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.

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 \[23\]. 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-\(\rho\)POMDP \[10\,11\]. In contrast to standard Dec-POMDPs, the reward function in a Dec-\(\rho\)POMDP 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 \[11,12\]...
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. Active perception is facilitated by rewarding the agent for correctly predicting the true underlying state. The equivalence of $\rho$POMDP and POMDP-IR model was later established [20]. Recently, Satsangi et al. [21] propose a reinforcement learning method to solve $\rho$POMDPs taking advantage of the equivalence. In this paper we prove an analogous equivalence for Dec-POMDPs, by converting a Dec-$\rho$POMDP 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 $\mathcal{M} = (h, I, S, b_0, A, Z, T, R)$, 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 $t = 0$,
- $A$ 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 $t$,
- $Z$ 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 $P(z_{t+1}, s_{t+1} \mid s_t, a_t)$, and
- $R$ 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 $\pi$, i.e., a tuple $\langle \pi_1, \ldots , \pi_n \rangle$ where the individual policy $\pi_i$ of each agent $i$ maps individual observation sequences $\vec{z}_{i,t} = \langle z_{i,1}, \ldots , z_{i,t} \rangle$ to an individual action $\hat{a}_{i,t}$. An individual policy is a sequence $\pi_t = \langle \delta_{i,0} , \ldots , \delta_{i,h-1} \rangle$ 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 $\vec{z}_t = \langle z_{1,t} , \ldots , z_{n,t} \rangle$ 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 $i$.\(^1\)

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

3
The objective is to find an optimal joint policy $\pi^*$ that maximizes the expected sum of rewards, that is, $\pi^* = \arg\max_{\pi} \mathbb{E} \left[ \sum_{t=0}^{h-1} R_t(s_t, \delta_t(\vec{z}_t)) \right]$, where the expectation is with respect to the distribution of states and joint observation sequences induced under $\pi$, 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-$\rho$POMDP [1] 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 $f: \Delta(S) \rightarrow \mathbb{R}$. 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 $\nabla$ 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 $\Gamma$ as the set of corresponding tangent hyperplanes $\alpha_j$. 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 $\rho$POMDPs, and has a bounded error [1]. The approximation is also used in [2], where error bounds for cases such as 0-1 rewards are provided. The Dec-$\rho$POMDP problem we consider is as follows.

**Definition 2** (Dec-$\rho$POMDP). A Dec-$\rho$POMDP is a pair $(\mathcal{M}, \Gamma)$, where $\mathcal{M}$ is a Dec-POMDP and $\Gamma$ is a set of tangent hyperplanes that determine the centralized prediction reward $\rho: \Delta(S) \rightarrow \mathbb{R}$ defined as $\rho(b) \triangleq \max_{\alpha \in \Gamma} \sum_{s} b(s) \alpha(s)$.

The standard Dec-POMDP is the special case where $\Gamma$ contains a single all-zero element. The objective in the Dec-$\rho$POMDP 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: $\varphi_{i,t} = \langle \delta_{i,0,t}, \ldots, \delta_{i,t-1,t} \rangle$. The sufficient plan-time statistic for initial state distribution $b_0$ and a past joint policy $\varphi_t$ is defined as

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

Sufficient plan-time statistics are updateable to any extension of the past joint policy $\varphi_t$. The extension of an individual past policy $\varphi_{i,t}$ by an individual decision rule $\delta_{i,t}$ is defined $\varphi_{i,t+1} = \varphi_{i,t} \circ \delta_{i,t} \triangleq \langle \varphi_{1,t+1}, \ldots, \varphi_{n,t+1} \rangle$. We defer discussion on how we select the linearization points to Section 5.
The extension of a past joint policy \( \varphi_t \) by \( \delta_t = (\delta_{1,t}, \ldots, \delta_{n,t}) \) is defined \( \varphi_{t+1} = \varphi_t \circ \delta_t \). Let \( \bar{z}_{t+1} = (\bar{z}_t, z_{t+1}) \) be a joint observation sequence that extends \( \bar{z}_t \) with \( z_{t+1} \). Given the statistic \( \sigma_t \) for \( \varphi_t \) and the next joint decision rule \( \delta_t \), the updated statistic for \( \varphi_{t+1} = \varphi_t \circ \delta_t \) is \( \sigma_{t+1} = U_{as}(\sigma_t, \delta_t) \), where the update operator \( U_{as} \) is defined as

\[
[U_{as}(\sigma_t, \delta_t)](s_{t+1}, \bar{z}_{t+1}) \triangleq \sum_{s_t} T(z_{t+1}, s_{t+1} \mid s_t, \delta_t(\bar{z}_t))\sigma_t(s_t, \bar{z}_t).
\]

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

\[
\hat{R}_t(\sigma_t, \delta_t) = \sum_{\bar{z}_t, s_t} \sigma_t(s_t, \bar{z}_t) R_t(s_t, \delta_t(\bar{z}_t)).
\]

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

\[
\hat{\rho}(\sigma_h) = \sum_{\bar{z}_h} \sigma_h(\bar{z}_h) \max_{\alpha} \sum_{s_h} \sigma_h(s_h \mid \bar{z}_h) \alpha(s_h).
\]

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

\[
Q^\pi_t(\sigma_t, \delta_t) = \begin{cases} 
\hat{R}_t(\sigma_t, \delta_t) + \hat{\rho}(\sigma_{t+1}) & \text{if } t = h - 1, \\
\hat{R}_t(\sigma_t, \delta_t) + Q^\pi_{t+1}(\sigma_{t+1}, \delta_{t+1}) & \text{if } 0 \leq t < h - 1,
\end{cases}
\]

with the shorthand \( \sigma_{t+1} = U_{as}(\sigma_t, \delta_t) \). The value function \( V^\pi \) of a joint policy \( \pi \) is defined as the sum of expected rewards when the agents act according to \( \pi \), \( V^\pi(\sigma_0) \triangleq Q^\pi_0(\sigma_0, \delta_0) \). The value function of an optimal policy \( \pi^* \) is denoted \( V^* \), and it satisfies \( V^*(\sigma_0) \geq V^\pi(\sigma_0) \) for all \( \pi \). We write \( V^\pi_M \) and \( V^\pi_{(M, \Gamma)} \) for the value function in a standard Dec-POMDP and a Dec-\( \rho \)POMDP, respectively.

## 4 Conversion to standard Dec-POMDP

In this section, we show that any Dec-\( \rho \)POMDP 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 [I]. 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-\( \rho \)POMDP 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-\( \rho \)POMDP 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-\( \rho \)POMDP 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-\( \rho \)POMDP \((M, \Gamma)\) with \( M = (h, I, S, b_0, A, Z, T, R) \), convert it to a standard Dec-POMDP \( M^+ = (h + 1, I, S, b_0, A^+, Z^+, T^+, R^+) \) where the horizon is incremented by one and

- in \( A^+ \), 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 \( A \),
We now show that the difference between the optimal values of the standard Dec-POMDP and the decentralized prediction reward never exceeds the centralized prediction reward as we show next.

The decentralized prediction reward is a sum of individual prediction rewards.

Lemma 1. Let \((M, \Gamma)\) be a Dec-\(p\)POMDP and define \(M^+\) as above. Then, for any past joint policy \(\varphi_t\) with \(t \leq h\), the respective plan-time sufficient statistics \(\sigma_t\) in \((M, \Gamma)\) and \(\sigma_t^+\) in \(M^+\) are equivalent, i.e., \(\sigma_t \equiv \sigma_t^+\).

In \(M^+\) the horizon is incremented, so that each agent takes one more action than in \((M, \Gamma)\). 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 \(\phi_i\) maps the individual observation sequence \(\vec{z}_{i,h}\) and the plan-time sufficient statistic \(\sigma_h\) to an individual prediction action. A joint prediction rule \(\phi = (\phi_1, \ldots, \phi_n)\) is a tuple of individual prediction rules, and maps a joint observation sequence \(\vec{z}_h\) and \(\sigma_h\) to a joint prediction action. The expected decentralized prediction reward given \(\sigma_h\) and \(\phi\) is defined analogously to Eq. (3) as

\[
\hat{R}_h(\sigma_h, \phi) = \sum_{\bar{z}_{i,h}, \bar{z}_{-i,h}} (1 - \rho) \hat{R}_h(s_h, \bar{z}_{i,h}, \sigma_h) R_i(h, s_h, \phi(\bar{z}_h, \sigma_h))
\]

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-\(p\)POMDP. The key property required is that the decentralized prediction reward is a sum of individual prediction rewards.

Lemma 2 (Optimal joint prediction rule). Let \((M, \Gamma)\) be a Dec-\(p\)POMDP and define \(M^+\) as above, and let \(\sigma_t\) be a plan-time sufficient statistic for any past joint policy. Then the joint prediction rule \(\phi^* = (\phi_1^*, \ldots, \phi_n^*)\) where each individual prediction rule \(\phi_i^*\) is defined as

\[
\phi_i^*(\vec{z}_{i,h}, \sigma_h) = \arg\max_{a_i,h} \sum_{\bar{z}_{-i,h}, s_h} \sigma_h(s_h, \bar{z}_{-i,h}) R_i(h, s_h, a_i,h)
\]

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

In the Dec-\(p\)POMDP \((M, \Gamma)\), 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 \(M^+\), 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 \(M^+\) is at most equal to the expected centralized prediction reward \(\hat{\rho}(\sigma_h)\) in \((M, \Gamma)\), i.e., \(\hat{R}_h(\sigma_h, \phi^*) \leq \hat{\rho}(\sigma_h)\).

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

We now show that the difference between the optimal values of the standard Dec-POMDP \(M^+\) and the Dec-\(p\)POMDP \((M, \Gamma)\) is bounded. We call the error the loss due to decentralization.

\[
\text{loss due to decentralization}
\]
Theorem 1 (Loss due to decentralization). Consider a Dec-\(\rho\)POMDP \(\langle M, \Gamma \rangle\) with the optimal value function \(V^{*}_{(M, \Gamma)}\). Let \(\pi\) be an optimal policy for the standard Dec-POMDP \(M^+\) created as in Definition 2 and denote by \(\varphi_0\) the past joint policy consisting of the first \(h\) decision rules of \(\pi\). Then the difference of \(V^{*}_{(M, \Gamma)}\) and the value function \(V^{\varphi_0}_{(M, \Gamma)}\) of applying \(\varphi_0\) to \(\langle M, \Gamma \rangle\) is bounded by
\[
|V^{*}_{(M, \Gamma)}(\sigma_0) - V^{\varphi_0}_{(M, \Gamma)}(\sigma_0)| \leq 2 \max_{\sigma_h} |\hat{\rho}(\sigma_h) - \hat{R}_h(\sigma_h, \phi^*)|,
\]
where \(\phi^*\) is the optimal joint prediction rule.

Proof. For clarity, in the following we omit the argument \(\sigma_0\) from the value functions. Suppose the optimal policy \(\pi^*\) in \(\langle M, \Gamma \rangle\) consists of the joint decision rules \(\delta_0^*, \ldots, \delta_{h-1}^*\). For \(M^+\), 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 \(\pi^*\) of \(M^+\) is an extension of some past joint policy \(\varphi_h^*\) by \(\phi^*\), i.e., \(\pi = \varphi_h^* \circ \phi^*\). Because \(\varphi_h^* \circ \phi^*\) is optimal in \(M^+\), we have that \(V^{\varphi_h^* \circ \phi^*}_{M^+} \leq V^{\varphi_h^*}_{M^+}\). Finally, because \(\pi^*\) is optimal in \(\langle M, \Gamma \rangle\), we have \(V^{\varphi_h^*}_{(M, \Gamma)} \leq V^{\varphi_h^*}_{(M, \Gamma)}\). Next, we apply the triangle inequality.

The theorem shows that given a Dec-\(\rho\)POMDP \(\langle M, \Gamma \rangle\), we may solve the standard Dec-POMDP \(M^+\) and apply its optimal policy to the Dec-\(\rho\)POMDP 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 \(\rho\)POMDP 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 \(\sigma^*_h\) at time \(h\) under an optimal policy \(\pi^*\) of \(\langle M, \Gamma \rangle\). 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 \(i\), \(\sigma_h(\tilde{z}_i) = \sigma_h(\tilde{z}_{-i,h})\sigma_h(\tilde{z}_{-i,h})\). Then \(\pi^*\) is an optimal joint policy for \(\langle M, \Gamma \rangle\) if and only if \(\varphi_h^* \circ \phi^*\) is an optimal joint policy for \(M^+\).

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 \(\rho\) is obtained by approximating a continuous convex function \(f\) by a set of \(\alpha\)-vectors. The approximation is more accurate the more \(\alpha\)-vectors
Algorithm 1 Adaptive prediction action search (APAS) for Dec-$\rho$POMDP planning

\begin{algorithm}
\begin{algorithmic}
\State \textbf{Input:} Dec-$\rho$POMDP $<M, \Gamma>$, convex function $f: \Delta(S) \rightarrow \mathbb{R}$, number of linearization points $K$
\State \textbf{Output:} Best joint policy found, $\pi_{\text{best}}$
\State $V_{\text{best}} \leftarrow -\infty$, $\pi_{\text{best}} \leftarrow \emptyset$
\State \textbf{repeat}
\State \hspace{1em} // Policy optimization phase
\State $M^+ \leftarrow \text{CONVERTDECPOMDP}(M, \Gamma)$ \Comment{Apply Definition 4}
\State $\pi \leftarrow \text{PLAN}(M^+)$ \Comment{Use any Dec-POMDP planner}
\State $V \leftarrow \text{EVALUATE}(\pi)$ \Comment{Output: $V_{\text{best}}$}
\State \hspace{1em} if $V > V_{\text{best}}$ then $V_{\text{best}} \leftarrow V$, $\pi_{\text{best}} \leftarrow \pi$
\State \hspace{1em} // Adaptation phase
\State $\Gamma \leftarrow \emptyset$
\State \hspace{1em} for $k = 1, \ldots, K$ do
\State \hspace{2em} $\tilde{z}_k \sim \sigma_h(\tilde{z}_k)$ \Comment{Sample joint observation sequence $\tilde{z}_k$ using $\pi_{\text{best}}$}
\State \hspace{2em} $b_k \leftarrow \sigma_b(\cdot | \tilde{z}_k)$ \Comment{Final state estimate corresponding to $\tilde{z}_k$}
\State \hspace{2em} $\alpha_k \leftarrow \nabla f(b_k) - f^*(\nabla f(b_k))$ \Comment{Tangent hyperplane of $f$ at $b_k$}
\State \hspace{2em} $\Gamma \leftarrow \Gamma \cup \{\alpha_k\}$
\State \hspace{1em} end for
\State \hspace{1em} until converged
\State \Return $\pi_{\text{best}}$
\end{algorithmic}
\end{algorithm}

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 $\alpha$-vector adaptation phase. In the policy optimization phase, the best joint policy for the current set $\Gamma$ of $\alpha$-vectors is found. On the first iteration we initialize $\Gamma$ randomly as explained in the next section. In the adaptation phase, the $\alpha$-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-$\rho$POMDP 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 $\pi$ for the converted Dec-POMDP. If the value of $\pi$ exceeds the value of the best policy so far, $\pi_{\text{best}}$, 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 $\Gamma$. This corresponds to sampling a point on the horizontal axis in Figure 1 and adding the corresponding tangent hyperplane to $\Gamma$. This is repeated $K$ times. Specifically, we simulate a trajectory under the current best policy $\pi_{\text{best}}$ to sample a joint observation sequence (Line 11), use Bayesian filtering to compute the corresponding state estimate $b_k$ (Line 12), and compute corresponding $\alpha$-vector (Line 13). The updated set $\Gamma$ of $\alpha$-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-$\rho$POMDP we target is computationally more challenging (NEXP-complete [3]) than centralized POMDP, POMDP-IR, or $\rho$POMDP (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-$\rho$POMDP 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...
We showed that multi-agent active perception modelled as a Dec-POMDP 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-POMDP 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.

### Table 1: Average policy values ± standard error in the MAV (left) and the rovers domains (right).

<table>
<thead>
<tr>
<th></th>
<th>MAV</th>
<th>Rovers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>APAS (ours)</td>
<td>APAS (no adapt.)</td>
</tr>
<tr>
<td>2</td>
<td>-2.006 ± 0.005</td>
<td>-2.112 ± 0.002</td>
</tr>
<tr>
<td>3</td>
<td>-1.936 ± 0.004</td>
<td>-2.002 ± 0.002</td>
</tr>
<tr>
<td>4</td>
<td>-1.879 ± 0.004</td>
<td>-1.943 ± 0.002</td>
</tr>
<tr>
<td>5</td>
<td>-1.842 ± 0.005</td>
<td>-1.918 ± 0.002</td>
</tr>
<tr>
<td>6</td>
<td>-1.814 ± 0.004</td>
<td>-1.898 ± 0.002</td>
</tr>
<tr>
<td>7</td>
<td>-1.789 ± 0.004</td>
<td>-1.892 ± 0.003</td>
</tr>
<tr>
<td>8</td>
<td>-1.820 ± 0.005</td>
<td>-1.885 ± 0.006</td>
</tr>
</tbody>
</table>

**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 21-25. Instead, at the start of each iteration of the while loop we randomly sample \( K \) linearization points using [23] to create \( \Gamma \) 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 \( \alpha \)-vector adaptation clearly improves performance in both domains.

### 7 Conclusion

We showed that multi-agent active perception modelled as a Dec-POMDP 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-POMDP 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).

References


