Multi-agent active perception with prediction rewards

Mikko Lauri
Department of Computer Science
Universität Hamburg
Hamburg, Germany
lauri@informatik.uni-hamburg.de

Frans A. Oliehoek
Department of Computer Science
TU Delft
Delft, the Netherlands
f.a.oliehoek@tudelft.nl

Abstract

Multi-agent active perception is a task where a team of agents cooperatively gathers observations to compute a joint estimate of a hidden variable. The task is decentralized and the joint estimate can only be computed after the task ends by fusing observations of all agents. The objective is to maximize the accuracy of the estimate. The accuracy is quantified by a centralized prediction reward determined by a centralized decision-maker who perceives the observations gathered by all agents after the task ends. In this paper, we model multi-agent active perception as a decentralized partially observable Markov decision process (Dec-POMDP) with a convex centralized prediction reward. We prove that by introducing individual prediction actions for each agent, the problem is converted into a standard Dec-POMDP with a decentralized prediction reward. The loss due to decentralization is bounded, and we give a sufficient condition for when it is zero. Our results allow application of any Dec-POMDP solution algorithm to multi-agent active perception problems, and enable planning to reduce uncertainty without explicit computation of joint estimates. We demonstrate the empirical usefulness of our results by applying a standard Dec-POMDP algorithm to multi-agent active perception problems, showing increased scalability in the planning horizon.

1 Introduction

Active perception, collecting observations to reduce uncertainty about a hidden variable, is one of the fundamental capabilities of an intelligent agent [2]. In *multi-agent active perception* a team of autonomous agents cooperatively gathers observations to infer the value of a hidden variable. Application domains include search and rescue robotics, sensor networks, and distributed hypothesis testing. A multi-agent active perception task often has a finite duration: after observations have been gathered, they are collected to a central database for inference. While the inference phase is centralized, the observation gathering phase is decentralized: each agent acts independently, without knowing the observations collected by the other agents nor guaranteed communication to the other agents.

The key problem in multi-agent active perception is to determine how each agent should act during the decentralized phase to maximize the informativeness of the collected observations, evaluated afterwards during the centralized inference phase. The problem can be formalized as a decentralized partially observable Markov decision process (Dec-POMDP) [3, 15], a general model of sequential multi-agent decision-making under uncertainty. At each time step in a Dec-POMDP, each agent in the team takes an individual action. The next state is determined according to a Markov chain conditional on the current hidden state and all individual actions. Each agent then perceives an individual observation correlated with the next state and the individual actions. The agents should act so as to maximize the expected sum of shared rewards, accumulated at each time step over a finite horizon.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In the decentralized phase, the per-step reward depends on the hidden state and the individual actions of all agents. In typical Dec-POMDPs, the reward on all time steps is of this form. Active perception problems are modelled by a reward that is a convex function of the team's joint estimate of the hidden state, for example the negative entropy [11]. This encourages agents to act in ways that lead to joint state estimates with low uncertainty. In analogy to the centralized inference phase, the reward at the final time step can be thought of as a *centralized prediction reward* where a centralized decision-maker perceives the individual action-observation histories of all agents and determines a reward based on the corresponding joint state estimate. Due to the choice of reward function, algorithms for standard Dec-POMDPs are not applicable to such active perception problems. Despite the pooling of observations after the end of the task, the problem we target is decentralized. We design a strategy for each agent to act independently during the observation gathering phase, without knowing for sure how the others acted or what they perceived. Consequently, the joint state estimate is not available to any agent during the observation gathering phase. Strategies executable in a centralized manner are available only if all-to-all communication during task execution is possible, which we do not assume. As decentralized strategies are a strict subset of centralized strategies, the best decentralized strategy is at most as good as the best centralized strategy [16].

In this paper, we show that the convex centralized prediction reward can be converted to a *decentralized prediction reward* that is a function of the hidden state and so-called *individual prediction actions*. This converts the Dec-POMDP with a convex centralized prediction reward into a standard Dec-POMDP with rewards that depend on the hidden state and actions only. This enables solving multi-agent active perception problems without explicit computation of joint state estimates applying *any* standard Dec-POMDP algorithm. We show that the error induced when converting the centralized prediction reward into a decentralized one, the *loss due to decentralization*, is bounded. We also give a sufficient condition for when the loss is zero, meaning that the problems with the centralized, respectively decentralized, prediction rewards are equivalent. We prove the empirical usefulness of our results by applying standard Dec-POMDP solution algorithms to active perception problems, demonstrating improved scalability over the state-of-the-art.

The remainder of the paper is organized as follows. We review related work in Section 2, and give preliminary definitions for Dec-POMDPs in Section 3. In Section 4, we introduce our proposed conversion of a centralized prediction reward to a decentralized prediction reward. We propose a method to apply standard Dec-POMDP algorithms to multi-agent active perception in Section 5, and empirically evaluate the method in Section 6. Section 7 concludes the paper.

2 Related work

We briefly review possible formulations of multi-agent active perception problems, and then focus on the Dec-POMDP model that provides the most general formulation.

Multi-agent active perception has been formulated as a distributed constraint optimization problem (DCOP), submodular maximization, or as a specialized variant of a partially observable Markov decision process (POMDP). Probabilistic DCOPs with partial agent knowledge have been applied to signal source localization [9, 26]. DCOPs with Markovian dynamics have been proposed for target tracking by multiple sensors [13]. DCOPs are a simpler model than Dec-POMDPs, as a fixed communication structure is assumed or the noise in the sensing process is not modelled. Submodular maximization approaches assume the agents' reward can be stated as a submodular set function, and apply distributed greedy maximization to obtain an approximate solution [22, 8, 7]. Along with the structure of the reward function, inter-agent communication is typically assumed. Specialized variants of POMDPs may also be applied. If all-to-all communication without delay during task execution is available, centralized control is possible and the problem can be solved as a multi-agent POMDP [24]. Auctioning of POMDP policies can facilitate multi-agent cooperation when agents can communicate [6]. Best et al. [4] propose a decentralized Monte Carlo tree search planner where agents periodically communicate their open-loop plans to each other.

Multi-agent active perception without implicit communication with uncertainty on state transitions and the agents' perception may be modelled as a Dec-ρPOMDP [10, 11]. In contrast to standard Dec-POMDPs, the reward function in a Dec-ρPOMDP is a convex function of the joint state estimate, for example the negative entropy. Unfortunately, because of the convex reward, standard Dec-POMDP solution algorithms are not applicable. Furthermore, the heuristic algorithm proposed in [11, 12] requires explicit computation of reachable joint state estimates. Computing and storing these joint state estimates adds significant memory overhead to the already high computational burden of solving Dec-POMDPs [3]. We address both shortcomings, showing that a Dec-ρPOMDP can be converted to a standard Dec-POMDP with linear rewards. This in turn enables applying standard planning algorithms that do not require computation of reachable joint state estimates, leading to improved scalability.

Our work in the decentralized setting draws inspiration from approaches using prediction rewards in single-agent and centralized scenarios. Araya-López et al. [1] proposed to tackle single-agent active perception as a ρPOMDP with a convex reward. The related POMDP with information rewards (POMDP-IR) was proposed in [25]. The POMDP-IR model adds prediction actions that the agent selects in addition to the usual actions. Active perception is facilitated by rewarding the agent for correctly predicting the true underlying state. The equivalence of ρPOMDP and POMDP-IR model was later established [20]. Recently, Satsangi et al. [21] propose a reinforcement learning method to solve ρPOMDPs taking advantage of the equivalence. In this paper we prove an analogous equivalence for Dec-POMDPs, by converting a Dec-ρPOMDP to a standard Dec-POMDP with individual prediction actions and a decentralized prediction reward. Unlike in the POMDP case, the conversion does not always result in a perfect equivalence, but is associated with a loss due to decentralization.

3 Multi-agent active perception as a Dec-POMDP

In this section, we review how active perception problems are modelled as Dec-POMDPs. We also review plan-time sufficient statistics, which allow us to concisely express joint state estimates and value functions of a Dec-POMDP. We concentrate on the practically relevant active perception problem where a prediction reward at the final time step depends on the joint state estimate.

3.1 Decentralized POMDPs

We define a Dec-POMDP where the action and observation spaces along with the reward function are time-dependent, as this will be convenient for our subsequent introduction of prediction actions.

Definition 1. *A Dec-POMDP is a tuple* $\mathcal{M} = \langle h, I, S, b_0, \mathcal{A}, \mathcal{Z}, T, \mathcal{R}\rangle$*, where*

- $h \in \mathbb{N}$ *is the time horizon of the problem,*
- $I = \{1, 2, \ldots, n\}$ *is a set of* n *agents,*
- S *is the finite set of states* s*,*
- $b_0 \in \Delta(S)$ *is the initial state distribution at time step* $t = 0$*,*
- $\mathcal{A}$ *is the collection of individual action spaces* $A_{i,t}$ *for each agent* $i \in I$ *and time step* $t = 0, \ldots, h-1$*. The tuple* $a_t = \langle a_{1,t}, a_{2,t}, \ldots, a_{n,t}\rangle$ *of individual actions is called the joint action at time step* t*,*
- $\mathcal{Z}$ *is the collection of individual observation spaces* $Z_{i,t}$ *for each agent* $i \in I$ *and time step* $t = 1, \ldots, h$*. The tuple* $z_t = \langle z_{1,t}, z_{2,t}, \ldots, z_{n,t}\rangle$ *of individual observations is called the joint observation at time step* t*,*
- T *is the dynamics function specifying conditional probabilities* $\mathbb{P}(z_{t+1}, s_{t+1} \mid s_t, a_t)$*, and*
- $\mathcal{R}$ *is the collection of reward functions* $R_t(s_t, a_t)$ *for time steps* $t = 0, \ldots, h-1$*.*

An admissible solution of a Dec-POMDP is a decentralized joint policy π, i.e., a tuple $\langle \pi_1, \ldots, \pi_n\rangle$ where the individual policy π_i of each agent i maps individual observation sequences $\vec{z}_{i,t} = (z_{i,1}, \ldots, z_{i,t})$ to an individual action.[1] An individual policy is a sequence $\pi_i = (\delta_{i,0}, \ldots, \delta_{i,h-1})$ of individual decision rules that map length-t individual observation sequences to an individual action $\delta_{i,t}(\vec{z}_{i,t}) = a_{i,t}$. A joint decision rule is a tuple $\delta_t = \langle \delta_{1,t}, \ldots, \delta_{n,t}\rangle$ that maps a length-t joint observation sequence $\vec{z}_t = \langle \vec{z}_{1,t}, \ldots, \vec{z}_{n,t}\rangle$ to a joint action $\delta_t(\vec{z}_t) = a_t$. We shall use notation $\vec{z}_{-i,t}$ to denote the individual observation sequences of all agents *except* i.

[1] $\vec{z}_{i,0} = \emptyset$ as there is no observation at time $t = 0$.

The objective is to find an optimal joint policy π^* that maximizes the expected sum of rewards, that is, $\pi^* = \operatorname{argmax}_\pi \mathbb{E}\left[\sum_{t=0}^{h-1} R_t(s_t, \delta_t(\vec{z}_t))\right]$, where the expectation is with respect to the distribution of states and joint observation sequences induced under π, i.e., $\mathbb{P}(s_0, \ldots, s_h, \vec{z}_h \mid b_0, \pi) \triangleq b_0(s_0) \prod_{t=0}^{h-1} T(z_{t+1}, s_{t+1} \mid s_t, \delta_t(\vec{z}_t))$.

In this paper, we are interested in cooperative active perception. We assume that after the task terminates at time step h, the joint observation sequence $\vec{z}_h$ is used to compute the conditional distribution $b_h \in \Delta(S)$ over the final state s_h, that is, $b_h(s_h) \triangleq \mathbb{P}(s_h \mid \vec{z}_h, b_0, \pi)$. We seek policies that, instead of only maximizing the expected sum of rewards $R_t(s_t, a_t)$, also maximize the informativeness of b_h.

The Dec-ρPOMDP [11] models active perception by maximizing a convex function of b_h, e.g., the negative entropy. This convex function may be though of as a *centralized prediction reward* that is independent of the agents' individual actions. Conceptually, the centralized prediction reward is determined by a virtual centralized decision-maker that perceives the agents' individual action and observation sequences at time h, computes b_h, and then determines the final reward. We approximate the centralized prediction reward using a piecewise linear and convex function as illustrated in Figure 1. Consider a bounded, convex, and differentiable function $f\colon \Delta(S) \to \mathbb{R}$. The tangent hyperplane $\alpha \in \mathbb{R}^{|S|}$ of f at $b \in \Delta(S)$ is $\alpha = \nabla f(b) - f^*(\nabla f(b))$, where ∇ denotes the gradient and f^* is the convex conjugate of f [5]. We select a finite set of linearization points $b_j \in \Delta(S)$, and define Γ as the set of corresponding tangent hyperplanes α_j.[2] Then, we obtain a lower approximation as $f(b) \geq \max_{\alpha\in\Gamma} \sum_s b(s)\alpha(s)$. This approximation is also used in ρPOMDPs, and has a bounded error [1]. The approximation is also used in [21], where error bounds for cases such as 0-1 rewards are provided. The Dec-ρPOMDP problem we consider is as follows.

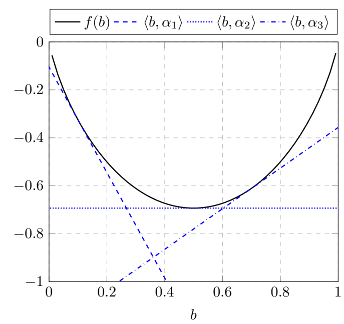

Figure 1: A convex function f is approximated by $\max_i \langle b, \alpha_i \rangle$ where α_i are tangent hyperplanes and $\langle \cdot, \cdot \rangle$ is the inner product. The horizontal axis depicts $b(s)$ in a system with two states s and s', where $b(s') = 1 - b(s)$. A prediction action corresponds to selecting a single α_i.

Definition 2 (Dec-ρPOMDP). *A Dec-ρPOMDP is a pair $\langle \mathcal{M}, \Gamma \rangle$, where $\mathcal{M}$ is a Dec-POMDP and Γ is a set of tangent hyperplanes that determine the centralized prediction reward $\rho\colon \Delta(S) \to \mathbb{R}$ defined as* $\rho(b) \triangleq \max\limits_{\alpha\in\Gamma} \sum\limits_s b(s)\alpha(s)$.

The standard Dec-POMDP is the special case where Γ contains a single all-zero element. The objective in the Dec-ρPOMDP is to find an optimal joint policy that maximizes the expected sum of rewards and the centralized prediction reward, i.e., $\mathbb{E}\left[\sum_{t=0}^{h-1} R_t(s_t, \delta_t(\vec{z}_t)) + \rho(b_h)\right]$.

3.2 Sufficient plan-time statistics and the optimal value function

Sufficient plan-time statistics are probability distributions over states and joint observation sequences given the past joint policy followed by the agents [14]. A past joint policy at time t is a joint policy specified until time t, denoted $\varphi_t = \langle \varphi_{1,t}, \ldots, \varphi_{n,t} \rangle$, where each individual past policy is a sequence of individual decision rules: $\varphi_{i,t} = (\delta_{i,0}, \ldots, \delta_{i,t-1})$. The sufficient plan-time statistic for initial state distribution b_0 and a past joint policy φ_t is defined as

$$\sigma_t(s_t, \vec{z}_t) \triangleq \mathbb{P}(s_t, \vec{z}_t \mid b_0, \varphi_t), \tag{1}$$

and at the starting time, $\sigma_0(s_0) \triangleq b_0(s_0)$. The conditional $\sigma_t(\cdot \mid \vec{z}_t)$ is the state distribution after perceiving the joint observation sequence $\vec{z}_t$ when executing policy φ_t with initial state distribution b_0. The marginal $\sigma_t(\vec{z}_t)$ is the probability of the joint observation sequence $\vec{z}_t$ given φ_t and b_0.

Sufficient plan-time statistics are updateable to any extension of the past joint policy φ_t. The extension of an individual past policy $\varphi_{i,t}$ by an individual decision rule $\delta_{i,t}$ is defined $\varphi_{i,t+1} = \varphi_{i,t} \circ \delta_{i,t} \triangleq$

[2] We defer discussion on how we select the linearization points to Section 5.

$(\delta_{i,0}, \ldots, \delta_{i,t-1}, \delta_{i,t})$. The extension of a past joint policy φ_t by $\delta_t = \langle \delta_{1,t}, \ldots, \delta_{n,t} \rangle$ is defined $\varphi_{t+1} = \varphi_t \circ \delta_t \triangleq \langle \varphi_{1,t+1}, \ldots, \varphi_{n,t+1} \rangle$. Let $\vec{z}_{t+1} = (\vec{z}_t, z_{t+1})$ be a joint observation sequence that extends $\vec{z}_t$ with z_{t+1}. Given the statistic σ_t for φ_t and the next joint decision rule δ_t, the updated statistic for $\varphi_{t+1} = \varphi_t \circ \delta_t$ is $\sigma_{t+1} = U_{ss}(\sigma_t, \delta_t)$, where the update operator U_{ss} is defined as

$$[U_{ss}(\sigma_t, \delta_t)](s_{t+1}, \vec{z}_{t+1}) \triangleq \sum_{s_t} T(z_{t+1}, s_{t+1} \mid s_t, \delta_t(\vec{z}_t))\sigma_t(s_t, \vec{z}_t). \tag{2}$$

The statistic σ_t is sufficient to predict the expected immediate reward $\hat{R}_t(\sigma_t, \delta_t)$ of choosing the next joint decision rule δ_t:

$$\hat{R}_t(\sigma_t, \delta_t) = \sum_{\vec{z}_t, s_t} \sigma_t(s_t, \vec{z}_t) R_t(s_t, \delta_t(\vec{z}_t)). \tag{3}$$

Furthermore, σ_h is sufficient to predict the expected centralized prediction reward $\hat{\rho}(\sigma_h)$:

$$\hat{\rho}(\sigma_h) = \sum_{\vec{z}_h} \sigma_h(\vec{z}_h) \max_{\alpha \in \Gamma} \sum_{s_h} \sigma_h(s_h \mid \vec{z}_h)\alpha(s_h). \tag{4}$$

In contrast to the reward in earlier stages, $\hat{\rho}$ does not depend on decision rules at all. We interpret $\hat{\rho}$ as the expectation of the prediction reward of a centralized decision-maker who selects the best $\alpha \in \Gamma$ after perceiving the full history of joint actions and observations at task termination.

Let π be a joint policy consisting of the joint decision rules $\delta_0, \delta_1, \ldots, \delta_{h-1}$. The value-to-go of π starting from a sufficient plan-time statistic σ_t at time step t is

$$Q_t^{\pi}(\sigma_t, \delta_t) = \begin{cases} \hat{R}_t(\sigma_t, \delta_t) + \hat{\rho}(\sigma_{t+1}) & \text{if } t = h-1, \\ \hat{R}_t(\sigma_t, \delta_t) + Q_{t+1}^{\pi}(\sigma_{t+1}, \delta_{t+1}) & \text{if } 0 \leq t < h-1, \end{cases} \tag{5}$$

with the shorthand $\sigma_{t+1} = U_{ss}(\sigma_t, \delta_t)$. The value function V^{π} of a joint policy π is defined as the sum of expected rewards when the agents act according to π, $V^{\pi}(\sigma_0) \triangleq Q_0^{\pi}(\sigma_0, \delta_0)$. The value function of an optimal policy π^* is denoted V^*, and it satisfies $V^*(\sigma_0) \geq V^{\pi}(\sigma_0)$ for all π. We write $V_{\mathcal{M}}^{\pi}$ and $V_{\langle \mathcal{M}, \Gamma \rangle}^{\pi}$ for the value function in a standard Dec-POMDP and a Dec-ρPOMDP, respectively.

4 Conversion to standard Dec-POMDP

In this section, we show that any Dec-ρPOMDP can be converted to a standard Dec-POMDP by adding *individual prediction actions* for each agent, and by introducing a *decentralized prediction reward*. Each individual prediction action corresponds to a selection of a tangent hyperplane of the centralized prediction reward, as illustrated in Figure 1. As the individual prediction actions are chosen in a decentralized manner, the decentralized prediction reward never exceeds the centralized prediction reward. Consequently, we prove that an optimal policy of the standard Dec-POMDP may be applied to the Dec-ρPOMDP with bounded error compared to the true optimal policy. We give a sufficient condition for when this loss due to decentralization is zero and a Dec-ρPOMDP is equivalent to a standard Dec-POMDP.

A major implication of our results is that it is possible to apply any Dec-POMDP solver to a Dec-ρPOMDP problem. We approximate a centralized prediction reward by a decentralized prediction reward that only depends on states and actions. This allows planning for active perception problems without explicit computation of joint state estimates.

We first introduce our proposed conversion and prove its properties, including the error bound. We conclude the section by giving a sufficient condition for when the two problems are equivalent. All omitted proofs are found in the supplementary material.

Definition 3. *Given a Dec-ρPOMDP $\langle \mathcal{M}, \Gamma \rangle$ with $\mathcal{M} = \langle h, I, S, b_0, \mathcal{A}, \mathcal{Z}, T, \mathcal{R} \rangle$, convert it to a standard Dec-POMDP $\mathcal{M}^+ = \langle h+1, I, S, b_0, \mathcal{A}^+, \mathcal{Z}^+, T^+, \mathcal{R}^+ \rangle$ where the horizon is incremented by one and*

- *in $\mathcal{A}^+$, the individual action space $A_{i,h}$ for each agent $i \in I$ at time h is a set of* ***individual prediction actions*** *$a_{i,h}$, with one individual prediction action for each tangent hyperplane $\alpha \in \Gamma$; for other time steps $A_{i,t}$ are as in $\mathcal{A}$,*

- *in $\mathcal{Z}^+$, the individual observation space $Z_{i,h+1}$ for each agent $i \in I$ at time $h+1$ contains a single null observation; for other time steps $Z_{i,t}$ are as in $\mathcal{Z}$,*
- *$T^+(z_{h+1}, s_{h+1} \mid s_h, a_h)$ has probability of one for the joint null observation and $s_{h+1} = s_h$, and is otherwise zero; for other time steps T^+ is equal to T, and*
- *in $\mathcal{R}^+$, the reward function R_h at time step h is a linear combination of **individual prediction rewards** $R_{i,h}$ of each agent, such that for a joint prediction action $a_h = \langle a_{1,h}, \ldots, a_{n,h}\rangle$ the **decentralized prediction reward** is*

$$R_h(s_h, a_h) = \frac{1}{n}\sum_{i=1}^{n} R_{i,h}(s_h, a_{i,h}), \tag{6}$$

with $R_{i,h}(s_h, a_{i,h}) = \alpha_{a_{i,h}}(s_h)$, where $\alpha_{a_{i,h}} \in \Gamma$ is the tangent hyperplane corresponding to the individual prediction action $a_{i,h}$; for other time steps R_t are as in $\mathcal{R}$.

As the action and observation sets in $\mathcal{M}$ and $\mathcal{M}^+$ are the same until $t = h$, past joint policies are interchangeable, and plan-time sufficient statistics are identical.

Lemma 1. *Let $\langle \mathcal{M}, \Gamma\rangle$ be a Dec-ρPOMDP and define $\mathcal{M}^+$ as above. Then, for any past joint policy φ_t with $t \leq h$, the respective plan-time sufficient statistics σ_t in $\langle \mathcal{M}, \Gamma\rangle$ and σ_t^+ in $\mathcal{M}^+$ are equivalent, i.e., $\sigma_t \equiv \sigma_t^+$.*

In $\mathcal{M}^+$ the horizon is incremented, so that each agent takes one more action than in $\langle \mathcal{M}, \Gamma\rangle$. The additional action is one of the newly added individual prediction actions at time step h. To select an individual prediction action, an agent may use 1) individual information, that is, the agent's individual observation sequence, and 2) plan-time information common to all agents, that is, the plan-time sufficient statistic. An *individual prediction rule* ϕ_i maps the individual observation sequence $\vec{z}_{i,h}$ and the plan-time sufficient statistic σ_h to an individual prediction action. A *joint prediction rule* $\phi = \langle \phi_1, \ldots, \phi_n\rangle$ is a tuple of individual prediction rules, and maps a joint observation sequence $\vec{z}_h$ and σ_h to a joint prediction action. The *expected decentralized prediction reward* given σ_h and ϕ is defined analogously to Eq. (3) as $\hat{R}_h(\sigma_h, \phi) \triangleq \sum_{\vec{z}_h, s_h} \sigma_h(s_h, \vec{z}_h) R_h(s_h, \phi(\vec{z}_h, \sigma_h))$.

An optimal joint prediction rule maximizes the expected decentralized prediction reward. We prove that an optimal joint prediction rule consists of individual prediction rules that maximize the expected individual prediction reward. We then show that the expected decentralized prediction reward is at most equal to the centralized prediction reward, and that a similar relation holds between value functions in the respective Dec-POMDP and Dec-ρPOMDP. The key property required is that the decentralized prediction reward is a sum of individual prediction rewards.

Lemma 2 (Optimal joint prediction rule)**.** *Let $\langle \mathcal{M}, \Gamma\rangle$ be a Dec-ρPOMDP and define $\mathcal{M}^+$ as above, and let σ_h be a plan-time sufficient statistic for any past joint policy. Then the joint prediction rule $\phi^* = \langle \phi_1^*, \ldots, \phi_n^*\rangle$ where each individual prediction rule ϕ_i^* is defined as*

$$\phi_i^*(\vec{z}_{i,h}, \sigma_h) \triangleq \operatorname*{argmax}_{a_{i,h} \in A_{i,h}} \sum_{\vec{z}_{-i,h}, s_h} \sigma_h(s_h, \vec{z}_{-i,h} \mid \vec{z}_{i,h}) R_{i,h}(s_h, a_{i,h}) \tag{7}$$

maximizes expected decentralized prediction reward, that is, $\hat{R}_h(\sigma_h, \phi^*) = \max_\phi \hat{R}_h(\sigma_h, \phi)$.

In the Dec-ρPOMDP $\langle \mathcal{M}, \Gamma\rangle$, the centralized prediction reward is determined by a centralized virtual agent that has access to the observation histories of all agents. In our converted standard Dec-POMDP $\mathcal{M}^+$, each agent individually takes a prediction action, leading to the decentralized prediction reward. The decentralized prediction reward never exceeds the centralized prediction reward as we show next.

Lemma 3. *The expected decentralized prediction reward $\hat{R}_h(\sigma_h, \phi^*)$ in $\mathcal{M}^+$ is at most equal to the expected centralized prediction reward $\hat{\rho}(\sigma_h)$ in $\langle \mathcal{M}, \Gamma\rangle$, i.e., $\hat{R}_h(\sigma_h, \phi^*) \leq \hat{\rho}(\sigma_h)$.*

Lemma 4. *Let $\langle \mathcal{M}, \Gamma\rangle$ and $\mathcal{M}^+$ be as defined above. Let φ_h be a past joint policy for $\mathcal{M}^+$, and let ϕ^* be the optimal joint prediction rule. Then, the value of $\varphi_h \circ \phi^*$ in $\mathcal{M}^+$ is at most equal to the value of φ_h in $\langle \mathcal{M}, \Gamma\rangle$, i.e., $V_{\mathcal{M}^+}^{\varphi_h \circ \phi^*}(\sigma_0) \leq V_{\langle \mathcal{M}, \Gamma\rangle}^{\varphi_h}(\sigma_0)$.*

We now show that the difference between the optimal values of the standard Dec-POMDP $\mathcal{M}^+$ and the Dec-ρPOMDP $\langle \mathcal{M}, \Gamma\rangle$ is bounded. We call the error the *loss due to decentralization*.

Theorem 1 (Loss due to decentralization). *Consider a Dec-ρPOMDP $\langle \mathcal{M}, \Gamma \rangle$ with the optimal value function $V^*_{\langle \mathcal{M}, \Gamma \rangle}$. Let π be an optimal policy for the standard Dec-POMDP $\mathcal{M}^+$ created as in Definition 3, and denote by φ_h the past joint policy consisting of the first h decision rules of π. Then the difference of $V^*_{\langle \mathcal{M}, \Gamma \rangle}$ and the value function $V^{\varphi_h}_{\langle \mathcal{M}, \Gamma \rangle}$ of applying φ_h to $\langle \mathcal{M}, \Gamma \rangle$ is bounded by*

$$|V^*_{\langle \mathcal{M}, \Gamma \rangle}(\sigma_0) - V^{\varphi_h}_{\langle \mathcal{M}, \Gamma \rangle}(\sigma_0)| \leq 2 \max_{\sigma_h} |\hat{\rho}(\sigma_h) - \hat{R}_h(\sigma_h, \phi^*)|, \tag{8}$$

where ϕ^* is the optimal joint prediction rule.

Proof. For clarity, in the following we omit the argument σ_0 from the value functions. Suppose the optimal policy π^* in $\langle \mathcal{M}, \Gamma \rangle$ consists of the joint decision rules $\delta^*_0, \ldots, \delta^*_{h-1}$. For $\mathcal{M}^+$, define the partial joint policy $\varphi^*_h = (\delta^*_0, \ldots, \delta^*_{h-1})$ and its extension $\varphi^*_h \circ \phi^*$. By Lemma 2, an optimal joint policy π of $\mathcal{M}^+$ is an extension of some past joint policy φ_h by ϕ^*, i.e., $\pi = \varphi_h \circ \phi^*$. Because $\varphi_h \circ \phi^*$ is optimal in $\mathcal{M}^+$, we have that $V^{\varphi^*_h \circ \phi^*}_{\mathcal{M}^+} \leq V^{\varphi_h \circ \phi^*}_{\mathcal{M}^+}$. By Lemma 4, $V^{\varphi_h \circ \phi^*}_{\mathcal{M}^+} \leq V^{\varphi_h}_{\langle \mathcal{M}, \Gamma \rangle}$. Finally, because π^* is optimal in $\langle \mathcal{M}, \Gamma \rangle$, we have $V^{\varphi_h}_{\langle \mathcal{M}, \Gamma \rangle} \leq V^*_{\langle \mathcal{M}, \Gamma \rangle}$. Now,

$$\begin{aligned}
|V^*_{\langle \mathcal{M}, \Gamma \rangle} - V^{\varphi_h}_{\mathcal{M}^+}| &\leq |V^*_{\langle \mathcal{M}, \Gamma \rangle} - V^{\varphi^*_h \circ \phi^*}_{\mathcal{M}^+}| + |V^{\varphi^*_h \circ \phi^*}_{\mathcal{M}^+} - V^{\varphi_h}_{\mathcal{M}^+}| && \text{(9a)} \\
&\leq |V^*_{\langle \mathcal{M}, \Gamma \rangle} - V^{\varphi^*_h \circ \phi^*}_{\mathcal{M}^+}| + |V^{\varphi^*_h \circ \phi^*}_{\mathcal{M}^+} - V^*_{\langle \mathcal{M}, \Gamma \rangle}| = 2 \cdot |V^*_{\langle \mathcal{M}, \Gamma \rangle} - V^{\varphi^*_h \circ \phi^*}_{\mathcal{M}^+}| && \text{(9b)} \\
&= 2 \cdot \left| \left(\sum_{t=0}^{h-1} \hat{R}_t(\sigma^*_t, \delta^*_t) + \hat{\rho}(\sigma^*_h) \right) - \left(\sum_{t=0}^{h-1} \hat{R}_t(\sigma^*_t, \delta^*_t) + \hat{R}_h(\sigma^*_h, \phi^*) \right) \right| && \text{(9c)} \\
&= 2 \cdot |\hat{\rho}(\sigma^*_h) - \hat{R}_h(\sigma^*_h, \phi^*)| \leq 2 \cdot \max_{\sigma_h} |\hat{\rho}(\sigma_h) - \hat{R}_h(\sigma_h, \phi^*)|. && \text{(9d)}
\end{aligned}$$

We first apply the triangle inequality. The second inequality holds as $V^{\varphi^*_h \circ \phi^*}_{\mathcal{M}^+} \leq V^{\varphi_h}_{\langle \mathcal{M}, \Gamma \rangle} \leq V^*_{\langle \mathcal{M}, \Gamma \rangle}$. The next equality is by symmetry of absolute difference. The third line follows from the definition of the value function and Lemma 1. Note that we refer by σ^*_t to the sufficient plan-time statistics reached under partial joint policies φ^*_t of π^*. The final inequality follows by maximizing the difference of the expected centralized and decentralized prediction rewards. □

The theorem shows that given a Dec-ρPOMDP $\langle \mathcal{M}, \Gamma \rangle$, we may solve the standard Dec-POMDP $\mathcal{M}^+$ and apply its optimal policy to the Dec-ρPOMDP with bounded error. If there is only a single agent, that agent is equivalent to the conceptualized centralized agent, which implies $\hat{\rho} \equiv \hat{R}_h$ and that the loss is zero. The equivalence of a ρPOMDP with a convex prediction reward and a POMDP with information rewards first shown in [20] is therefore obtained as a special case.

The proof indicates that the error is at most twice the difference of the expected centralized prediction reward and the expected decentralized prediction reward at the sufficient plan-time statistic σ^*_h at time h under an optimal policy π^* of $\langle \mathcal{M}, \Gamma \rangle$. This suggests that the final bound given as maximum over all sufficient plan-time statistics may be overly pessimistic for some cases. While a complete characterization of such cases is beyond the scope of this paper, we give below a sufficient condition for when the error is zero in the multi-agent case. The proof is in the supplementary material.

Observation 1. *Consider the setting of Theorem 1. Assume that the observation sequence of each agent is conditionally independent of the observation sequences of all other agents given the past joint policy and initial state distribution, i.e., for every agent i, $\sigma_h(\vec{z}_h) = \sigma_h(\vec{z}_{i,h})\sigma_h(\vec{z}_{-i,h})$. Then π^* is an optimal joint policy for $\langle \mathcal{M}, \Gamma \rangle$ if and only if $\pi^* \circ \phi^*$ is an optimal joint policy for $\mathcal{M}^+$.*

The sufficient condition above is restrictive. Informally, it requires that each agent executes its own independent active perception task. However, the loss due to decentralization may be small if a multi-agent active perception task *almost* satisfies the requirement. It might be possible to derive less restrictive sufficient conditions in weakly-coupled problems by building on influence-based abstractions [17] Such weakly-coupled cases may arise, e.g., in distributed settings where two agents deployed in different regions far away from each other have limited influence on each other.

5 Adaptive prediction action search

Recall that the centralized prediction reward function ρ is obtained by approximating a continuous convex function f by a set of α-vectors. The approximation is more accurate the more α-vectors

Algorithm 1 Adaptive prediction action search (APAS) for Dec-ρPOMDP planning

Input: Dec-ρPOMDP $\langle \mathcal{M}, \Gamma \rangle$, convex function $f\colon \Delta(S) \to \mathbb{R}$, number of linearization points K
Output: Best joint policy found, π_{best}

1: $V_{best} \leftarrow -\infty, \pi_{best} \leftarrow \emptyset$
2: **repeat**
3:     *// Policy optimization phase*
4:     $\mathcal{M}^+ \leftarrow$ CONVERTDECPOMDP($\mathcal{M}, \Gamma$) ▷ Apply Definition 3
5:     $\pi \leftarrow$ PLAN($\mathcal{M}^+$) ▷ Use any Dec-POMDP planner
6:     $V \leftarrow$ EVALUATE(π)
7:     **if** $V > V_{best}$ **then** $V_{best} \leftarrow V, \pi_{best} \leftarrow \pi$
8:     *// Adaptation phase*
9:     $\Gamma \leftarrow \emptyset$
10:     **for** $k = 1, \ldots, K$ **do**
11:         $\vec{z}_h \sim \sigma_h(\vec{z}_h)$ ▷ Sample joint observation sequence $\vec{z}_h$ using π_{best}
12:         $b_k \leftarrow \sigma_h(\cdot \mid \vec{z}_h)$ ▷ Final state estimate corresponding to $\vec{z}_h$
13:         $\alpha_k \leftarrow \nabla f(b_k) - f^*(\nabla f(b_k))$ ▷ Tangent hyperplane of f at b_k
14:         $\Gamma \leftarrow \Gamma \cup \{\alpha_k\}$
15:     **end for**
16: **until** converged
17: **return** π_{best}

are used: for error bounds, see [1, 21]. We propose a planning algorithm for Dec-ρPOMDPs called Adaptive Prediction Action Search (APAS) that dynamically updates the α-vectors during planning. This avoids the need to a priori create a large number of α-vectors some of which may turn out to not be useful. Moreover, APAS allows application of any standard Dec-POMDP algorithm to solve Dec-ρPOMDPs by using the conversion proposed in Section 4. This can avoid the computation of all reachable state estimates required by earlier algorithms for Dec-ρPOMDPs [11, 12].

The pseudocode for APAS is shown in Algorithm 1. APAS consists of two phases that are repeated until convergence: the policy optimization phase and the α-vector adaptation phase. In the policy optimization phase, the best joint policy for the current set Γ of α-vectors is found. On the first iteration we initialize Γ randomly as explained in the next section. In the adaptation phase, the α-vectors are then modified such that they best approximate the final reward for joint state estimates that are most likely reached by the current joint policy. In the optimization phase the Dec-ρPOMDP is converted to a standard Dec-POMDP as described in Section 4. We then apply any standard Dec-POMDP algorithm to plan a joint policy π for the converted Dec-POMDP. If the value of π exceeds the value of the best policy so far, π_{best}, the best policy is updated. In the adaptation phase, we sample a final state distribution b_k under the currently best policy, and insert the tangent hyperplane at b_k into Γ. This corresponds to sampling a point on the horizontal axis in Figure 1 and adding the corresponding tangent hyperplane to Γ. This is repeated K times. Specifically, we simulate a trajectory under the current best policy π_{best} to sample a joint observation sequence (Line 11), use Bayesian filtering to compute the corresponding state estimate b_k (Line 12), and compute corresponding α-vector (Line 13). The updated set Γ of α-vectors provides the best approximation of the final reward for the sampled state estimates. The time complexity of APAS is determined by the complexity of the Dec-POMDP planner called in PLAN. A reference implementation is available at `https://github.com/laurimi/multiagent-prediction-reward`.

6 Experiments

The Dec-ρPOMDP we target is computationally more challenging (NEXP-complete [3]) than centralized POMDP, POMDP-IR, or ρPOMDP (PSPACE-complete [19]). Immediate all-to-all communication during task execution would be required to solve the problem we target as a centralized problem, and the resulting optimal value would be higher [16]. Therefore, a comparison to centralized methods is neither fair nor necessary. We compare APAS to the NPGI algorithm of [11] that solves a Dec-ρPOMDP by iterative improvement of a fixed-size policy graph. As the PLAN subroutine of APAS, we use the finite-horizon variant of the policy graph improvement method of [18]. This method is algorithmically similar to NPGI which helps isolate the effect of applying the proposed conversion to a standard Dec-POMDP from the effect of algorithmic design choices to the extent

Table 1: Average policy values $\pm$ standard error in the MAV (left) and the rovers domains (right).

	MAV					Rovers			
h	APAS (ours)	APAS (no adapt.)	NPGI [12]	Optimal	h	APAS (ours)	APAS (no adapt.)	NPGI [12]	Optimal
2	-2.006 $\pm$ 0.005	-2.112 $\pm$ 0.002	-1.931	-1.919	2	-3.484 $\pm$ 0.002	-3.800 $\pm$ 0.006	-3.495	-3.479
3	-1.936 $\pm$ 0.004	-2.002 $\pm$ 0.002	-1.833	-1.831	3	-3.402 $\pm$ 0.008	-3.680 $\pm$ 0.006	-3.192	-3.189
4	-1.879 $\pm$ 0.004	-1.943 $\pm$ 0.002	-1.768	—	4	-3.367 $\pm$ 0.011	-3.640 $\pm$ 0.006	-3.036	—
5	-1.842 $\pm$ 0.005	-1.918 $\pm$ 0.002	-1.725	—	5	-3.293 $\pm$ 0.012	-3.591 $\pm$ 0.006	-2.981	—
6	-1.814 $\pm$ 0.004	-1.898 $\pm$ 0.002	—	—	6	-3.333 $\pm$ 0.012	-3.631 $\pm$ 0.007	—	—
7	-1.789 $\pm$ 0.004	-1.892 $\pm$ 0.003	—	—	7	-3.375 $\pm$ 0.014	-3.673 $\pm$ 0.008	—	—
8	-1.820 $\pm$ 0.005	-1.885 $\pm$ 0.006	—	—	8	-3.496 $\pm$ 0.014	-3.860 $\pm$ 0.010	—	—

possible. Any other Dec-POMDP algorithm may also be used with APAS. Details on the algorithms and parameter settings are provided in the supplementary material.

We evaluate on the Dec-ρPOMDP domains from [11]: the micro air vehicle (MAV) domain and information gathering rovers domain. In the MAV domain two agents cooperatively track a moving target to determine if the target is friendly or hostile. In the rovers domain, two agents collect samples of scientific data in a grid environment. The reward at the final time step is the negative entropy of the joint state estimate b_h, that is, $f(b_h) = \sum_{s_h} b_h(s_h) \ln b_h(s_h)$. For the MAV problem, we use $K = 2$, and for the rovers problem $K = 5$ individual prediction actions. For APAS, we initialize Γ by randomly sampling linearization points $b_k \in \Delta(S)$, $k = 1, \dots, K$ using [23]. The corresponding α-vectors for the negative entropy are $\alpha_k(s) = \ln b_k(s)$. We only observed minor benefits from using a greater number of α-vectors and only for long horizons, details are reported in the supplementary material. We run 100 repetitions using APAS and report the average policy value and its standard error. Results for NPGI are as reported in [12]. To ensure comparability, all reported policy values are computed using f as the final reward. As a baseline, we report the optimal value if known.

Comparison to the state-of-the-art. The results are shown in Table 1. While APAS does not exceed state-of-the-art performance in terms of policy quality, the major advantage of APAS is that it is able to scale to greater horizons than NPGI. The number of joint state distributions reachable under the current policy grows exponentially in the horizon h. Computation of these state distributions is required by NPGI, causing it to run out of memory with $h > 5$.

While running APAS, we compute the value of the policies exactly (Alg. 1, Line 6). For long horizons, exact evaluation requires a significant fraction of the computation budget. We expect that further scaling in terms of planning horizon is possible by switching to approximate evaluation, at the cost of noisy value estimates. Results on computation time as well as results for horizons up to 10 for the Rovers problem are reported in the supplementary material.

Benefit of adaptive prediction action selection. The adaptation phase of APAS is disabled by not running Lines 9-15. Instead, at the start of each iteration of the while loop we randomly sample K linearization points using [23] to create Γ and solve the corresponding Dec-POMDP. We repeat this procedure 1000 times and report the results in the column "APAS (no adapt.)" of Table 1. We conclude that the α-vector adaptation clearly improves performance in both domains.

7 Conclusion

We showed that multi-agent active perception modelled as a Dec-ρPOMDP can be reduced to a standard Dec-POMDP by introducing individual prediction actions. The difference between the optimal solution of the standard Dec-POMDP and the Dec-ρPOMDP is bounded. Our reduction enables application of any standard Dec-POMDP solver to multi-agent active perception problems, as demonstrated by our proposed APAS algorithm.

Our results allow transferring advances in scalability for standard Dec-POMDPs to multi-agent active perception tasks. In multi-agent reinforcement learning, rewards typically depend on the underlying state of the system. Therefore, our reduction result also enables further investigation into learning for multi-agent active perception. An investigation of the necessary conditions for when the loss due to decentralization is zero is another future direction.

Broader Impact

This work is theoretical in nature, and therefore does not present any foreseeable societal consequence.

Acknowledgments

This project had received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758824 —INFLUENCE).

European Research Council
Established by the European Commission

References

[1] Mauricio Araya-López, Olivier Buffet, Vincent Thomas, and Francois Charpillet. A POMDP Extension with Belief-dependent Rewards. In *Advances in Neural Information Processing Systems*, pages 64–72, 2010.

[2] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42(2):177–196, 2018.

[3] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[4] Graeme Best, Oliver M Cliff, Timothy Patten, Ramgopal R Mettu, and Robert Fitch. Dec-MCTS: Decentralized planning for multi-robot active perception. *The International Journal of Robotics Research*, 38(2-3):316–337, 2019.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] Jesus Capitan, Matthijs T.J. Spaan, Luis Merino, and Anibal Ollero. Decentralized multi-robot cooperation with auctioned POMDPs. *The International Journal of Robotics Research*, 32(6):650–671, 2013.

[7] Micah Corah and Nathan Michael. Distributed matroid-constrained submodular maximization for multi-robot exploration: Theory and practice. *Autonomous Robots*, 43(2):485–501, 2019.

[8] Bahman Gharesifard and Stephen L Smith. Distributed submodular maximization with limited information. *IEEE Transactions on Control of Network Systems*, 5(4):1635–1645, 2017.

[9] Manish Jain, Matthew Taylor, Milind Tambe, and Makoto Yokoo. DCOPs meet the real world: Exploring unknown reward matrices with applications to mobile sensor networks. In *Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 181–186, 2009.

[10] Mikko Lauri, Eero Heinänen, and Simone Frintrop. Multi-robot active information gathering with periodic communication. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 851–856, 2017.

[11] Mikko Lauri, Joni Pajarinen, and Jan Peters. Information Gathering in Decentralized POMDPs by Policy Graph Improvement. In *Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1143–1151, 2019.

[12] Mikko Lauri, Joni Pajarinen, and Jan Peters. Multi-agent active information gathering in discrete and continuous-state decentralized POMDPs by policy graph improvement. *Autonomous Agents and Multi-Agent Systems*, 34(42):1–44, 2020. Issue no. 2.

[13] Duc Thien Nguyen, William Yeoh, Hoong Chuin Lau, Shlomo Zilberstein, and Chongjie Zhang. Decentralized multi-agent reinforcement learning in average-reward dynamic DCOPs. In *AAAI Conference on Artificial Intelligence*, pages 1447–1455, 2014.

[14] Frans A. Oliehoek. Sufficient plan-time statistics for decentralized POMDPs. In *Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 302–308, 2013.

[15] Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

[16] Frans A. Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, 32:289–353, 2008.

[17] Frans A. Oliehoek, Stefan J. Witwicki, and Leslie P. Kaelbling. Influence-based abstraction for multiagent systems. In *AAAI Conference on Artificial Intelligence*, pages 1422–1428, 2012.

[18] Joni K Pajarinen and Jaakko Peltonen. Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning. In *Advances in Neural Information Processing Systems*, pages 2636–2644. 2011.

[19] Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[20] Yash Satsangi, Shimon Whiteson, Frans A. Oliehoek, and Matthijs T. J. Spaan. Exploiting submodular value functions for scaling up active perception. *Autonomous Robots*, 42(2):209–233, 2018.

[21] Yash Satsangi, Sungsu Lim, Shimon Whiteson, Frans A. Oliehoek, and Martha White. Maximizing information gain in partially observable environments via prediction rewards. In *Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1215–1223, 2020.

[22] Amarjeet Singh, Andreas Krause, Carlos Guestrin, and William J Kaiser. Efficient informative sensing using multiple robots. *Journal of Artificial Intelligence Research*, 34:707–755, 2009.

[23] Noah A Smith and Roy W Tromble. Sampling uniformly from the unit simplex. *Johns Hopkins University, Techical Report*, 2004.

[24] Matthijs TJ Spaan and Pedro U Lima. A decision-theoretic approach to dynamic sensor selection in camera networks. In *Intl. Conf. on Automated Planning and Scheduling (ICAPS)*, pages 297–304, 2009.

[25] Matthijs TJ Spaan, Tiago S Veiga, and Pedro U Lima. Decision-theoretic planning under uncertainty with information rewards for active cooperative perception. *Autonomous Agents and Multi-Agent Systems*, 29 (6):1157–1185, 2015.

[26] Matthew E Taylor, Manish Jain, Yanquin Jin, Makoto Yokoo, and Milind Tambe. When should there be a “me” in a “team”?: Distributed multi-agent optimization under uncertainty. In *Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 109–116, 2010.