From Policies to Influences: A Framework for Nonlocal Abstraction in Transition-Dependent Dec-POMDP Agents

(Extended Abstract)

Stefan J. Witwicki and Edmund H. Durfee
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109
{witwicki,durfee}@umich.edu

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*

General Terms

Theory, Algorithms

Keywords

Multiagent Systems, Stochastic Planning, Coordination, Decentralized Partially-Observable Markov Decision Processes, Structured Interactions, Policy Abstraction

1. INTRODUCTION

Decentralized Partially-Observable Markov Decision Processes (Dec-POMDPs) are powerful theoretical models for deriving optimal coordination policies of agent teams in environments with uncertainty. Unfortunately, their general NEXP solution complexity [3] presents significant challenges when applying them to real-world problems, particularly those involving teams of more than two agents. Inevitably, the policy space becomes intractably large as agents coordinate joint decisions that are based on dissimilar beliefs about an uncertain world state and that involve performing actions with stochastic effects. Our work directly confronts the policy space explosion with the intuition that instead of coordinating all policy decisions, agents need only coordinate abstractions of their policies that constitute the essential influences that they exert on each other.

As a running example, consider the problem shown in Figure 1, involving two interacting rover agents (among a team of several others) that are exploring the surface of Mars. As shown, the agents perform various tasks (constrained to take place within a *window* of execution) with nondeterministic duration (D) and quality (Q) outcomes, and in performing their tasks may alter the outcomes of other agents' tasks. Here, agent 1 may choose to visit and prepare research site C, which will (in expectation) make agent 2's analysis of site C quicker and more valuable. This problem can be expressed as a Dec-POMDP whose world state includes features such as task statuses (each observed by the respective performing agent) and the rovers' positions (each partially-observed by the respective agent). Relevant environmental variables such as *time* or *sunlight* (jointly-observed) may also be included. A joint policy that maximizes the team's expected accumulation of task qualities constitutes a solution.

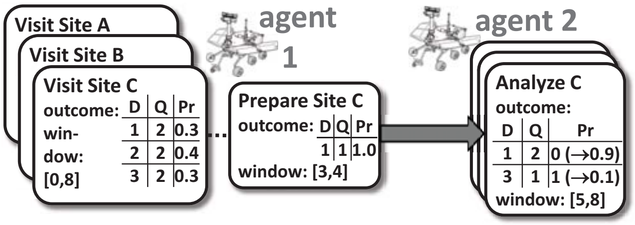

Figure 1: Rover Exploration Example.

(Rover 2)

To scale Dec-POMDPs to teams of many agents, researchers have adopted an approach that decomposes the joint policy formulation into a series of local policy formulations [1, 2, 4, 5, 6, 7]. The team searches the joint policy space by forming a set of candidate policies, to which each individual agent computes its **best-response** policy that (approximately) optimizes only its local behavior. This decomposed policy-space search has been effective in generating optimal and near-optimal solutions for restricted classes of Dec-POMDPs wherein agents can affect each others' rewards but cannot affect each others' observations nor action outcome transitions (commonly referred to as *transition-independent*, *observation-independent*, *reward-dependent* problems) [2, 6]. When applied to more general transition-dependent problems, however, the approach has only been shown to scale to two-agent teams [1] or to result in joint policies with no guarantees of optimality or near-optimality [4, 5, 7].

We have developed a framework for scaling up the aforementioned solution approach to teams of more agents in the context of a general subclass of transition-dependent Dec-POMDPs *without* sacrificing optimality. The key idea is to isolate and explicitly coordinate the agents' transition influences. With a well-defined characterization of what it means for one agent to influence another, agents can form compact models of nonlocal influence and compute best-responses to candidate influences instead of candidate policies. And instead of searching through an intractably-large policy space, agents can search through a more tractable influence space.

Cite as: From Policies to Influences: A Framework for Nonlocal Abstraction in Transition-Dependent Dec-POMDP Agents (Extended Abstract), S. Witwicki and E. Durfee, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1397-1398

2. TRANSITION-DECOUPLED POMDPS

Among the difficulties involved in solving transition-dependent problems is decoupling the joint model into compact, efficient local best-response models. In the case that agents' transitions are independent, it is easy to factor the world state s into independent *local states* $\{s_i\}$, each composed of features affected by, observed by, and affecting the transitions of agent i alone [2]. Transition dependencies like the one depicted in Figure 1 cause agents' action consequences to propagate to other agents' local states. Agent 1's preparation of site C causes changes in the transitions of agent 2's analysis task. So in order to effectively predict the consequences of its own actions, agent 2 needs to reason about agent 1's observed state and expected actions.

By explicitly acknowledging the structured dependence between individual features, we can reframe our problem as a collection of POMDP models, each representing a **local state** comprised of locally-observable features, whose feature sets may overlap. For instance, features such as *site-C-prepared* that are controlled by agent 1 but that directly affect agent 2's own action consequences are included in both agents' models. From the perspective of agent 2, these will be referred to as **nonlocally-controlled** (but locally-modeled) features. In essence, the Dec-POMDP has been decoupled into a set of local POMDPs tied to one another by the transition-dependence of their nonlocally-controlled features: *Transition-Decoupled POMDPs* (*TD-POMDPs*). Aside from being more general than related models (e.g. ED-DEC-MDPs[1]), the TD-POMPD provides a natural representation for exploiting locality of transition-dependent interaction.

3. INFLUENCE-BASED ABSTRACTION

With the TD-POMDP model structure we have defined, interagent **influence** may be characterized quite simply as the expected transition probabilities of nonlocally-controlled features. Since these probabilities are the only components of an agent's local model that may vary with the behavior of its peers, entire peer policies can be abstractly summarized by the influences they entail, and the corresponding probabilities incorporated into the agent's local POMDP for the purposes of best-response computation. The transition probabilities associated with a particular influence can be encoded with a probability distribution $Pr\left(\bar{n}'|f_1, f_2, ...\right)$, where $\bar{n}'$ are new values of nonlocally-controlled features conditioned on previous values of various state features $\bar{f} = \{f_1, f_2, ...\}$. In Figure 1, agent 1 influences agent 2 through the transitions of *site-C-prepared*. Assuming that agent 1 and agent 2 do not share any other state features (apart from a global clock signal which we call *time*), the influence distribution becomes $\{Pr\left(\textit{site-C-prepared} = \textit{true}|\textit{time} = t\right), \forall t\}$.

Our influence-based characterization is motivated by the fact that, while an agent may have many policies to choose from, it may only be able to exert a small number of unique influences on its peers. This is undoubtedly true for the example in Figure 1, where agent 1 is constrained such that it can only prepare site C between times 3 and 4. As such, any two policies that differ only in the decisions made after time 3 will result in identical influences. For some problems, the feasible **influence space** is substantially smaller than the policy space and more efficient to explore. In verifying that this property holds for a more general set of problems, we have developed an optimal influence-space search algorithm and performed empirical comparisons with state-of-the-art optimal policy search methods on randomly-generated instances. Initial results demonstrate up to two orders of magnitude speed-up and scalability to 4-agent problems that could previously only be approximately-solved.

One can envision variations of the example in which the probability of *site-C-prepared* would need to be conditioned on other jointly-observed state features like *sunlight*. The probability that agent 1 prepares site C could also be dependent on *past* values of shared features (since POMDP policy decisions may be based on entire histories of observations). We have proven that, for any TD-POMDP, the influences for the system of agents can be jointly specified with a Dynamic Bayesian Network (DBN) containing only (variables representing the past and present values of) shared state features. With such a representation, our influence-based abstraction framework has several important benefits:

- **Compactness.** The size of the influence representation is a function of the degree to which agents are dependent on their peers, *not* of the number of agents in the system.
- **Flexibility for Approximation.** Since we are representing influences with probability distributions, there are a variety of common techniques that we can apply to approximate the representation of a single influence or to approximate the space of possible influences (in support of approximately-optimal solution methods).
- **Privacy.** Because the DBN contains only *shared* state features, these are the only variables whose values agents need to coordinate over. In contrast to exchanging full policies, agents can keep private those decisions that do not impact their peers' decisions.

4. ACKNOWLEDGEMENTS

This material is based upon work supported, in part, by AFOSR under Contract No. FA9550-07-1-0262.

5. REFERENCES

[1] R. Becker, S. Zilberstein, and V. Lesser. Decentralized Markov decision processes with event-driven interactions. *AAMAS*, pages 302–309, 2004.

[2] R. Becker, S. Zilberstein, V. Lesser, and C. Goldman. Solving transition independent decentralized Markov Decision Processes. *JAIR*, 22:423–455, 2004.

[3] D. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[4] R. Nair, M. Tambe, M. Yokoo, D. V. Pynadath, and S. Marsella. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI*, pages 705–711, 2003.

[5] P. Varakantham, J. Kwak, M. Taylor, J. Marecki, P. Scerri, and M. Tambe. Exploiting coordination locales in distributed POMDPs via social model shaping. In *ICAPS*, pages 313–320, 2009.

[6] P. Varakantham, J. Marecki, Y. Yabu, M. Tambe, and M. Yokoo. Letting loose a spider on a network of POMDPs: generating quality guaranteed policies. In *AAMAS*, pages 817–824, 2007.

[7] S. Witwicki and E. Durfee. Commitment-driven distributed joint policy search. *AAMAS*, 480–487, 2007.