Communicating with Unknown Teammates

(Extended Abstract)

Samuel Barrett
The University of Texas
Austin, TX, USA
sbarrett@cs.utexas.edu

Noa Agmon
Bar-Ilan University
Ramat Gan, Israel
agmon@cs.biu.ac.il

Noam Hazon
Ariel University
Ariel, Israel
noamh@ariel.ac.il

Sarit Kraus[1,2]
[1]Bar-Ilan University
Ramat Gan, Israel
[2]University of Maryland
College Park, MD, USA
sarit@cs.biu.ac.il

Peter Stone
The University of Texas
Austin, TX, USA
pstone@cs.utexas.edu

ABSTRACT

Past research has investigated a number of methods for coordinating teams of agents, but, with the growing number of sources of agents, it is likely that agents will encounter teammates that do not share their coordination methods. Therefore, it is desirable for agents to form an effective *ad hoc team*. This research tackles the problem of communication in ad hoc teams, introducing a minimal version of the multiagent, multi-armed bandit problem with limited communication between the agents. This abstract summarizes theoretical results that prove that this problem setting can be solved in polynomial time when the agent knows the set of possible teammates, and the empirical results that show that the problems can be solved in practice.

1. INTRODUCTION

Given the growing number of both software and robotic agents, effective teamwork is becoming vital to many tasks. With this increase in numbers of agents, their interactions with other agents also increases, as does the number of companies and laboratories creating these agents. Therefore, there is a growing need for agents to be able to cooperate with a variety of different teammates. This need is addressed in the area of *ad hoc teamwork*, where agents are evaluated based on their ability to cooperate with a variety of teammates. Stone et al. [3] define ad hoc teamwork as problems in which a team cannot pre-coordinate its actions and introduce an algorithm for evaluating ad hoc team agents.

Past work on ad hoc teamwork has focused on the case where the ad hoc agent cannot (or does not) directly communicate to its teammates and can only coordinate by observing its teammates' actions. However, in an increasingly interconnected world, it is likely that agents will at least have some limited communication using a common language. This abstract summarizes research that introduces a minimal domain for investigating teammate communication, proves that finding optimal behavior for ad hoc teamwork is tractable in this domain, and shows that these problems are tractable in an empirical setting.

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

2. PROBLEM DESCRIPTION

This abstract summarizes a new multiagent, multi-armed bandit problem that allows limited communication. The multi-armed bandit setting is a fundamental problem in reinforcement learning, and it has been used to study ad hoc teamwork without communication [4]. While general multiagent research focuses on creating a coordinated team to perform tasks, in ad hoc teamwork, the goal is to create agents that can cooperate with a variety of possible teammates. Specifically, we assume that there are several existing teams and an ad hoc agent should fit into any of these teams.

We formally define our bandit problem as the tuple $G = (\mathbb{A}, \mathbb{C}, \mathbb{P}, R)$ where $\mathbb{A}$ is a set of two arms $\{arm_0, arm_1\}$ with Bernoulli payoff distributions, $c \in \mathbb{C}$ is a set of possible messages and their costs $cost(\mathrm{c})$, $\mathbb{P}$ are the players with $|\mathbb{P}| = n+1$ with n of the agents being a pre-designed team, and R is the number of rounds. Each round in the problem involves two phases: (1) a communication phase followed by (2) an action phase. In both phases, all agents act simultaneously. In the communication phase, each agent can broadcast a message of each type to its teammates:

- **obs** – Send the agent's last selected arm and payoff
- **mean**$_i$ – Send the agent's observed mean and number of pulls for arm_i
- **suggest**$_i$ – Suggest that the teammates pull arm_i

These message types are understood by all of the agents. In the action phase, each agent chooses an arm and receives a payoff. The team's goal is to maximize the sum of payoffs minus the communication costs. Note that the results in this work can be generalized to any number of fixed arms, other discrete distributions, and other messages.

We assume that the ad hoc agent's n teammates form an existing team, and therefore are tightly coordinated. Specifically, the team's behavior can be described as a function of the team's knowledge, pooled using the message types provided above. The team also uses the pulls and successes that the ad hoc agent has communicated.

3. MODELING THE PROBLEM

When the ad hoc agent knows its teammates' behaviors, it can model the bandit problem as a Markov Decision Process (MDP). The MDP's state is composed of the pulls and observations of the ad hoc agent's teammates as well as the messages it has sent. Let $K = (p_0, s_0, p_1, s_1)$ be the knowledge about the arms where p_i and s_i are the number of pulls and successes of arm_i. Then, the state is given by the vector $(K_t, K_a, K_c, r, phase, sugg)$, where K_t is the team's knowledge from their pulls, K_a is the ad hoc agent's knowledge from its pulls, K_c is the knowledge that the ad hoc agent has communicated, r is the current round number, *phase* is the phase of the round (either communication or action), and *sugg* is the ad hoc agent's most recent suggestion. As the n agents on the team are coordinated, their actions depend on K_t and K_c and *not* directly on K_a. Given that there are R rounds and n teammates, we know that p_i and s_i in K_t are each bounded by nR and p_i and s_i in both K_a and K_c are each bounded by R. The round r is bounded by R, there are 2 possible phases of a round, and 3 values for *sugg*. Therefore, the state space has at most $(nR)^4 \cdot R \cdot R^4 \cdot R^4 \cdot 2 \cdot 3 = 6n^4R^{13}$ states.

Actions are the ad hoc agent's selected arms and its messages. The transition function is composed of the teammates' decisions, the payoff distributions of the arms, and the effects of the ad hoc agent's messages. Rewards are a sum of the arms' payoffs and the communication costs.

4. THEORETICAL ANALYSIS

In this section, we investigate the complexity of planning to optimally cooperate with teammates that are drawn from a continuous set of stochastic behaviors. We consider a small number of possible behaviors, specifically ε-greedy and UCB(c). For these behaviors, ε is the probability of taking a random action, and c is the scaling factor of the confidence bound. Therefore, the ad hoc agent must maintain a belief distribution over values of ε, values of c, and p the probability of the teammates being ε-greedy. The ad hoc agent knows that ε, c are uniformly distributed over $[0, 1]$, and it starts with an initial estimate of p.

To analyze this problem, we model it as a POMDP based on the MDP described in Section 3. In this setting, the belief space has three partially observed values: ε, c, and p. The belief distribution over these values can be represented succinctly. The distribution of c can be represented using a minimum and maximum value, updated using linear programming, ε can be represented using a beta distribution, and p can be represented using a single real. As shown in [1], a POMDP can be solved approximately in polynomial time given a covering set. Lemma 1 states that the covering set can be calculate and is polynomial, so Theorem 1 follows.

LEMMA 1. *The belief space of the resulting POMDP has a δ-covering with size poly$(R, n, 1/\delta)$.*

THEOREM 1. *Consider an ad hoc agent that can observe its teammates' actions, knows the true arm distributions, and knows that its teammates are drawn from a known, continuous set of ε-greedy and UCB teammates. This agent can calculate an η-optimal behavior in poly$(n, R, b, 1/\eta)$ time.*

5. EMPIRICAL EVALUATION

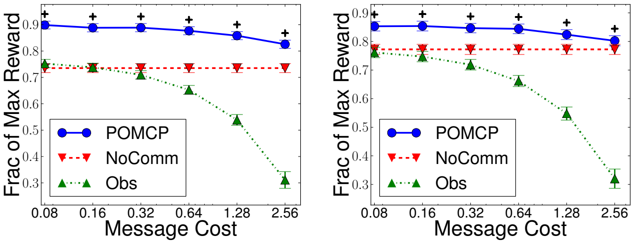
This section investigates whether the problem is empirically tractable in addition to being theoretically tractable. Calculating the exact optimal behavior becomes impractical as the number of rounds and arms grow, so we approximate the optimal behavior using Partially Observable Monte-Carlo Planning (POMCP) [2].

The evaluations use 100 trials with teams where ε, c, and the arms' success probabilities are selected randomly uniformly between 0 and 1. This randomness is fixed across the settings to allow for paired statistical tests. As the ad hoc agent does not know its teammates' behaviors, it initializes its beliefs by sampling both types of behaviors with random parameter values. The results are normalized by the average reward if every agent continuously pulled the best arm. Statistical significance is tested with a paired Student-T test with $p < 0.05$. Points where POMCP is significantly better than all other methods are denoted with "+".

We compare three behaviors of the ad hoc agent:

- **NoComm** - Always pulls the best arm and does not communicate
- **Obs** - Always pulls the best arm and communicates its last observation
- **POMCP** - Plans using POMCP

These tests use 3 arms, 10 rounds, and 7 teammates to test how our approach scales to bigger problems than are theoretically proven. Furthermore, the costs for sending messages are randomly selected for each run, and all agents are informed of the costs. Costs of communicating are randomly selected from the range $[0, m|c|]$, where $|c|$ is size of the message (3 for mean, 2 for obs, and 1 for sugg).



(a) ε-greedy teammates *(b) UCB teammates*

Figure 1: Normalized rewards with varied message costs with a logarithmic x-axis. Significance is denoted by "+"

Figure 1 presents the results when the ad hoc agent cooperates with teams that are ε-greedy or UCB, with varied message costs. The results indicate that the agent can effectively plan its actions, significantly outperforming the baselines. The improvement of POMCP diminishes as the cost of messages rises because affecting the teammates becomes more costly. When the ad hoc agent knows the correct behavior type, the results are similar to knowing that either ε-greedy or UCB teams are possible.

6. CONCLUSION

Past research on ad hoc teamwork has largely focused on scenarios in which the ad hoc agent cannot directly communicate with its teammates. This work addresses this gap by investigating an agent that reasons about communicating in ad hoc teams. To this end, this abstract summarizes a new minimal domain with communication, shows that the problem can be optimally solved in polynomial time, and analyzes an empirical approach to solving the problem.

7. REFERENCES

[1] D. Hsu, W. S. Lee, and N. Rong. What makes some POMDP problems easy to approximate? In *NIPS*. 2007.

[2] D. Silver and J. Veness. Monte-Carlo planning in large pomdps. In *NIPS '10*. 2010.

[3] P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI*, July 2010.

[4] P. Stone and S. Kraus. To teach or not to teach? Decision making under uncertainty in ad hoc teams. In *AAMAS*, May 2010.