Team Behavior in Interactive Dynamic Influence Diagrams with Applications to Ad Hoc Teams

(Extended Abstract)

Muthukumaran Chandrasekaran
University of Georgia, USA
mkran@uga.edu

Prashant Doshi
University of Georgia, USA
pdoshi@cs.uga.edu

Yifeng Zeng
Teesside University, UK
y.zeng@tees.ac.uk

Yingke Chen
Queen's University Belfast, UK
y.chen@qub.ac.uk

ABSTRACT

Planning for ad hoc teamwork is challenging because it involves individual agents collaborating with others without any prior coordination. However, individual decision making in multiagent settings faces the task of having to reason about other agents' actions, which in turn involves reasoning about others. An established approximation that operationalizes this approach is to bound the infinite nesting from below by introducing level 0 models, which results in suboptimal team solutions in cooperative settings. We demonstrate this limitation and mitigate it by integrating learning into planning. The augmented framework is applied to ad hoc teamwork.

Categories and Subject Descriptors

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

Keywords

multiagent planning; ad hoc teamwork; reinforcement learning

1. INTRODUCTION

Ad hoc teamwork involves a team of agents coming together to cooperate without any prior coordination or communication protocols [3]. The preclusion of prior commonality makes planning in ad hoc settings challenging thereby making frameworks such as DEC-POMDPs unsuitable for ad hoc teamwork. Other approaches such as online planning in ad hoc teams (OPAT) [3] assume perfect observability of physical states and others' actions, which often may not apply. Our focus is on how an inividual agent should behave online as an ad hoc teammate in partially observable settings with minimal prior assumptions. Frameworks such as interactive dynamic influence diagrams (I-DIDs) [1] are recognized to be suitable for ad hoc teamwork but their complexity is challenging.

While recent advances on model equivalence [4] allow frameworks such as I-DIDs to scale, another significant challenge that merits attention is due to the finitely-nested modeling used in these frameworks, which assumes the presence of level 0 models that do not explicitly reason about others. By augmenting I-DIDs by additionally attributing a new type of level 0 model that utilizes reinforcement learning (RL), we show a principled emergence of team behavior for the first time. We demonstrate the applicability of the *augmented* I-DIDs to ad hoc settings and show its effectiveness for varying types of teammates. We experiment with multiple cooperative domains and perform a baseline comparison with a generalized version of OPAT [3] that accounts for the partial observability.

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

2. TEAMWORK IN INTERACTIVE DIDS

Teamwork involves agents collaborating to optimize the team reward; ad hoc teamwork imposes no pre-coordination. We begin by showing that the finitely-nested hierarchy in I-DID (and I-POMDP) does not facilitate teamwork. Figure 1 shows a setting of two agents, i and j, in a *grid meeting* problem. If each agent deliberates at its own level, agent i modeled at level 0 chooses to move left while a level 0 agent j chooses to move down. Each agent obtains a reward of 15 and the team gets 30. Agent i modeled at *level 1* and modeling j at level 0 thinks that j will move down, and its own best response to predicted j's behavior is to move left. Analogously, a level 1 agent j would choose to move down. A level 2 agent i will predict that a level 1 j moves down as mentioned previously, due to which it decides to move left. Analogously, a level 2 agent j continues to decide to move down. We may apply this reasoning inductively to conclude that finite level $l \geq 0$ agents i and j would move left and down, respectively, earning a joint reward of 30. However, the optimal team behavior in this setting is for i to move right and j to move up obtaining a team reward of 40. Clearly, these finite hierarchical systems preclude the agents' optimal teamwork due to the bounded reasoning introduced by the lowest level (level 0) agents.

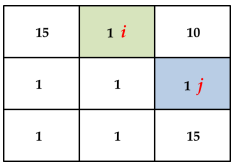

Figure 1: Multiagent grid domain. Numbers denote rewards.

Notice that an offline specification of level 0 models in cooperative settings is necessarily incomplete. This is because the true benefit of cooperative actions often hinges on others performing supporting actions, which by themselves may not be highly rewarding to the agent. Thus, despite solving the level 0 models optimally, the agent may not reliably engage in optimal team behavior.

While it is difficult to *a priori* discern the benefit of moving up for agent j in Fig. 1, it could be *experienced* by the agent. Specifi-

cally, it may explore moving in different directions including moving up and learn about its benefit from the ensuing, possibly indirect, team reward. Subsequently, we may expect an agent to learn policies that are consistent with optimal teammate behavior because the corresponding actions provide large reinforcements. For example, given that agent i moves right in Fig. 1, j may choose to move up in its exploration, and thereby receive a large reinforcing reward. This observation motivates formulating level 0 models that utilize RL to generate the predicted policy for the modeled agent. Essentially, we expect that RL with its explorations would compensate for the lack of teamwork caused by bounded reasoning in finitely-nested I-DIDs.

We therefore augment the level 0 model space, $\mathcal{M}'_{j,0}$, by additionally attributing a new type of level 0 model to the other agent j:$m'_{j,0} = \langle b_{j,0}, \hat{\theta}'_j \rangle$, where $b_{j,0}$ is j's belief and $\hat{\theta}'_{j,0}$ is the frame of the learning model. The frame, $\hat{\theta}'_{j,0}$, includes the learning rate, α; seed policy, π'_j, of planning horizon, T, with a fair amount of exploration; and the chance and utility nodes of the DID along with a candidate policy of agent i, which could be an arbitrary policy from i's policy space, Π_i, as agent i's actual behavior is not known.

Level 0 agent j learns its policy while agent i's actions are a part of the environment (hidden). As j's model space is inclusive of i's policy space, it is intractable to learn a policy for all i's policies. Considering that few of i's policies are actually collaborative, we formulate a principled way to reduce the full space to those policies of i that could be collaborative. We may further reduce agent j's policy space by keeping *top-K* policies of j in terms of their expected utilities. Agent j's policy space will be additionally reduced because behaviorally equivalent models – learning and other models with identical solutions – will be clustered [4].

3. EXPERIMENTAL RESULTS

We adapt Perkin's Monte Carlo Exploring Starts for POMDPs (MCESP) [2], which learn good policies in fewer iterations while making no prior assumptions about the agent's models to perform the RL. We empirically evaluate the performance of Aug. I-DIDs in *three* well-known cooperative domains: 3×3 grid meeting (Grid), box-pushing (BP), and multi-access broadcast channel (MABC).

In the first set of experiments, we show that Aug. I-DIDs facilitate team behavior, which was traditionally implausible (see Table 1). We observe that Aug. I-DID's solutions approach the globally optimal team behavior as generated by GMAA*-ICE. We observe that the larger weights on the learned policies lead to better quality i's policies. The small gap from the optimal DEC-POMDP value is due to the uncertainty over different models of j. *Furthermore, Aug. I-DID generates the optimal team behavior identical to that of GMAA*-ICE when i's belief places probability 1 on the true model of j, as in Dec-POMDPs.*

Next, we apply the Aug. I-DIDs in an ad hoc setting similar to the one used by Wu *et al.* [3] (Table 2) involving different teammate types including teammate policies that may not be most effective in advancing the joint goal. Aug. I-DID's better performance is in part due to the sophisticated belief update that gradually increases the probability on the true model if it is present in j's model space. Consequently, they allow better adaptability than OPAT which focuses on a single optimal behavior of teammates during planning. On the other hand, OPAT takes significantly less time because it approximates the problem by solving a series of stage games modeling the other agent using a single type. Further experiments on the robustness of Aug. I-DIDs in dynamic settings showed that agent i obtained significantly better average rewards compared to OPAT for the setting where the other agent is of type *predefined* and after 15 steps is substituted by an *optimal* type for the remaining 15 steps in the MABC domain.

Table 1: Performance comparison between the trad. I-DID, aug. I-DID, and GMAA*-ICE (shown only for largest horizon)

		Aug. I-DID		Trad I-DID
Domain	**K**	**Uniform**	**Diverse**	**Uniform**
Grid (T=4)	100	37.15	53.26	21.55
	64	35.33	53.26	
	32	35.33	53.26	
	Dec-POMDP(GMAA*-ICE): 58.75			
BP (T=3)	32	73.45	76.51	4.75
	16	73.45	76.51	
	8	71.36	76.51	
	Dec-POMDP(GMAA*-ICE): 85.18			
MABC (T=5)	64	4.08	4.16	3.29
	32	3.99	4.16	
	16	3.99	4.16	
	Dec-POMDP(GMAA*-ICE): 4.79			

Table 2: Baseline Comparison with OPAT with different types of teammates. Each datapoint is the average of 10 runs.

Ad Hoc Teammate	OPAT	Aug. IDID
Grid *T=20, look-ahead=3*		
Random	12.25 ± 1.26	14.2 ± 0.84
Predefined	11.7 ± 1.63	16.85 ± 1.35
Optimal	28.35 ± 2.4	27.96 ± 1.92
BP *T=20, look-ahead=3*		
Random	29.26 ± 2.17	36.15 ± 1.95
Predefined	41.1 ± 1.55	54.43 ± 3.38
Optimal	52.11 ± 0.48	59.2 ± 1.55
MABC *T=20, look-ahead=3*		
Random	9.68 ± 1.37	12.13 ± 1.08
Predefined	12.8 ± 0.65	13.22 ± 0.21
Optimal	16.64 ± 0.28	15.97 ± 1.31

4. CONCLUSION

Self-interested individual decision makers face hierarchical belief systems in their multiagent planning. We explicated a negative consequence of bounding the hierarchy: the agent may not behave as an optimal teammate. By integrating learning in planning, we show the emergence of team behavior. This facilitates a natural application to ad hoc teamwork with no pre-coordination for which they are well suited. By allowing models formalized as I-DIDs or DIDs to vary in the beliefs and frames, we considered an exhaustive and general space of models during planning. Aug. I-DIDs provide a bridge between multiagent planning frameworks such as DEC-POMDPs and joint learning for cooperative domains.

Acknowledgement We acknowledge support from an ONR grant, N000141310870, and a NSF CAREER grant, IIS-0845036.

5. REFERENCES

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

[2] T. J. Perkins. Reinforcement learning for pomdps based on action values and stochastic optimization. In *AAAI*, pages 199–204, 2002.

[3] F. Wu, S. Zilberstein, and X. Chen. Online planning for ad hoc autonomous agent teams. In *IJCAI*, pages 439–445, 2011.

[4] Y. Zeng and P. Doshi. Exploiting model equivalences for solving interactive dynamic influence diagrams. *JAIR*, 43:211–255, 2012.