{N} S_i$. A joint strategy profile $s^* \in S$ is a *Nash equilibrium* (NE) if for all agents, $u_i(s_i^*, s_{-i}^*) - u_i(s_i, s_{-i}^*) \geq 0 \; \forall \; s_i$. *Bayesian games* model situations where agents have to act without knowing the true state of the world These are noncooperative games with the addition of a *state space* Ω, and for each player $i \in N$: a set of possible *types* Θ_i, a *signal function* $\zeta_i : \Omega \rightarrow \Theta_i$, and a *prior belief* about the state of the world and the payoffs for each action. Here, $\omega \in \Omega$ is interpreted as a particular "state of nature", and associated with each state is a particular *stage game*, defining all the agents' types. Then, an agent's utility function, $u_i : S \times \Omega \rightarrow \mathbb{R}$, maps from strategy profiles and states (types) to its payoffs. The signal function maps from states to types, such that $\zeta_i(\omega) = \theta_i$ is the type of player i in state ω. Finally the conditional probability $p_i(\omega|\theta_i)$ summarises what i believes about the state of nature given its own type.

Following this, we are particularly interested in agent coordination, so we focus on *coordination games*. In these games, coordination results in a high payoff to the agents, while any mis–coordination leads to a low payoff. Specifically: (i) Each agent has the same size strategy space $m_i = m$ for all i; (ii) Strategies can be ordered such that $s^l = (l,\dots,l)$ is a strict NE for all $l = 1,\dots,m$; (iii) For all $i, j \in N$ and all $h, l = 1,...,m$, $u_i(s^h) \geq u_i(s^l)$ if and only if $u_j(s^h) \geq u_j(s^l)$; (iv) $u(s^j) >> u(s)$ for all $s \in S/\{s^1,\dots,s^m\}$. Importantly, these constraints imply that in a Bayesian coordination game, different states only define different rankings of the NE.

2.1 Iterated Bayesian Coordination Games

As mentioned above, our domain differs from the standard model of Bayesian games in two important ways, both of which allow agents to coordinate by achieving a similar view of the world. First, agents can explicitly choose to make observations of the world's state, which causes their beliefs to converge because they access the same



	L	R	O	C
U	a,a	c,c	δa,0	δa,0
D	c,c	b,b	c,0	c,0
O	0,δa	0,c	0,0	0,0
C	0,δa	0,c	0,0	0,0

	L	R	O	C
U	b,b	c,c	δa,0	δa,0
D	c,c	a,a	c,0	c,0
O	0,δa	0,c	0,0	0,0
C	0,δa	0,c	0,0	0,0

Figure 1: A two–player, two state Bayesian coordination game with explicit observation and communication, where $a \geq b > c \leq d$. When the state is l (left), the NE $\{U,L\}$ is preferred over $\{D,R\}$, while when $\omega = r$ (right), the opposite is the case.

observation function and are more likely to take the high payoff action. Second, the agents can directly communicate (broadcast) their beliefs over the state of the world to each other. Furthermore, in our model, both of these actions take time. This is a key feature, and allows us to model more general problems in which communication consumes resources like any other action. Now, because there are only a finite number of time steps in the repeated game, the choice to observe or communicate must be made at the expense of forgoing a payoff–earning action. That is, the value of observing the state or communicating one's beliefs must be traded–off against the value of taking a less informed action.

Formally, we consider a Bayesian coordination game with the addition of explicit, time–consuming observing (O) and communicating (C) actions, repeated a finite number of times. An agent's utility function is the sum of its payoffs from each stage–game. In the stage–games, the payoffs to O and C are zero, regardless of the actions of other players in the game. In the two–player version of the game, if one agent plays O or C, we define the payoff for the second agent that takes the payoff–dominant equilibrium policy (e.g. U or L in $\omega = l$) as some fraction, $0 < \delta < 1$, of its equilibrium payoff. If the second agent takes a different policy, it receives the payoff for mis–coordinating. For the two–agent two–state case, these stage game payoffs are summarised in Fig 1 for $\omega = l$ (corresponding to the payoff–dominant equilibrium at $\{U,L\}$), where $a > 0 > c$, $a \geq b$ and $0 < \delta < 1$. Note that when the column player plays O or C, the payoff to the row player for playing the payoff–dominant equilibrium policy U is δa, and when it plays D it is c.

Finally, we introduce the concept of time by considering the finite iterated version of the game. This is a structured way of formally defining the opportunity costs of actions — each action takes a time–step and it is not possible to conduct several actions in parallel. The level of noise in the signal function, ε (i.e. $Pr(\zeta = \omega) = 1 - \varepsilon$, with $0 < \varepsilon < 1$) is constant throughout. Now, in a finitely repeated game, the appropriate payoff function is the undiscounted sum of agents' payoffs, which is equivalent to maximising the average payoff per time–step. As such, we reduce the problem of finding a payoff–dominant equilibrium in the repeated game to finding one in the stage–game.

2.2 The Structure of the Auxiliary Game

Here, we describe how to build an auxiliary game representing the value of different communication protocols from the Bayesian coordination game. To begin, we can considerably collapse the set of policies admitted for all cases. This is done by, first, defining expected rewards for policies in terms of whether an agent's beliefs tend towards the true state of the world or not. Second, we assume that when an agent takes a payoff generating action (i.e. not O or C) it takes the action with the highest expected reward given its beliefs. This allows us to reason over all the payoff generating actions as one, abstract 'act' action, which we write as A. For example, in Fig 1, A for the row player is the act of moving up or down in response to its beliefs, and not specifically U or D. Third, in general, an agent's policy may be any combination of O and C actions followed by A, with the game resetting after this action. As such, we consider only policies which conclude with an A and do not contain multiple As, as all other strategies can be constructed by combining these strategies, so they are redundant. Furthermore, we do not allow the agents to make any additional observations after communicating — they always act immediately after communicating because communicating more than once makes earlier Cs redundant and for a fixed strategy length, communicating later always dominates communicating earlier, because more information is transferred. Therefore, a single C immediately before A dominates all other combinations containing one or more Cs. Thus, we can restrict the agent's policy to the following combinations of actions: (i) Observe m times and then act, (e.g. A or OOA), or (ii) Observe m times, communicate and then act (e.g. OCA). Observing m times means using a search strategy of length m.

	A	OA	OCA	OOA	$\cdots$	O^mCA	O^mOA
A	$\pi(A,A)$	$\pi\binom{A,}{OA}$	$\pi\binom{A,}{OCA}$	$\pi\binom{A,}{OOA}$	$\cdots$	$\pi\binom{A,}{O^mCA}$	$\pi\binom{A,}{O^mOA}$
OA	0	$\pi\binom{OA,}{OA}$	$\pi\binom{OA,}{OCA}$	$\pi\binom{OA,}{OOA}$	$\cdots$	$\pi\binom{OA,}{O^mCA}$	$\pi\binom{OA,}{O^mOA}$
OCA	0	0	$\pi\binom{OCA,}{OCA}$	$\pi\binom{OCA,}{OOA}$	$\cdots$	$\pi\binom{OCA,}{O^mCA}$	$\pi\binom{OCA,}{O^mOA}$
OOA	0	0	$\pi\binom{OOA,}{OCA}$	$\pi\binom{OOA,}{OOA}$	$\cdots$	$\pi\binom{OOA,}{O^mCA}$	$\pi\binom{OOA,}{O^mOA}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
O^mCA	0	0	0	0	$\cdots$	$\pi\binom{O^mCA,}{O^mCA}$	$\pi\binom{O^mCA,}{O^mOA}$
O^mOA	0	0	0	0	$\cdots$	$\pi\binom{O^mOA,}{O^mCA}$	$\pi\binom{O^mOA,}{O^mOA}$

Figure 2: Generic payoff table for row player in the 2–player auxiliary game

Now that we have reduced the set of policies that need to be considered, we can derive the expected rewards to agents for following combinations of these policies. Specifically, the interaction of agents' policies is described as a normal form *auxiliary game* (a higher level game describing a game). Each outcome of this auxiliary game defines a combination of the agents' A, O and C policies. The value of the payoff to an agent for an outcome in the auxiliary game is the average expected payoff per time–step that the agent receives in the underlying Bayesian coordination stage game. We denote this value as $\pi_i(a_i, a_j)$, and it is given by the expected reward $\mathbb{E}[u_i(a_i, a_j)]$ (derived in the coming section) divided by the length of its corresponding policy:

$$\pi_i(a_i, a_j) = \frac{\mathbb{E}[u_i(a_i, a_j)]}{\min\{|a_i|, |a_j|\}} \qquad (1)$$

where $|a_i|$ is the length of agent i's policy. Furthermore, we can drop the agent index on π_i because the expected payoffs to all agents are symmetric. In the case of two agents, Fig 2 illustrates the generic payoff matrix to the row agent.

3. FUTURE WORK

By comparing the relative costs and benefits of communicating, observing and acting given an agent's beliefs, we hope to show that the optimal communication policy is a symmetric Nash equilibrium. Moreover, the framework described here allows us to specify and analyse general information gathering strategies, and to do so independently of the environment they operate in. In future work, we will use this framework to derive a general method for generating optimal communication policies in iterated Bayesian coordination games with explicit communication and observation, using different observation–gathering strategies.

4. REFERENCES

[1] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella. Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *IJCAI '03*, 2003.

[2] S. A. Williamson, E. H. Gerding, and N. R. Jennings. Reward shaping for valuing communications during multi-agent coordination. In *AAMAS '09*, pages 641–648, 2009.

[3] P. Xuan, V. Lesser, and S. Zilberstein. Communication decisions in multi-agent cooperation: Model and experiments. In *ICAA' 01*, pages 616–623. ACM Press, 2001.