Bimodal Switching for Online Planning in Multiagent Settings*

Ekhlas Sonu and **Prashant Doshi**
THINC Lab, Department of Computer Science
University of Georgia
Athens, GA 30602 USA
esonu@uga.edu, pdoshi@cs.uga.edu

Abstract

We present a bimodal method for online planning in partially observable multiagent settings as formalized by a finitely-nested interactive partially observable Markov decision process (I-POMDP). An agent planning in an environment shared with another updates beliefs both over the physical state and the other agents' models. In problems where we do not observe other's action explicitly but must infer it from sensing its effect on the state, observations are more informative about the other when the belief over the state space has reduced uncertainty. For typical, uncertain initial beliefs, we model the agent as if it were acting alone and utilize fast online planning for POMDPs. Subsequently, the agent switches to online planning in multiagent settings. We maintain tight lower and upper bounds at each step, and switch over when the difference between them reduces to less than ϵ.

1 Introduction

Online planning involves deliberating on optimal actions to perform in a limited amount of time given beliefs that evolve as the agent acts and observes. The limited-time aspect puts an emphasis on reduced planning time, which is achieved by trading off optimality and relying on approximations. Attention to online planning has predominantly focused on single-agent settings [Ross *et al.*, 2008] albeit isolated approaches for cooperative problems do exist [Wu *et al.*, 2011].

We attend to planning in the presence of an interacting agent, which complicates the process. In this space, the finitely-nested interactive partially observable Markov decision process (I-POMDP) [Gmytrasiewicz and Doshi, 2005; Doshi, 2012] is a recognized approach for performing the deliberations by modeling the other agent. Approaches such as the interactive particle filtering [Doshi and Gmytrasiewicz, 2009] offer a way to plan online given a belief. However, this is time consuming and does not scale well.

Within the I-POMDP framework, an agent simultaneously updates its beliefs over both the physical state and the models of other agent based on its actions and observations. The distribution over the models is updated based on the predicted action, whose effect on the state is observed by the agent. These observations become more informative when the agent's belief over the physical state has reduced uncertainty (entropy). For example, in the context of the well-known multiagent tiger problem [Gmytrasiewicz and Doshi, 2005], let the agent strongly believe that the tiger is on the left but on listening hear a growl from the right. If the observation is reliable with a high probability, the agent infers that the other agent likely opened the door causing the tiger to change its location.

We present a novel two-stage approach that focuses first on online planning as if the agent is alone, in order to reduce uncertainty in its belief over the physical state. In this mode, the agent is modeled as a POMDP and utilizes a fast POMDP-based planning technique, SARSOP [Kurniawati *et al.*, 2008], that takes orders of magnitude less time to execute as compared to the I-POMDP solver. Subsequently, the agent switches to the I-POMDP model combining its updated belief over the state and the initial belief over the models. It now performs online planning using interactive particle filtering.

A key question is when should the agent switch from the POMDP to the I-POMDP mode? In order to answer this, the agent at every step computes lower and upper bounds on the optimal decision at that step. The agent switches to the latter mode when the fractional difference between the lower and upper bounds at any step become less than a parameter, ϵ. Because of the convexity property of the lower-bound value function, the difference between the two typically reduces as beliefs become less uncertain. Nevertheless, the bounds may not converge, and very small ϵ values may not cause a switch.

We estimate the error that this approach entails. The computational savings result because during the initial steps of online planning, a fast and scalable single-agent approach is utilized. We illustrate this approach on the multiagent tiger problem and show that the total time elapsed over several steps is significantly less compared to using an I-POMDP model only at the expense of reduced cumulative reward.

2 Overview of Finitely-Nested I-POMDPs

A finitely-nested I-POMDP [Gmytrasiewicz and Doshi, 2005] for an agent i with strategy level, l, interacting with another agent j is defined using the tuple:

*We acknowledge support from NSF CAREER grant, #IIS-0845036.

$$\text{I-POMDP}_{i,l} = \langle IS_{i,l}, A, T_i, \Omega_i, O_i, R_i, OC_i \rangle$$

- $IS_{i,l}$ denotes the set of *interactive states* at strategy level, l, defined as, $IS_{i,l} = S \times M_{j,l-1}$, where S is the set of physical states, and $M_{j,l-1}$ is the set of models ascribed to the other agent.
- $A = A_i \times A_j$ is the set of joint actions of all agents.

The remaining parameters – T_i (transition function), Ω_i and O_i (observation space and function respectively), R_i (reward function), and OC_i (optimality criterion) – are defined analogously to POMDPs with the generalization that the functions are conditioned on the joint actions, and the states are the physical states only. We focus on optimizing the summation of the discounted reward over a finite number of remaining steps, called the horizon.

Agent i's 0-th level belief is a probability distribution over the physical states only, and the 0-th level models, $M_{j,0}$, consist of the set of computable intentional models of level 0, $\Theta_{j,0}$, and the subintentional models, SM_j. An intentional model, $\theta_{j,0} = \langle b_{j,0}, \hat{\theta}_j \rangle$, where $b_{j,0}$ is j's level 0 belief, $b_{j,0} \in \triangle(S)$ – $\triangle(S)$ is the set of all distributions over S – and $\hat{\theta}_j = \langle A, T_j, \Omega_j, O_j, R_j, OC_j \rangle$, is j's frame. Here, j is assumed to be Bayes-rational. 0-th level intentional models are POMDPs. An agent's level 1 beliefs are probability distributions over the physical states and level 0 models of the other. First-level models include level 1 intentional models and level 0 models of the agent, and so on up to level l.

We limit our attention to intentional models only, and simplify the interactive state space to include intentional models of the other agent that are of level one less than that of the modeling agent. We assume that the frame of agent j is known and remains fixed; it need not be the same as that of i.

An agent's belief over its interactive states, $b_{i,l}$, fully summarizes the agent's observation history. We may factor the agent's belief as: $b_{i,l}(is) = b_{i,l}(s)\ b_{i,l}(\theta_{j,l-1}|s)$. Beliefs are updated after the agent's action and observation using Bayes rule. First, as the state of the physical environment depends on the actions performed by both agents, the prediction of how it changes has to be made based on the probabilities of various actions of the other agent. Probabilities of other's actions are obtained by solving its models. Second, changes in the models of the other agent have to be included in the update. The changes reflect the other's observations and, when modeled intentionally, the update of other agent's beliefs.

Given the generalized belief update, solution to an I-POMDP$_{i,l}$ is a *policy*, analogous to that in a POMDP. Gmytrasiewicz and Doshi (2005) provide additional details on I-POMDPs, and the Bellman operator for backup, H.

One approach for online planning in settings formalized by I-POMDP$_{i,l}$ involves sampling the interactive state space and projecting the samples, called particles, across time in order to simulate the belief update. The *interactive particle filter* (I-PF) [Doshi and Gmytrasiewicz, 2009] propagates, weights and resamples a nested set of particles of level l. A reachability tree is generated where the sampled representations of beliefs at the nodes are obtained by running the I-PF. Value iteration using sample sets on the reachability tree is performed to obtain the approximately-optimal action at the given belief.

3 Bimodal Online Planning

Let $b^0_{i,l}$ be agent i's initial belief over the interactive state space. Agent i initially views the problem as a single-agent POMDP and its belief is the distribution over the physical states only, $b^0_{i,l}(s)$. Conditional beliefs over the models given the state, $b^0_{i,l}(\cdot|s)$, are held fixed. Subsequently, the agent switches to updating the conditionals as well after some steps.

We show that in problems where direct (including noisy) observations about the other agent's actions are not available and its actions are inferred by sensing the next state, perhaps noisily, knowing the current state provides greater information about the true model of the other agent.

Definition 1 (Unobservable actions)**.** *Actions of agent j are directly unobservable to agent i iff observations of i are conditionally independent of j's action, $O_i(s, a_i, a_j, o_i) = O_i(s, a_i, o_i)$; $\forall\ a_j \in A_j, o_i \in \Omega_i, a_i \in A_i, s \in S$.*

Let $Pr(\dot{\theta}^{t+1}_{j,l-1}|o^{t+1}_i, a^t_i, b^t_{i,l})$ be i's probability of the true model of the other agent, $\dot{\theta}^{t+1}_{j,l-1}$, on observing, o^{t+1}_i, and performing action, a^t_i, given belief, $b^t_{i,l}$. We may write it as,

$$Pr(\dot{\theta}^{t+1}_{j,l-1}|o^{t+1}_i, a^t_i, b^t_{i,l}) = \sum_{s^{t+1}} Pr(s^{t+1}, \dot{\theta}^{t+1}_{j,l-1}|o^{t+1}_i, a^t_i, b^t_{i,l})$$

Notice that the term on the right is agent i's updated belief over a state and true model of j, which is obtained using the I-POMDP belief update. Under Def. 1 applied to both agents,

$$Pr(\dot{\theta}^{t+1}_{j,l-1}|o^{t+1}_i, a^t_i, b^t_{i,l}) = \sum_{s^{t+1}} \sum_{s^t} b^t_{i,l}(s^t) \sum_{\theta^t_{j,l-1}} b^t_{i,l}(\theta^t_{j,l-1}|s^t)$$
$$\sum_{a^t_j} Pr(a^t_j|\theta^t_{j,l-1})\ T_i(s^t, a^t_i, a^t_j, s^{t+1})\ O_i(s^{t+1}, a^t_i, o^{t+1}_i)$$
$$\sum_{o^{t+1}_j} O_j(s^{t+1}, a^t_j, o^{t+1}_j)\ \tau(\theta^t_{j,l-1}, a^t_j, o^{t+1}_j, \dot{\theta}^{t+1}_{j,l-1}) \tag{1}$$

where τ is an indicator function that is 1 when model, $\theta^t_{j,l-1}$, updates to $\dot{\theta}^{t+1}_{j,l-1}$ on performing action, a^t_j, and receiving observation, o^{t+1}_j; otherwise 0. Equation 1 shows that i's belief over j's true model at $t+1$ is influenced by j's predicted actions from its models, the observations that j may likely receive and agent i's transition and observation functions.

Let agents i and j perform actions, a^t_i and a^t_j respectively, due to which the state transitions from s^t to s^{t+1}. As j's actions are unobservable, state transitions allow valuable inference of j's actions.

Definition 2 (Maximally informative transition)**.** *The above state transition is maximally informative about j's action iff $T_i(s^t, a^t_i, a^t_j, s^{t+1}) \geq T_i(\bar{s}^t, a^t_i, a^t_j, s^{t+1})$ for any other state, $\bar{s}^t$, and $T(s^t, a^t_i, a^t_j, s^{t+1}) > T(s^t, a^t_i, \bar{a}^t_j, s^{t+1})$ for all other actions, $\bar{a}^t_j$, of j.*

Consider domains where the other agent's actions are unobservable (Def. 1). In such settings, observations are more informative about the other agent's models if the uncertainty over the physical state is mitigated. In order to show this, let the transition that occurs from the current state, $\dot{s}^t$, due to joint actions be maximally informative (Def. 2). Let j's performed action solely lead to its true model. Agent i then receives the observation, o^{t+1}_i, that is most likely. Then, the following proposition holds:

Proposition 1. *If $b^{t,1}_{i,l}$ is a belief which assigns probability 1 on the current state, $\dot{s}^t$, then, $Pr(\dot{\theta}^{t+1}_{j,l-1}|o^{t+1}_i, a^t_i, b^{t,1}_{i,l}) \geq Pr(\dot{\theta}^{t+1}_{j,l-1}|o^{t+1}_i, a^t_i, b^{t,2}_{i,l})$, for any other belief, $b^{t,2}_{i,l}$, under the assumption that the conditional distributions over the models of j in both beliefs are identical and do not change with state.*

Proof. For convenience, we may rewrite Eq. 1 as,

$$Pr(\dot{\theta}^{t+1}_{j,l-1}|o^{t+1}_i, a^t_i, b^t_{i,l}) = \sum_{s^t} b^t_{i,l}(s^t)\ \mathcal{X}(s^t, a^t_i, o^{t+1}_i, \dot{\theta}^{t+1}_{j,l-1})$$

where $\mathcal{X}(s^t, a^t_i, o^{t+1}_i, \dot{\theta}^{t+1}_{j,l-1})$ denotes the remaining terms of Eq. 1 as a function of s^t, a^t_i, o^{t+1}_i, and $\dot{\theta}^{t+1}_{j,l-1}$. Under the assumption that i's conditional distribution over the models is identical for any state, if the current state is $\dot{s}^t$, then $\mathcal{X}(\dot{s}^t, a^t_i, o^{t+1}_i, \dot{\theta}^{t+1}_{j,l-1})$ is greater than $\mathcal{X}(\bar{s}^t, a^t_i, o^{t+1}_i, \dot{\theta}^{t+1}_{j,l-1})$ for any other state, $\bar{s}^t$. This is because, $\sum_{s^{t+1}} T_i(\bar{s}^t, a^t_i, a^t_j, s^{t+1})\ O_i(s^{t+1}, a^t_i, o^{t+1}_i) \leq \sum_{s^{t+1}} T_i(\dot{s}^t, a^t_i, a^t_j, s^{t+1})\ O_i(s^{t+1}, a^t_i, o^{t+1}_i)$, as any transition to a possible next state, s^{t+1}, is maximally informative about j's action given i's action, from the current state, $\dot{s}^t$. While some other action of j could result in a next state, s^{t+1}, from some $\bar{s}^t$, it does not lead to the true model of j as per τ. As $b^{t,1}_{i,l}$ puts a probability 1 on $\dot{s}^t$, it follows that $\mathcal{X}(\dot{s}^t, a^t_i, o^{t+1}_i, \dot{\theta}^{t+1}_{j,l-1})$ is greater than $\sum_{s^t} b^{t,2}_{i,l}(s^t)\ \mathcal{X}(s^t, a^t_i, o^{t+1}_i, \dot{\theta}^{t+1}_{j,l-1})$, for any other belief, $b^{t,2}_{i,l}$, which differs from $b^{t,1}_{i,l}$ in its distribution over the physical states only. □

While we sought to reduce some of the possibilities due to uncertainty, Proposition 1 formalizes the intuition that in domains where behavioral information about the other agent must be inferred by sensing the dynamic state, received observations (that are not noise) tend to be more informative about the other's model when the uncertainty over the current physical state is as less as possible.

If agent i's conditional belief over models is decoupled from its belief over the physical state, the proposition becomes useful as the beliefs, $b^{t,2}_{i,l}$ and $b^{t,1}_{i,l}$, could be those that are in the sequence of beliefs that i may have as it acts and observes. Consequently, in problems where j's actions are unobservable, it motivates that we update the distributions over the models as late as possible in the game. Of course, the trade off is that the conditional distributions over models are held fixed for a longer time affecting early predictions.

3.1 POMDP model with noise

Our formulation of the POMDP model [Kaelbling *et al.*, 1998] for agent i, POMDP$_i$, uses as its state space the set of physical states, S, in the I-POMDP$_{i,l}$. The action and observation spaces are identical to those in I-POMDP$_{i,l}$. The transition, observation, and reward functions are marginals of those in I-POMDP$_{i,l}$, obtained by summing out j's actions from the functions. The optimality criterion remains a sum of discounted rewards over the finite horizon.

In order to sum out j's actions from the functions, we begin by mapping j's model space to a set of *nodes*, $\mathcal{N}_{j,l-1}$. Each node in this set corresponds to a distribution over j's actions. If j is at level 0, we obtain these nodes by performing bounded policy iteration (BPI) [Poupart and Boutilier, 2003] on the level 0 POMDP, which is a technique for obtaining a controller of fixed size. Our BPI begins with a single node that corresponds to a random action followed by one step of full backup. Subsequently, the controller is improved using BPI until convergence. Let $\mathcal{N}_{j,l-1}$ be the set of nodes of this converged controller. Associated with each node is also a value vector that gives the expected reward of performing the action(s) corresponding to the node from each state, and then following the remaining controller until the values converge.

Nodes in $\mathcal{N}_{j,l-1}$ partition the continuous model space: For a belief, $b_{j,0}$, in each model, compute the inner product between the belief and the value vector for each node. Then, map the model to the node with the largest value breaking ties randomly. Note that multiple models may be mapped to a single node. For j's models at a level, $l-1$, greater than 0, we may perform *interactive* BPI [Sonu and Doshi, 2012] resulting in a (nested) controller of level $l-1$. A benefit of this approach of mapping models to nodes is that the countably infinite model space is reduced to a finite set of nodes, $\mathcal{N}_{j,l-1}$. We may obtain the distribution over j's actions for summing out as:

$$Pr(a_j|s) = \sum_{n_{j,l-1} \in \mathcal{N}_{j,l-1}} b^0_{i,l}(n_{j,l-1}|s)\ Pr(a_j|n_{j,l-1}) \quad (2)$$

where $Pr(a_j|n_{j,l-1})$ is the probability assigned to action, a_j, by the node, $n_{j,l-1}$, and $b^0_{i,l}(n_{j,l-1}|s)$ is the conditional probability mass in i's initial belief over models that get transferred to $n_{j,l-1}$ due to the mapping: $b^0_{i,l}(n_{j,l-1}|s) = \sum_{\theta_{j,l-1} \mapsto n_{j,l-1}} b^0_{i,l}(\theta_{j,l-1}|s)$.

Equation 2 provides the distribution over the other agent's actions used for formulating the marginal functions. This distribution is static and the resulting POMDP$_i$ models the other agent as noise in the environment. Solution of this POMDP is obtained by using SARSOP [Kurniawati *et al.*, 2008], which is a fast and scalable POMDP solution technique that produces a *policy graph*.

Next, we show that the expected reward from our formulation of POMDP$_i$ is a lower bound to the expected reward from I-POMDP$_{i,l}$ in which the model space, $\Theta_{j,l-1}$ has been substituted with the space, $\mathcal{F}_{j,l-1}$. Here, $f_{j,l-1} \in \mathcal{F}_{j,l-1}$ is, $f_{j,l-1} = \langle n_{j,l-1}, \hat{f}_{j,l-1}, \hat{\theta}_j \rangle$, where $n_{j,l-1}$ is a node in the set of nodes in j's level $l-1$ controller, $n_{j,l-1} \in \mathcal{N}_{j,l-1}$; $\hat{f}_{j,l-1}$ includes the set of edge labels, distributions of actions for each node and the edge transition function of the controller; and $\hat{\theta}_j$ is j's fixed frame. In other words, treating the other agent as noise is less valuable than correctly modeling it.

In order to compare between the two value functions, we obtain the value of i's marginal belief over the physical state in the I-POMDP$_{i,l}$ as: $V(b_{i,l}(s)) = \max_{\hat{\alpha}} \sum_{s \in S} b_{i,l}(s)\hat{\alpha}(s)$.

Here,

$$\hat{\alpha}(s) = \sum_{n_{j,l-1}} b^0_{i,l}(n_{j,l-1}|s)\ \alpha(s, n_{j,l-1}) \tag{3}$$

where $b^0_{i,l}(n_{j,l-1}|s)$ is the initial conditional belief over nodes as defined previously, and $\alpha(s, n_{j,l-1})$ is an alpha vector that composes the value function for I-POMDP$_{i,l}$.

Proposition 2 (Lower bound)**.** *Let V denote the value function of I-POMDP$_{i,l}$. Let $\underline{H}$ and H be the backup operators for POMDP$_i$ and I-POMDP$_{i,l}$, respectively. Then, it holds that $HV \geq \underline{H}V$.*

Proof. Let $b_{i,l}(s)$ be a belief over the physical states, and $b_{i,l}(s, n_{j,l-1}) = b_{i,l}(s)\ b^0_{i,l}(n_{j,l-1}|s)$, where $b^0_{i,l}(n_{j,l-1}|s)$ is the initial conditional distribution.

We begin by showing that for horizon 1, for any $b_{i,l}(s)$, value of this belief given by POMDP$_i$ is the same as the value of the belief, $b_{i,l}(s, n_{j,l-1})$, shown previously:

$$\underline{V}(\hat{b}_{i,l}) = \max_{a_i \in A_i} \sum_s b_{i,l}(s) \hat{R}_i(s, a_i)$$

where $\hat{R}_i(s, a_i)$ is agent i's reward function in POMDP$_i$. Introducing a_j gives,

$$\begin{aligned}
&\underline{V}(b_{i,l}) = \max_{a_i \in A_i} \sum_s b_{i,l}(s) \textstyle\sum_{a_j} R_i(s, a_i, a_j) Pr(a_j|s) \\
&= \max_{a_i \in A_i} \sum_s b_{i,l}(s) \sum_{a_j} R_i(s, a_i, a_j) \sum_{n_{j,l-1}} b^0_{i,l}(n_{j,l-1}|s) \\
&\ Pr(a_j|n_{j,l-1}) \qquad \text{(from Eq. 2)} \\
&= \max_{a_i \in A_i} \sum_{s, n_j} b_{i,l}(s, n_{j,l-1}) \sum_{a_j} R_i(s, a_i, a_j)\ Pr(a_j|n_{j,l-1}) \\
&= V(b_{i,l})
\end{aligned}$$

Next, we move to the case where the horizon is 2, and apply the I-POMDP$_{i,l}$ backup operator to the value function, V:

$$\begin{aligned}
&HV(b_{i,l}) = \max_{a_i \in A_i} \sum_{s, n_{j,l-1}} b_{i,l}(s, n_{j,l-1}) ER_i(s, a_i) \\
&+ \sum_{o_i} Pr(o_i|a_i, b_{i,l}) \max_{\alpha} b'_{i,l} \cdot \alpha \\
&= \max_{a_i \in A_i} \sum_s \sum_{n_{j,l-1}} b_{i,l}(s) b^0_{i,l}(n_{j,l-1}|s) \sum_{a_j} Pr(a_j|n_{j,l-1}) \\
&\left\{ R_i(s, a_i, a_j) + \sum_{o_i} \sum_{s'} Pr(s', o_i|s, a_i, a_j) \sum_{o_j} O_j(s', a_j, o_j) \right. \\
&\left. \sum_{n'_{j,l-1}} Pr(n'_{j,l-1}|n_{j,l-1}, a_i, o_j)\ \alpha^k(s', n'_{j,l-1}) \right\}
\end{aligned}$$

where k is the index of the alpha vector that provides the maximal value at the updated belief, $b'_{i,l}$. We may rewrite the above dynamic programming update by noting that $\sum_{o_j} O_j(s', a_j, o_j) \sum_{n'_{j,l-1}} Pr(n'_{j,l-1}|n_{j,l-1}, a_i, o_j)$ represents the updated belief over the models conditioned on the updated state, $b'_{j,l-1}(n'_{j,l-1}|s')$.

$$\begin{aligned}
&HV(b_{i,l}) = \max_{a_i \in A_i} \sum_s \sum_{n_j} b_{i,l}(s) b^0_{i,l}(n_{j,l-1}|s) \sum_{a_j} Pr(a_j|n_{j,l-1}) \\
&\left\{ R_i(s, a_i, a_j) + \sum_{o_i} \sum_{s'} Pr(s', o_i|s, a_i, a_j) \sum_{n'_j} b'_{j,l-1}(n'_{j,l-1}|s') \right. \\
&\left. \alpha^k(s', n'_{j,l-1}) \right\}
\end{aligned}$$

$$\begin{aligned}
&\geq \max_{a_i \in A_i} \sum_s \sum_{n_{j,l-1}} b_{i,l}(s) b^0_{i,l}(n_{j,l-1}|s) \sum_{a_j} Pr(a_j|n_{j,l-1}) \\
&\left\{ R_i(s, a_i, a_j) + \sum_{o_i} \sum_{s'} Pr(s', o_i|s, a_i, a_j) \sum_{n_{j'}} b^0_{j,l-1}(n'_{j,l-1}|s') \right. \\
&\left. \hat{\alpha}^k(s', n'_{j,l-1}) \right\}
\end{aligned}$$

The above holds because $\alpha^k(s', n'_{j,l-1})$ is maximal at $b'_{i,l} = b'_{i,l}(s') b'_{i,l}(n'_{j,l-1}|s')$. For the belief in the above equation, $b'_{i,l}(s') b^0_{i,l}(n'_{j,l-1}|s')$, it may continue to remain maximal if $b'_{i,l}(n'_{j,l-1}|s')$ and $b^0_{i,l}(n'_{j,l-1}|s')$ are identical, otherwise it's suboptimal. Using Eq. 3, above equation may be rewritten.

$$\begin{aligned}
&HV(b_{i,l}) \geq \max_{a_i \in A_i} \sum_s \sum_{n_j} b_{i,l}(s) b^0_{i,l}(n_{j,l-1}|s) \sum_{a_j} Pr(a_j|n_{j,l-1}) \\
&\left\{ R_i(s, a_i, a_j) + \sum_{o_i} \sum_{s'} Pr(s', o_i|s, a_i, a_j) \hat{\alpha}^k(s') \right\} \\
&= \max_{a_i \in A_i} \sum_s b_{i,l}(s) \sum_{a_j} \sum_{n_{j,l-1}} b^0_{i,l}(n_{j,l-1}|s) Pr(a_j|n_{j,l-1}) \\
&\left\{ R_i(s, a_i, a_j) + \sum_{o_i} \sum_{s'} Pr(s', o_i|s, a_i, a_j) \hat{\alpha}^k(s') \right\} \\
&= \max_{a_i \in A_i} \sum_s b_{i,l}(s) \sum_{a_j} Pr(a_j|s) \left\{ R_i(s, a_i, a_j) \right. \\
&\left. + \sum_{o_i} \sum_{s'} Pr(s', o_i|s, a_i, a_j) \hat{\alpha}^k(s') \right\} \qquad \text{(Using Eq. 2)} \\
&= \max_{a_i \in A_i} \sum_s b_{i,l}(s) \left\{ \sum_{a_j} R_i(s, a_i, a_j) Pr(a_j|s) \right. \\
&\left. + \sum_{o_i} \sum_{s'} \sum_{a_j} Pr(s', o_i|s, a_i, a_j) Pr(a_j|s) \hat{\alpha}^k(s') \right\} \\
&= \max_{a_i \in A_i} \sum_s b_{i,l}(s) \left\{ \hat{R}_i(s, a_i) + \sum_{o_i} \sum_{s'} Pr(s', o_i|s, a_i) \hat{\alpha}^k(s') \right\} \\
&= \underline{H}V(b_{i,l})
\end{aligned}$$

Here, $\hat{R}$ is the reward function in POMDP$_i$ as defined previously, $Pr(s', o_i|s, a_i)$ is the joint observation and transition functions, and $\underline{H}$ is the corresponding backup operator.

As $V = \underline{V}$, we also get, $H\underline{V} \geq \underline{H}\underline{V}$ from the above proof, and furthermore, $H(\underline{H}\underline{V}) \geq \underline{H}(\underline{H}\underline{V})$. Because the I-POMDP$_{i,l}$ backup operator is isotonic, $H(H\underline{V}) \geq H(\underline{H}\underline{V})$. This implies, $H(HV) \geq \underline{H}(\underline{H}\underline{V})$. Thus, repeatedly applying the two backup operators maintains the lower bound. □

This is intuitive and demonstrates the benefit of closely tracking the other agent's dynamic models.

3.2 I-POMDP$_{i,l}$ with perfectly observable state

As we mentioned previously, agent i utilizes the policy graph resulting from solving POMDP$_i$ using a fast and scalable technique. The value of the action given by POMDP$_i$ for any i's belief over the physical states is guaranteed to be a lower bound to the optimal approach of using I-POMDP$_{i,l}$. Eventually, the agent switches to using an online solution of I-POMDP$_{i,l}$ for acting. In order to facilitate the switching, an upper bound value for its belief over the physical states is additionally needed.

If an observation reveals the physical state perfectly to agent i in an I-POMDP$_{i,l}$, the proposition below shows that the resulting value of $b_{i,l}$ is an upper bound to the general value. Notice that despite the physical state being perfectly observable, the resulting model does not collapse into an MDP because the model of the other agent continues to remain uncertain. Let $\overline{V}$ be the value function of this model, which we denote as I-POMDP$^S_{i,l}$. The value function is composed of possibly multiple vectors for each state, s. For each state, we obtain the maximal value of the inner product between the initial conditional belief, $b_{i,l}(n_{j,l-1}|s)$, and the updated alpha vectors for that state. These values form a single alpha vector over the physical states.

$$\begin{aligned}
&\overline{V}(b_{i,l}) = \textstyle\sum_s b_{i,l}(s) \max_{\alpha^s} \sum_{n_{j,l-1}} b_{i,l}(n_{j,l-1}|s)\alpha^s(n_{j,l-1}) \\
&= \textstyle\sum_s b_{i,l}(s) \max_{a_i \in A_i} \sum_{n_{j,l-1}} b_{i,l}(n_{j,l-1}|s) \sum_{a_j} Pr(a_j|n_{j,l-1}) \\
&\Big\{ R_i(s,a_i,a_j) + \textstyle\sum_{s'} Pr(s'|s,a_i,a_j) \sum_{o_j} O_j(s',a_j,o_j) \\
&\sum_{n'_{j,l-1}} Pr(n'_{j,l-1}|n_{j,l-1},a_i,a_j) \alpha^{s',k'}(n'_{j,l-1}) \Big\}
\end{aligned} \tag{4}$$

Proposition 3 (Upper bound)**.** *Let $\overline{H}$ be the backup operator for the value function of I-POMDP$^S_{i,l}$ as defined in Eq. 4. Then, it holds that $H\overline{V} \leq \overline{HV}$, where H is the backup operator for I-POMDP$_{i,l}$ as defined previously.*

Proof. Value of a belief, $b_{i,l}$, for horizon 1 in I-POMDP$_{i,l}$ is,

$$\begin{aligned}
&V(b_{i,l}) = \max_{a_i \in A_i} \sum_{s,n_j} b_{i,l}(s,n_{j,l-1}) R_i(s,a_i,a_j) Pr(a_j|n_{j,l-1}) \\
&= \max_{a_i \in A_i} \sum_s b_{i,l}(s) \sum_{n_j} b_{i,l}(n_{j,l-1}|s) R_i(s,a_i,a_j) Pr(a_j|n_{j,l-1}) \\
&\leq \sum_s b_{i,l}(s) \max_{a_i \in A_i} \sum_{n_j} b_{i,l}(n_{j,l-1}|s) R_i(s,a_i,a_j) Pr(a_j|n_{j,l-1}) \\
&= \overline{V}(b_{i,l})
\end{aligned}$$

For a horizon greater than one, we obtain,

$$\begin{aligned}
&H\overline{V}(b_{i,l}) = \max_{a_i \in A_i} \textstyle\sum_s b_{i,l}(s) \sum_{n_{j,l-1}} b_{i,l}(n_{j,l-1}|s) \\
&\sum_{a_j} Pr(a_j|n_{j,l-1}) \Big\{ R_i(s,a_i,a_j) + \textstyle\sum_{o_i} \sum_{s'} Pr(s',o_i|s,a_i,a_j) \\
&\textstyle\sum_{o_j} O_j(s',a_j,o_j) \sum_{n'_{j,l-1}} Pr(n'_{j,l-1}|n_{j,l-1},a_i,a_j) \\
&\alpha^{s',k'}(n'_{j,l-1}) \Big\}
\end{aligned}$$

The alpha vector, $\alpha^{s',k'}(n'_{j,l-1})$, is the one in the set of vectors for the next state, s', that gives the largest value for the updated conditional belief over the models. Its selection from the set does not depend on the observation, o_i, unlike in I-POMDP$_{i,l}$. Therefore, the equation above simplifies to,

$$\begin{aligned}
&H\overline{V}(b_{i,l}) = \max_{a_i \in A_i} \textstyle\sum_s b_{i,l}(s) \sum_{n_{j,l-1}} b_{i,l}(n_{j,l-1}|s) \\
&\textstyle\sum_{a_j} Pr(a_j|n_{j,l-1}) \Big\{ R_i(s,a_i,a_j) + \sum_{s'} Pr(s'|s,a_i,a_j) \\
&\textstyle\sum_{o_j} O_j(s',a_j,o_j) \sum_{n'_{j,l-1}} Pr(n'_{j,l-1}|n_{j,l-1},a_i,a_j) \\
&\alpha^{s',k'}(n'_{j,l-1}) \Big\} \\
&\leq \textstyle\sum_s b_{i,l}(s) \max_{a_i \in A_i} \sum_{n_{j,l-1}} b_{i,l}(n_{j,l-1}|s) \\
&\textstyle\sum_{a_j} Pr(a_j|n_{j,l-1}) \Big\{ R_i(s,a_i,a_j) + \sum_{s'} Pr(s'|s,a_i,a_j) \\
&\textstyle\sum_{o_j} O_j(s',a_j,o_j) \sum_{n'_{j,l-1}} Pr(n'_{j,l-1}|n_{j,l-1},a_i,a_j) \\
&\alpha^{s',k'}(n'_{j,l-1}) \Big\} \\
&= \overline{HV}
\end{aligned}$$

Under the isotonicity property of the I-POMDP backup operator and $V \leq \overline{V}$, we obtain, $H(HV) \leq H(H\overline{V})$. Similarly, $H(H\overline{V}) \leq H(\overline{HV})$. We may reapply the above proof and assert that, $H(\overline{HV}) \leq \overline{H}(\overline{HV})$. This implies that, $H(HV) \leq \overline{H}(\overline{HV})$, which means that the upper bound is maintained over any number of applications of the two backup operators to their respective value functions. □

3.3 Bimodal switching

Let the difference between the upper and lower bound values for a belief, $b_{i,l}$, over the physical states be, $\mathsf{Diff} = \overline{V}_i(b_{i,l}) - \underline{V}_i(b_{i,l})$. Let R_{min} and R_{max} be the smallest and highest rewards in agent i's reward function, R_i. Subsequently, $R_{min}\frac{1-\gamma^H}{1-\gamma}$ and $R_{max}\frac{1-\gamma^H}{1-\gamma}$ are the minimum and maximum rewards that agent i could obtain over a finite horizon of H with a discount factor of γ. These may be easily calculated from the model definition.

Because of the piecewise linear and convexity property of the value function of POMDP$_i$, and the relatively flat value function of I-POMDP$^S_{i,l}$, we expect Diff to reduce as the uncertainty in agent i's belief over the state space reduces and the belief approaches the edges of the belief simplex. Our approach switches from online planning using POMDP$_i$ to planning using I-POMDP$_{i,l}$ when $\frac{\mathsf{Diff}\cdot(1-\gamma)}{(1-\gamma^H)\cdot(R_{max}-R_{min})}$ drops to below a parameter, $\epsilon \in [0,1]$. In other words, ϵ is the fraction of the largest possible difference in value, which triggers the switch. However, not all values of ϵ may be reached. Specifically, there is no guarantee that the upper and lower bounds converge near the edges of the belief simplex. Consequently, small values of ϵ may not cause a switch.

3.4 Computational Savings and Error Bound

Instead of solving a multiagent planning problem from the start, our approach exploits single-agent planning in the early stages subsequently switching to multiagent planning on reaching a belief distribution that admits reduced error bound. Consequently, computational savings occur due to the steps when single-agent planning is performed. In order to obtain an estimate of the savings, let us suppose that we use an exact POMDP-based approach for online planning that generates a perfect reachability tree of $H-1$ steps whose branching factor is $(|A_i||\Omega_i|)$, from a belief. The tree contains $(|A_i||\Omega_i|)^H - 1$ nodes each of which is associated with a single real number whose computation takes time $\mathcal{O}(|A_i||S|)$ if the node is a leaf node, otherwise it takes $\mathcal{O}(|A_i||\Omega_i||S|^2)$. On the other hand, let us suppose that we use an exact I-POMDP$_{i,l}$-based approach for the planning. The reachability tree from a belief over the interactive state space would continue to have a branching factor of $(|A_i||\Omega_i|)$ and as many nodes as mentioned previously. However, calculating the value of the belief associated with each leaf node takes time $\mathcal{O}(|A_i||S||N_{j,l-1}||A_j|)$ and the time for calculating the value at a non-leaf node takes $\mathcal{O}(|A_i||S|^2|N_{j,l-1}||A_j|\Omega_i||\Omega_j|)$. The difference in computation time at each node is due to modeling the other agent, and the savings at all the nodes accumulates over the number of steps for which planning uses POMDP$_i$, which may vary.

Our choice of SARSOP – a state of the art approach – minimizes the time taken to perform the POMDP-based planning. SARSOP generates a policy graph that is near-optimal for a given horizon from any belief, and the initial computation time is amortized over the multiple steps for which the graph is used. The I-POMDP$_{i,l}$-based online planning utilizes I-PF and generates a reachability tree for a given horizon from a given belief resulting in an approximate action.

If an exact approach is utilized for online multiagent planning after the switch, error is incurred until the approach switches when ϵ is achieved. At this point, the difference between the upper and lower bounds is, $\epsilon \cdot \frac{(1-\gamma^H)\cdot(R_{max}-R_{min})}{1-\gamma}$, which bounds the error as well. If T steps were performed before switching, then the error is at least, $T \cdot \epsilon \cdot \frac{(1-\gamma^H)\cdot(R_{max}-R_{min})}{1-\gamma}$. This also serves as a reasonable estimate of the error because our bounds are tight resulting in small ϵ values, as we illustrate next.

4 Experimental Evaluation

In Figure 1, we illustrate the lower and upper bound values for changing beliefs of agent i over the physical states for the multiagent tiger problem [Gmytrasiewicz and Doshi, 2005]. We modify this problem by removing creaks thereby making j's actions unobservable (Def. 1) and let the tiger persist with a probability of 0.75 in its original location after a door is opened. Furthermore, we mitigate the amount of noise in j's observations of the state as modeled by i to 0.05.

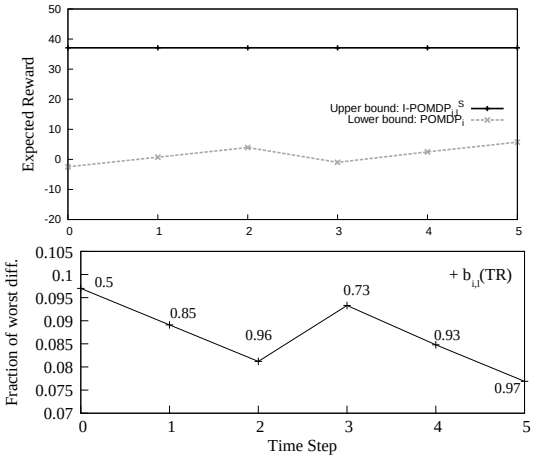

Figure 1: **(top)** Beginning with $b_{i,l}(TR) = 0.5$, we show the lower and upper bound values obtained from POMDP$_i$ and I-POMDP$^S_{i,l}$, respectively, for a run of the multiagent persistent tiger problem. **(bottom)** The fraction of the largest difference in bounds is shown as agent i acts, observes and its beliefs update, in simulation.

Beginning at a belief of $b_{i,1}(TR) = 0.5$ indicating that the tiger is believed to be behind the right door with a probability of 0.5, the beliefs are updated as the agent listens and receives observations. We obtained j's controller using BPI, which has 5 nodes. Agent i's initial distribution over these nodes is obtained by mapping a distribution over level 0 models of j to a distrbution over the nodes. Notice that the fraction, ϵ, becomes smaller as the beliefs show less uncertainty although not monotonically. The increase from steps 2 to 3 occurs due to i opening the left door causing the uncertainty in its beliefs to increase and reducing the lower bound value.

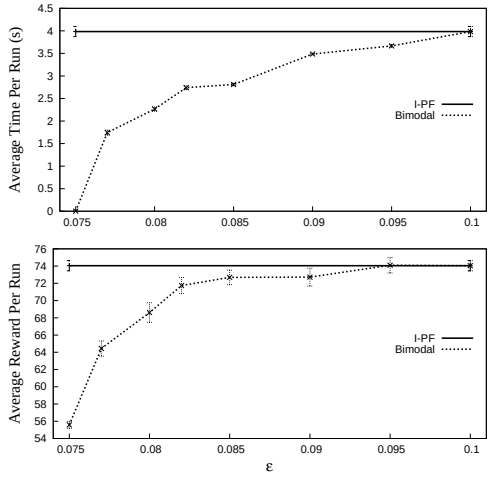

Figure 2: **(top)** Average time taken per simulation run for different values of ϵ (Xeon i3 2.6GHz, 4GB with Linux). **(bottom)** Cumulative rewards averaged over the 300 runs with differing ϵ.

Taking an online planning horizon of $H = 5$, we simulated agent i's play of the tiger problem for 30 steps. We vary ϵ and show the time and cumulative reward averaged over 5 trials of 300 simulation runs each. Notice the low values of ϵ indicating that our bounds are tight. The feasible range of ϵ is [0.075,0.1], with the approach unable to reach $\epsilon < 0.075$ and degenerating into POMDP$_i$ for all the steps, while satisfying $\epsilon < 0.1$ at the first step itself thereby running using I-PF for all the steps. For the smallest ϵ value which causes a bimodal switch, the average time is about 50% less than I-PF albeit obtaining average reward that is significantly lower. However, as ϵ increases, POMDP$_i$ runs for less steps and both the time and rewards increase approaching that of I-PF.

5 Conclusion

We presented a new approach for online planning in multiagent settings where actions of the other agent are not directly observable and must be inferred from the state transitions. For typical initial beliefs with high uncertainty over the physical states, our approach utilizes POMDP-based planning and switches over to online multiagent planning. The mode changes when the fraction of the difference between the upper and lower bounds reduces to less than a parameter. This technique of utilizing bounds is analogous to previous approaches such as HSVI [Smith and Simmons, 2004] and SARSOP, although these compute bounds relevant to single agent settings. Our demonstration on a toy problem domain indicates that the bimodal approach is flexible and significantly saves on time while obtaining improved rewards.

References

[Doshi and Gmytrasiewicz, 2009] Prashant Doshi and Piotr Gmytrasiewicz. Monte carlo sampling methods for approximating interactive pomdps. *Journal of Artificial Intelligence Research*, 34:297–337, 2009.

[Doshi, 2012] Prashant Doshi. Decision making in complex mulitiagent contexts: A tale of two frameworks. *AI Magazine*, 33(4):82–95, 2012.

[Gmytrasiewicz and Doshi, 2005] Piotr Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multiagent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[Kaelbling *et al.*, 1998] Leslie Kaelbling, Michael Littman, and Anthony Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[Kurniawati *et al.*, 2008] H. Kurniawati, D. Hsu, and W.S. Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*, 2008.

[Poupart and Boutilier, 2003] Pascal Poupart and Craig Boutilier. Bounded finite state controllers. In *Neural Information Processing Systems*, 2003.

[Ross *et al.*, 2008] S. Ross, J. Pineau, S. Paquet, and B. Chaib-draa. Online planning algorithms for pomdps. *Journal of Artificial Intelligence Research (JAIR)*, 32:663–704, 2008.

[Smith and Simmons, 2004] Trey Smith and Reid Simmons. Heuristic search value iteration for pomdps. In *Uncertainty in Artificial Intelligence (UAI)*, pages 520–527, 2004.

[Sonu and Doshi, 2012] Ekhlas Sonu and Prashant Doshi. Generalized and bounded policy iteration for interactive pomdps: Scaling up. In *Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1039–1046, 2012.

[Wu *et al.*, 2011] Feng Wu, Shlomo Zilberstein, and Xiaoping Chen. Online planning for multi-agent systems with bounded communication. *Artificial Intelligence Journal (AIJ)*, 175(2):487–511, 2011.