To appear in *Autonomous Agents and Multiagent Systems - Adaptive Learning Agents Workshop* (AAMAS - ALA), Taipei, Taiwan, May 2011.

Ad Hoc Teamwork Modeled with Multi-armed Bandits: An Extension to Discounted Infinite Rewards

Samuel Barrett
Dept. of Computer Science
University of Texas at Austin
Austin, TX 78712 USA
sbarrett@cs.utexas.edu

Peter Stone
Dept. of Computer Science
University of Texas at Austin
Austin, TX 78712 USA
pstone@cs.utexas.edu

ABSTRACT

Before deployment, agents designed for multiagent team settings are commonly developed together or are given standardized communication and coordination protocols. However, in many cases this pre-coordination is not possible because the agents do not know what agents they will encounter, resulting in *ad hoc team* settings. In these problems, the agents must learn to adapt and cooperate with each other on the fly. We extend existing research on ad hoc teams, providing theoretical results for handling cooperative multi-armed bandit problems with infinite discounted rewards.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]

General Terms

Algorithms, Theory

Keywords

Ad Hoc Teams, Agent Cooperation: Teamwork, coalition formation, coordination, Agent Reasoning: Planning (single and multi-agent), Agent Cooperation: Implicit Cooperation

1. INTRODUCTION

Autonomous agents are becoming increasingly prevalent in society, both as robots and as software agents. As this trend progresses, there is a growing need for agents to interact and cooperate with other agents. In many situations, these interactions can be specified ahead of time, as in many multiagent team settings. However, agents are also becoming more robust and reliable, so it is likely that they will also encounter agents that are unknown during development. In these cases, the agents should be able to adapt and cooperate with these unknown teammates.

In a recent AAAI challenge paper, Stone et al. [13] formally introduced the *ad hoc team setting* and described it as a problem in which strategies for team coordination cannot be specified a priori. As autonomous agents proliferate in our society, it is important that they are capable of handling ad hoc team settings. Specifically, we study the effectiveness of an *individual* ad hoc team agent's strategy to cooperate with a teammate.

The remainder of the paper is organized as follows. Section 2 provides a motivating example for this research, and Section 3 specifies the formal framework that will be used in this paper, specifically a cooperative multi-armed bandit with infinite discounted rewards. Then, Section 4 presents the main contribution, namely theoretical results considering a three armed bandit with arbitrary distributions of the arms. Next, Section 5 extends these results for many arms. Section 6 situates our contribution in the literature, and Section 7 concludes.

2. MOTIVATING EXAMPLE

Consider two robots tasked with picking up as much trash as possible from two beaches. Each robot must recharge its batteries daily, and between recharging, the travel times to the beaches, and the tides, each robot is only able to clean one beach a day. The tides wash away trash that a robot does not pick up, so the trash does not build up. Therefore, the robots are set to pick up trash during alternating tides. Each robot should choose to clean the beach with the highest amount of trash, but the amount of trash is random, depending on the weather and popularity of the beaches as well as additional factors. The robots communicate to each other about how much trash they found at the beaches they cleaned. By trying both beaches and tracking the average amount of trash picked up, the robots can learn to clean the messier beach with high probability. The robots try to maximize the trash picked up over time, but they value immediately cleaning over future cleaning.

Suppose that several years have passed and one of the robots has broken, and original developers no longer work on the project. Therefore, another robot has been built to help clean the beaches. The new robot has an internet connection and can gather information about the popularities of each beach from a municipal website. Also, a new, more popular beach has been created, but the old robot does not know the path to this beach. Unfortunately, this path cannot be added to the old robot's memory because the original developers are not available. The new robot can still communicate the amount of trash it finds at each beach, but the old robot cannot receive other information. The new robot's goal is still to maximize the amount of trash the robots pick up. If the new robot were acting alone, it could pick up the most trash at the new beach, but since it is on a team, it can also affect what beach the old robot chooses. The old robot cannot go to the new beach, so the new robot should use its additional information help guide the old one to clean the more popular of the older beaches. Another robot is being built to replace the old robot, but its completion time is unknown.

The above fictional setting can be formalized as a cooperative multi-armed bandit [12] with infinite discounted rewards because the robots are interested in their long term rewards, but value immediate rewards more than later rewards. Immediate rewards are more valuable because there is a chance that the episode will end before the robots receive any future rewards. This problem is similar to the one described by Stone and Kraus [15], except that we consider infinite discounted rewards. This formulation is a commonly studied problem in reinforcement learning [16]. This problem is a simple form of the ad hoc team problem since the behavior of the teammate is fixed and known. Despite these limitations, this problem raises interesting questions about how a knowledgeable agent can teach a novice without explicit communication while operating embedded in the domain.

3. MULTI-ARMED BANDIT

The multi-armed bandit (MAB) problem [12] is well studied in sequential decision making. The problem is modeled after slot machines (often referred to as one-armed bandits), where an agent must choose between a set of arms to pull. Each arm has a payoff distribution that is usually unknown to the agent, and the agent wants to maximize its sum of payoffs over time. An important problem that comes up from the multi-armed bandit domain is that of exploration vs. exploitation, where the agent must decided whether to pull the arm with the best observed sample mean or pull other arms to gain more information about their distributions. The multi-armed bandit problem is a stateless action selection problem, which is a fundamental problem for reinforcement learning theory [16].

This research adopts Stone and Kraus's [15] formulation of a multi-agent version of the MAB problem. The agents share payoffs and want to maximize this shared payoff. Specifically, there are two agents: a teacher and a learner. The teacher has complete information about the arm distributions and the behavior of the learner. The learner has no prior information and estimates the arm distributions by observing the results of pulls, and greedily pulling the arm with the highest sample mean. Importantly, the teacher is embedded in the environment as a part of the team and its rewards count towards the team reward, so it cannot focus on teaching without considering what other rewards it could achieve. Stone and Kraus consider the case in which there are a finite number of pulls remaining, with undiscounted rewards. They give several interesting results for this case, but do not handle the case where there are an infinite number of pulls, which is a common formulation of the MAB problem. We address this gap, considering infinite sequences of pulls discounted by a multiplicative factor, γ. We extend the results from Section 3 of their paper to the infinite play with discounted rewards scenario.

Intuitively, γ can be seen as either an interest rate or as a chance of the problem ending. Viewing it as an interest rate, immediate rewards are more valuable as you can invest the reward and earn interest over time. On the other hand, if the episode has a chance of ending, immediate rewards are more valuable because it is uncertain whether future rewards will be received. When the number of remaining pulls is known, γ can be set to 1 because there is no uncertainty about the episode ending and the maximum reward is bounded. In the case of infinite pulls, the episode will not end, and setting $\gamma < 1$ is necessary to bound the maximum cumulative reward achievable from any state.

We consider the case with three arms, where the teacher may pull any arm, but the learner is constrained to the two arms with lower expected values. Therefore, the teacher must sacrifice some reward to show the learner a pull from a relevant arm. We will refer to these arms as Arm_*, Arm_1, and Arm_2, where Arm_* is the arm only pullable by the teacher. Let the true expected value of these arms be μ_*, μ_1, and μ_2 with $\mu_* > \mu_1 > \mu_2$ w.l.o.g. Similarly, let the observed sample means of Arm_1 and Arm_2 be $\bar{x}_1$ and $\bar{x}_2$. Note that if μ_* is not the largest, the teacher should always pull the arm with the highest expected payoff. For this paper, we assume that the teacher and learner alternate pulls and the discount factor is applied after a pair of pulls, one by the teacher and one by the learner. Furthermore, we assume that the learner follows a greedy policy, pulling the arm with the highest observed sample mean and optimistically pulling previously unseen arms.

4. THREE ARMS WITH ARBITRARY DISTRIBUTIONS

This section presents theoretical results that apply regardless of the distributions of the payoffs for the arms. For these proofs, we assume that the payoff of each arm only depends on the underlying distribution and the number of pulls of that arm, and not on time. In other words, each arm has a fixed sequence of payoffs that is only moved through when that arm is pulled.

4.1 The teacher should consider pulling Arm_1

It is sometimes beneficial for the teacher to teach, sacrificing its pull of Arm_* to pull Arm_1. We know that in any configuration, the maximum expected value achievable is $(\mu_* + \mu_1)\frac{1}{1-\gamma}$, which occurs when the teacher always pulls Arm_* and the learner always pulls Arm_1. Similarly, the minimum expected value achievable is $(\mu_2 + \mu_2)\frac{1}{1-\gamma}$. Consider the situation when $\mu_* = 10$, $\mu_1 = 9$, $\mu_2 = 5$, $\bar{x}_1 = 6$, $\bar{x}_2 = 7$, and $n_1 = n_2 = 1$. Suppose that the distribution of payoffs is known, and the probability of Arm_1 obtaining a value ≥ 8 is $\eta > \frac{1}{2}$. Therefore, if the teacher pulls Arm_1, $\bar{x}_1$ will be greater than $\bar{x}_2$ with probability η. After this pull, the teacher will play arbitrarily. Let us call this pull and the following ones S. In the worst case scenario, all remaining pulls of each agent are of Arm_2. Therefore, we know that $E[V(S)] \geq \mu_1 + \eta\mu_1 + (1-\eta)\mu_2 + \gamma(\mu_2 + \mu_2)\frac{1}{1-\gamma}$.

If the teacher instead chooses to pull Arm_*, the learner has seen only a single, low pull from Arm_1, so it will greedily pull Arm_2. Afterwards, the teacher plays arbitrarily, resulting in sequence T. The best case scenario is that remaining teacher's pulls are of Arm_*, and the learner's are of Arm_1. Then, $E[V(T)] \leq \mu_* + \mu_2 + \gamma(\mu_* + \mu_1)\frac{1}{1-\gamma}$.

By comparing these two expected values, we get that if $\gamma \leq 0.1$, $E[V(S)] > E[V(T)]$. For example, if $\eta = 0.6$ and $\gamma = 0.05$, then $E[V(S)] \geq 16.92$ and $E[V(T)] \leq 16.0$. Therefore, there are situations in which the teacher should teach, pulling Arm_1 instead of Arm_*.

4.2 If the learner is going to pull Arm_i, the teacher should not pull Arm_i

If the sample mean $\bar{x}_i$ is the highest, the learner will pull Arm_i if the teacher's pull does not change the relative values of the arms. Let a be the value obtained by pulling Arm_i. If the teacher pulls Arm_i, it will obtain a_i and then the learner will pull Arm_j, obtaining the value a_j. Afterwards, the teacher follows the optimal policy and the learner continues to play greedily with respect to the sample means, resulting in the sequence OPT. So the sequence, S, that occurs if the teacher pulls Arm_i is given in Table 1. This gives a total value of

$$V(S) = a_i + a_j + \gamma V(\mathrm{OPT})$$

n	0	1	...
Teacher	a_i		OPT
Learner		a_j	

Table 1: The sequence, S, resulting from the teacher pulling Arm_i

n	0	1	2	3	...
Teacher	a_*		a'_*		OPT
Learner		a_i		a_j	

Table 2: The sequence, T, resulting from the teacher pulling Arm_* twice instead of Arm_i

Now, consider an alternative sequence, T, where the teacher instead pulls Arm_* twice, and then follows the optimal policy. If the teacher instead pulls Arm_*, then the learner will pull Arm_i and obtain a_i. If the teacher then pulls Arm_* again, the learner will pull Arm_j and obtain a_j. Then, the optimal policy after these pulls will be the same as in sequence S as the learner has seen the same pulls of Arm_1 and Arm_2. Let us call the values obtained by pulling Arm_* a_* and a'_* respectively. Therefore, the sequence T is given in Table 2 This gives a total value of $V(T) = a_* + a_i + \gamma a'_* + \gamma a_j + \gamma^2 V(\mathrm{OPT})$.

Let us look at the expected values of these sequences: $E[V(S)]$ and $E[V(T)]$. We know that $E[a_i] = \mu_i \le \mu_1$, $E(a_j) = \mu_j \le \mu_1$, and $E(a_*) = E(a'_*) = \mu_*$. So $E[V(S)] = \mu_i + \mu_j + \gamma E[V(OPT)]$, and $E[V(T)] = \mu_* + \mu_i + \gamma\mu_* + \gamma\mu_j + \gamma^2 E[V(OPT)]$. By the definition of OPT, we know

$$\begin{aligned} E[V(\mathrm{OPT})] &\le (\mu_* + \mu_1)\frac{1}{1-\gamma} \\ (1-\gamma)E[V(\mathrm{OPT})] &\le (\mu_* + \mu_1) \end{aligned}$$

In the following calculations, for the sake of brevity, let $\mathrm{EO} = E[V(\mathrm{OPT})]$. We know that $\mu_* > \mu_i$ and $\mu_* > \mu_j$, so

$$\begin{aligned} \mu_* &> (1-\gamma)\mu_j + \gamma\mu_i \\ \mu_* + \gamma\mu_* &> (1-\gamma)\mu_j + \gamma(\mu_i + \mu_*) \\ \mu_* + \gamma\mu_* &> (1-\gamma)\mu_j + \gamma(1-\gamma)\mathrm{EO} \\ \mu_* + \gamma\mu_* + \gamma^2\mathrm{EO} &> (1-\gamma)\mu_j + \gamma\mathrm{EO} \\ \mu_* + \gamma\mu_* + \gamma\mu_j + \gamma^2\mathrm{EO} &> \mu_j + \gamma\mathrm{EO} \\ \mu_* + \mu_i + \gamma\mu_* + \gamma\mu_j + \gamma^2\mathrm{EO} &> \mu_i + \mu_j + \gamma\mathrm{EO} \\ E[V(T)] &> E[V(S)] \end{aligned}$$

The expected value of sequence T is greater than that of S. Therefore, it is desirable to follow sequence T over S, so the teacher can achieve higher reward without pulling Arm_i. This reasoning shows that pulling Arm_i is not optimal in this scenario, so the teacher should not pull Arm_i if the learner would currently pull Arm_i.

4.3 The teacher should never pull Arm_2

If $\bar{x}_2 > \bar{x}_1$, we know that the teacher should not pull Arm_2 from Section 4.2. Therefore, we only need to consider the case when $\bar{x}_1 > \bar{x}_2$.

The intuition of this proof is that the teacher can follow a policy that either 1) makes its history match up with the one achieved by pulling Arm_2 at least once or 2) if the histories do not match, the new policy is better. To this end, we use the idea of simulating another series of pulls, as do Stone and Kraus [15]. The idea is that if the teacher has seen enough pulls of Arm_1 and Arm_2, it can tell what it and the learner would have done in other situations. For example, if the teacher has seen 5 pulls of Arm_1 and 3 pulls of Arm_2, it can reason about any sequence of pulls that would have had ≤ 5 pulls of Arm_1 and ≤ 3 pulls of Arm_2. Note that pulls of Arm_* are irrelevant as they do not affect the teacher or learner because the teacher already knows the payoff distribution of Arm_* and the learner does not consider Arm_*.

Definition 1. *$S_i(n)$ is the number of pulls of Arm_i in sequence S after the first n pulls. Therefore, $S_i(n)$ of the first n pulls by the teacher and learner were of arm Arm_i.*

Definition 2. *$Sim(n)$ is the greatest round number r such that $T_1(n) \ge S_1(r)$ and $T_2(n) \ge S_2(r)$. This corresponds to the number of pulls of S that the teacher can simulate after following n pulls of T.*

Definition 3. *$T(n) = S(m)$ iff $T_1(n) = S_1(m)$ and $T_2(n) = S_2(m)$.*

Definition 4. *$T(n) > S(m)$ iff $T_1(n) \ge S_1(m)$ and $T_2(n) \ge S_2(m)$ and at least one of the inequalities is strict.*

Let us consider the sequence S that occurs from the teacher pulling Arm_2 and then acting arbitrarily. Then, let T be the sequence resulting from using the following policy:

1. If $n = 0$, $T(n) > S(\mathrm{Sim}(n))$, or $\mathrm{Sim}(n)$ is odd, choose Arm_*.
2. Else (if $T(n) = S(\mathrm{Sim}(n))$ and $\mathrm{Sim}(n)$ is even), choose the next action of S.

The idea is that the teacher should pull Arm_* until its history matches up to S, and then follow the same policy as used in S. We want to show that $E[V(S)] < E[V(T)]$. This would establish that every policy starting with Arm_2 is dominated by some other policy, so it is not optimal to pull Arm_2.

n	0	1	2	3	...
Teacher	Arm_2		Arm_1		
Learner		Arm_1		Arm_2	

Table 3: A possible sequence of pulls, S.

n	0	1	2	3	4	5	...
Teacher	Arm_*		Arm_*		Arm_1		
Learner		Arm_1		Arm_2		Arm_2	

Table 4: Another possible sequence of pulls, T.

For example, consider the sequences in Table 3 and 4. Note that $S_2(1) = 1$, $S_1(1) = 1$, $T_2(3) = 1$, and $T_1(3) = 1$. So $\text{Sim}(3) = 1$, but $\text{Sim}(2) = 0$. Therefore, for pull 4, the teacher in T will do the same thing as it would for pull 2 of S (i.e. pull Arm_1).

We know that at every point in time, if T has more pulls of Arm_* than S and fewer pulls of Arm_2 than S, it must have a higher expected value. Note that all remaining pulls in both sequences must be of Arm_1. We do not condition on the values of the pulls or on the policy of S since the requirements of the following lemmas hold in all cases. Therefore, we can consider the expected values of each arm independently. Therefore, all pulls of Arm_2 will have expected value μ_2, etc. So if these conditions hold, we know that the low pulls of Arm_2 will be more discounted in T than in S, and the high pulls of Arm_* will be less discounted in T than in S. Therefore, the $E[V(T)] > E[V(S)]$ if these conditions hold.

Now, we will describe these conditions more exactly and prove that they hold for these sequences, but first we will reason about the policy for sequence T. Note that the teacher will start by following the first part of its policy, when $n = 0$. If the teacher follows the second part of its policy, there is at least one n, call it n', such that $T(n') = S(\text{Sim}(n'))$ and $\text{Sim}(n')$ is even. Once the teacher switches to the second part of its policy, it will take the same actions as the teacher in S, and the learner will take similar actions. Therefore, after the teacher switches to the second part of its policy, $T(n)$ and $S(n)$ will increment similarly, and the teacher will remain in this part of the policy.

LEMMA 1. *$Sim(n') < n'$*

PROOF. After n' steps, there are exactly $\frac{n'}{2}$ pulls of Arm_1 and Arm_2 ($T_1(n') + T_2(n') = \frac{n'}{2}$) because all the teacher's pulls have been of Arm_* until now. But after n' steps, there are *at least* $\frac{n'}{2} + 1$ pulls of Arm_1 and Arm_2 in sequence S ($S_1(n') + S_2(n') \geq \frac{n'}{2} + 1$) because the teacher pulled Arm_2 at least once, and all the learner's actions are pulls of Arm_1 or Arm_2. Thus the simulation of S always lags behind T in the number of steps simulated: $\text{Sim}(n') < n'$. □

LEMMA 2. $\forall n > 0, T_2(n) \leq S_2(n)$.

PROOF. We will show that $T_2(n) = S_2(\text{Sim}(n))$, and from Lemma 1, $\text{Sim}(n) < n$, so $T_2(n) \leq S_2(n)$.
Case 1: $T(n) > S(\text{Sim}(n))$ or $\text{Sim}(n)$ is odd.
Proof by induction on the number of steps, i, in T.
When $i = 2$, $T_2(2) = 0$ because the teacher pulls Arm_* and the learner pulls Arm_1. The first step of S is a pull of Arm_2, so $\text{Sim}(2) = 0$ and $S_2(\text{Sim}(2)) = 0$.
Assume that $T_2(i-1) = S_2(\text{Sim}(i-1)$. Look at the next action in T; if it is a pull of Arm_* or Arm_1, then $T_2(i) = T_2(i-1)$ and $\text{Sim}(i) = \text{Sim}(i-1) \Rightarrow S_2(\text{Sim}(i)) = S_2(\text{Sim}(i-1))$. If the next action is a pull of Arm_2, then $T_2(i) = T_2(i-1) + 1$ and $S_2(\text{Sim}(i)) = S_2(\text{Sim}(i-1)) + 1$, because the new pull of Arm_2 can be used to simulate S at least one more step, but only one more pull of Arm_2 can be simulated. Therefore $T_2(i) = S_2(\text{Sim}(i))$.
Case 2: $T(n) = S(\text{Sim}(n))$ and $Sim(n)$ is even.
$T_2(n) = S_2(\text{Sim}(n))$ by the case assumptions. □

LEMMA 3. $\forall n > 0, T_*(n) > S_*(n)$.

PROOF. The proof progresses by reasoning about the possible histories that the teacher can simulate.
Case 1: $T(n) > S(\text{Sim}(n))$ or $\text{Sim}(n)$ is odd.
The teacher in T has only pulled Arm_*, and the teacher in S has pulled Arm_2 at least once, so $T_*(n) > S_*(n)$.
Case 2: $T(n) = S(\text{Sim}(n))$ and $Sim(n)$ is even.
Let n' be the first pull for which these conditions hold. At step n', the only difference between S and T is $n' - \text{Sim}(n')$ extra pulls of Arm_* in T. Afterwards, there are $n - n'$ steps in which S and T are identical, with x pulls of Arm_* in this period. The final $n' - \text{Sim}(n')$ steps of S include at least one pull of Arm_1 or Arm_2 (the learner's first action and any of its later actions). So $T_*(n) = n' - \text{Sim}(n') + x$ and $S_*(n) \leq x + n' - \text{Sim}(n') - 1$. Therefore, $T_*(n) > S_*(n)$. □

From Lemmas 1-3, we know that for all time steps, T has more pulls of Arm_* than S and fewer pulls of Arm_2 than S. Since the lemmas hold regardless of the values of the pulls, we consider the expected values of each pull independently. So the expected value of each pull is just the expected value of the arm. We know that the pulls of Arm_2 must happen later in T, so they will be more discounted. Similarly, the pulls of Arm_* will occur sooner in T, and will therefore be less discounted. Therefore, the low pulls are more discounted and the high pulls are less discounted, so $E[V(T)] > E[V(S)]$. So the teacher should never pull Arm_2.

4.4 The teacher should not teach when $n_1 = 0$ and/or $n_2 = 0$

At the beginning of a task, the learner has no experience with any of its arms, so it will explore its world optimistically, pulling each of the arms. From Section 4.2, we know that the teacher should not pull any arm that the learner is going to pull. Therefore, the teacher should not pull the arms that the learner is going to explore.

5. MORE THAN THREE ARMS

Until this point, we have focused on the case where there are three arms for the agents to pull. However, these results generalize to the case where there are many arms.

First, notice that adding additional arms that are only available to the teacher changes nothing. The teacher has complete knowledge, so it should only consider the arm with the greatest expected value. Therefore, we can continue to call this arm Arm_* and ignore these other arms.

We will focus on the case where there are arms Arm_1, Arm_2, ..., Arm_z and w.l.o.g. assume that $\mu_1 > \mu_2 > \ldots > \mu_z$. The following conclusions follow quite simply.

- It can be beneficial for the teacher to pull Arm_1 - Arm_z. Examples similar to those in Section 4.1 can be constructed for this setting.
- The teacher should not teach with Arm_i when $\bar{x}_i > \bar{x}_j, \forall j \neq i$. Similar to Section 4.2, if the agent is going to pull Arm_i, the teacher should not pull Arm_i.

- Do not teach if $\exists i$ s.t. $n_i = 0$. The same reasoning from Section 4.4 applies here, as the learner will optimistically explore its world.
- The teacher should never pull Arm_z. If we consider Arm_1–$\text{Arm}_z - 1$ as one arm with a complex distribution, its mean will still be higher than that of Arm_z. Therefore, the reasoning from Section 4.3 applies if we consider this complex arm as Arm_1 and Arm_z as Arm_2; thus, the teacher should always avoid pulling Arm_z.

We hypothesize that it can also be advantageous to teach with Arm_j for $j < k$ even when $\exists i < j$ s.t. $\bar{x}_i > \bar{x}_j$, similar to Stone and Kraus's result [15]. However, this result is left for future research.

6. RELATED WORK

The formal description of ad hoc team problems was proposed by Stone et al. [13]. This research builds on work by Stone and Kraus [15]. They introduced this formulation of a cooperative multi-armed bandit with a teacher and a learner. However, they consider the case with a known, finite number of rounds. This research extends their results into the case of infinite, discounted rewards. The trash collecting robots in our motivating example in Section 2 was taken from Stone and Kraus, who were inspired by ad hoc *human* teams such as [7].

Stone et al. [14] studied an ad hoc team setting involving cooperating with a best response teammate on a repeated normal-form game. They provide several interesting theoretical results as well proposing an efficient empirical algorithm for handling teammates with short memories. Barrett et al. [2] also investigate ad hoc teams, but in the pursuit (or predator-prey) domain. They take an empirical approach and develop an agent that plans using Monte Carlo Tree Search (MCTS) using a set of known models of possible teammates.

Other investigations of ad hoc teams include Brafman and Tennenholtz's work [5] in which one agent teaches another while engaging in a repeated joint task. However, they mainly focus on the case where teaching is not costly, and the teacher's goal is to help the learner maximize the times it chooses the best action. We consider the case where teaching has a cost, and the teacher's goal is to maximize the shared payoffs. Another domain that has been investigated is that of simulated robot soccer. Bowling and McCracken [4] investigate the effectiveness of ad hoc agents, comparing them to inoperative and absent players. Their ad hoc team agent has a playbook different from that of its teammates and tries to independently choose plays that perform well with its team.

Jones et al. [9] investigate pickup teams working in the treasure hunt domain. These teams can consist of heterogeneous robots, but they coordinate by using a communication protocol that they use to bid on desired roles. Another empirical approach is given by Knudson and Tumer [11]. However, all of their agents are adaptive and each is given a clear metric of how each of its actions affect the teams' performance.

A large body of work exists for coordinating teams of agents using standard protocols for communication and coordination such as SharedPlans [8], STEAM [17], and GPGP [6]. Our work does not assume that such a protocol is known by all the agents.

The multi-armed bandit problem has been studied extensively [3], and several variations have been considered in which there are multiple agents that can observe the actions or outcomes of each other. Keller and Rady [10] investigate a two-armed bandit with multiple players cooperating. In this scenario, there is a risky arm that distributes lump-sum payoffs according to a Poisson distribution. The agents share a common cut-off for their belief about the expected reward of the risky arm and either all pull the risky arm or all choose the other arm. Aoyagi [1] focuses on a two-armed bandit problem with multiple players that can only observe the actions on other players rather than the outcome of these actions. Under some restrictions of the arms' payoff distributions, he proves that all players will settle on the same arm. Our research indicates that learning from other agents is possible without explicit communication.

7. CONCLUSIONS AND FUTURE WORK

This paper presents an extension of theory to the cooperative multi-armed bandit problem with infinite, discounted rewards. We have studied in detail the case where a teacher knows the true payoff distribution of all of the arms, and, while embedded in the domain, it interacts with a teammate that lacks this information. In this setting, teaching has a cost, and we give insight into the trade-off between teaching and exploitation. We show that teaching can be advantageous, but that there are some guidelines that the teacher should follow, such as not teaching by pulling the worst arm.

This paper opens up several avenues for future research. It motivates research into stateful, infinite reward problems, such as those commonly faced in reinforcement learning. In addition, it spurs research into the trade-offs between teaching, exploration, and exploitation. Furthermore, more research into teammates with more information and the possibility of limited communication is needed. From a high level, we view these results as a small step towards the long-term goal of fully general and robust ad hoc team agents.

Acknowledgments

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by grants from the National Science Foundation (IIS-0917122), ONR (N00014-09-1-0658), and the Federal Highway Administration (DTFH61-07-H-00030). Samuel Barrett is supported by a NDSEG fellowship.

8. REFERENCES

[1] M. Aoyagi. Mutual observability and the convergence of actions in a multi-person two-armed bandit model. *Journal of Economic Theory*, 82:405–424, 1998.

[2] S. Barrett, P. Stone, and S. Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *AAMAS '11*, May 2010. To appear.

[3] D. Bergemann and J. Valimaki. Bandit problems. Technical report, Cowles Foundation Discussion Paper, 2006.

[4] M. Bowling and P. McCracken. Coordination and adaptation in impromptu teams. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI)*, pages 53–58, 2005.

[5] R. I. Brafman and M. Tennenholtz. On partially controlled multi-agent systems. *JAIR*, 4:477–507, 1996.

[6] K. S. Decker and V. R. Lesser. Designing a family of coordination algorithms. In *ICMAS '95*, pages 73–80, June 1995.

[7] J. A. Giampapa, K. Sycara, and G. Sukthankar. Toward identifying process models in ad hoc and distributed teams. In K. V. Hindriks and W.-P. Brinkman, editors, *Proceedings of the First International Working Conference on Human Factors and Computational Models in Negotiation (HuCom 2008)*, pages 55–62, Mekelweg 4, 2628 CD Delft, The Netherlands, December 2008. Delft University of Technology.

[8] B. Grosz and S. Kraus. Collaborative plans for complex group actions. *Artificial Intelligence*, 86:269–368, 1996.

[9] E. Jones, B. Browning, M. B. Dias, B. Argall, M. M. Veloso, and A. T. Stentz. Dynamically formed heterogeneous robot teams performing tightly-coordinated tasks. In *International Conference on Robotics and Automation*, pages 570 – 575, May 2006.

[10] G. Keller and S. Rady. Strategic experimentation with poisson bandits. Technical report, Free University of Berlin, Humboldt University of Berlin, University of Bonn, University of Mannheim, University of Munich, 2009. Discussion Papers 260.

[11] M. Knudson and K. Tumer. Robot coordination with ad-hoc team formation. In *AAMAS '10*, pages 1441–1442, 2010.

[12] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

[13] P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI '10*, July 2010.

[14] P. Stone, G. A. Kaminka, and J. S. Rosenschein. Leading a best-response teammate in an ad hoc team. In *Agent-Mediated Electronic Commerce: Designing Trading Strategies and Mechanisms for Electronic Markets.* November 2010.

[15] P. Stone and S. Kraus. To teach or not to teach? Decision making under uncertainty in ad hoc teams. In *AAMAS '10*, May 2010.

[16] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, USA, 1998.

[17] M. Tambe. Towards flexible teamwork. *JAIR*, 7:81–124, 1997.