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 individual 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.

2. TEAMWORK IN INTERACTIVE DIDS

Teamwork involves agents collaborating to optimize the team reward; ad hoc teamwork imposes no precoordination. 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 \( 1 \leq 0 \) 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.

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.

![Figure 1: Multiagent grid domain. Numbers denote rewards.](image)
cally, it may explore moving in different directions including mov-
ing up and learn about its benefit from the ensuing, possibly indi-
rect, team reward. Subsequently, we may expect an agent to learn
policies that are consistent with optimal teammate behavior be-
cause 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 com-
pensate 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 addi-
tionally attributing a new type of level 0 model to the other agent
$j: \mathcal{M}^i_0 = (b_i, \theta^i_0, \theta^j_0)$, where $b_i$ is $j$’s belief and $\theta^j_0$ is the frame of
the learning model. The frame, $\theta^j_0$, includes the learning rate, $\alpha$;
seed policy, $\pi^j_0$, of planning horizon, $\mathcal{T}$, with a fair amount of ex-
dploration; 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, $\Pi_i$, 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 poli-
cies 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 re-
duced 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: $3 \times 3$ grid meeting (Grid),
box-pushing (BP), and multi-access broadcast channel (MABC).
In the first set of experiments, we show that Aug. I-DIDs facil-
itate team behavior, which was traditionally imimplausible (see Ta-
table 1). We observe that Aug. I-DID’s solutions approach the glo-
ally optimal team behavior as generated by GMAA*-ICE. We ob-
serve 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 $j$. Furth-
ere more, Aug. 1-IDID generates the optimal team behavior identical to
that of GMAA*-ICE when it’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. 1-IDID’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 fo-
cuses 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 mod-
eling 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.

4. CONCLUSION
Self-interested individual decision makers face hierarchical be-
ief 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 ap-
lication 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 exhaust-
ive 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
interactive pomdps: Representations and solutions. JAAMAS,
action values and stochastic optimization. In AAAI, pages
solving interactive dynamic influence diagrams. JAIR,