Approximating Behavioral Equivalence of Models Using Top-K Policy Paths

(Extended Abstract)

Yifeng Zeng
Dept. of Computer Science
Aalborg University
DK-9220 Aalborg, Denmark
yfzeng@cs.aau.edu

Yingke Chen
Dept. of Computer Science
Aalborg University
DK-9220 Aalborg, Denmark
ykchen@cs.aau.dk

Prashant Doshi
Dept. of Computer Science
University of Georgia
Athens, GA 30602, USA
pdoshi@cs.uga.edu

ABSTRACT

Decision making and game play in multiagent settings must often contend with behavioral models of other agents in order to predict their actions. One approach that reduces the complexity of the unconstrained model space is to group models that tend to be behaviorally equivalent. In this paper, we seek to further compress the model space by introducing an approximate measure of behavioral equivalence and using it to group models.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

General Terms

Algorithms, Experimentation

Keywords

decision making, agent modeling, behavioral equivalence

1. INTRODUCTION

Several areas of multiagent systems such as decision making and game playing benefit from modeling other agents sharing the environment, in order to predict their actions. In the absence of constraining assumptions about the behaviors of other agents, the general space of these models is very large. Multiple researchers have proposed grouping together *behaviorally equivalent (BE)* models [2, 6, 7] to reduce the number of possible models. Models that are BE prescribe identical behavior, and these may be grouped because it is the prescriptive aspects of the models and not the descriptive that matter to the decision maker. The basic idea is to cluster behaviorally equivalent models of the other agents and select representative models for each cluster. By doing this, we are able to limit the model space of the other agents while maintaining the solution optimality of the modeling agent. One particular decision making framework in which BE has received much attention is the interactive dynamic influence diagram (I-DID) [5].

I-DIDs are 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 possibly similar ways. Previous I-DID solutions, including both exact and approximate ones, mainly exploit the concept of BE to reduce the dimensionality of the state space. For example, Doshi and Zeng [4] minimize the model space by updating only those models that lead to behaviorally distinct models at the next time step. While this approach speeds up solutions of I-DIDs considerably and is the state of the art, it doesn't scale desirably to large horizons. This is because: (a) models are compared for BE using their solutions which tend to be policy trees. As the horizon increases, the size of the policy tree increases exponentially; (b) the condition for BE is quite strict: entire policy trees of two models must match exactly. While this can be done bottom up [4], the complexity of this depends on the size of the policy tree.

Progress in the context of BE is possible by grouping models that are likely to be BE. Because this will potentially result in more models being clustered, the model space is partitioned into less number of classes. In this paper, we introduce a way to identify models that are approximately BE by limiting attention to paths in a policy tree that are most likely. Models are approximately BE and may be grouped together if these K most likely policy paths are identical. Because we focus on a subset of the policy tree for comparison, more models may be included in a single approximate BE group. However, computing the probability of an action-observation path in a multiagent setting requires knowledge of the actions of the modeling agent as well [3]. We address this fundamental barrier by utilizing a more probabilistic choice model for the other agent instead of using the traditional maximum utility action(s). Specifically, we employ the *quantal response* model [1] – fast emerging as a viable alternative choice model for agents – to compute the policy. Our hypothesis is that by allowing for more actions (not just those that have maximum utility) we consider a larger number of possible paths and select the likely paths among these. In computing the probability of a path, we do not consider actions of the modeling agent, but those of the other agent only or those of the subject agent modeled at a lower level by the other.

2. TOP K POLICY PATHS

We label the sequence of actions and observations experienced by an agent participating in an interaction as a *path*. Formally, let $h_j^q = \{a_j^t, o_j^{t+1}\}_{t=1}^q$ be the q-length path for an agent j where o_j^{T+1} is null for a T horizon problem ($q \leq T$). If $a_j^t \in A_j$ and $o_j^{t+1} \in \Omega_j$, where A_j and Ω_j are agent j's action and observation sets respectively, then the set of all q-length paths is, $H_j^q = \Pi_1^q(A_j \times \Omega_j)$. In a two-agent interaction, the probability of j experiencing an observation depends on actions of both

Cite as: Approximating Behavioral Equivalence of Models Using Top-K Policy Paths (Extended Abstract), Yifeng Zeng, Yingke Chen and Prashant Doshi, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1229-1230.

agents. Because an agent's optimal actions are obtained from its model ($m_{j,l-1}^t$ for j and $m_{i,l}^t$ for the subject agent i), we define the probability of a q-length path in a factored form as shown below:

$$\begin{aligned} Pr(h_j^q) = \; & \Pi_{t=1}^q Pr(a_j^t|m_{j,l-1}^t) \textstyle\sum_{a_i \in A_i} Pr(o_j^{t+1}|h_j^{t-1}, a_j^t, a_i^t) \\ & \times Pr(a_i^t|m_{i,l}^t) \end{aligned} \tag{1}$$

We then define the most probable path of T horizon below.

Definition 1 (Most Probable Path). *Define the most probable path, h_j^T, for the level $l-1$ agent j as:*

$$\begin{aligned} h_j^T = \; & \underset{h_j^T \in H_j^T}{argmax} \; \Pi_{t=1}^q Pr(a_j^t|m_{j,l-1}^t) \textstyle\sum_{a_i \in A_i} Pr(o_j^{t+1}|h_j^{t-1}, \\ & a_j^t, a_i^t) Pr(a_i^t|m_{i,l}^t) \end{aligned}$$

Intuitively, K-most probable paths are then those K paths that have the largest probabilities among all the paths of T horizon.

Although Eq. 1 provides us with a way to compute path probabilities, it requires the solution of the subject agent i's model (in the term, $Pr(a_i^t|m_{i,l}^t)$). This is a fundamental barrier to using the exact path probabilities because agent i's level l solution is what we seek and is not known. Clearly, exact path probabilities may not be available for use in any approach for solving I-DIDs (or other such frameworks). Another challenge is that the number of paths grows exponentially with time. However, we address this issue by focusing on K paths only at every time step.

One way around the problem of computing exact path probabilities is to utilize a quick but inexact solution for i's model with the guarantee that optimal actions are given higher utility in the inexact solution as well. To the best of our knowledge, we are unaware of such an approximation technique. Instead, we utilize a more probabilistic solution of j's models that would allow for more paths considered plausible while continuing to assign higher probabilities to optimal actions, thereby compensating for not knowing i's action probabilities. We utilize the *quantal response* [1] model, which assigns a probability to each action in proportion to its utility. Formally, the quantal response is defined in Eq. 2:

$$Pr(a_j^t|m_{j,l-1}^t) = \frac{e^{\lambda EU(a_j^t)}}{\sum_{a_j^t \in A_j} e^{\lambda EU(a_j^t)}} \tag{2}$$

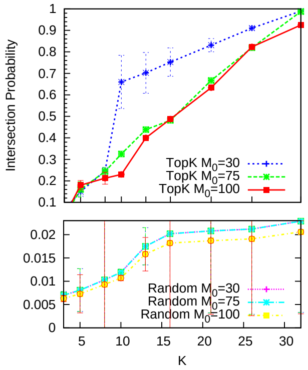
Non-negative parameter λ quantifies the rationality of the actions.

In order to identify the top K paths, we replace the decision nodes in j's level $l-1$ I-DID (or DID) with the corresponding chance nodes effectively turning the DID into a dynamic Bayesian network (DBN). In order to avoid searching over an exponential number of policy paths, $|A_j||\Omega_j|^{T-1}$ where T is the horizon, we identify exactly K paths at every time step. Specifically, at time $t=0$, we compute the probabilities for $|A_j||\Omega_j|$ action-observation combinations and select K-most probable ones. Thereafter, at any time step until $T-1$, we compute the probabilities of $K|A_j||\Omega_j|$ paths and select K most probable paths (as per Def. 1) among them. Consequently, we obtain K most probable paths while avoiding an exponential number of path probability computations.

Models that have identical top K paths are grouped together. We pick a representative model from each group and prune all other models in the group. All the representative models are retained and updated. We point out that unlike exact BE, we compare just a subset of the policy paths in order to group the models. On the other hand, because we use the quantal response the top K paths are not necessarily the most probable paths in the original policy tree obtained when the maximum expected utility is used. Consequently, models that were originally BE may not be grouped together. As a result, we are unable to precisely characterize the error in predicting j's actions due to this approach.

3. RESULTS

We implemented this approach (**TopK**) within the framework of I-DIDs. In order to demonstrate the suitability of using the quantal response model for j's actions, we implemented a baseline approach that selects top K paths using randomized response for j's actions. In Fig. 1(a), we show that TopK maintains a relatively high chance of fully intersecting the actual K most probable paths. Increasing K improves the likelihood as we may expect.



(a)

Level 1	T	Time (s)	
		DMU	TopK
Tiger	10	2.7	1.4
	14	77	45.6
	20	*	238
	25	*	697
MM	8	0.6	0.2
	10	3.1	1.9
	14	91	60.3
	20	*	805
UAV	6	6.5	4.9
	8	166.6	111
	10	*	462

(b)

Figure 1: (a) **TopK captures the K-most probable paths with a large probability in the multiagent tiger problem.** (b) **TopK scales significantly better than DMU to larger horizons. All experiments are run on a dual processor Xeon 2.0GHz, 2GB memory and WinXP platform.**

In Fig. 1(b), we show the reduced running times and improved scalability of TopK compared with the DMU approach [4] over three domains. We were able to solve I-DIDs over more than 25 horizon using TopK. More significantly, for the large UAV domain we achieved solutions to I-DIDs for horizon of more than 10.

Acknowledgment Yifeng Zeng acknowledges partial support from the Obel Family Foundation, Denmark and NSFC (#60974089 and #60975052). Prashant Doshi acknowledges partial support from an NSF CAREER grant, #IIS-0845036.

4. REFERENCES

[1] C. F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

[2] E. Dekel, D. Fudenberg, and S. Morris. Topologies on types. *Theoretical Economics*, 1:275–309, 2006.

[3] P. Doshi, M. Chandrasekaran, and Y. Zeng. Epsilon-subjective equivalence of models for interactive dynamic influence diagrams. In *WIC/ACM/IEEE WI-IAT*, pages 165–172, 2010.

[4] P. Doshi and Y. Zeng. Improved approximation of interactive dynamic influence diagrams using discriminative model updates. In *AAMAS*, pages 907–914, 2009.

[5] P. Doshi, Y. Zeng, and Q. Chen. Graphical models for interactive pomdps: Representations and solutions. *Journal of AAMAS*, 18(3):376–416, 2009.

[6] D. Pynadath and S. Marsella. Minimal mental models. In *AAAI*, pages 1038–1044, Vancouver, Canada, 2007.

[7] B. Rathnas., P. Doshi, and P. J. Gmytrasiewicz. Exact solutions to interactive pomdps using behavioral equivalence. In *AAMAS*, pages 1025–1032, 2006.