Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Bayes-Adaptive Interactive POMDPs

Brenda Ng and **Kofi Boakye** and **Carol Meyers**
Lawrence Livermore National Laboratory
Livermore, CA 94550

Andrew Wang
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We introduce the Bayes-Adaptive Interactive Partially Observable Markov Decision Process (BA-IPOMDP), the first multiagent decision model that explicitly incorporates model learning. As in I-POMDPs, the BA-IPOMDP agent maintains beliefs over interactive states, which include the physical states as well as the other agents' models. The BA-IPOMDP assumes that the state transition and observation probabilities are unknown, and augments the interactive states to include these parameters. Beliefs are maintained over this augmented interactive state space. This (necessary) state expansion exacerbates the curse of dimensionality, especially since each I-POMDP belief update is already a recursive procedure (because an agent invokes belief updates from other agents' perspectives as part of its own belief update, in order to anticipate other agents' actions). We extend the interactive particle filter to perform approximate belief update on BA-IPOMDPs. We present our findings on the multiagent Tiger problem.

1 Introduction

Within the last decade, the body of work in multiagent sequential decision-making methods has grown substantially, both in terms of theory and practical feasibility. For these methods to be applicable in real-world settings, we must account for the fact that agents usually lack perfect knowledge about their environments, with regard to (1) the state of the world, and (2) the consequences of their interactions, in terms of model parameters such as transition and observation probabilities. Thus, agents must infer their state from the history of actions and observations, *and* concurrently learn their model parameters via trial and error. The development of such a framework is the objective of this paper.

Our particular focus is on adversarial agents that are intelligent, and actively seek to "game" against each other during the course of repeated interactions. In such a setting, each agent must anticipate its adversary's responses to its actions, which entails also anticipating the adversary's observations and beliefs about the state of the world. We feel that this type of study is highly relevant to realistic security problems, since these intelligent agents abound in the form of cyber intruders, money launderers, material smugglers, etc. While each "attack" might be launched by one individual, it is reasonable to treat an entire class of attackers as a single adversary, as similar tactics are adopted by multiple individuals to exploit the vulnerabilities of the target agent.

Many sequential multiagent decision-making frameworks are extensions of the single-agent Partially Observable Markov Decision Process (POMDP) model. In a POMDP, a single agent with limited knowledge of its environment attempts to optimize a discrete sequence of actions to maximize its expected rewards. Because the agent does not fully know the state of the environment, it infers a state distribution through a series of noisy observations. Solution algorithms for POMDPs have been studied extensively (Kaelbling, Littman, and Cassandra 1998), and POMDPs have been applied to real-world problems including the assistance of patients with dementia (Hoey et al. 2007).

Among the multiagent frameworks that have been studied, the largest body of literature has been on decentralized POMDPs (DEC-POMDPs) (Bernstein et al. 2002), which generalize POMDPs to multiple decentralized agents and are used to model multiagent teams (Seuken and Zilberstein 2008). While algorithms have been developed to solve such problems (Seuken and Zilberstein 2008), DEC-POMDPs are not suitable for modeling adversarial agents because the framework assumes common rewards among agents. The related framework of Markov Team Decision Problems (MTDPs) (Pynadath and Tambe 2002) has the same issue. Partially Observable Stochastic Games (POSGs) (Hansen, Bernstein, and Zilberstein 2004) avoid this problem by allowing for different agent rewards, but exact POSG algorithms have been so far limited to small problems (Guo and Lesser 2006), and approximate POSG algorithms have only been developed for the common rewards case (Emery-Montemerlo et al. 2004).

A suitable framework for modeling multiagent adversarial interactions is that of interactive POMDPs (I-POMDPs) (Doshi and Gmytrasiewicz 2005). The I-POMDP is a multiagent extension of the POMDP, in which each agent maintains beliefs about both the physical states of the world and the decision process models of the other agents. An I-POMDP incorporates nested intent into agent beliefs, which potentially allows for modeling of "gaming" agents. There are approximate algorithms for solving I-POMDPs that do not impose common agent rewards. (Ng et al. 2010) demonstrates an attempt to apply I-POMDPs to money laundering.

Although I-POMDPs can be used to model adversarial agents, they are not amenable to real-world applications because transition and observation probabilities need to be specified as part of the model. In most cases, these parameters are not known exactly, and must be approximated a priori or learned during the interaction. Reinforcement learning

(Kaelbling, Littman, and Moore 1996) provides a methodology by which these parameters may be estimated sequentially, thus avoiding potentially non-optimal solutions associated with poor a priori approximations.

Bayes-Adaptive POMDPs (BA-POMDPs) (Ross, Chaib-draa, and Pineau 2007) enable parameter learning in POMDPs. In a BA-POMDP, the agent's state is augmented to include the agent's counts of state transitions and observations, and these counts are used to estimate parameters in the transition and observation functions. Parameter estimates are improved through interactions with the environment, and optimal actions converge over time to the true optimal solution.

To date, work in multiagent learning has been focused mainly on fully observable domains (Busoniu, Babuska, and Schutter 2008; Melo and Ribeiro 2010) or cooperative, partially observable domains (Peshkin et al. 2000; Zhang and Lesser 2011; Oliehoek 2012), where policy learning is emphasized over model learning.

The contribution of this work is the *Bayes-Adaptive I-POMDP (BA-IPOMDP)*, that incorporates elements of I-POMDPs and BA-POMDPs, to achieve the first multiagent adversarial learning framework that explicitly learns model parameters. The BA-IPOMDP allows for imperfect knowledge of both the world state and the agents' transition and observation probabilities, thus bringing theory closer to human agent modeling. Our preliminary results show that the BA-IPOMDP *learning* agent achieves better rewards than the I-POMDP *static* agent, when the two start from the same prior model. We cover technical background in Section 2, explain our BA-IPOMDP model in Section 3 and our belief update algorithm in Section 4. We present results in Section 5 and conclude in Section 6.

2 Preliminaries

2.1 Bayes-Adaptive POMDPs (BA-POMDPs)

In a BA-POMDP (Ross, Chaib-draa, and Pineau 2007), the state, action, and observation spaces are assumed to be finite and known, but the state transition and observation probabilities are unknown and must be inferred. It extends the POMDP $\langle \mathcal{S}, \mathcal{A}, T, \mathcal{R}, \mathcal{Z}, O \rangle$, by allowing uncertainty to be associated with $T(s,a,s')$ and $O(s',a,z)$.

The uncertainties are parametrized by Dirichlet distributions defined over experience counts. The count $\phi^a_{ss'}$ denotes the number of times that transition (s,a,s') has occurred, and the count $\psi^a_{s'z}$ denotes the number of times observation z was made in state s' after performing action a. Given these counts, the expected transition probabilities $T_\phi(s,a,s')$ and the expected observation probabilities $O_\psi(s',a,z)$ are:

$$T_\phi(s,a,s') = \frac{\phi^a_{ss'}}{\sum_{s'' \in \mathcal{S}} \phi^a_{ss''}} \tag{1}$$

$$O_\psi(s',a,z) = \frac{\psi^a_{s'z}}{\sum_{z' \in \mathcal{Z}} \psi^a_{s'z'}} \tag{2}$$

The BA-POMDP incorporates the count vectors ϕ and ψ as part of the state, so functions need to be augmented accordingly to capture the evolution of these vectors. Let $\delta^a_{ss'}$ be a vector of zeros with a 1 for the count $\phi^a_{ss'}$, and $\delta^a_{s'z}$ be a vector of zeros with a 1 for the count $\psi^a_{s'z}$. Formally, a BA-POMDP is parametrized by $\langle \mathcal{S}', \mathcal{A}, T', \mathcal{R}', \mathcal{Z}, O' \rangle$, where the differences from a POMDP are:

- $\mathcal{S}' = \mathcal{S} \times \mathcal{T} \times \mathcal{O}$ is the augmented state space, where $\mathcal{T} = \{\phi \in \mathbb{N}^{|\mathcal{S}|^2|\mathcal{A}|} | \forall (s,a), \sum_{s' \in \mathcal{S}} \phi^a_{ss'} > 0\}$, and $\mathcal{O} = \{\psi \in \mathbb{N}^{|\mathcal{S}||\mathcal{A}||\mathcal{Z}|} | \forall (s,a), \sum_{z \in \mathcal{Z}} \psi^a_{sz} > 0\}$;
- $T' : \mathcal{S}' \times \mathcal{A} \times \mathcal{S}' \to [0,1]$ is the (joint) state transition function, defined as $T'((s,\phi,\psi),a,(s',\phi',\psi')) = T_\phi(s,a,s')O_\psi(s',a,z)$ if $\phi' = \phi + \delta^a_{ss'}$ and $\psi' = \psi + \delta^a_{s'z}$, and 0 otherwise;
- $\mathcal{R}' : \mathcal{S}' \times \mathcal{A} \to \mathbb{R}$ is the reward function, defined as $\mathcal{R}'((s,\phi,\psi),a) = \mathcal{R}(s,a)$;
- $O' : \mathcal{S}' \times \mathcal{A} \times \mathcal{S}' \times \mathcal{Z} \to [0,1]$ is the observation function, defined as $O'((s,\phi,\psi),a,(s',\phi',\psi'),z) = 1$ if $\phi' = \phi + \delta^a_{ss'}$ and $\psi' = \psi + \delta^a_{s'z}$, and 0 otherwise.

The belief update in BA-POMDPs is analogous to that of POMDPs, but inference is performed over a larger state space, as beliefs are maintained over the (unobserved) counts, ϕ and ψ, in addition to the physical states. Monte Carlo sampling along with online look-ahead search have been applied to solve BA-POMDPs.

2.2 Interactive POMDPs (I-POMDPs)

I-POMDPs (Doshi and Gmytrasiewicz 2005) generalize POMDPs to multiple agents with different (and possibly conflicting) objectives. In an I-POMDP, beliefs are maintained over *interactive* states, which include the physical states *and* the models of other agents' behaviors.

In the case of two intentional agents, i and j, agent i's I-POMDP with l levels of nesting is specified by:

$$\langle \mathcal{IS}_{i,l}, \mathcal{A}, T_i, \mathcal{R}_i, \mathcal{Z}_i, O_i \rangle$$

where:

- $\mathcal{IS}_{i,l} = \mathcal{S} \times \Theta_{j,l-1}$ is the finite set of i's interactive states, with $\mathcal{IS}_{i,0} = \mathcal{S}$; $\Theta_{j,l-1}$ as the set of intentional models of j, where each model $\theta_{j,l-1} \in \Theta_{j,l-1}$ consists of j's belief $b_{j,l-1}$ and frame $\hat{\theta}_j = \langle \mathcal{A}, T_j, \mathcal{R}_j, \mathcal{Z}_j, O_j \rangle$;
- $\mathcal{A} = \mathcal{A}_i \times \mathcal{A}_j$ is the finite set of joint actions;
- $T_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is i's model of the joint transition function;
- $\mathcal{R}_i : \mathcal{IS}_i \times \mathcal{A} \to \mathbb{R}$ is i's reward function;
- $\mathcal{Z}_i$ is the finite set of i's observations; and
- $O_i : \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_i \to [0,1]$ is i's observation function.

At each time step, agent i maintains a belief state:

$$\begin{aligned} b^t_{i,l}(is^t) = \beta \textstyle\sum_{is^{t-1}} b^{t-1}_{i,l}(is^{t-1}) \\ \textstyle\sum_{a_j^{t-1}} P(a_j^{t-1} | \theta^{t-1}_{j,l-1}) T_i(s^{t-1}, a^{t-1}, s^t) O_i(s^t, a^{t-1}, z^t_i) \\ \textstyle\sum_{z_j^t} P(b^t_{j,l-1} | b^{t-1}_{j,l-1}, a^{t-1}_j, z^t_j) O_j(s^t, a^{t-1}, z^t_j) \end{aligned} \tag{3}$$

where β is a normalizing factor and $P(a_j^{t-1} | \theta^{t-1}_{j,l-1})$ is the probability that a_j^{t-1} is Bayes rational for an agent modeled by $\theta^{t-1}_{j,l-1}$. Let $OPT(\theta_j)$ denote the set of j's optimal actions computed from a planning algorithm that maximizes rewards accrued over an infinite horizon (i.e., $E(\sum_{t=0}^{\infty} \gamma^t r_t)$

where $0 < \gamma < 1$ is the discount factor and r_t is the reward achieved at time t. $P(a_j^{t-1}|\theta_{j,l-1}^{t-1})$ is set to $\frac{1}{|OPT(\theta_{j,l-1}^{t-1})|}$ if $a_j^{t-1} \in OPT(\theta_{j,l-1}^{t-1})$ and 0 otherwise.

The belief update procedure for I-POMDPs is more complicated than that of POMDPs, because physical state transitions depend on both agents' actions. To predict the next physical state, i must update its beliefs about j's behavior based on its anticipation of how j updates its belief (note $P(b_{j,l-1}^t|b_{j,l-1}^{t-1}, a_j^{t-1}, z_j^t)$ in the belief update equation). This can lead to infinite nesting of beliefs, for which a finite level of nesting l is imposed in practice. (Doshi et al. 2010) has shown that most humans act using up to two levels of reasoning in general-sum strategic games, suggesting the nesting level of $l = 2$ as an upper bound for modeling human agents.

A popular method for solving I-POMDPs is the interactive particle filter (I-PF) in conjunction with reachability tree sampling (RTS). Other methods include dynamic influence diagrams (Doshi, Zeng, and Chen 2009) and generalized point-based value iteration (Doshi and Perez 2008).

3 Bayes-Adaptive I-POMDPs

While other cooperative multiagent frameworks can exploit common rewards to simplify the learning problem into multiple parallel instances of single-agent learning, the learning process in an adversarial multiagent framework is intrinsically more coupled. This is because each agent can no longer rely on its own model as a baseline for modeling others; each agent is less informed about the other (adversarial) agents because agents have different rewards and potentially different dynamics stemming from their own actions and observations.

Since state transitions depend on joint actions from all agents, it is imperative that the adversarial agent consider other agents' perspectives before taking action. The I-POMDP offers a vehicle for recursive modeling to address this, which the BA-IPOMDP augments with the additional capability for learning. Thus, the BA-IPOMDP can model learning about self, learning about other agents, and learning about other agents learning about self, etc.

Like the BA-POMDP, the BA-IPOMDP assumes that the state, action, and observation spaces are finite and known a priori. Each agent is trying to learn a $|\mathcal{S}| \times |\mathcal{A}| \times |\mathcal{S}|$ matrix T of state transition probabilities, where $T(s^{t-1}, a^{t-1}, s^t) = P(s^t|s^{t-1}, a^{t-1})$, and a $|\mathcal{S}| \times |\mathcal{A}| \times |\mathcal{Z}_i|$ matrix O of observation probabilities, where $O(s^t, a^t, z_i^t) = P(z_i^t|s^t, a^t)$.

Each agent's physical state is augmented to include the transition counts and *all* agents' observation counts. We denote this state as $s' = (s, \phi, \psi_i, \psi_j) \in \mathcal{S}' = \mathcal{S} \times \mathcal{T} \times \mathcal{O}_i \times \mathcal{O}_j$, where s is the physical state, ϕ is the transition counts (over $\mathcal{S}$), ψ_i is agent i's observation counts (over $\mathcal{Z}_i$), and ψ_j is agent j's observation counts (over $\mathcal{Z}_j$). Note that this does not require access to other agents' observations. We treat the other agent's observations like the physical states, as partially observable and maintain beliefs over them. (If each agent only maintained its *individual* observation counts, there would be insufficient information to infer the *joint* transition function.) Given ϕ, ψ_i, and ψ_j, the expected probabilities, $T_{\phi^{t-1}}^{s^{t-1}a^{t-1}s^t}$, $O_{\psi_i^{t-1}}^{s^t a^{t-1} z_i^t}$ and $O_{\psi_j^{t-1}}^{s^t a^{t-1} z_j^t}$ are defined similarly as in BA-POMDPs (cf. Equations (1) and (2)).

We construct the BA-IPOMDP $\langle \mathcal{IS}'_{i,l}, \mathcal{A}, T'_i, \mathcal{R}'_i, \mathcal{Z}_i, O'_i \rangle$ from the I-POMDP $\langle \mathcal{IS}_{i,l}, \mathcal{A}, T_i, \mathcal{R}_i, \mathcal{Z}_i, O_i \rangle$ as follows. Agent i's *augmented interactive state*, $is'^t_{i,l} = \{(s^t, \phi^t, \psi_i^t, \psi_j^t), \theta_{j,l-1}^t\}$, combines the augmented state, $s'^t = (s^t, \phi^t, \psi_i^t, \psi_j^t)$, and its knowledge of agent j's model, $\theta_{j,l-1}^t = \{b_{j,l-1}^t, \hat{\theta}_j\}$, which consists of j's belief $b_{j,l-1}^t$ and frame $\hat{\theta}_j = \langle A_j, T'_j, \mathcal{R}'_j, \mathcal{Z}_j, O'_j \rangle$. Let $[\phi]$ denote $\{\phi^{t-1} : \phi^t = \phi^{t-1} + \delta_{s^{t-1}s^t}^{a^{t-1}}\}$ and $[\psi_k]$ denote $\{\psi_k^{t-1} : \psi_k^t = \psi_k^{t-1} + \delta_{s^t z_k^t}^{a^{t-1}}\}$ for $k \in \{i, j\}$. Recall that $\delta_{s^{t-1}s^t}^{a^{t-1}}$ and $\delta_{s^t z_k^t}^{a^{t-1}}$ are each a vector of zeros with a 1 for the counts that correspond to the "$s^{t-1} \rightarrow s^t$" transition and the "$s^t \rightarrow z_k^t$" observation respectively. The conditions $[\phi]$ and $[\psi_k]$ ensure that ϕ and ψ_k are properly incremented by the state transition or observation that manifests after each action. Subsequently:

$$T'_i(s'^{t-1}, a^{t-1}, s'^t) = T_\phi^{s^{t-1}a^{t-1}s^t} O_{\psi_i}^{s^t a^{t-1} z_i^t} O_{\psi_j}^{s^t a^{t-1} z_j^t} \quad (4)$$

$$O'_i(s'^{t-1}, a^{t-1}, s'^t, z_i^{t-1}) = 1 \quad (5)$$

$$O'_j(s'^{t-1}, a^{t-1}, s'^t, z_j^{t-1}) = 1 \quad (6)$$

if conditions $[\phi]$, $[\psi_i]$, $[\psi_j]$ hold, and 0 otherwise. As in I-POMDPs, the *joint* transition function T' can differ between agents, reflecting different degrees of knowledge about the environment. In contrast, the *individual* observation function O' is deterministic; it is parametrized by the previous *and* current augmented states. Lastly, the *individual* reward function is $\mathcal{R}'_i((s^t, \phi^t, \psi_i^t, \psi_j^t), a^t) = \mathcal{R}_i(s^t, a^t)$.

At each time step, agent i maintains beliefs over the augmented interactive states, which include agent j's beliefs ($b_{j,l-1}^t$) and frame ($\hat{\theta}_j$). As part of the l-th level belief update, i invokes j's $(l-1)$-th level belief update:

$$\tau_{\theta_{j,l-1}^t}(b_{j,l-1}^{t-1}, a_j^{t-1}, z_j^t, b_{j,l-1}^t) = P(b_{j,l-1}^t|b_{j,l-1}^{t-1}, a_j^{t-1}, z_j^t)$$

inducing recursion. The recursion ends at $l = 0$ when the BA-IPOMDP reduces to a BA-POMDP. Unlike the usual I-POMDP assumption of static frames, the BA-IPOMDP can estimate components of these frames, so they need not be completely fixed.

Theorem 1. *The belief update for the BA-IPOMDP*

$$\langle \mathcal{IS}'_{i,l}, \mathcal{A}, T'_i, \mathcal{R}'_i, \mathcal{Z}_i, O'_i \rangle$$

is:

$$b_{i,l}^t(is'^t_{i,l}) = \beta \sum_{**} b_{i,l}^{t-1}(is'^{t-1}_{i,l}) \sum_{a_j^{t-1}} P(a_j^{t-1}|\theta_{j,l-1}^{t-1}) \\ T_{\phi^{t-1}}^{s^{t-1}a^{t-1}s^t} O_{\psi_i^{t-1}}^{s^t a^{t-1} z_i^t} O_{\psi_j^{t-1}}^{s^t a^{t-1} z_j^t} \\ \tau_{\theta_{j,l-1}^t}(b_{j,l-1}^{t-1}, a_j^{t-1}, z_j^t, b_{j,l-1}^t) \quad (7)$$

where $**$ represents the set of interactive states $is'^{t-1}_{i,l}$ such that $[\phi]$, $[\psi_i]$, $[\psi_j]$ hold.

Proof. (Sketch) Start with Bayes' theorem and expand:

$$b^t_{i,l}(is^{'t}_{i,l}) = P(is^{'t}_{i,l}|z^t_i, a^{t-1}_i, b^{t-1}_{i,l}) = \frac{P(is^{'t}_{i,l}, z^t_i|a^{t-1}_i, b^{t-1}_{i,l})}{P(z^t_i|a^{t-1}_i, b^{t-1}_{i,l})}$$
$$= \beta \sum_{is^{'t-1}_{i,l}} b^{t-1}_{i,l}(is^{'t-1}_{i,l}) P(is^{'t}_{i,l}, z^t_i|a^{t-1}_i, is^{'t-1}_{i,l})$$
$$= \beta \sum_{is^{'t-1}_{i,l}} b^{t-1}_{i,l}(is^{'t-1}_{i,l}) \sum_{a^{t-1}_j} P(a^{t-1}_j|\theta^{t-1}_{j,l-1})$$
$$P(z^t_i|is^{'t}_{i,l}, a^{t-1}, is^{'t-1}_{i,l}) P(is^{'t}_{i,l}|a^{t-1}, is^{'t-1}_{i,l})$$

From $is^{'t}_{i,l}$ and $is^{'t-1}_{i,l}$, the count vectors, $(\psi^{t-1}_i, \psi^{t-1}_j)$ and (ψ^t_i, ψ^t_j), determine the agents' observations (thus eliminating the need to marginalize over z^t_j as is traditionally done in the I-POMDP belief update procedure).

Under $[\phi]$, $[\psi_i]$ and $[\psi_j]$, the following terms simplify:

$$P(z^t_i|is^{'t}_{i,l}, a^{t-1}, is^{'t-1}_{i,l}) = 1$$
$$P(is^{'t}_{i,l}|a^{t-1}, is^{'t-1}_{i,l}) =$$
$$P(b^t_{j,l-1}|s^{'t}, \hat{\theta}^t_j, a^{t-1}, is^{'t-1}_{i,l}) T'_i(s^{'t-1}, a^{t-1}, s^{'t})$$
$$P(b^t_{j,l-1}|s^{'t}, \hat{\theta}^t_j, a^{t-1}, is^{'t-1}_{i,l}) = \tau_{\theta^t_{j,l-1}}(b^{t-1}_{j,l-1}, a^{t-1}_j, z^t_j, b^t_{j,l-1})$$

Furthermore, $T'_i(s^{'t-1}, a^{t-1}, s^{'t})$ can be simplified by Equation (4), where the constituent expected probabilities T_ϕ and O_{ψ_k} can be computed from the counts in the same fashion as Equations (1) and (2).

4 Solving BA-IPOMDPs

For a problem with $|\mathcal{S}|$ physical states and $|\mathcal{Z}|$ observations, an I-POMDP formulation for two symmetric agents will contain up to $|\mathcal{S}|^2$ (non-interactive) states, and a single-agent BA-POMDP formulation will contain $|\mathcal{S}|\binom{|z|+|\mathcal{Z}|-1}{|\mathcal{Z}|-1}$ possible augmented states, where $|z|$ is the total number of observations received during the episode. The corresponding (one-level) BA-IPOMDP formulation will contain up to $|\mathcal{S}|^2\binom{|z_i|+|\mathcal{Z}_i|-1}{|\mathcal{Z}_i|-1}^2$ augmented (non-interactive) states, which is exponentially larger than either of the previous quantities. Thus, the extension from either BA-POMDPs or I-POMDPs to BA-IPOMDPs is nontrivial.

4.1 Bayes Adaptive Interactive Particle Filter

To perform inference on BA-IPOMDPs, we approximate the belief as a set of samples over the augmented interactive states, and update the agent's beliefs via an extension of the interactive particle filter (I-PF) (Doshi and Gmytrasiewicz 2005). Our algorithm, the *Bayes-Adaptive interactive particle filter* (BA-IPF), is presented in Figure 1. Each sample is a possible interactive state. For each sample, an approximate value iteration algorithm, using a sparse reachability tree (to be explained in the next subsection), is applied to compute the set of approximately optimal actions for the opposing agent. Each action in this set is then uniformly weighted so each of these actions is equally likely to be sampled. Then, enumerating over physical states from the next time step, we sample the opposing agent's action and, for all possible opposing agent's observations, we apply this action to update the model of the opposing agent and the counts in the sample. As part of updating the opposing agent's model, which includes its belief, the procedure recurses until level one is reached, when the standard BA-POMDP update is invoked instead. After the samples are propagated forward in time as prescribed, we weigh the samples and normalize them so their weights sum to one. Lastly, the samples are resampled with replacement to avoid sample degeneracy.

Function BA-IPF($\tilde{b}^{t-1}_{k,l}, a^{t-1}_k, z^t_k, l > 0$)
returns $\tilde{b}^t_{k,l}$

$\tilde{b}^{tmp}_{k,l} \leftarrow \emptyset, \tilde{b}^t_{k,l} \leftarrow \emptyset$
Importance Sampling
for all $is^{(n),t-1}_k = \langle (s^{t-1}, \phi^{t-1}, \psi^{t-1}_k, \psi^{t-1}_{-k})^{(n)}, \theta^{(n),t-1}_{-k} \rangle \in \tilde{b}^{t-1}_{k,l}$ **do**
  $P(A_{-k}|\theta^{(n),t-1}_{-k}) \leftarrow$ APPROXPOLICY($\theta^{(n),t-1}_{-k}, l-1$)
  for all $s^t \in S$ **do**
    Sample $a^{t-1}_{-k} \sim P(A_{-k}|\theta^{(n),t-1}_{-k})$
    for all $z^t_{-k} \in Z_{-k}$ **do**
      if $l = 1$ **then**
        $\tilde{b}^{(n),t}_{-k,0} \leftarrow$ BAPOMDP-UPDATE($\tilde{b}^{(n),t-1}_{-k,0}, a^{t-1}_{-k}, z^t_{-k}$)
        $\theta^{(n),t}_{-k} \leftarrow \langle \tilde{b}^{(n),t}_{-k,0}, \hat{\theta}^{(n)}_{-k} \rangle$
        $is^{(n),t}_k \leftarrow \langle (s^t, \phi^{t-1} + \delta^{a^{t-1}}_{s^{t-1}s^t}, \psi^{t-1}_k + \delta^{a^{t-1}}_{s^t z^t_k}, \psi^{t-1}_{-k} + \delta^{a^{t-1}}_{s^t z^t_{-k}})^{(n)}, \theta^{(n),t}_{-k} \rangle$
      else
        $\tilde{b}^{(n),t}_{-k,l-1} \leftarrow$ BA-IPF($\tilde{b}^{(n),t-1}_{-k,l-1}, a^{t-1}_{-k}, z^t_{-k}, l-1$)
        $\theta^{(n),t}_{-k} \leftarrow \langle \tilde{b}^{(n),t}_{-k,l}, \hat{\theta}^{(n)}_{-k} \rangle$
        $is^{(n),t}_k \leftarrow \langle (s^t, \phi^{t-1} + \delta^{a^{t-1}}_{s^{t-1}s^t}, \psi^{t-1}_k + \delta^{a^{t-1}}_{s^t z^t_k}, \psi^{t-1}_{-k} + \delta^{a^{t-1}}_{s^t z^t_{-k}})^{(n)}, \theta^{(n),t}_{-k} \rangle$
      end if
      $w^{(n)}_t \leftarrow T^{s^{t-1}a^{t-1}s^t}_{\phi^{t-1}} O^{s^t a^{t-1} z^t_k}_{\psi^{t-1}_k} O^{s^t a^{t-1} z^t_{-k}}_{\psi^{t-1}_{-k}}$
      $\tilde{b}^{tmp}_{k,l} \overset{\cup}{\leftarrow} (is^{(n),t}_k, w^{(n)}_t)$
    end for
  end for
end for
Normalize all $w^{(n)}_t$ so that $\sum_{n=1}^{N \times |S| \times |Z_{-k}|} w^{(n)}_t = 1$
Selection
Resample with replacement N particles from the set $\tilde{b}^{tmp}_{k,l}$ according to the importance weights; store these (unweighted) samples as $is^{(n),t}_k, n = 1, \ldots, N$
$\tilde{b}^t_{k,l} \leftarrow is^{(n),t}_k, n = 1, \ldots, N$
return $\tilde{b}^t_{k,l}$

Figure 1: The BA-IPF algorithm for approximate BA-IPOMDP belief update. n is the particle index and k is the agent index. If k denotes i, then $-k$ denotes j, and vice versa.

Compared to I-PF, the BA-IPF consists of additional steps

to update the counts (Lines 10 and 14). We have also chosen to enumerate the physical states (Line 4) instead of sampling, to achieve higher accuracy for our test problem where the small number of physical states allowed this to be feasible. As a result, our samples of interactive states are weighted by the BA-IPOMDP transition function (Line 16) instead of the uniform weighting suggested in (Ross, Chaibdraa, and Pineau 2007) in which physical states were sampled rather than enumerated.

4.2 Reachability Tree Sampling

While I-PF addresses the curse of dimensionality due to the complexity of the belief state, the curse of history can also be problematic, because the search space for policies increases with the horizon length. To perform this policy search, a reachability tree is constructed, and with increasing horizons, this tree grows exponentially to account for every possible sequence of actions and observations. To address this issue, (Doshi and Gmytrasiewicz 2005) proposed reachability tree sampling (RTS) as a way to reduce the tree branching factor. In RTS, observations are sampled according to $z_k^t \sim P(\mathcal{Z}_k | a_k^{t-1}, \tilde{b}_{k,l}^{t-1})$ and a partial reachability tree is built based on the sampled observations and the complete set of actions.

In solving BA-IPOMDPs, the curse of history requires approximations beyond the standard RTS to address an *additional* computational bottleneck: the construction of the opposing agent's reachability tree. In order for agent k to behave optimally, it must anticipate what action $-k$ might take; thus, in solving for k's optimal policy, it must also construct $-k$'s reachability tree and use it to find $-k$'s optimal action. As the tree size grows as $\mathcal{O}((|\mathcal{A}_{-k}||\mathcal{Z}_{-k}|)^l)$, it becomes large quickly. Consequently, we follow (Ng et al. 2010) to prune the opposing agent's reachability tree in addition to the agent's reachability tree.

5 Empirical Results

In our evaluation, we are interested in (1) the effect of the approximate belief update from BA-IPF, and (2) the effect of learning. We applied the multiagent Tiger problem (Doshi and Gmytrasiewicz 2009) to study these effects. In our experiments, we limit the nesting to one level and the planning horizon to two. For each scenario, we solved both agents as level-1 BA-IPOMDPs (since each models the other agent) independently and present results obtained from simulating their behaviors against each other. Our experiments were performed on a 2.53GHz dual quad core Intel Xeon processor with 24GB of RAM.

In what follows, Sections 5.1 and 5.2 present results for learning the observation probabilities, which entails estimating the 12 observation probabilities associated with the joint action of $\langle$*Listen*, *Listen*$\rangle$ (six for *TigerLeft* and six for *TigerRight*). Section 5.3 briefly discusses the results for learning the transition probabilities concurrently. This involves estimating the 32 probabilities associated with either agent opening either door (16 for *TigerLeft* and 16 for *TigerRight*).

# Particles	Avg. Reward	Final KL Div.	Avg. Time (s)
4	-15.9	1.21	0.22
8	-6.3	0.52	1.09
16	-1.8	0.26	4.53

Table 1: Comparison of average rewards, final KL divergence, and average overall time, for varying numbers of particles.

Scenario	Agent 0		Agent 1	
	Self	Opp.	Self	Opp.
1	Learn	Correct	Correct	Correct
2	Learn	Learn	Correct	Correct
3	Learn	Correct	Learn	Correct
4	Learn	Incorrect	Learn	Incorrect
5	Learn	Learn	Learn	Learn

Table 2: Parameter learning scenarios. The uniform distribution is used as (1) the prior for when the agent is learning, and (2) the static incorrect parameter for when the agent is not learning.

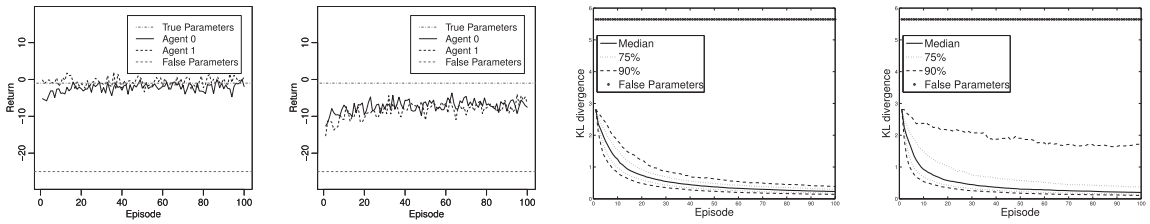
5.1 Analysis of BA-IPF

The quality of approximation in BA-IPF is parametrized by the number of particles. Table 1 shows, as a function of particle number, Agent 0's (1) average reward per episode; (2) KL divergence between the actual and estimated observation distributions at the end of each episode; and (3) average time for planning and execution per episode. The results are averaged over 200 simulations of 100 episodes each. In this scenario, Agent 1 is using the correct observation probabilities and both agents are using the correct observation probabilities for their respective opponent models. Hence, Agent 0 is only learning its own observation probabilities. The prior for Agent 0's observation probabilities is set to uniform.
In general, as the number of particles increases, reward and time increase while KL divergence decreases. For subsequent experiments, the number of particles is set to 16.

5.2 Analysis of Observation Parameter Learning

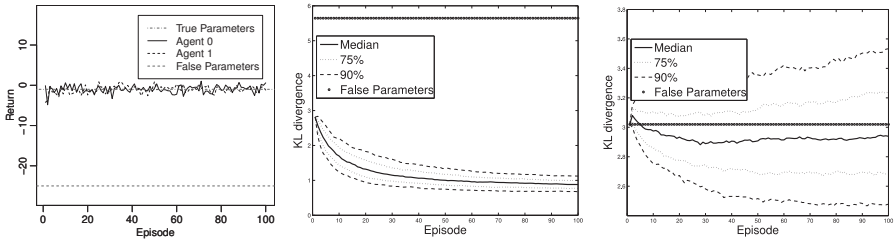
In a given simulation of the two-agent Tiger game, parameter learning can occur for each agent and/or the agent's model of its opponent. Furthermore, when an agent is not learning, parameter values can be set correctly to the actual values or incorrectly to the uniform distribution. We explored a variety of simulation scenarios and report on the select ones that show interesting trends (cf. Table 2). In each scenario, we have two agents, each of which could either be (1) not learning and using correct model parameters; (2) learning its own parameters while assuming correct or incorrect values for its opponent's parameters; or (3) learning both its own and the opponent's parameters. In these scenarios, learning takes place only over the observation probabilities, which uses the uniform distribution as the prior.

One notable trend is that agents accrue less rewards when they are both learning compared to when only when one is learning. This is shown by comparing Scenario 1, in which only Agent 0 is learning, with Scenario 3, where both agents are learning. Figure 2 shows the results for these two scenarios, averaged over 500 simulations of 100 episodes each. It also shows the "baseline" *static* performance, for when the



(a) Scenario 1: Rewards (b) Scenario 3: Rewards (c) Scenario 1: KL div (d) Scenario 3: KL div

Figure 2: Plots comparing agent rewards (Figures 2(a) and 2(b)) and KL divergences (Figures 2(c) and 2(d)) in Scenarios 1 and 3.



(a) Agent rewards (b) Observation KL divergences (c) Transition KL divergences

Figure 3: Plots of agent rewards and observation/transition KL divergences for a case of biased doors.

agents are not learning, but are using either the correct or incorrect observation probabilities. Each baseline is averaged over 50,000 simulations (here episodes have no significance as no learning occurs).

For Scenario 1, Figure 2(a) shows the learning agent—Agent 0—to be initially at a disadvantage, accruing smaller returns than its opponent. This gap in performance is narrowed over time, reflecting the positive effects of successful learning. This is further supported by Figure 2(c), which reveals a dramatic reduction in KL divergence for the learning agent.

For Scenario 3, when Agent 1 is also learning, it too yields improved rewards over time, as evidenced in Figure 2(b). The average reward obtained for the two agents is less than the correct parameter baseline, but is still higher than the incorrect parameter baseline, showing benefits provided by the BA-IPOMDP over the I-POMDP with static incorrect parameters. When both agents are learning, each agent's learning success is reduced, as reflected in the KL divergence in Figure 2(d). Only one agent's KL divergence is shown since both exhibited similar convergence. Nonetheless, as shown in the percentile plots, most of the simulations are still in close convergence to the correct parameters.

Figure 4 summarizes the results for all five scenarios from Table 2. First, we see that the boxplots confirm the trend in reward reductions evident in Scenarios 1 and 3. Interestingly, in the case of Scenario 2, this phenomenon manifests itself in a different way. Here, Agent 0 both learns and models its opponent as learning; Agent 1, in contrast, is using the correct parameters for both itself and its opponent model. The result is that Agent 0 experiences a substantial *relative* reward reduction. This highlights another interesting trend we have observed in our various scenario analyses: a "stronger" opponent model leads to higher rewards.

This trend may be explained as follows. In Scenario 1, Agent 0's opponent model possesses the correct parameter values and thus represents a strong opponent. A strong opponent will typically pursue exploitation over exploration, adopting a more aggressive strategy. Thus, Agent 0 will utilize a exploitation-dominant strategy, and will consequently reap higher rewards on average. Contrast this with the "weaker" learning opponent model in Scenario 2, in which Agent 0 opts more for exploration, performing the listening action more frequently to improve parameter estimates (assuming its opponent is likely to do the same). The ultimate result is that Agent 0's average accrued reward is reduced in Scenario 2 compared to Scenario 1.

The boxplot results for Scenarios 4 and 5 illustrate the effect of fixed incorrect parameters versus learned parameters for the opponent model. At least for Agent 0, the comparison between the two scenarios suggests that, for the specific choice of prior (i.e., uniform), an opponent model whose parameters are fixed incorrectly to the prior yields comparable rewards to an opponent model that starts with the prior and evolves with learning.

5.3 Analysis of Joint Parameter Learning

Unfortunately, the multiagent Tiger problem is not very well-suited for analyzing the learning of transition probabilities, because the only state transition is the resetting of the tiger's location to the left or right door with equal probability, triggered by the door opening action. To investigate the impact of learning both transition and observation probabilities, we considered an alternative version of the Tiger problem, one with *biased* doors, in which the tiger resets to the left door with a probability of 0.75 and the right with a probability of 0.25.

Figure 3 presents our results for Scenario 3. (Like before, only one agent's KL divergence is shown since both exhibited similar convergence.) Under this biased-door version of the problem, concurrent learning (of observation and transition probabilities) leads to rewards comparable to the correct parameter baseline. Compared to the learning agent in the unbiased multiagent Tiger problem, this learning agent

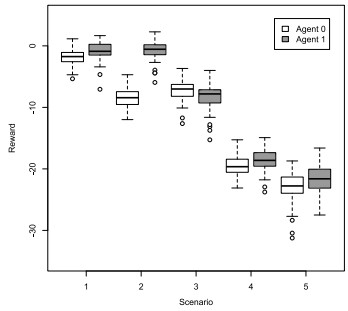


Figure 4: Boxplots of rewards in an episode for each agent and across all 100 simulation episodes. Each boxplot pair represents one of the five scenarios in Table 2.

accrues higher rewards. However, the reduction in observation KL divergence is somewhat less than in the unbiased problem. The transition KL divergence also reflects the fact that this problem does not offer sufficient information to adequately learn the transition probabilities, since the true state is revealed only after a door is opened.

6 Discussion

Our results demonstrate that the BA-IPOMDP framework can successfully be used to model multiple agents with uncertainties about (1) the current state of the world, and (2) their associated transition and observation probabilities. In particular, we observe that the reward for agents employing learning strategies improves over time, suggesting that learning is beneficial (Figure 2); moreover, the speed of this convergence is heavily affected by whether one or more agents are learning. Generally, scenarios in which more learning takes place (see Table 2) take longer to converge than scenarios with less learning, and consequently the average rewards over the same timespan are lower (Figure 4). This effect is also impacted by whether non-learning agents have a fixed and incorrect model of their opponent.

Finally, we note that recent work (Doshi-Velez 2009) has addressed learning in POMDP models where the state space itself is not fully known. This approach uses infinite hidden Markov models in learning the size of the state space. Incorporating such a framework within the BA-IPOMDP would produce more computational challenges, but it might have the benefit of broadening BA-IPOMDP's applicability.

Acknowledgements

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. The authors are tremendously grateful to Prashant Doshi and Stéphane Ross for their helpful advice and starter code; Ekhlas Sonu, Finale Doshi-Velez and anonymous reviewers for valuable feedback on the paper.

References

Bernstein, D.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research* 27(4):819–840.

Busoniu, L.; Babuska, R.; and Schutter, B. D. 2008. A comprehensive survey of multiagent reinforcement learning. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 38(2):156–172.

Doshi, P., and Gmytrasiewicz, P. 2005. Approximating state estimation in multiagent settings using particle filters. In *Proceedings of the 4th AAMAS Conference*, 320–327. Utrecht, the Netherlands: ACM.

Doshi, P., and Gmytrasiewicz, P. 2009. Monte Carlo sampling methods for approximating Interactive POMDPs. *Journal of AI Research* 34:297–337.

Doshi, P., and Perez, D. 2008. Generalized point-based value iteration for interactive POMDPs. In *Proceedings of the 23rd AAAI Conference*, 63–68. Chicago, IL: AAAI.

Doshi, P.; Qu, X.; Goodie, A.; and Young, D. 2010. Modeling recursive reasoning by humans using empirically informed interactive POMDPs. In *Proceedings of the 9th AAMAS Conference*, 1223–1230. Toronto, Canada: ACM.

Doshi-Velez, F. 2009. The infinite partially observable Markov decision process. *Neural Information Processing Systems* 22:477–485.

Doshi, P.; Zeng, Y.; and Chen, Q. 2009. Graphical models for interactive POMDPs: representations and solutions. *Autonomous Agents and Multi-Agent Systems* 18(3):376–416.

Emery-Montemerlo, R.; Gordon, G.; Schneider, J.; and Thrun, S. 2004. Approximate solutions for partially observable stochastic games with common payoffs. In *Proceedings of the 3rd AAMAS Conference*, 136–143. New York: IEEE.

Guo, A., and Lesser, V. 2006. Stochastic planning for weakly coupled distributed agents. In *Proceedings of the 5th AAMAS Conference*, 326–328. Hakodate, Japan: ACM.

Hansen, E.; Bernstein, D.; and Zilberstein, S. 2004. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th AAAI Conference*, 709–715. San Jose, CA: AAAI.

Hoey, J.; von Bertoldi, A.; Poupart, P.; and Mihailidis, A. 2007. Assisting persons with dementia during handwashing using a partially observable Markov decision process. In *Proceedings of the 5th ICVS Conference*.

Kaelbling, L.; Littman, M.; and Cassandra, A. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101:99–134.

Kaelbling, L.; Littman, M.; and Moore, A. 1996. Reinforcement learning: a survey. *Journal of AI Research* 4:237–285.

Melo, F., and Ribeiro, I. 2010. Coordinated learning in multiagent MDPs with infinite state spaces. *Journal of Autonomous Agents and Multiagent Systems* 21(3):321–367.

Ng, B.; Meyers, C.; Boakye, K.; and Nitao, J. 2010. Towards applying interactive POMDPs to real-world adversary modeling. In *Proceedings of the 22nd IAAI Conference*, 1814–1820. Atlanta, GA: AAAI.

Oliehoek, F. A. 2012. Decentralized POMDPs. In Wiering, M., and van Otterlo, M., eds., *Reinforcement Learning: State of the Art*. Springer.

Peshkin, L.; Kim, K.-E.; Meuleau, N.; and Kaelbling, L. P. 2000. Learning to cooperate via policy search. In *Proceedings of the 16th UAI Conference*, 489–496. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Pynadath, D., and Tambe, M. 2002. The communicative multiagent team decision problem: analyzing teamwork theories and models. *Journal of AI Research* 16:389–423.

Ross, S.; Chaib-draa, B.; and Pineau, J. 2007. Bayes-adaptive POMDPs. In *Proceedings of the 21st NIPS Conference*. Vancouver, Canada: MIT Press.

Seuken, S., and Zilberstein, S. 2008. Formal models and algorithms for decentralized decision making under uncertainty. *Autonomous Agents and Multi-Agent Systems* 17(2):1387–2532.

Zhang, C., and Lesser, V. R. 2011. Coordinated multi-agent reinforcement learning in networked distributed POMDPs. In *Proceedings of the 25th AAAI Conference*, 764–770. San Francisco, CA: AAAI.