

ϵ -Behavioral Equivalence for Interactive Dynamic Influence Diagrams

Technical Report

Abstract

Interactive dynamic influence diagrams (I-DID) are graphical models for sequential decision making in uncertain settings shared by other agents. Algorithms for solving I-DIDs face the challenge of an exponentially growing space of candidate models ascribed to other agents, over time. Pruning the behaviorally equivalent models is one way toward identifying a *minimal* model set. We seek to further reduce the complexity by additionally pruning models that are approximately behaviorally equivalent. Toward this, we redefine behavioral equivalence in terms of the distribution over the subject agent’s future action-observation paths, and introduce the notion of ϵ -behavioral equivalence. We present a new approximation technique that reduces the candidate model space by removing models that are ϵ -behaviorally equivalent with representative ones. We empirically demonstrate that our approach improves significantly in performance on previous techniques, with some limitations.

1 Introduction

Interactive dynamic influence diagrams (I-DID) [2] are recognized graphical models for sequential decision making in uncertain multiagent settings. I-DIDs concisely represent the problem of how an agent should act in an uncertain environment shared with others who may act in sophisticated ways. I-DIDs may be viewed as graphical counterparts of interactive POMDPs (I-POMDPs) [5], providing a way to model and exploit the embedded structure often present in real-world decision-making situations. They generalize DIDs [13], which are graphical representations of POMDPs, to multiagent settings analogously to how I-POMDPs generalize POMDPs.

As we may expect, I-DIDs acutely suffer from both the curses of dimensionality and history. This is because the state space in I-DIDs includes the models of other agents in addition to the traditional physical states. These models encompass the agents’ beliefs, action and sensory capabilities, and preferences, and may themselves be formalized as I-DIDs. The nesting is terminated at the 0^{th} level where the other agents are modeled using DIDs. As the agents act, observe, and update beliefs, I-DIDs must track the evolution of the models over time. Consequently, I-DIDs not only suffer from the curse of history that afflicts the modeling agent, but more so from that exhibited by the modeled agents. The exponential growth in the number of models over time also further contributes to the dimensionality of the state space. This is

complicated by the nested nature of the space.

Previous approaches for approximating I-DIDs [1, 14] focus on reducing the dimensionality of the state space by limiting the number of candidate models of other agents. Using the insight that beliefs that are spatially close are likely to be behaviorally equivalent [10], Zeng et al. [14] cluster the models of other agents and select representative models from each cluster. Intuitively, a cluster contains models that are likely to be behaviorally equivalent and hence may be replaced by a subset of representatives without a significant loss in the optimality of the decision maker. However, this approach often retains more models than needed. Doshi and Zeng ([1]) formalize the concept of a *minimal set* of models, previously discussed by Pynadath and Marsella [9]. At each time step, only those models are updated which will result in predictive behaviors that are distinct from others in the updated model space. The initial set of models are solved and merged to obtain a policy graph, which assists in discriminating between model updates.

In this paper, we aim to reduce the model space by additionally pruning models that are approximately behaviorally equivalent. Toward this objective, we introduce the concept of ϵ -behavioral equivalence among candidate models. In doing so, we redefine behavioral equivalence as the class of models of the other agents that induce an identical distribution over the subject agent’s future action-observation paths in the interaction. Subsequently, models that induce distributions over the paths, which are no more than $\epsilon \geq 0$ apart are termed as being ϵ -behaviorally equivalent. Intuitively, this results in a lesser number of equivalence classes in the partition. If we pick a single representative model from each class, we typically end up with no more models than in the minimal set which need be solved thereby improving on approaches that utilize exact behavioral equivalence.

We begin by selecting a model at random and grouping together ϵ -behaviorally equivalent models with it. We repeat this procedure for the remaining models until all models have been grouped. The retained model set consists of the representative model from each equivalence class. In the worst case ($\epsilon = 0$), our approach identifies exact behavioral equivalence and the model set consists of all the behaviorally unique models. We theoretically analyze the error introduced by this approach in the optimality of the solution. More importantly, we experimentally evaluate our approach on I-DIDs formulated for benchmark problem domains and show significant improvement in performance, although with limitations.

2 Background: Interactive DID

We briefly describe interactive influence diagrams (I-IDs) for two-agent interactions followed by their extensions to dynamic settings, I-DIDs, and refer the reader to [2] for more details.

2.1 Syntax

In addition to the usual chance, decision, and utility nodes, I-IDs include a new type of node called the *model node* (hexagonal node, $M_{j,l-1}$, in Fig. 1(a)). We note that

the probability distribution over the chance node, S , and the model node together represents agent i 's belief over its *interactive state space*. In addition to the model node, I-IDs differ from IDs by having a chance node, A_j , that represents the distribution over the other agent's actions, and a dashed link, called a *policy link*.

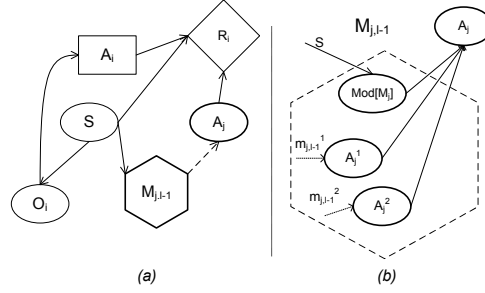


Figure 1: (a) A generic level $l > 0$ I-ID for agent i situated with one other agent j . The hexagon is the model node ($M_{j,l-1}$) and the dashed arrow is the policy link. (b) Representing the model node and policy link using chance nodes and dependencies.

The model node contains as its values the alternative computational models ascribed by i to the other agent. We denote the set of these models by $\mathcal{M}_{j,l-1}$. A model in the model node may itself be an I-ID or ID, and the recursion terminates when a model is an ID or a simple probability distribution over the actions. Formally, we denote a model of j as, $m_{j,l-1} = \langle b_{j,l-1}, \hat{\theta}_j \rangle$, where $b_{j,l-1}$ is the level $l-1$ belief, and $\hat{\theta}_j$ is the agent's *frame* encompassing the action, observation, and utility nodes. We observe that the model node and the dashed policy link that connects it to the chance node, A_j , could be represented as shown in Fig. 1(b). The decision node of each level $l-1$ I-ID is transformed into a chance node. Specifically, if OPT is the set of optimal actions obtained by solving the I-ID (or ID), then $Pr(a_j \in A_j^1) = \frac{1}{|OPT|}$ if $a_j \in OPT$, 0 otherwise. The conditional probability table (CPT) of the chance node, A_j , is a *multiplexer*, that assumes the distribution of each of the action nodes (A_j^1, A_j^2) depending on the value of $Mod[M_j]$. In other words, when $Mod[M_j]$ has the value $m_{j,l-1}^1$, the chance node A_j assumes the distribution of the node A_j^1 , and A_j assumes the distribution of A_j^2 when $Mod[M_j]$ has the value $m_{j,l-1}^2$. The distribution over $Mod[M_j]$, is i 's belief over j 's models given the state. For more than two agents, we add a model node and a chance node representing the distribution over an agent's action linked together using a policy link, for each other agent.

I-DIDs extend I-IDs to allow sequential decision making over several time steps (see Fig. 2). In addition to the model nodes and the dashed policy link, what differentiates an I-DID from a DID is the *model update link* shown as a dotted arrow in Fig. 2. We briefly explain the semantics of the model update next.

The update of the model node over time involves two steps: First, given the models at time t , we identify the updated set of models that reside in the model node at time $t+1$. Because the agents act and receive observations, their models are updated to reflect their changed beliefs. Since the set of optimal actions for a model could include

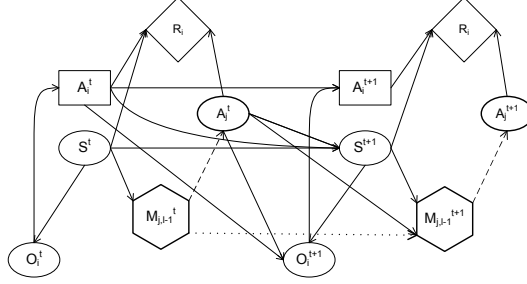


Figure 2: A generic two time-slice level l I-DID for agent i .

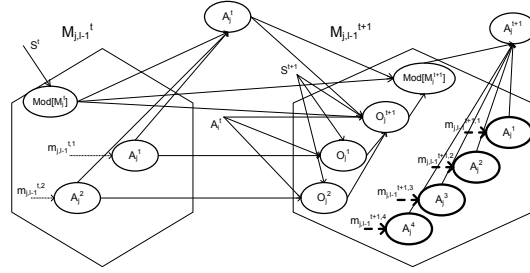


Figure 3: The semantics of the model update link. Notice the growth in the number of models at $t + 1$ shown in bold.

all the actions, and the agent may receive any one of $|\Omega_j|$ possible observations, the updated set at time step $t + 1$ will have up to $|\mathcal{M}_{j,l-1}^t| |A_j| |\Omega_j|$ models. Here, $|\mathcal{M}_{j,l-1}^t|$ is the number of models at time step t , $|A_j|$ and $|\Omega_j|$ are the largest spaces of actions and observations respectively, among all the models. The CPT of $Mod[M_{j,l-1}^{t+1}]$ encodes the function, $\tau(b_{j,l-1}^t, a_j^t, o_j^{t+1}, b_{j,l-1}^{t+1})$ which is 1 if the belief $b_{j,l-1}^t$ in the model $m_{j,l-1}^t$ using the action a_j^t and observation o_j^{t+1} updates to $b_{j,l-1}^{t+1}$ in a model $m_{j,l-1}^{t+1}$; otherwise it is 0. Second, we compute the new distribution over the updated models, given the original distribution and the probability of the agent performing the action and receiving the observation that led to the updated model. The dotted model update link in the I-DID may be implemented using standard dependency links and chance nodes, as shown in Fig. 3 transforming it into a flat DID.

2.2 Behavioral Equivalence and Solution

Although the space of possible models is very large, not all models need to be considered in the model node. Models that are *behaviorally equivalent* [9, 10] – whose behavioral predictions for the other agent are identical – could be pruned and a single representative model considered. This is because the solution of the subject agent’s I-DID is affected by the predicted behavior of the other agent only; thus we need not distinguish between behaviorally equivalent models. Let **BehavioralEq**($\mathcal{M}_{j,l-1}$) be

the procedure that prunes the behaviorally equivalent models from $\mathcal{M}_{j,l-1}$ returning the set of representative models.

The solution of an I-DID (and I-ID) proceeds in a bottom-up manner, and is implemented recursively as shown in Fig. 4. We start by solving the level 0 models, which may be traditional DIDs. Their solutions provide probability distributions which are entered in the corresponding action nodes found in the model node of the level 1 I-DID. The solution method uses the standard look-ahead technique, projecting the agent’s action and observation sequences forward from the current belief state, and finding the possible beliefs that i could have in the next time step. Because agent i has a belief over j ’s models as well, the look-ahead includes finding out the possible models that j could have in the future. Consequently, each of j ’s level 0 models represented using a standard DID in the first time step must be solved to obtain its optimal set of actions. These actions are combined with the set of possible observations that j could make in that model, resulting in an updated set of candidate models (that include the updated beliefs) that could describe the behavior of j . $SE(b_j^t, a_j, o_j)$ is an abbreviation for the belief update. The updated set is minimized by excluding the behaviorally equivalent models. Beliefs over these updated set of candidate models are calculated using the standard inference methods through the dependency links between the model nodes (Fig. 3). The algorithm in Fig. 4 may be realized using the standard implementations of DIDs.

3 Redefining Behavioral Equivalence

We assume that the models of j have identical frames and differ only in their beliefs. As we mentioned previously, two models of the other agent are behaviorally equivalent (BE)¹ if they produce identical behaviors for the other agent. More formally, models $m_{j,l-1}, \hat{m}_{j,l-1} \in \mathcal{M}_{j,l-1}$ are BE if and only if $OPT(m_{j,l-1}) = OPT(\hat{m}_{j,l-1})$, where $OPT(\cdot)$ denotes the solution of the model that forms the argument. If the model is a DID or an I-DID, its solution is a policy tree.

Our aim is to identify models that are *approximately* BE. While a pair of policy trees may be checked for equality, disparate policy trees do not directly permit intuitive behavioral comparisons. This makes it difficult to define a measure of approximate BE, motivating investigations into a more rigorous formalization of BE.

Recall that BE models impact the decision-making of the modeling agent similarly, thereby motivating interest in grouping such models together. We utilize this insight toward introducing a new definition of BE. Let $h = \{a_i^t, o_i^{t+1}\}_{t=1}^T$ be the action-observation path for the modeling agent i , where o_i^{T+1} is null for a T horizon problem. If $a_i^t \in A_i$ and $o_i^{t+1} \in \Omega_i$, where A_i and Ω_i are i ’s action and observation sets respectively, then the set of all paths is, $H = \Pi_1^T(A_i \times \Omega_i)$, and the set of action-observation histories up to time t is $H^t = \Pi_1^{t-1}(A_i \times \Omega_i)$. The set of future action-observation paths is, $H_{T-t} = \Pi_t^T(A_i \times \Omega_i)$, where t is the current time step. We show example action-observation paths in Fig. 5.

¹We will use BE as an acronym for both, behaviorally equivalent (adjective form) and behavioral equivalence (noun form). Appropriate usage will be self-evident.

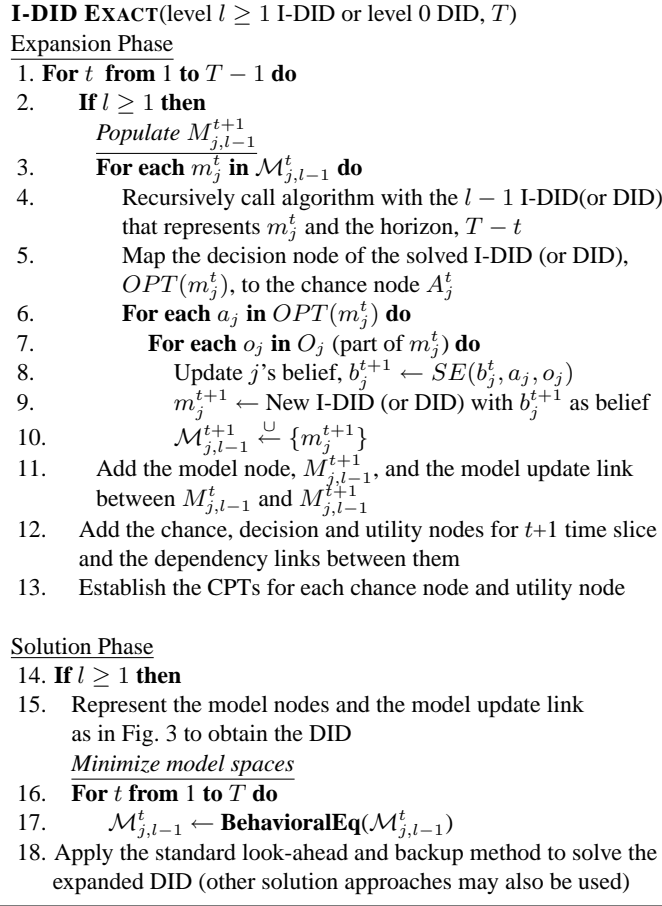


Figure 4: Algorithm for exactly solving a level $l \geq 1$ I-DID or level 0 DID expanded over T time steps.

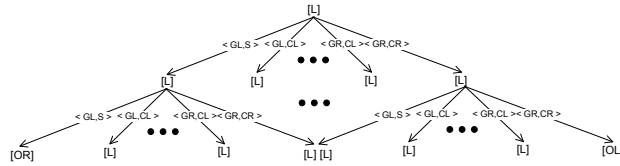


Figure 5: Future action-observation paths of agent i in a 2-horizon multiagent tiger problem. The nodes represent i 's action, while the edges are labeled with the possible observations. This example starts with i listening. Agent i may receive one of six observations conditional on j 's action, and performs an action that optimizes its resulting belief.

We observe that agent j 's model together with agent i 's perfect knowledge of its own model and its action-observation history induces a predictive distribution over i 's future action-observation paths such as those in Fig. 5. This distribution plays a critical role in our approach and we denote it as, $Pr(H_{T-t}|h^t, m_{i,l}, m_{j,l-1}^t)$, where $h^t \in H^t$, $m_{i,l}$ is i 's level l I-DID and $m_{j,l-1}^t$ is the level $l-1$ model of j in the model node at time t . For the sake of brevity, we rewrite the distribution term as, $Pr(H_{T-t}|m_{i,l}^t, m_{j,l-1}^t)$, where $m_{i,l}^t$ is i 's horizon $T-t$ I-DID with its initial belief updated given the actions and observations in h^t . We define BE below:

Definition 1 (Behavioral Equivalence). *Two models of agent j , $m_{j,l-1}^t$ and $\hat{m}_{j,l-1}^t$, are behaviorally equivalent if and only if $Pr(H_{T-t}|m_{i,l}^t, m_{j,l-1}^t) = Pr(H_{T-t}|m_{i,l}^t, \hat{m}_{j,l-1}^t)$, where H_{T-t} and $m_{i,l}^t$ are as defined previously.*

In other words, BE models are those that induce an identical distribution over agent i 's future action-observation history. This reflects the fact that such models impact agent i 's behavior similarly.

Let h_{T-t} be some future action-observation path of agent i , $h_{T-t} \in H_{T-t}$. In Proposition 1, we provide a recursive way to arrive at the probability, $Pr(h_{T-t}|m_{i,l}^t, m_{j,l-1}^t)$. Of course, the probabilities over all possible paths sum to 1.

Proposition 1. $Pr(h_{T-t}|m_{i,l}^t, m_{j,l-1}^t) = Pr(a_i^t, o_i^{t+1}|m_{i,l}^t, m_{j,l-1}^t)$
 $\sum_{a_j^t, o_j^{t+1}} Pr(h_{T-t-1}|a_i^t, o_i^{t+1}, m_{i,l}^t, a_j^t, o_j^{t+1}, m_{j,l-1}^t) Pr(a_j^t, o_j^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t)$
 $= Pr(a_i^t, o_i^t|m_{i,l}^t, m_{j,l-1}^t) \sum_{a_j^t, o_j^{t+1}} Pr(h_{T-t-1}|m_{i,l}^{t+1}, m_{j,l-1}^{t+1})$
 $\times Pr(a_j^t, o_j^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t)$

where

$$\begin{aligned} Pr(a_i^t, o_i^{t+1}|m_{i,l}^t, m_{j,l-1}^t) &= Pr(a_i^t|OPT(m_{i,l}^t)) \sum_{a_j^t} Pr(a_j^t| \\ &OPT(m_{j,l-1}^t)) \sum_{s^{t+1}} O_i(s^{t+1}, a_i^t, a_j^t, o_i^{t+1}) \\ &\times \sum_{s, m_j} T_i(s, a_i^t, a_j^t, s^{t+1}) b_{i,l}^t(s, m_j) \end{aligned} \quad (1)$$

and

$$\begin{aligned} Pr(a_j^t, o_j^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t) &= Pr(a_j^t|OPT(m_{j,l-1}^t)) \sum_{s^{t+1}} \\ &O_j(s^{t+1}, a_j^t, a_i^t, o_j^{t+1}) \sum_{s, m_j} T_i(s, a_i^t, a_j^t, s^{t+1}) b_{i,l}^t(s, m_j) \end{aligned} \quad (2)$$

In Eq. 1, $O_i(s^{t+1}, a_i^t, a_j^t, o_i^{t+1})$ is i 's observation function contained in the CPT of the chance node, O_i^{t+1} , in the I-DID, $T_i(s, a_i^t, a_j^t, s^{t+1})$ is i 's transition function contained in the CPT of the chance node, S^{t+1} , $Pr(a_i^t|OPT(m_{i,l}^t))$ is obtained by solving agent i 's I-DID, $Pr(a_j^t|OPT(m_{j,l-1}^t))$ is obtained by solving j 's model and appears in the CPT of node, A_j^t . In Eq. 2, $O_j(s^{t+1}, a_j^t, a_i^t, o_j^{t+1})$ is j 's observation function contained in the CPT of the chance node, O_j^{t+1} , given j 's model is $m_{j,l-1}^t$. We give the proof of Proposition 1 in the Appendix.

Now that we have a way of computing the distribution over the future paths, we may relate Definition 1 to our previous understanding of BE models:

Proposition 2 (Correctness). *$Pr(H_{T-t}|m_{i,l}^t, m_{j,l-1}^t) = Pr(H_{T-t}|m_{i,l}^t, \hat{m}_{j,l-1}^t)$ if and only if $OPT(m_{j,l-1}^t) = OPT(\hat{m}_{j,l-1}^t)$, where $m_{j,l-1}^t$ and $\hat{m}_{j,l-1}^t$ are j 's models.*

Sketch. The proof is reducible to showing the above for some individual path, $h_{T-t} \in H_{T-t}$.

(\Rightarrow) From Proposition 1, we observe that the probability over i 's path depends on the policy trees of i and j , i and j 's observation functions, i 's transition function and i 's belief over the chance node S and model node. Given that the probabilities over the path are identical for two different models of j (identical frames), let us assume that the solutions of these models are distinct, $OPT(m_{j,l-1}^t) \neq OPT(\hat{m}_{j,l-1}^t)$. Therefore, there exists a_j^t for which, $Pr(a_j^t | OPT(m_{j,l-1}^t)) \neq Pr(a_j^t | OPT(\hat{m}_{j,l-1}^t))$. Because all other terms in Eqs. 1 and 2 either pertain to i or to j 's frame, and remain unchanged for both models, we deduce that the two probabilities over the path must be different. However, this is a contradiction of our initial premise, and therefore the two models must have the same solution.

(\Leftarrow) Given $OPT(m_{j,l-1}^t) = OPT(\hat{m}_{j,l-1}^t)$, we may write, $Pr(a_j^t | OPT(m_{j,l-1}^t)) = Pr(a_j^t | OPT(\hat{m}_{j,l-1}^t))$ for all a_j^t . Because all other terms in Eqs. 1 and 2 are identical, it follows that $Pr(h_{T-t} | m_{i,l}^t, m_{j,l-1}^t)$ must be same as $Pr(h_{T-t} | m_{i,l}^t, \hat{m}_{j,l-1}^t)$. ■

A simple method for computing the distribution over the paths given models of i and j is to replace agent i 's decision nodes in the I-DID with chance nodes so that $Pr(a_i \in A_i^t) = \frac{1}{|OPT(m_{i,l}^t)|}$ and remove the utility nodes, thereby transforming the I-DID into a dynamic Bayesian network (DBN). The desired distribution is then the marginal over the chance nodes that represent i 's actions and observations with j 's model entered as evidence in the Mod node at t .

4 ϵ -Behavioral Equivalence

Rethinking of BE in Section 3 formalizes the intuition that BE models impact the modeling agent identically. In addition to being more rigorous, it has the additional advantage that it permits us to measure the degree to which models are BE, thereby allowing us to introduce *approximate BE*.

4.1 Definition

We introduce the notion of ϵ -behavioral equivalence (ϵ -BE) and define it as follows:

Definition 2 (ϵ -BE). *Given $\epsilon \geq 0$, two models, $m_{j,l-1}^t$ and $\hat{m}_{j,l-1}^t$, are ϵ -BE if the divergence between the distributions $Pr(H_{T-t} | m_{i,l}^t, m_{j,l-1}^t)$ and $Pr(H_{T-t} | m_{i,l}^t, \hat{m}_{j,l-1}^t)$ is no more than ϵ .*

Here, the distributions over i 's future paths are computed as shown in Proposition 1. While multiple ways to measure the divergence between distributions exist, we utilize the well-known Kullback-Leibler (KL) divergence [7] in its symmetric form, in this paper. Consequently, the models are ϵ -BE if,

$$D_{KL}(Pr(H_{T-t} | m_{i,l}^t, m_{j,l-1}^t) || Pr(H_{T-t} | m_{i,l}^t, \hat{m}_{j,l-1}^t)) \leq \epsilon$$

where $D_{KL}(p||p')$ denotes the symmetric KL divergence between distributions, p and p' , and is calculated as:

$$D_{KL}(p||p') = \frac{1}{2} \sum_k \left(p(k) \log \frac{p(k)}{p'(k)} + p'(k) \log \frac{p'(k)}{p(k)} \right)$$

If $\epsilon = 0$, ϵ -BE collapses into exact BE. Sets of models exhibiting ϵ -BE for some non-zero but small ϵ do not differ significantly in how they impact agent i 's decision making. As we mention in the next section, these models could be candidates for pruning.

4.2 Approach

We proceed by picking a model of j at random, $m_{j,l-1}^{t=1}$, from the model node in the first time step, which we call the representative. All other models in the model node that are ϵ -BE with $m_{j,l-1}^{t=1}$ are grouped together with it. Of the remaining models, another representative is picked at random and the previous procedure is repeated. The procedure terminates when no more models remain to be grouped. We illustrate the process in Fig. 6. We point out that for $\epsilon > 0$, in general, more models will likely be grouped together than if we considered exact BE. This will result in a fewer number of classes in the partition.

We first observe that the outcome is indeed a partition of the model set into ϵ -BE classes. This is because we continue to pick representative models and build classes until no model remains ungrouped. There is no overlap between classes since new ones are built only from the models that did not get previously grouped. We observe that the representatives of different classes are ϵ -behaviorally distinct, otherwise they would have been grouped together. However, this set is not unique and the partition could change with different representatives. Furthermore, let \mathcal{M}_j be the largest set of behaviorally distinct models, also called the minimal set [1]. Then, the following proposition holds:

Proposition 3 (Cardinality). *The ϵ -BE approach results in at most $|\hat{\mathcal{M}}_j|$ models after pruning.*

Intuitively, the Proposition follows from the fact that in the worst case, $\epsilon = 0$, resulting in behaviorally distinct models.

Transfer of probability mass From each class in the partition, the previously picked representative is retained and all other models are pruned. The representatives are distinguished in that all models in its group are ϵ -BE with it. Unlike exact BE, ϵ -BE relation is not necessarily transitive. Consequently, we may not select any model from each class as the representative since others may not be ϵ -BE with it.

Recall that agent i 's belief assigns some probability mass to each model in the model node. A consequence of pruning some of the models is that the mass assigned to the models would be lost. Disregarding this probability mass may introduce further error in the optimality of the solution. We avoid this error by transferring the probability mass over the pruned models in each class to the ϵ -BE representative that is retained in the model node (see Fig. 6).

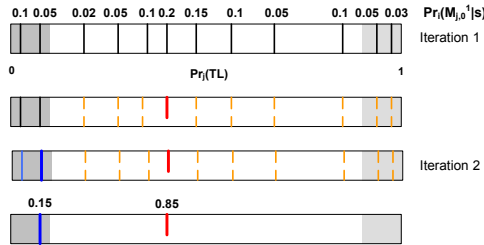


Figure 6: Illustration of the iterative ϵ -BE model grouping using the tiger problem. Black vertical lines denote the beliefs contained in different models of agent j included in the initial model node, $M_{j,0}^1$. Decimals on top indicate i 's probability distribution over j 's models. We begin by picking a representative model (red line) and grouping models that are ϵ -BE with it. Unlike exact BE, models in a different behavioral (shaded) region get grouped as well. Of the remaining models, another is selected as representative. Agent i 's distribution over the representative models is obtained by summing the probability mass assigned to the individual models in each class.

Sampling actions and observations Recall that the predictive distribution over i 's future action-observation paths, $Pr(H_{T-t}|h^t, m_{i,t}, m_{j,l-1}^t)$, is conditioned on the history of i 's observations, h^t , as well. Because the model grouping is performed while solving the I-DID when we do not know the actual history, we obtain a likely h^t by sampling i 's actions and observations for subsequent time steps in the I-DID.

Beginning with the first time step, we pick an action, a_i^t , at random assuming that each action is equally likely. An observation is then sampled from the distribution given i 's sampled action and belief, $o_i^{t+1} \sim Pr(\Omega_i|a_i^t, b_{i,l}^t)$, where $b_{i,l}^t$ is the prior belief. We utilize this sampled action and observation pair as the history, $h^t \leftarrow \langle a_i^t, o_i^{t+1} \rangle$. We may implement this procedure by entering as evidence i 's action in the chance node, A_i^t , of the DBN (mentioned in Section 3) and sampling from the inferred distribution over the chance node, O_i^{t+1} .

Finally, we note that in computing the distribution over the paths, solution to agent i 's I-DID is needed as well ($Pr(a_i^t|OPT(m_{i,l}^t))$ term in Eq. 1). As we wish to avoid this, we observe that ϵ -BE is based on the *comparative impact* that j 's models have on i , which is independent of i 's decisions. Therefore, we assume a uniform distribution over i 's actions, $Pr(a_i^t|OPT(m_{i,l}^t)) = \frac{1}{|A_i|}$, which does not change the ϵ -BE of models.

5 Algorithm

We present the algorithm for partitioning the models in the model node of the I-DID at each time step according to ϵ -BE, in Fig. 7. The procedure, ϵ -**BehaviorEquivalence** replaces the procedure, **BehaviorEq**, in the algorithm in Fig. 4. The procedure takes as input, the set of j 's models, \mathcal{M}_j , agent i 's DID, m_i , current time step and horizon, and the approximation parameter, ϵ . The algorithm begins by computing the distribution

over the future paths of i for each model of j . If the time step is not the initial one, the prior action-observation history is first sampled. We may compute the distribution by transforming the I-DID into a DBN as mentioned in Section 3 and entering the model of j as evidence – this implements Eqs. 1 and 2.

We then pick a representative model at random, and using the cached distributions group together models whose distributions exhibit a divergence less than ϵ from the distribution of the representative model. We iterate over the models left ungrouped until none remain. Each iteration results in a new class of models including a representative. In the final selection phase, all models except the representative are pruned from each class in the partition. The set of representative models, which are ϵ -behaviorally distinct, are returned.

ϵ -**BEHAVIOREQUIVALENCE** (Model set \mathcal{M}_j , DID m_i , current time step tt , horizon T , ϵ) **returns** \mathcal{M}'_j

1. Transform DID m_i into DBN by replacing i 's decision nodes with chance nodes having uniform distribution
2. **For** t **from** 1 **to** tt **do**
3. Sample, $a_i^t \sim Pr(A_i^t)$
4. Enter a_i^t as evidence into chance node, A_i^t , of DBN
5. Sample, $o_i^{t+1} \sim Pr(O_i^{t+1})$
6. $h^t \leftarrow \cup \langle a_i^t, o_i^{t+1} \rangle$
7. **For each** m_j^k **in** \mathcal{M}_j **do**
8. Compute the distribution, $P[k] \leftarrow Pr(H_{T-t}|h^t, m_i, m_j^k)$, obtained from the DBN by entering m_j^k as evidence (Proposition 1)

Clustering Phase

9. **While** \mathcal{M}_j not empty
10. Select a model, $m_j^{\hat{k}} \in \mathcal{M}_j$, at random as representative
11. Initialize, $\mathcal{M}_j^{\hat{k}} \leftarrow \{m_j^{\hat{k}}\}$
12. **For each** m_j^k **in** \mathcal{M}_j **do**
13. **If** $D_{KL}(P[\hat{k}]||P[k]) \leq \epsilon$
14. $\mathcal{M}_j^{\hat{k}} \leftarrow \cup m_j^k, \mathcal{M}_j \leftarrow \setminus m_j^k$

Selection Phase

15. **For each** $\mathcal{M}_j^{\hat{k}}$ **do**
16. Retain the representative model, $\mathcal{M}'_j \leftarrow \cup \mathcal{M}_j^{\hat{k}}$
17. **Return** \mathcal{M}'_j

Figure 7: Algorithm for partitioning j 's model space using ϵ -BE. This function replaces **BehaviorEq()** in Fig. 4.

6 Computational Savings and Error Bound

As with previous approaches, the primary complexity of solving I-DIDs is due to the large number of models that must be solved over T time steps. At time step t , there

could be $|\mathcal{M}_j^0|(|A_j||\Omega_j|)^t$ many models of the other agent j , where $|\mathcal{M}_j^0|$ is the number of models considered initially. The nested modeling further contributes to the complexity since solution of each model at level $l - 1$ requires solving the lower level $l - 2$ models, and so on recursively up to level 0. In an $N+1$ agent setting, if the number of models considered at each level for an agent is bound by $|\mathcal{M}|$, then solving an I-DID at level l requires the solutions of $\mathcal{O}((N|\mathcal{M}|)^l)$ many models. As we mentioned in Proposition 3, the ϵ -BE approximation reduces the number of agent models at each level to at most the size of the minimal set, $|\hat{\mathcal{M}}^t|$. In doing so, it solves $|\mathcal{M}_j^0|$ many models initially and incurs the complexity of performing inference in a DBN for computing the distributions. This complexity while significant is less than that of solving DIDs. Consequently, we need to solve at most $\mathcal{O}((N|\hat{\mathcal{M}}^*|)^l)$ number of models at each non-initial time step, typically less, where $\hat{\mathcal{M}}^*$ is the largest of the minimal sets, in comparison to $\mathcal{O}((N|\mathcal{M}|)^l)$. Here \mathcal{M} grows exponentially over time. In general, $|\hat{\mathcal{M}}| \ll |\mathcal{M}|$, resulting in a substantial reduction in the computation. Additionally, a reduction in the number of models in the model node also reduces the size of the state space, which makes solving the upper-level I-DID more efficient.

Given that lower-level models of the other agent are solved exactly, we analyze the conditional error bound of this approach. In the trivial case, $\epsilon = 0$, and there is no optimality error in the solution. If we limit the pruning of ϵ -BE models to the initial model node, the error is due to transferring the probability mass of the pruned model to the representative, effectively replacing the pruned model with the representative. Our definition of BE provides us with a unique opportunity to bound the error for i . We observe that the expected value of the I-DID could be obtained as the expected reward of following each path weighted by the probability of that path. Let $\rho_{b_{i,l}}(H_T)$ be the vector of expected rewards for agent i given it's belief when each path in H_T is followed. Here, T is the horizon of the I-DID. The expected value for i is:

$$EV_i = Pr(H_T | m_{i,l}, m_{j,l-1}) \cdot \rho_{b_{i,l}}(H_T)$$

where $m_{j,l-1}$ is the model of j .

If the above model of j is pruned in the Mod node, let model $\hat{m}_{j,l-1}$ be the representative that replaces it. Then $\hat{b}_{i,l}$ is i 's belief in which model $m_{j,l-1}$ is replaced with the representative. Expected value for i , $E\hat{V}_i$, is:

$$E\hat{V}_i = Pr(H_T | m_{i,l}, m_{j,l-1}) \cdot \rho_{\hat{b}_{i,l}}(H_T)$$

Then, the effective error bound is:

$$\begin{aligned}
\Delta &= \|\hat{E}V_i - EV_i\|_\infty \\
&= \|\Pr(H_T|m_{i,l}, m_{j,l-1}) \cdot \rho_{\hat{b}_{i,l}}(H_T) \\
&\quad - \Pr(H_T|m_{i,l}, m_{j,l-1}) \cdot \rho_{b_{i,l}}(H_T)\|_\infty \\
&= \|\Pr(H_T|m_{i,l}, m_{j,l-1}) \cdot \rho_{\hat{b}_{i,l}}(H_T) \\
&\quad - \Pr(H_T|m_{i,l}, \hat{m}_{j,l-1}) \cdot \rho_{b_{i,l}}(H_T) \\
&\quad + \Pr(H_T|m_{i,l}, \hat{m}_{j,l-1}) \cdot \rho_{b_{i,l}}(H_T) \\
&\quad - \Pr(H_T|m_{i,l}, m_{j,l-1}) \cdot \rho_{b_{i,l}}(H_T)\|_\infty \quad (\text{add zero}) \\
&\leq \|\Pr(H_T|m_{i,l}, m_{j,l-1}) \cdot \rho_{\hat{b}_{i,l}}(H_T) \\
&\quad - \Pr(H_T|m_{i,l}, \hat{m}_{j,l-1}) \cdot \rho_{\hat{b}_{i,l}}(H_T) \\
&\quad + \Pr(H_T|m_{i,l}, \hat{m}_{j,l-1}) \cdot \rho_{b_{i,l}}(H_T) \\
&\quad - \Pr(H_T|m_{i,l}, m_{j,l-1}) \cdot \rho_{b_{i,l}}(H_T)\|_\infty \quad (|\rho_{\hat{b}_{i,l}}| \leq |\rho_{b_{i,l}}|) \\
&\leq \|\rho_{\hat{b}_{i,l}}(H_T) - \rho_{b_{i,l}}(H_T)\|_\infty \cdot \|\Pr(H_T|m_{i,l}, m_{j,l-1}) \\
&\quad - \Pr(H_T|m_{i,l}, \hat{m}_{j,l-1})\|_1 \quad (\text{Hölder's inequality}) \\
&\leq (R_i^{max} - R_i^{min})T \times 2\epsilon \quad (\text{Pinsker's inequality})
\end{aligned}$$

Matters become more complex when we additionally prune models in the subsequent model nodes as well. This is because rather than comparing over distributions given each history of i , we sample i 's action-observation history. Consequently, additional error incurs due to the sampling, which is difficult to bound. Finally, Doshi and Zeng [1] show that it is difficult to usefully bound the error if lower-level models are themselves solved approximately.

In summary, error in i 's behavior due to pruning ϵ -BE models in the initial model node may be bounded, but we continue to investigate how to usefully bound the error due to multiple additional approximations.

7 Experimental Evaluation

We implemented the algorithms in Figs. 4 and 7 utilizing the Hugin API for DIDs and show results for the well-known two-agent *tiger problem* ($|S|=2$, $|A_i|=|A_j|=3$, $|\Omega_i|=6$, $|\Omega_j|=3$) [5] and the multiagent version of the machine maintenance (MM) problem ($|S|=3$, $|A_i|=|A_j|=4$, $|\Omega_i|=2$, $|\Omega_j|=2$) [11]. We formulate level 1 I-DIDs of increasing time horizons for the problems, and solve it approximately for varying ϵ . We show that, (i) the quality of the solution generated using our approach (ϵ -BE) improves as we reduce ϵ for given numbers of initial models of the other agent, M_0 , and approaches that of the exact solution. This is indicative of the flexibility of the approach; (ii) in comparison to the approach of updating models discriminatively (DMU) [1], which is the current efficient technique, ϵ -BE is able to obtain larger rewards for an identical number of initial models. This indicates a more informed clustering and pruning using ϵ -BE in comparison to DMU, although it is less efficient in doing so.

In Fig. 8(a, b), we show the average rewards gathered by executing the policies obtained from solving level 1 I-DIDs approximately within a simulation of the problem domain. Each data point is the average of 300 runs where the true model of j is picked randomly according to i 's belief. Notice that as we reduce ϵ the policies tend to converge to the exact (denoted by flat lines) and this remains true for different numbers

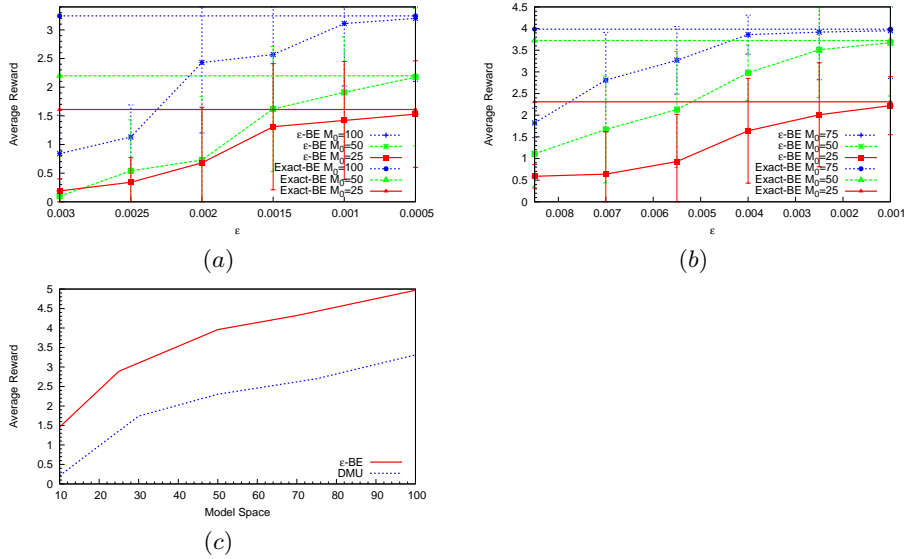


Figure 8: Performance profile obtained by solving a level 1 I-DID for the multiagent tiger problem using the ϵ -BE approach for (a) 3 horizons and (b) 4 horizons. As ϵ reduces, quality of the solution improves and approaches that of the exact. (c) Comparison of ϵ -BE and DMU in terms of the rewards obtained given identical numbers of models in the initial model node after clustering and pruning.

of initial models and across horizons. Values of these policies increase as i considers greater numbers of models thereby improving it's chances of modeling j correctly.

Next, we compare the performance of this approach with that of DMU. While both approaches cluster and prune models, DMU does so only in the initial model node, thereafter updating only those models which on update will be behaviorally distinct. Thus, we compare the average rewards obtained by the two approaches when an identical number of models remain in the initial model node after clustering and selection. This is done by varying ϵ in both approaches until the desired number of models are retained. This allows us to compare between the clustering and selection techniques of the two approaches. From Fig. 8(c), we observe that ϵ -BE results in better quality policies that obtain significantly higher average reward. This indicates that the models pruned by DMU were more valuable than those pruned by ϵ -BE, thereby testifying to the more informed way in which we compare between models by directly gauging the impact on i 's history. DMU's approach of measuring simply the closeness of beliefs in models for clustering results in significant models being pruned. However, the trade off is the increased computational cost in calculating the distributions over future paths. To illustrate, ϵ -BE consumed an average of 9.1 secs in solving a 4 horizon I-DID with 25 initial models and differing ϵ , which represents approximately a three-fold increase compared to DMU.

We show a similar set of results for the MM problem in Fig. 9. The MM problem differs in having one more physical state and action in comparison to the tiger problem,

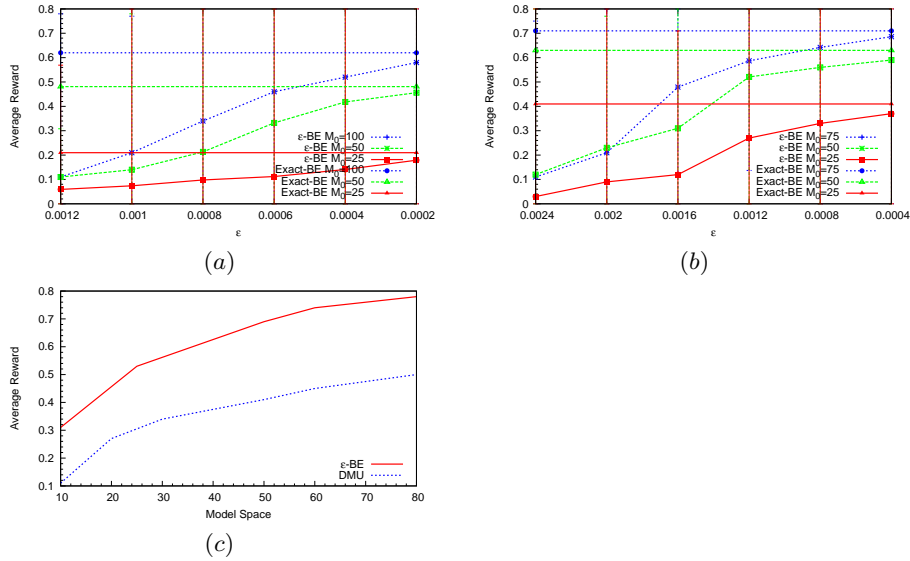


Figure 9: Performance profile for the multiagent MM problem obtained by solving level 1 I-DIDs approximately using ϵ -BE for (a) 3 horizon and (b) 4 horizon. Reducing ϵ results in better quality solutions. (c) Significant increase in rewards obtained for ϵ -BE given identical numbers of retained models in the initial model node. This is indicative of the more accurate way of judging behavioral similarity between models that we employ here.

and less observations. We observe a similar convergence toward the performance of the exact solution as we gradually reduce ϵ . This affirms the flexibility in selecting ϵ provided by the approach. Furthermore, we again note the significant increase in average reward exhibited by ϵ -BE compared to DMU given an identical number of retained models. This clearly illustrates the improvement in clustering models that are truly approximately similar, in comparison to using heuristics such as closeness of beliefs. Thus, models retained by ϵ -BE are significantly more valuable than those retained by DMU translating into greater reward, albeit at the cost of efficiency. The approach exhibited a five-fold increase in time taken compared to DMU in order to solve a horizon 4 I-DID with 25 initial models.

In summary, experiments on two multiagent problem domains indicate that the ϵ -BE approach models behavioral similarity between models of the other agent more accurately resulting in favorable performance in terms of quality of the solutions, but at the expense of computational efficiency.

8 Related Work

Suryadi and Gmytrasiewicz [12] in an early piece of related work, proposed modeling other agents using IDs. The approach proposed ways to modify the IDs to better reflect

the observed behavior. However, unlike I-DIDs, other agents did not model the original agent and the distribution over the models was not updated over time based on the actions and observations.

I-DIDs contribute to a growing line of work that includes multiagent influence diagrams (MAIDs) [6], and more recently, networks of influence diagrams (NIDs) [3, 4]. These formalisms seek to explicitly and transparently model the structure that is often present in real-world problems by decomposing the situation into chance and decision variables, and the dependencies between the variables. MAIDs objectively analyze the game, efficiently computing the Nash equilibrium profile by exploiting the independence structure. NIDs extend MAIDs to include agents’ uncertainty over the game being played and over models of the other agents.

Both MAIDs and NIDs provide an analysis of the game from an external viewpoint, and adopt Nash equilibrium as the solution concept. However, equilibrium is not unique – there could be many joint solutions in equilibrium with no clear way to choose between them – and incomplete – the solution does not prescribe a policy when the policy followed by the other agent is not part of the equilibrium. Specifically, MAIDs do not allow us to define a distribution over non-equilibrium behaviors of other agents. Furthermore, their applicability is limited to static single play games. Interactions are more complex when they are extended over time, where predictions about others’ future actions must be made using models that change as the agents act and observe. I-DIDs seek to address this gap by offering an intuitive way to extend sequential decision making as formalized by DIDs to multiagent settings. They allow the explicit representation of other agents’ models as the values of a special *model node*. Other agents’ models and the original agent’s beliefs over these models are then updated over time.

As we mentioned, a dominating cause of the complexity of I-DIDs is the exponential growth in the candidate models over time. Using the heuristic that models whose beliefs are spatially close are likely to be behaviorally equivalent, Doshi et al. [2] utilized a k -means approach to cluster models together and select K models closest to the means in the model node at each time step. While this approach requires all models to be expanded before clustering is applied, a recent approach [1] preemptively avoids expanding models that will turn out to be behaviorally equivalent to others. By discriminating between model updates, the approach guarantees a minimal set of models in each non-initial model node. To avoid solving all of the initial models, the approach clusters models whose beliefs are within a parameter, ϵ . Minimal sets of models were previously discussed by Pynadath and Marsella in [9], which used the concept of behavioral equivalence, introduced earlier [10], to form the space.

9 Discussion

I-DIDs provide a graphical formalism for modeling the sequential decision making of an agent in an uncertain multiagent setting. Our results demonstrate the potential for obtaining flexible approximations of I-DIDs by pruning models that are approximately BE. Redefining BE by explicitly focusing on the impact that the other agents’ models have on the subject agent in the interaction allows us to better identify behavioral

similarity. This translates into solutions of better quality given a limit on the number of models that could be held in memory. Consequently, other approaches would need more models to achieve comparable quality, which could translate into better efficiencies for our approach.

However, we face the challenge of computing distributions over a number of paths that grow exponentially with horizon. We are investigating ways of reducing the computational cost, for example, by directly computing the distributions instead of using the DBN and preemptively discriminating between model updates. Additionally, mitigating the impact of the curse of history of the subject agent on this approach is crucial. We are optimistic that this can be done in a systematic way, and this will facilitate application to larger multiagent problem domains.

References

- [1] P. Doshi and Y. Zeng. Improved approximation of interactive dynamic influence diagrams using discriminative model updates. In *AAMAS*, pages 907–914, 2009.
- [2] P. Doshi, Y. Zeng, and Q. Chen. Graphical models for interactive pomdps: Representations and solutions. *Journal of Autonomous Agents and Multiagent Systems*, 18(3):376–416, 2009.
- [3] K. Gal and A. Pfeffer. Networks of influence diagrams: A formalism for representing agents’ beliefs and decision-making processes. *Journal of Artificial Intelligence Research*, 33:109–147, 2008.
- [4] Y. Gal and A. Pfeffer. A language for modeling agent’s decision-making processes in games. In *AAMAS*, pages 265–272, 2003.
- [5] P. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multiagent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.
- [6] D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. In *IJCAI*, pages 1027–1034, 2001.
- [7] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [8] J. Pineau, G. Gordon, and S. Thrun. Anytime point-based value iteration for large pomdps. *Journal of Artificial Intelligence Research*, 27:335–380, 2006.
- [9] D. Pynadath and S. Marsella. Minimal mental models. In *AAAI*, pages 1038–1044, 2007.
- [10] B. Rathnas., P. Doshi, and P. J. Gmytrasiewicz. Exact solutions to interactive pomdps using behavioral equivalence. In *AAMAS*, pages 1025–1032, 2006.
- [11] R. Smallwood and E. Sondik. The optimal control of partially observable markov decision processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

- [12] D. Suryadi and P. Gmytrasiewicz. Learning models of other agents using influence diagrams. In *User Modeling*, pages 223–232, 1999.
- [13] J. A. Tatman and R. D. Shachter. Dynamic programming and influence diagrams. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2):365–379, 1990.
- [14] Y. Zeng, P. Doshi, and Q. Chen. Approximate solutions of interactive dynamic influence diagrams using model clustering. In *AAAI*, pages 782–787, 2007.

Appendix

of Proposition 1. $Pr(h_{T-t}|m_{i,l}^t, m_{j,l-1}^t) = Pr(h_{T-t-1}, a_i^t, o_i^{t+1}|m_{i,l}^t, m_{j,l-1}^t) = Pr(h_{T-t-1}|a_i^t, o_i^{t+1}, m_{i,l}^t, m_{j,l-1}^t) Pr(a_i^t, o_i^{t+1}|m_{i,l}^t, m_{j,l-1}^t)$ (using Bayes rule)

We focus on the first term next:

$$\begin{aligned} Pr(h_{T-t-1}|a_i^t, o_i^{t+1}, m_{i,l}^t, m_{j,l-1}^t) &= \sum_{a_j^t, o_j^{t+1}} Pr(h_{T-t-1}|a_i^t, o_i^{t+1}, m_{i,l}^t, a_j^t, o_j^{t+1}, m_{j,l-1}^t) Pr(a_j^t, o_j^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t) \\ &= Pr(h_{T-t-1}|m_{i,l}^{t+1}, m_{j,l-1}^{t+1}) Pr(a_j^t, o_j^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t) \end{aligned}$$

In the above equation, the first term results due to an update of the models at time step t with actions and observations. This term is computed recursively. For the second term, j 's level $l-1$ actions and observations are independent of i 's observations.

We now focus on the term, $Pr(a_i^t, o_i^{t+1}|m_{i,l}^t, m_{j,l-1}^t)$:

$$\begin{aligned} Pr(a_i^t, o_i^{t+1}|m_{i,l}^t, m_{j,l-1}^t) &= Pr(o_i^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t) \\ &\times Pr(a_i^t|OPT(m_{i,l}^t)) \quad (i\text{'s action is conditionally independent of } j \text{ given its model)} \\ &= Pr(a_i^t|OPT(m_{i,l}^t)) \sum_{a_j^t} Pr(o_i^{t+1}|a_i^t, a_j^t, m_{i,l}^t, m_{j,l-1}^t) \\ &\times Pr(a_j^t|OPT(m_{j,l-1}^t)) \\ &= Pr(a_i^t|OPT(m_{i,l}^t)) \sum_{a_j^t} Pr(o_i^{t+1}|a_i^t, a_j^t, m_{i,l}^t) \\ &\times Pr(a_j^t|OPT(m_{j,l-1}^t)) \quad (i\text{'s observation is conditionally independent of } j\text{'s model)} \\ &= Pr(a_i^t|OPT(m_{i,l}^t)) \sum_{a_j^t} Pr(a_j^t|OPT(m_{j,l-1}^t)) Pr(o_i^{t+1}|a_i^t, a_j^t, b_{i,l}^t) \quad (b_{i,l}^t \text{ is } i\text{'s belief in model, } m_{i,l}^t) \\ &= Pr(a_i^t|OPT(m_{i,l}^t)) \sum_{a_j^t} Pr(a_j^t|OPT(m_{j,l-1}^t)) \\ &\times \sum_{s^{t+1}} Pr(o_i^{t+1}|s^{t+1}, a_i^t, a_j^t) Pr(s^{t+1}|a_i^t, a_j^t, b_{i,l}^t) \\ &= Pr(a_i^t|OPT(m_{i,l}^t)) \sum_{a_j^t} Pr(a_j^t|OPT(m_{j,l-1}^t)) \sum_{s^{t+1}} O_i(s^{t+1}, a_i^t, a_j^t, o_i^{t+1}) \sum_{s, m_j} T_i(s, a_i^t, a_j^t, s^{t+1}) b_{i,l}^t(s, m_j) \end{aligned}$$

where O_i and T_i are i 's observation and transition functions respectively, in the I-DID denoted by model, $m_{i,l}^t$. This proves Eq. 1 in Proposition 1.

Finally, we move to the term, $Pr(a_j^t, o_j^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t)$, to obtain Eq. 2:

$$\begin{aligned} Pr(a_j^t, o_j^{t+1}|a_i^t, m_{i,l}^t, m_{j,l-1}^t) &= Pr(o_j^{t+1}|a_j^t, a_i^t, m_{i,l}^t, m_{j,l-1}^t) \\ &\times Pr(a_j^t|a_i^t, m_{i,l}^t, m_{j,l-1}^t) \\ &= Pr(o_j^{t+1}|a_j^t, a_i^t, m_{i,l}^t, m_{j,l-1}^t) Pr(a_j^t|OPT(m_{j,l-1}^t)) \quad (j\text{'s action is conditionally independent of } i \text{ given its model)} \end{aligned}$$

$$\begin{aligned}
&= Pr(a_j^t | OPT(m_{j,l-1}^t)) \sum_{s^{t+1}} Pr(o_j^{t+1} | a_j^t, a_i^t, s^{t+1}) \\
&\times Pr(s^{t+1} | a_j^t, a_i^t, m_{i,l}^t, m_{j,l-1}^t) \\
&= Pr(a_j^t | OPT(m_{j,l-1}^t)) \sum_{s^{t+1}} O_j(s^{t+1}, a_j^t, a_i^t, o_j^{t+1}) \sum_{s, m_j} \\
&Pr(s^{t+1} | a_j^t, a_i^t, s) b_{i,l}^t(s, m_j) \quad (b_{i,l}^t \text{ is } i\text{'s belief in } m_{i,l}^t) \\
&= Pr(a_j^t | OPT(m_{j,l-1}^t)) \sum_{s^{t+1}} O_j(s^{t+1}, a_j^t, a_i^t, o_j^{t+1}) \sum_{s, m_j} \\
&T_i(s, a_i^t, a_j^t, s^{t+1}) b_{i,l}^t(s, m_j) \quad (\text{agent } i\text{'s I-DID is used})
\end{aligned}$$

where O_j is j 's observation function in model $m_{j,l-1}^t$, which is a part of i 's I-DID. ■