Exact Solutions of Interactive POMDPs Using Behavioral Equivalence

Bharanee Rathnasabapathy
Dept. of Computer Science
University of Illinois at Chicago
brathnas@cs.uic.edu

Prashant Doshi
Dept. of Computer Science
University of Georgia
pdoshi@cs.uga.edu

Piotr Gmytrasiewicz
Dept. of Computer Science
University of Illinois at Chicago
piotr@cs.uic.edu

ABSTRACT

We present a method for transforming the infinite interactive state space of interactive POMDPs (I-POMDPs) into a finite one, thereby enabling the computation of exact solutions. I-POMDPs allow sequential decision making in multi-agent environments by modeling other agents' beliefs, capabilities, and preferences as part of the interactive state space. Since beliefs are allowed to be arbitrarily nested and are continuous, it is not possible to compute optimal solutions using value iteration as in POMDPs. We present a method that transforms the original state space into a finite one by grouping the other agents' behaviorally equivalent models into equivalence classes. This enables us to compute the *complete* optimal solution for the I-POMDP, which may be represented as a policy graph. We illustrate our method using the multi-agent Tiger problem and discuss features of the solution.

1. INTRODUCTION

Interactive partially observable Markov decision processes (I-POMDPs) [8] provide a framework for sequential decision making in partially observable multi-agent environments. Along with the physical states of the environment they include other agents' computable models in the state space. The models encompass all information influencing the corresponding agent's behavior, which includes its preferences, capabilities, and beliefs. They are analogous to *types* in Bayesian games[10]. This augmented state space is called an *interactive state space*. Because beliefs in I-POMDPs are arbitrarily nested, the interactive state space is not the traditional one – it is a complex infinite space that contains beliefs over others' beliefs.

Traditional methods for solving POMDPs [4, 11, 16] focus on finite state spaces and exploit the piecewise linearity and convexity (PWLC) property of the value function. For POMDPs with infinite state spaces exact closed form solutions exist only in linear Gaussian systems [13, 17]. I-POMDPs, which generalize POMDPs to multi-agent domains, have a infinite state space as well, making it impossible to compute the exact solution by applying value iteration based methods in a straightforward manner.

In this paper, we present a method for computing the exact solutions to finitely nested I-POMDPs introduced in [8]. Our method transforms the infinite interactive state space into a finite space by grouping models of other agents into a finite set of behavioral equivalence classes. They are defined so that the optimal action(s) of the modeled agents are the same for all models inside a class. Given such behavioral equivalence classes, and the agents' optimal actions in these classes, we show that it is sufficient to reason about the other agents' equivalence classes, rather than the infinity of individual models, to compute a solution. When the set of actions and observations of each agent is finite, the set of behavioral equivalence classes for a finite horizon is also finite, and at most countably infinite for an infinite horizon. In practice though the number of behavioral equivalence classes will be finite for the infinite horizon because we usually settle for ϵ-optimal policies. Once we have transformed the interactive state space using the equivalence classes, we modify the I-POMDP belief update and value iteration to operate over the transformed interactive state space. Consequently, we can extend the exact POMDP solution methods in a straightforward manner to solve I-POMDPs exactly.

Related work in the area of multi-agent and multi-body planning like DEC-POMDPs [1, 2] and MTDP [12] only allow us to compute equilibrium based solutions under the assumption of common knowledge of the agents' prior beliefs. Further, dynamic programming approaches to find optimal solutions to DEC-POMDPs [9] are only possible in special co-operative cases where all agents share the same payoff function. In the context of state equivalence and aggregation, Given et al. [7] provide a nice discussion of using stochastic bisimilarity for model minimization in Markov decision processes and Poupart and Boutilier [14] propose an approach to get linear lossy compression of value function in POMDPs. Our method, in contrast, exploits the piecewise linearity and convexity of the value function of agent models to find the behavioral equivalence classes. Additionally, unlike [7, 14] our approach yields an optimal solution. Methods to solve I-POMDPs and DEC-POMDPs approximately have also appeared. In [5, 6], a sampling based approximation technique is presented, while in [3], bounded finite state controllers are used to derive the joint policy.

The rest of the paper is structured in the following manner. In Section 2 we give a brief overview of the I-POMDP framework. Section 3 introduces the idea of behavioral equivalence and its use in solving I-POMDPs exactly. In Section 4, we illustrate our method, using an example problem and discuss several solutions. We conclude in Section 5 and outline some avenues of future work.

2. OVERVIEW OF I-POMDP

Interactive POMDPs extend POMDPs to multi-agent settings by including other agents' models as part of the state space [8]. Since other agents might also be reasoning about others, the interactive

AAMAS'06 May 8–12 2006, Hakodate, Hokkaido, Japan.
Copyright 2006 ACM 1-59593-303-4/06/0005 ...$5.00.

state space can be arbitrarily nested (belief about other agents' beliefs about other agents' beliefs and so on). For simplicity of presentation we will consider an agent, i, that is interacting with only one other agent, j.

DEFINITION 1 (I-POMDP$_{i,l}$). *A finitely nested interactive POMDP of agent i with a belief nesting level l referred to as I-POMDP$_{i,l}$ is defined as the tuple*

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

where,

- $IS_{i,l}$ denotes a set of interactive states defined as, $IS_{i,l} = S \times \{\Theta_{j,l-1} \cup SM_j\}$, for $l \geq 1$, and $IS_{i,0} = S$, where S is the set of states of the physical environment, $\Theta_{j,l-1}$ is the set of computable *intentional models* of agent j at belief nesting level $l-1$: $\theta_{j,l-1} = \langle b_{j,l-1}, \hat{\theta}_{j,l-1} \rangle$ where the *frame*, $\hat{\theta}_{j,l-1} = \langle A, \Omega_j, T_j, O_j, R_j, OC_j \rangle$. Here, j is Bayes rational and OC_j is j's optimality criterion that influences its decision making process. SM_j is the set of sub-intentional models of j. We give a recursive bottom-up construction of the interactive state space below,

$$\begin{aligned} IS_{i,0} &= S, \\ &\quad \Theta_{j,0} = \{\langle b_{j,0}, \hat{\theta}_{j,0} \rangle \mid b_{j,0} \in \Delta(IS_{j,0})\} \\ IS_{i,1} &= S \times \{\Theta_{j,0} \cup SM_j\}, \\ &\quad \Theta_{j,1} = \{\langle b_{j,1}, \hat{\theta}_{j,1} \rangle \mid b_{j,1} \in \Delta(IS_{j,1})\} \\ &\vdots \qquad \vdots \\ IS_{i,l} &= S \times \{\Theta_{j,l-1} \cup SM_j\}, \\ &\quad \Theta_{j,l} = \{\langle b_{j,l}, \hat{\theta}_{j,l} \rangle \mid b_{j,l} \in \Delta(IS_{j,l})\} \end{aligned}$$

- A is the set of joint actions of all agents in the environment, $A = A_i \times A_j$
- T_i describes the effect of the joint actions on the physical states of the environment, $T_i : S \times A \times S \rightarrow [0, 1]$
- Ω_i is the set of observations of agent i
- $O_i : S \times A \times \Omega_i \rightarrow [0, 1]$ gives the likelihood of the observations given the physical state and joint action
- $R_i : IS_i \times A \rightarrow \mathbb{R}$ describes agent i's preferences over it's interactive states. Usually only the physical states will matter

Agent i's policy is the mapping $\Omega_i^* \rightarrow \Delta(A_i)$ where Ω_i^* is the set of all observation histories of agent i. Since belief over the interactive states forms a sufficient statistic [8], the policy can also be represented as a mapping from the set of all belief's of agent i to a distribution over it's actions, $\Delta(IS_i) \rightarrow \Delta(A_i)$.

2.1 Belief Update

Analogous to POMDPs, an agent within the I-POMDP framework also updates its belief as it acts and observes. However, there are two differences that complicate a belief update in multi-agent settings when compared to single-agent ones. First, since the state of the physical environment depends on the actions performed by both agents, – in our example agents i and j – i's prediction of how the physical state changes has to be made based on it's prediction of j's actions. Second, changes in j's model have to be included in i's belief update. Specifically, if j is intentional then an update of j's beliefs due to its action and a new observation has to be included. In other words, i has to update its belief based on it's prediction of what j would observe and how j would update its belief. If j's model is sub-intentional, then j's probable observations are appended to the observation history contained in the model. Formally, following [8], we have:

$$\begin{aligned} &Pr(is^t|a_i^{t-1}, b_{i,l}^{t-1}) = \beta \textstyle\int_{IS^{t-1}:\widehat{m}_j^{t-1}=\widehat{\theta}_j^t} b_{i,l}^{t-1}(is^{t-1}) \\ &\times \textstyle\sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_{j,l-1}^{t-1}) O_i(s^t, a_i^{t-1}, a_j^{t-1}, o_i^t) \\ &\times T_i(s^{t-1}, a_i^{t-1}, a_j^{t-1}, s^t) \textstyle\sum_{o_j^t} O_j(s^t, a_i^{t-1}, a_j^{t-1}, o_j^t) \\ &\times \delta_D(SE_{\widehat{\theta}_j^t}(b_{j,l-1}^{t-1}, a_j^{t-1}, o_j^t) - b_{j,l-1}^t) d(is^{t-1}) \end{aligned} \tag{1}$$

where β is the normalizing constant, δ_D is the Dirac-delta function, $SE(\cdot)$ is an abbreviation for the belief update, and $Pr(a_j^{t-1}|\theta_{j,l-1}^{t-1})$ is the probability that a_j^{t-1} is Bayes rational for the agent described by model $\theta_{j,l-1}^{t-1}$. When j's model is sub-intentional, the integration is over $IS^{t-1} : \widehat{m}_j^{t-1} = \widehat{m}_j^t$, $Pr(a_j^{t-1}|\theta_{j,l-1}^{t-1})$ is replaced with $Pr(a_j^{t-1}|m_{j,l-1}^{t-1})$, and $\delta_D(SE_{\widehat{\theta}_j^t}(b_{j,l-1}^{t-1}, a_j^{t-1}, o_j^t) - b_{j,l-1}^t)$ is replaced with $\delta_K(APPEND(h_j^{t-1}, o_j^t) - h_j^t)$. δ_K is the Kronecker delta, and $APPEND$ returns a string with the second argument appended to the first. If j is also modeled as an I-POMDP, then i's belief update invokes j's belief update (via the term $SE_{\widehat{\theta}_j^t}$ ($b_{j,l-1}^{t-1}$, a_j^{t-1} , o_j^t)), which in turn could invoke i's belief update and so on. This recursion in belief nesting bottoms out at the 0^{th} level. At this level, belief update of the agent reduces to a POMDP belief update. [1] For an illustration of the belief update, additional details on I-POMDPs, and how they compare with other multi-agent planning frameworks, see [8].

2.2 Value Iteration

Each belief state in a finitely nested I-POMDP has an associated value reflecting the maximum payoff the agent can expect in this belief state:

$$\begin{aligned} U^n(\langle b_{i,l}, \widehat{\theta}_i \rangle) = \underset{a_i \in A_i}{max} \Bigg\{ \int\limits_{is \in IS_{i,l}} ER_i(is, a_i) b_{i,l}(is) d(is) + \\ \gamma \sum_{o_i \in \Omega_i} Pr(o_i|a_i, b_{i,l}) U^{n-1}(\langle SE_{\widehat{\theta}_i}(b_{i,l}, a_i, o_i), \widehat{\theta}_i \rangle) \Bigg\} \end{aligned} \tag{2}$$

where, $ER_i(is, a_i) = \sum_{a_j} R_i(is, a_i, a_j) Pr(a_j|m_{j,l-1})$ (since $is = (s, m_{j,l-1})$). Eq. 2 is a basis for value iteration in I-POMDPs.

Agent i's optimal action, a_i^*, for the case of finite horizon with discounting, is an element of the set of optimal actions for the belief state, $OPT(\theta_i)$, defined as:

$$\begin{aligned} OPT(\langle b_{i,l}, \widehat{\theta}_i \rangle) = \underset{a_i \in A_i}{argmax} \Bigg\{ \int\limits_{is \in IS_{i,l}} ER_i(is, a_i) b_{i,l}(is) d(is) \\ + \gamma \sum_{o_i \in \Omega_i} Pr(o_i|a_i, b_{i,l}) U^n(\langle SE_{\widehat{\theta}_i}(b_{i,l}, a_i, o_i), \widehat{\theta}_i \rangle) \Bigg\} \end{aligned} \tag{3}$$

3. BEHAVIORAL EQUIVALENCE

The main idea in this paper is to aggregate interactive states into a finite number of equivalence classes using behavioral equivalence. Instead of reasoning over the infinite set of interactive states, we operate over the finite set of equivalence classes of interactive states (ECIS). In the next subsection, we introduce the concept of behavioral equivalence and show that reasoning over the equivalence classes preserves the optimal policy.

3.1 Defining the Equivalence Classes

In order to illustrate the construction of the behavioral equivalence classes we look at a simple example – the classical *tiger*

[1] The 0^{th} level model is a POMDP: Other agent's actions are treated as exogenous events and folded into the T, O, and R functions.

problem introduced in [11]. The problem is that of an agent having to open either of two doors. Behind one of the doors is a tiger waiting to eat the agent and behind the other is a pot of gold. There is a reward of +10 to get the gold and -100 when the agent is eaten by the tiger. There are two states signifying the tiger's location behind the left (*TL*) door and the right (*TR*) door. The agent has three actions: open left door (*OL*), open right door (*OR*) and listen (*L*). Listening always incurs a cost of 1. When the agent listens, it can hear a growl on the left (*GL*) or right (*GR*) with 85% certainty. The value function gives the value of performing the optimal plan at any given belief. In Fig. 1, the value function is the envelope that provides the maximum value of any plan over all beliefs. We note that the agent opens the right door if it believes the probability that the tiger is behind the right door, *P(TR)*, is less than 0.1. It will listen if 0.1 < *P(TR)* < 0.9 and open left door if *P(TR)* > 0.9.

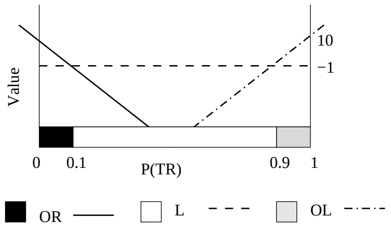


Figure 1: **Horizon-1 value function in the tiger game and the belief ranges corresponding to different optimal actions.**

In Fig. 1 we observe that each of the optimal plans spans over multiple belief points. For example, opening the right door is the optimal action for all beliefs in the set [0–0.1), i.e. OPT(P(TR)= [0–0.1)) = OR. Thus, beliefs in the set [0–0.1) are equivalent to each other in that they induce the same optimal behavior. We call such beliefs *behaviorally equivalent.* [2] The collection of the equivalence classes (five of them in the example in Fig. 1) forms a partition of the belief space. For finite horizons, and a finite number of actions and observations, the number of distinct plans and therefore the equivalence classes is also finite. Note that each equivalence class is a node in the policy graph in a POMDP.

We use the above insight to discretize the infinite interactive state space of I-POMDPs into a finite one. The optimal value function in I-POMDPs, like POMDPs, is also piecewise linear and convex (PWLC) [8]. Recall that the interactive state space, $IS_{i,l} = S \times \{\Theta_{j,l-1} \cup SM_j\}$, where $\Theta_{j,l-1}$ is the set of intentional models of agent j that includes its beliefs, and all other parameters sufficient to determine its behavior. SM_j is the set of j's sub-intentional models. We transform the infinite space into a finite one by grouping together the behaviorally equivalent models into the equivalence classes. As we mentioned before, for a finite horizon and finite number of actions and observations, there exist a finite number of equivalence classes, thereby facilitating the application of standard POMDP solution methods for solving I-POMDPs.

We point out that the method scales well to higher levels of nesting. Since at each level of nesting we operate over the equivalence classes only, we need not concern ourselves with the complex space of beliefs over beliefs of other agents and so on. Of course, in deriving the equivalence classes we must solve all lower level models in a recursive manner. We formally define the equivalence classes below:

DEFINITION 2 (ECIS). *Equivalence classes of interactive states (ECIS$_{i,l}$) of an agent i is a* partition *of $IS_{i,l}$ such that the behavior of agent j is the same for all interactive states in a given subset.*

$$\begin{aligned} ECIS_{i,l} = \quad & \{ \quad (s, M_{j,k}^{l-1}) \mid M_{j,k}^{l-1} \subseteq \{\Theta_{j,l-1} \cup SM_j\} and \\ & \forall_{m_{j,l-1}, m'_{j,l-1} \in M_{j,k}^{l-1}} OPT(m_{j,l-1}) = OPT(m'_{j,l-1}) \\ & \} \end{aligned}$$

Here, $M_{j,k}^{l-1}$, is the kth equivalence class in the partition. Note that OPT(.) takes a model of the other agent as a parameter and gives the optimal action of the agent (see Eq. 3).

3.2 Belief Update and Value Iteration over ECIS

The main difference between the belief update as defined in Eq. 1 over the original interactive state space and the belief update over the transformed one is in the manner in which agent i updates j's models. Since j's model space is partitioned into equivalence classes, i must update j's equivalence class based on j's action and observation. To understand how this is done, let us consider the simple case where i is uncertain only about j's beliefs. In this case, each equivalence class is a disjoint subset of j's beliefs that induces identical optimal behavior.

As we mentioned before, the optimal solution to an I-POMDP (and a POMDP) provides a conditional plan or a policy graph that dictates the action(s) that agent j must perform given the equivalence class of its beliefs. Agent i updates j's equivalence class by locating the node for the current equivalence class in j's policy graph, and using j's predicted action and possible observation to trace to the next one. Before presenting the belief update equations, we note that i's belief over the transformed interactive state space forms a sufficient statistic.

PROPOSITION 1. *(Sufficiency) In a finitely nested I-POMDP, a probability distribution over $ECIS_{i,l}$, $\tilde{b}_{i,l} \in \Delta(ECIS_{i,l})$, provides a sufficient statistic for the past history of i's observations.*

We outline the proof of this proposition in the Appendix. We give the belief update over ECIS for an arbitrary belief nesting level l below:

$$\begin{aligned} & Pr(ecis^t | o_i^t, a_i^{t-1}, \tilde{b}_{i,l}^{t-1}) = \beta \textstyle\sum_{ECIS^{t-1}} \tilde{b}_{i,l}^{t-1}(ecis^{t-1}) \\ & \times \textstyle\sum_{a_j^{t-1}} Pr(a_j^{t-1} \mid M_{j,k}^{t-1}) \, O_i(s^t, a^{t-1}, o_i^t) \\ & \times \; T_i(s^{t-1}, a^{t-1}, s^t) \textstyle\sum_{o_j^t} \tilde{\tau}(M_{j,k}^{t-1}, a_j^{t-1}, o_j^t, M_{j,k}^t) \\ & \times \; O_j(s^t, a^{t-1}, o_j^t) \end{aligned} \tag{4}$$

where β is the normalizing constant. Agent i's update of j's equivalence classes is captured using the $\tilde{\tau}$ operator which returns 1 if $Pr(M_{j,k}^t | M_{j,k}^{t-1}, a_j, o_j) = 1$, and 0 otherwise. The $\tilde{\tau}$ operator uses j's policy graph to determine the updated equivalence class. Note that $M_{j,k}^{t-1}$ contains j's equivalent models before j performs it's belief update and $M_{j,k}^t$ contains models that are equivalent after j's belief update.

Since i's beliefs are now distributions over the transformed interactive state space, the value iteration, as shown in Eq. 2 is revised to include the value function defined over the transformed belief space:

$$\begin{aligned} \tilde{U}^n(\langle \tilde{b}_{i,l}, \widehat{\theta}_i \rangle) = \underset{a_i \in A_i}{max} & \Bigg\{ \sum_{ecis \in ECIS_{i,l}} ER_i(ecis, a_i) \tilde{b}_{i,l}(ecis) \\ \times \gamma \sum_{o_i \in \Omega_i} & Pr(o_i | a_i, \tilde{b}_{i,l}) \tilde{U}^{n-1}(\langle SE_{\widehat{\theta}_i}(\tilde{b}_{i,l}, a_i, o_i), \widehat{\theta}_i \rangle) \Bigg\} \end{aligned} \tag{5}$$

[2] It is possible that some plans may be optimal for only a single belief. In this case, the equivalence class would be a singleton set containing the single belief.

In the next proposition, we show that the discretization of the interactive state space is a *lossless* one. In other words, for a given belief over the original interactive state space, the value of the optimal plan (and the optimal plan itself) remains unchanged for the corresponding belief over the discretized interactive state space. We define the concept of a value preserving transformation and formalize our result in Proposition 2.

DEFINITION 3 (VALUE PRESERVING TRANSFORMATION). *Define a transformation $F : B_1 \rightarrow B_2$, where B_1 is the space of original beliefs, and B_2 is the space of corresponding transformed beliefs. Let $Q : N \times B_1 \times A \rightarrow \mathbb{R}$ and $\tilde{Q} : N \times B_2 \times A \rightarrow \mathbb{R}$ be the N horizon action-value functions. F is said to be value preserving iff $\forall_{n \in [1,N], b \in B_1, a \in A} Q^n(b,a) = \tilde{Q}^n(F(b),a)$. In other words, the value of the transformed belief and action pair is same as the value of the original belief and action, for all horizons.*

PROPOSITION 2. *Define a mapping $CP : \Delta(IS_{i,l}) \longrightarrow \Delta(ECIS_{i,l})$ in a finitely nested I-POMDP such that,*

$$\tilde{b}_i(s, M^{l-1}_{j,k}) = \int_{m_j \in M^{l-1}_{j,k}} b_i(s, m_j) dm_j \qquad (6)$$

where $b_i \in \Delta(IS_{i,l})$, $\tilde{b}_i \in \Delta(ECIS_{i,l})$, and $M^{l-1}_{j,k}$ is some equivalence class. The mapping CP is value preserving.

The proof of proposition 2 is given in the Appendix.

COROLLARY 1. *The optimal policy of the transformed finitely nested I-POMDP remains unchanged.*

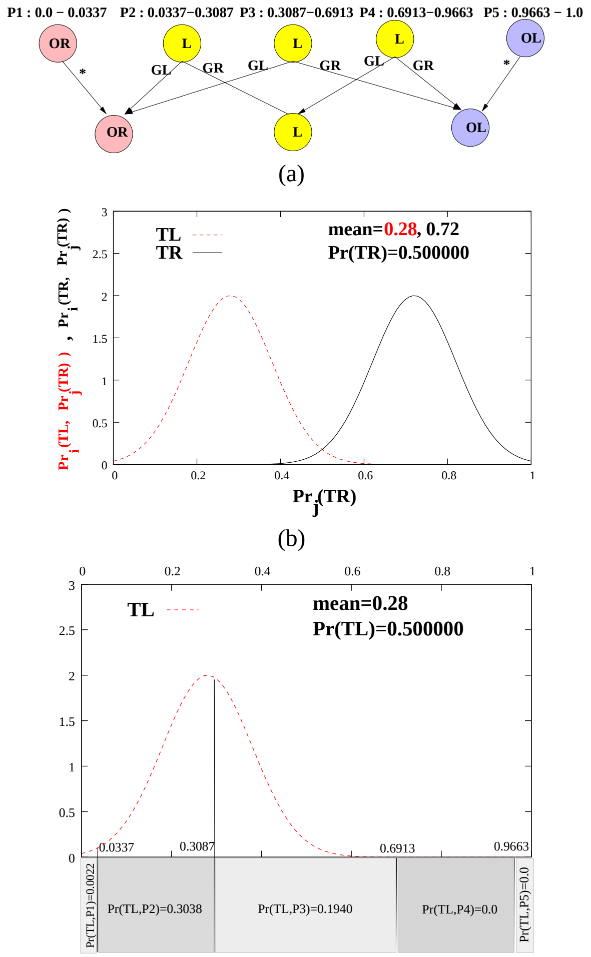
PROOF. *Proof of this corollary is inductive, and involves a straightforward application of Proposition 2. From Proposition 2, CP is a value-preserving transformation. For the basis step – horizon is 1 – $\forall_{b_i,a_i} Q^1(b_i,a_i) = \tilde{Q}^1(CP(b_i),a_i)$. Therefore, the optimal single step action remains unchanged. The inductive proof then follows for any arbitrary horizon n. Since $\forall_{b_i,a_i} Q^n(b_i,a_i) = \tilde{Q}^n(CP(b_i),a_i)$, therefore $\forall_{b_i \in \Delta(IS_i)} OPT(b_i) = \tilde{OPT}(CP(b_i))$, the optimal policy remains the same.* □

4. MULTI-AGENT TIGER PROBLEM

To illustrate our method we use a slightly modified version of the multi-agent tiger problem discussed in [8] (based on [15].) The problem has two agents, each of which can open the right door (OR), the left door (OL) or listen (L). In addition to observing growls when they listen, the agents can also hear creaks (creak on the left (CL), creak on the right (CR) or silence (S), i.e no creaks) indicating the other agent's opening one of the doors. When any door is opened, the tiger's location is switched with a probability of 5%. The agent's preferences are as in the single agent game discussed in Section 3. The transition, reward and observation functions are shown in Table 1. From the observation functions we see that, in this case, agent i hears growls with a reliability of 65% and creaks with a reliability of 95%. Agent j, on the other hand, hears growls with a reliability of 95%. Thus, the setting is such that agent i can hear agent j opening doors more reliably than the tiger's growls. This suggests that i could use j's actions as an indication of the location of the tiger, as we discuss below.

For the sake of simplicity we assume agent i's I-POMDP to be singly nested and that agent i ascribes only intentional models to j. Additionally, we assume that i is uncertain only about j's beliefs and not j's frame. Note that in a general I-POMDP, i can be uncertain about all parameters of j's model. In Fig. 2(a) we give the optimal solution of the level 0 I-POMDP of agent j. Since agent j's belief is not nested, j reasons about the tiger's location only and not about agent i's beliefs. Therefore, agent j's actions depend on hearing growls alone. The transformed horizon-2 interactive state space is: $ECIS_{i,1} = S \times \{P1, P2, P3, P4, P5\}$, where the classes $P1 - P5$ are shown in Fig. 2(a). Since $ECIS_{i,1}$ is finite, we can obtain the *complete* optimal value function using the modified value iteration shown in Eq. 5.

In the remainder of this section we visualize i's value function



(a)

(b)

(c)

Figure 2: (a) Agent j's horizon 2 policy showing optimal plans given different prior beliefs. Since j's belief is not nested, its prior belief is given as the probability that the tiger is behind the right door. (b) An example of agent i's level-1 belief. Note that we have 2 distributions, one when the tiger is on the left and other when tiger is on the right. In the belief shown, i thinks that j is likely to be somewhat informed about the tiger's true location. (c) Shows how agent i's belief $b_{i,1} \in \Delta(IS_{i,1})$ (at 2 steps to go) is transformed into a belief $\tilde{b}_{i,1} \in \Delta(ECIS_{i,1})$. For example, the probability that the tiger is behind the left door and j's belief is in region $P2$ is given as $\tilde{b}_{i,1}(TL, P2) = \int_{0.0337}^{0.3087} b_{i,1}(TL, Pr_j(TR))dPr_j(TR) = 0.3038$, where $b_{i,1}(TL, Pr_j(TR))$ is the Gaussian pdf describing i's belief about j's belief when the tiger is on the left. $\int_0^1 b_{i,1}(TL, Pr_j(TR))dPr_j(TR)$ gives i's belief that the tiger is behind the left door. Similarly, $\tilde{b}_{i,1}(TR, P1), \tilde{b}_{i,1}(TR, P2)...\tilde{b}_{i,1}(TR, P5)$ are computed from the Gaussian describing i's belief about j's belief when the tiger is behind the right door.

(a) Transition function for i and j

Action <i,j>	State	TL	TR
<OL,*> , <*,OL>	TL	0.95	0.05
<OL,*> , <*,OL>	TR	0.05	0.95
<OR,*> , <*,OR>	TL	0.95	0.05
<OR,*> , <*,OR>	TR	0.05	0.95
<L,*> , <*,L>	TL	1.0	0.0
<L,*> , <*,L>	TR	0.0	1.0

(b) Reward function for i and j

Action <i,j>	TL	TR
<OL,OL>	-100,-100	10,10
<OL,OR>	-100,10	10,-100
<OL,L>	-100,-1	10,-1
<OR,OL>	10,-100	-100,10
<OR,OR>	10,10	-100,-100
<OR,L>	10,-1	-100,-1
<L,OL>	-1,-100	-1,10
<L,OR>	-1,10	-1,-100
<L,L>	-1,-1	-1,-1

(c) Observation function for i with a growl reliability of 65% and creak reliability of 95%

Action <i,j>	State	<GL,CL>	<GL,CR>	<GL,S>	<GR,CL>	<GR,CR>	<GR,S>
<OL,*>	*	1/6	1/6	1/6	1/6	1/6	1/6
<OR,*>	*	1/6	1/6	1/6	1/6	1/6	1/6
<L,OL>	TL	0.65*0.95	0.65*0.025	0.65*0.025	0.35*0.95	0.35*0.025	0.35*0.025
<L,OL>	TR	0.35*0.95	0.35*0.025	0.35*0.025	0.65*0.95	0.65*0.025	0.65*0.025
<L,OR>	TL	0.65*0.025	0.65*0.95	0.65*0.025	0.35*0.025	0.35*0.95	0.35*0.025
<L,OR>	TR	0.35*0.025	0.35*0.95	0.35*0.025	0.65*0.025	0.65*0.95	0.65*0.025
<L,L>	TL	0.65*0.025	0.65*0.025	0.65*0.95	0.35*0.025	0.35*0.025	0.35*0.95
<L,L>	TR	0.35*0.025	0.35*0.025	0.35*0.95	0.65*0.025	0.65*0.025	0.65*0.95

(d) Observation function for j

Action <i,j>	State	GL	GR
<*, OL>	*	0.5	0.5
<*, OR>	*	0.5	0.5
<OL, *>	*	0.5	0.5
<OR, *>	*	0.5	0.5
<L, L>	TL	0.95	0.05
<L, L>	TR	0.05	0.95

Table 1: The transition, reward and observation functions of agents i and j

and discuss i's behavior in the presence of j. The value function is defined over a ten dimensional space since agent j has five behavioral equivalence classes and there are two physical states. While we can compute the complete value function and policy, it is not possible to visualize it. Hence, here we plot agent i's value function for a subset of beliefs it has about agent j.

An example of agent i's belief about the location of the tiger and about j's belief is shown in Fig. 2(b). We represent agent i's belief about its interactive state as a pair of Gaussian densities with a variance of 0.01. One Gaussian captures i's belief about j's belief when the tiger is on the left, and the other captures i's belief about j's belief when the tiger is on the right (see [8] for a more detailed exposition on representing nested beliefs in I-POMDPs). To enable visualization, we capture i's belief about j's using a single parameter. To do this, we only look at those beliefs of i in which the mean of the two probability densities sum to 1, as do the densities shown in Fig. 2(b). The value function is plotted against agent i's belief about the tiger's location ($Pr_i(TR)$) and the mean of the probability density function (describing $Pr_i(Pr_j(TR))$ given the tiger is on the left. Note that in considering the particular form of i's beliefs, we have reduced the dimensions over which the value function spans in a non-linear fashion, and hence the value function plot is not linear.

Figures 3(a), 3(b), 3(c) and 3(d) show the projections of i's horizon 2 value functions on the belief space of agent i. We varied the reliability of the creaks, in order to explore its effect on i's policy. Figure 3(e) describes the optimal policies for the corresponding belief regions.

When creaks are not very reliable i's behavior is analogous to that of a single agent POMDP as seen in Fig. 3(a), as should be expected. Here, the agent uses only growls to reason about the location of the tiger.

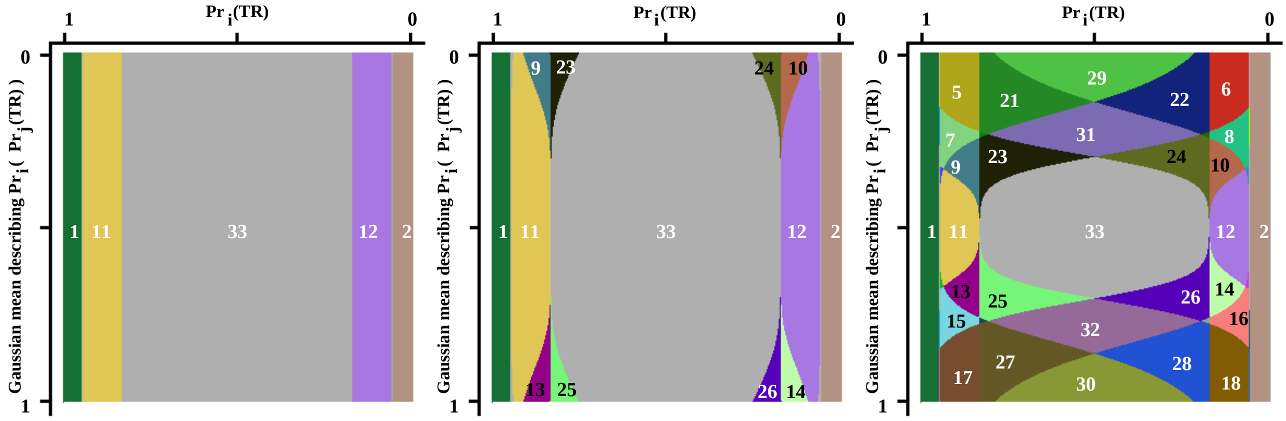
When creaks become more reliable, it is possible for i to use them to reinforce the estimates made using the growls. For example, in the belief region 9 [*Policy: L(); OL(GR,CL|GR,S) L(?)*] shown in Fig. 3(b), agent i *supplements* the information from growl (GR) with either a creak from the left (CL) or silence (S). This is because a creak from the left or silence indicates that j likely did not open the right door. The fact that CL or S are not likely to be the outcome of noise is due to i's belief that j believes with a fairly large probability that the tiger is behind the right door. When agent i is less certain about the location of the tiger (in region 23 [*Policy: L(); OL(GR,CL) L(?)*]), it has to observe a creak from the left, in addition to the growl from the right, to open the left door.

In regions of the policy (Fig. 3(b)) where the Gaussian mean is closer to 1, it is possible to get misleading information and still make a decision of opening doors. For example, in the region 25 [*Policy: L(); OL(GR,CR) L(?)*], the agent opens the left door even though it hears a growl from the right and a creak from the right. Because in this region the mean is closer to 1, i thinks that j is misinformed. In other words, for the physical state that the tiger is on the right, i likely believes that j believes the tiger to be on the left. Hence, the creak from the right is interpreted to be a result of this misinformation, thereby reinforcing i's belief that the tiger is on the right.

As the reliability of the creaks increases to 99.90%, we see that agent i starts exhibiting a more complex behavior. Sometimes it is guided only by creaks (regions 29 and 30), when it knows that j's opening the doors is a good predictor of the tiger's location. Sometimes, however, when i is quite sure about the tiger's location and it thinks that j is misinformed, i can ignore the creaks and base its decisions only on growls (regions 1, 2, 11, and 12).

Let us now turn our attention to regions 5 and 17 and their symmetric counterparts – regions 6 and 18 in Fig. 3(c). We focus on only the region 5, but the arguments for the other regions are analogous. For region 5, i believes j is informed and it believes that the tiger is likely on the right. For this region, the agent opens the left door if it hears a creak from the left (since it was fairly certain that the tiger was behind right door and the tiger persists after opening doors) or on hearing a growl from the right and no creaks. It opens the right door only if it hears a growl from the left along with creak from the right. Since i believes j is informed, a creak from the right indicates that j likely opened the right door which means that j believed that the tiger was on the left. An observation of GL further reinforces i's belief that the tiger is indeed on the left, causing i to open the right door.

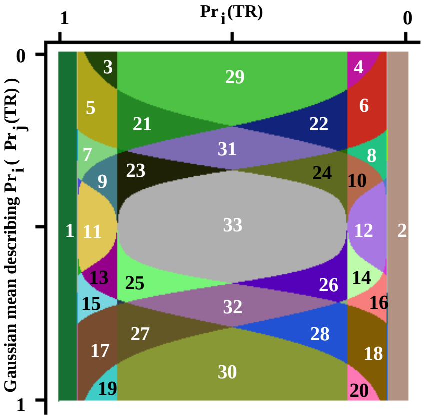
When the reliability of the creaks increases to 99.97%, the size of the regions that utilize only the creaks to guide the actions, increase



(a) Creak accuracy: 33%

(b) Creak accuracy: 70%

(c) Creak accuracy: 99.9%



(d) Creak accuracy: 99.97%

```
1   OL(); OL(*,*)                                  2   OR(); OR(*,*)
3    L(); OL(*,CL | GR,S) OR(*,CR) L(?)            4    L(); OL(*,CL) OR(*,CR | GL,S) L(?)
5    L(); OL(*,CL | GR,S) OR(GL,CR) L(?)           6    L(); OL(GR,CL) OR(*,CR | GL,S) L(?)
7    L(); OL(*,CL | GR,S) L(?)                     8    L(); OR(*,CR | GL,S) L(?)
9    L(); OL(GR,CL | GR,S) L(?)                    10   L(); OR(GL,CR | GL,S) L(?)
11   L(); OL(GR,*) L(?)                            12   L(); OR(GL,*) L(?)
13   L(); OL(GR,CR | GR,S) L(?)                    14   L(); OR(GL,CL | GL,S) L(?)
15   L(); OL(*,CR | GR,S) L(?)                     16   L(); OR(*,CL | GL,S) L(?)
17   L(); OL(*,CR | GR,S) OR(GL,CL) L(?)           18   L(); OL(GR,CR) OR(*,CL | GL,S) L(?)
19   L(); OL(*,CR | GR,S) OR(*,CL) L(?)            20   L(); OL(*,CR) OR(*,CL | GL,S) L(?)
21   L(); OL(*,CL) OR(GL,CR) L(?)                  22   L(); OL(GR,CL) OR(*,CR) L(?)
23   L(); OL(GR,CL) L(?)                           24   L(); OR(GL,CR) L(?)
25   L(); OL(GR,CR) L(?)                           26   L(); OR(GL,CL) L(?)
27   L(); OL(*,CR) OR(GL,CL) L(?)                  28   L(); OL(GR,CR) OR(*,CL) L(?)
29   L(); OL(*,CL) OR(*,CR) L(?)                   30   L(); OL(*,CR) OR(*,CL) L(?)
31   L(); OL(GR,CL) OR(GL,CR) L(?)                 32   L(); OL(GR,CR) OR(GL,CL) L(?)
33   L();  L(*,*)
```

(e) Policy of agent i. ? indicates all other observations and * indicates all possible values

Figure 3: Figures (a-d) show the projections of the value functions of agent i, with varying creak reliabilities. The belief regions are marked with numbers that index into the list of policy trees in table (e)

(see regions 29 and 30), as should be expected. Additionally, new regions (3, 4, 19, and 20) appear as part of the policy.

5. CONCLUSION

We presented a method that exploits behavioral equivalence among other agents' possible models to exactly solve finitely nested interactive POMDPs. By aggregating the behaviorally equivalent models into equivalence classes, we transformed the infinite interactive state space of I-POMDPs into a finite one, thereby allowing the application of standard POMDP solution techniques for solving I-POMDPs. Our method works because the discretization of the state space preserves the value of the plans. We applied our method to the multi-agent tiger problem, and presented a series of policies for various values of the parameters of the problem. Our aim is to understand the interplay between an agent's belief about other agents' beliefs and its observations, in arriving at its optimal plan of action. As we demonstrated, an agent's optimal policy changes drastically as its capability about observing others' behaviors improves. We believe it may be possible to combine different behavioral equivalence regions together to approximately represent an agent's behavior and reason about them. This is one possible avenue of future work.

Acknowledgements

This research was supported by the NSF CAREER award IRI-9702132 and NSF award IRI-0119270. Prashant Doshi would also like to acknowledge the support of a UIC University Fellowship.

APPENDIX

PROOF OF PROPOSITION 1.

$$\tilde{b}^t_{i,l}(ecis^t) = Pr(ecis^t \mid o^t_i, a^{t-1}_i, \tilde{b}^{t-1}_{i,l}) = \frac{Pr(ecis^t, o^t_i \mid a^{t-1}_i, \tilde{b}^{t-1}_{i,l})}{Pr(o^t_i \mid a^{t-1}_i, \tilde{b}^{t-1}_{i,l})}$$

$$= \beta \sum_{ecis^{t-1}} \tilde{b}^{t-1}_{i,l} Pr(ecis^t, o^t_i \mid a^{t-1}_i, ecis^{t-1})$$

$$= \beta \sum_{ecis^{t-1}, a^{t-1}_j} \tilde{b}^{t-1}_{i,l} Pr(a^{t-1}_j \mid ecis^{t-1}) \times Pr(ecis^t, o^t_i \mid a^{t-1}_i, a^{t-1}_j, ecis^{t-1})$$

$$
\begin{aligned}
&= \beta \sum_{ecis^{t-1}, a_j^{t-1}} \tilde{b}_{i,l}^{t-1} Pr(a_j^{t-1} \mid \langle s^{t-1}, M_{j,k}^{t-1} \rangle) \\
&\quad \times Pr(ecis^t, o_i^t \mid a^{t-1}, ecis^{t-1}) \\
&= \beta \sum_{ecis^{t-1}, a_j^{t-1}} \tilde{b}_{i,l}^{t-1} Pr(a_j^{t-1} \mid M_{j,k}^{t-1}) Pr(o_i^t \mid ecis^t, a^{t-1}, ecis^{t-1}) \\
&\quad \times Pr(ecis^t \mid a^{t-1}, ecis^{t-1}) \\
&= \beta \sum_{ecis^{t-1}, a_j^{t-1}} \tilde{b}_{i,l}^{t-1} Pr(a_j^{t-1} \mid M_{j,k}^{t-1}) O_i(s^t, a^{t-1}, o_i^t) \\
&\quad \times Pr(\langle s^t, M_{j,k}^t \rangle \mid a^{t-1}, \langle s^{t-1}, M_{j,k}^{t-1} \rangle) \\
&= \beta \sum_{ecis^{t-1}, a_j^{t-1}} \tilde{b}_{i,l}^{t-1} Pr(a_j^{t-1} \mid M_{j,k}^{t-1}) O_i(s^t, a^{t-1}, o_i^t) \\
&\quad \times T_i(s^t, a^{t-1}, s^{t-1}) \sum_{o_j^t} \tilde{\tau}(M_{j,k}^t, a_j^{t-1}, o_j^t, M_{j,k}^{t-1}) \\
&\quad \times O_j(o_j^t \mid a^{t-1}, s^t)
\end{aligned}
$$

Thus belief at time t can always be expressed in terms of belief and actions at time $t-1$ and the observations at time t. □

PROOF OF PROPOSITION 2 BY INDUCTION. To keep the proof short, we will assume that only the belief of the other agent is unknown, i.e. all other parameters of the model are known. Thus Eq. 6 can be written as $\tilde{b}_i(s, \mathbb{B}_{j,k}) = \int_{b_j \in \mathbb{B}_{j,k}} b_i(s, b_j) db_j$

Let b_i be an arbitrary belief of agent i. Let $\{\mathbb{B}_{j,1}, \mathbb{B}_{j,2}, ..., \mathbb{B}_{j,n}\}$ be the collection of equivalence classes of agent j's belief. Each class $\mathbb{B}_{j,k}$ is a set of beliefs of j such that the action a_j^k is optimal for each belief. Thus $\forall_{b_j \in \mathbb{B}_{j,k}} ER(s, b_j, a_i) = R(s, a_i, a_j^k)$, because a_j^k is optimal for all $b_j \in \mathbb{B}_{j,k}$.

Basis Step: We show that the horizon 1 value remains unchanged when i's original belief is replaced by its belief over the equivalence classes.

$$
\begin{aligned}
Q^1(b_i, a_i) &= \int_{is_i} b_i(is_i) ER(is_i, a_i) \\
&= \sum_s \int_{b_j \in B_j} b_i(s, b_j) ER(s, b_j, a_i)
\end{aligned}
$$

From the definition of equivalence classes,

$$
\begin{aligned}
Q^1(b_i, a_i) &= \sum_s \left\{ \int_{b_j \in \mathbb{B}_{j,1}} b_i(s, b_j) ER(s, b_j, a_i) \right. \\
&\quad \left. + ... + \int_{b_j \in \mathbb{B}_{j,n}} b_i(s, b_j) ER(s, b_j, a_i) \right\}
\end{aligned}
$$

Note that $ER(s, b_j, a_i) = \sum_{a_j} R(s, a_i, a_j) Pr(a_j | b_j)$, and for all $b_j \in \mathbb{B}_{j,n}$ the $Pr(a_j | b_j)$ will be 1 for some a_j^n (assuming a deterministic policy) and 0 for others. Therefore,

$$
\begin{aligned}
Q^1(b_i, a_i) &= \sum_s \left\{ \int_{b_j \in \mathbb{B}_{j,1}} b_i(s, b_j) R(s, a_i, a_j^1) \right. \\
&\quad \left. + ... + \int_{b_j \in \mathbb{B}_{j,n}} b_i(s, b_j) R(s, a_i, a_j^n) \right\} \\
&= \sum_s \left\{ R(s, a_i, a_j^1) \int_{b_j \in \mathbb{B}_{j,1}} b_i(s, b_j) + ... \right. \\
&\quad \left. + R(s, a_i, a_j^n) \int_{b_j \in \mathbb{B}_{j,n}} b_i(s, b_j) \right\}
\end{aligned}
$$

Using Eq. 6,

$$
\begin{aligned}
Q^1(b_i, a_i) &= \sum_s \left\{ R(s, a_i, a_{j,1}) \tilde{b}_i(s, \mathbb{B}_{j,1}) + ... \right. \\
&\quad \left. + R(s, a_i, a_{j,n}) \tilde{b}_i(s, \mathbb{B}_{j,n}) \right\} \\
&= \sum_{s,k} \tilde{b}_i(s, \mathbb{B}_{j,k}) R(s, a_i, a_{j,k}) \\
&= \tilde{Q}^1(\tilde{b}_i, a_i)
\end{aligned}
$$

Because the Q values remain unchanged, maximizing over them will also yield identical values.

Inductive Hypothesis: Let us assume that $\forall a_i, b_i \; Q^n(b_i, a_i) = \tilde{Q}^n(\tilde{b}_i, a_i)$ where $\tilde{b}_i$ is related to b_i using Eq. 6. Because the Q values are identical, the N horizon value function also remains unchanged.

Inductive Proof:

$$
\begin{aligned}
Q^{n+1}(b_i, a_i) &= \int_{is_i} b_i(is_i) ER(is_i, a_i) \\
&\quad + \gamma \sum_{o_i} Pr(o_i | b_i, a_i) V^N(SE(b_i, a_i, o_i)) \\
&= Q^1(b_i, a_i) + \gamma \sum_{o_i, is_i} Pr(o_i | is_i, a_i) \\
&\quad \times b_i(is_i) U^n(SE(b_i, a_i, o_i))
\end{aligned}
$$

From the basis step,

$$
\begin{aligned}
Q^{n+1}(b_i, a_i) &= \tilde{Q}^1(\tilde{b}_i, a_i) + \gamma \sum_{o_i, is_i, a_j} Pr(o_i | is_i, a_i, a_j) \\
&\quad \times Pr(a_j | b_j) b_i(is_i) U^n(SE(b_i, a_i, o_i)) \\
&= \tilde{Q}^1(\tilde{b}_i, a_i) + \gamma \sum_{o_i, s} \int_{b_j} \sum_{a_j} Pr(o_i | is_i, a_i, a_j) \\
&\quad \times Pr(a_j | b_j) b_i(s, b_j) U^n(SE(b_i, a_i, o_i))
\end{aligned}
$$

From the definition of equivalence classes,

$$
\begin{aligned}
Q^{N+1}(b_i, a_i) &= \tilde{Q}^1(\tilde{b}_i, a_i) + \\
&\quad \gamma \sum_{o_i, s, \mathbb{B}_{j,k}} \int_{b_j \in \mathbb{B}_{j,k}} \sum_{a_j} Pr(o_i | is_i, a_i, a_j) \\
&\quad \times Pr(a_j | b_j) b_i(s, b_j) U^n(SE(b_i, a_i, o_i))
\end{aligned}
$$

Using the BNM and BNO assumptions of the I-POMDP framework [8],

$$
\begin{aligned}
Q^{n+1}(b_i, a_i) &= \tilde{Q}^1(\tilde{b}_i, a_i) + \gamma \sum_{o_i, s, \mathbb{B}_{j,k}} \int_{b_j \in \mathbb{B}_{j,k}} Pr(o_i | s, a_i, a_j^k) \\
&\quad \times b_i(s, b_j) U^n(SE(b_i, a_i, o_i))
\end{aligned}
$$

From Eq. 6,

$$
\begin{aligned}
Q^{n+1}(b_i, a_i) &= \tilde{Q}^1(\tilde{b}_i, a_i) + \gamma \sum_{o_i, s, \mathbb{B}_{j,k}} Pr(o_i | s, a_i, a_j^k) \\
&\quad \times \int_{b_j \in \mathbb{B}_{j,k}} b_i(s, b_j) U^n(SE(b_i, a_i, o_i)) \\
&= \tilde{Q}^1(\tilde{b}_i, a_i) + \gamma \sum_{o_i, s, \mathbb{B}_{j,k}} Pr(o_i | s, a_i, a_j^k) \\
&\quad \times \tilde{b}_i(s, \mathbb{B}_{j,k}) U^n(SE(b_i, a_i, o_i))
\end{aligned}
$$

Using $Pr(o_i|s,a_i,\mathbb{B}_{j,k}) = \sum_{a_j} Pr(o_i|s,a_i,a_j,\mathbb{B}_{j,k})Pr(a_j|\mathbb{B}_{j,k}) = Pr(o_i|s,a_i,a_{j,k})$ and the inductive hypothesis,

$$\begin{aligned}
Q^{n+1}(b_i,a_i) &= \tilde{Q}^1(\tilde{b}_i,a_i) + \gamma \sum_{o_i,s,\mathbb{B}_{j,k}} Pr(o_i|s,a_i,\mathbb{B}_{j,k}) \\
&\quad \times \tilde{b}_i(s,\mathbb{B}_{j,k})\tilde{U}^n(SE(\tilde{b}_i,a_i,o_i)) \\
&= \tilde{Q}^1(\tilde{b}_i,a_i) + \gamma \sum_{o_i} Pr(o_i|a_i,\tilde{b}_i) \\
&\quad \times \tilde{U}^n(SE(\tilde{b}_i,a_i,o_i)) \\
&= \tilde{Q}^{n+1}(\tilde{b}_i,a_i)
\end{aligned}$$

Because the Q values are identical under the mapping, the values of the beliefs over the equivalence classes of interactive states remain unchanged. We have assumed in the proof that agent j's policies are deterministic, i.e., for every class $\mathbb{B}_{j,k}$ there is one optimal action $a_{j,k}$. The proof extends in a straightforward manner when there is more than one optimal action for a class. □

A. REFERENCES

[1] R. Becker, S. Zilberstein, and V. Lesser. Decentralized markov decision processes with event-driven interactions. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS-04)*, 2004.

[2] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of markov decision processes. In *Mathematics of Operations Research*, volume 27, No. 4, pages 819–840, 2002.

[3] D. S. Bernstein, E. A. Hansen, and S. Zilberstein. Bounded policy iteration for decentralized pomdps. In *Proceedings of the The Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005.

[4] A. R. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In *Uncertainty in Artificial Intelligence*, Rhode Island, Providence, 1997.

[5] P. Doshi and P. Gmytrasiewicz. Approximating state estimation in multiagent settings using particle filters. In *Proceedings of the Fourth International Conference on Autonomous Agents and Multi Agent Systems (AAMAS-05)*, 2005.

[6] P. Doshi and P. Gmytrasiewicz. A particle filtering based approach to approximating interactive pomdps. In *Proceedings of the The Twentieth National Conference on Artificial Intelligence (AAAI-05)*, 2005.

[7] R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in markov decision processes. In *Artificial Intelligence Journal*, 2003.

[8] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. In *Journal of AI Research*, 2005.

[9] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04)*, 2004.

[10] J. C. Harsanyi. Games with incomplete information played by 'bayesian' players. In *Management Science*, pages 14(3):159–182, Nov 2005.

[11] L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 2, 1998.

[12] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella. Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *Proceedings of International Joint Conference in Artificial Intelligence (IJCAI)*, 2003.

[13] J. M. Porta, M. T. Spaan, and N. Vlassis. Value iteration for continuous-state pomdps. In *IAS Technical report IAS-UVA-04-04*, 2004.

[14] P. Poupart and C. Boutilier. Value directed compression of pomdps. In *Seventeenth Annual Conference on Neural Information Processing Systems*, 2003.

[15] D. V. Pynadath and M. Tambe. Multiagent teamwork: Analyzing the optimality and complexity of key theories and models. In *Joint Conference on Autonomous Agents and Multi-Agent Systems*, 2002.

[16] R. Smallwood and E. Sondik. The optimal control of partially observable markov decision processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

[17] S. Thrun. Monte carlo POMDPs. In S. Solla, T. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1064–1070. MIT Press, 2000.