Teamwork in Distributed POMDPs: Execution-time Coordination Under Model Uncertainty

(Extended Abstract)

Jun-young Kwak, Rong Yang, Zhengyu Yin, Matthew E. Taylor*, Milind Tambe
University of Southern California, Los Angeles, CA, 90089
*Lafayette College, Easton, PA 18042
{junyounk,yangrong,zhengyuy,tambe}@usc.edu, *taylorm@lafayette.edu

Categories and Subject Descriptors

I.2.11 [**ARTIFICIAL INTELLIGENCE**]: Distributed Artificial Intelligence

General Terms

Algorithms

Keywords

Distributed POMDPs, Model Uncertainty, Teamwork

1. INTRODUCTION

Despite their NEXP-complete policy generation complexity [1], *Distributed Partially Observable Markov Decision Problems* (DEC-POMDPs) have become a popular paradigm for multiagent teamwork [2, 6, 8]. DEC-POMDPs are able to quantitatively express observational and action uncertainty, and yet optimally plan communications and domain actions.

This paper focuses on teamwork under *model uncertainty* (i.e., potentially inaccurate transition and observation functions) in DEC-POMDPs. In many domains, we only have an approximate model of agent observation or transition functions. To address this challenge we rely on execution-centric frameworks [7, 11, 12], which simplify planning in DEC-POMDPs (e.g., by assuming cost-free communication at plan-time), and shift coordination reasoning to execution time. Specifically, during planning, these frameworks have a standard single-agent POMDP planner [4] to plan a policy for the team of agents by assuming zero-cost communication. Then, at execution-time, agents model other agents' beliefs and actions, reason about when to communicate with teammates, reason about what action to take if not communicating, etc. Unfortunately, past work in execution-centric approaches [7, 11, 12] also assumes a correct world model, and the presence of model uncertainty exposes key weaknesses that result in erroneous plans and additional inefficiency due to reasoning over incorrect world models at every decision epoch.

This paper provides two sets of contributions. The first is a new execution-centric framework for DEC-POMDPs called MODERN (MOdel uncertainty in Dec-pomdp Execution-time ReasoNing). MODERN is the first execution-centric framework for DEC-POMDPs explicitly motivated by model uncertainty. It is based on three key ideas: (i) it maintains an exponentially smaller model of other agents' beliefs and actions than in previous work and then further reduces the computation-time and space expense of this model via bounded pruning; (ii) it reduces execution-time computation by exploiting BDI theories of teamwork, thus limiting communication to key trigger points; and (iii) it simplifies its decision-theoretic reasoning about communication over the pruned model and uses a systematic markup, encouraging extra communication and reducing uncertainty among team members at trigger points.

This paper's second set of contributions are in opening up model uncertainty as a new research direction for DEC-POMDPs and emphasizing the similarity of this problem to the *Belief-Desire-Intention* (BDI) model for teamwork [5, 9]. In particular, BDI teamwork models also assume inaccurate mapping between real-world problems and domain models. As a result, they emphasize robustness via execution-time reasoning about coordination [9]. Given some of the successes of prior BDI research in teamwork, we leverage insights from BDI in designing MODERN.

2. RELATED WORK

Related work includes DEC-POMDP planning that specifically focuses on optimal communication [2, 6]. In addition to its lack of investigation into model uncertainty, the policy generation problem remains NEXP-complete, given general communication costs. Although existing execution-centric approaches [7, 10, 11, 12] lead to a provably exponential improvement in worst-case complexity over optimal DEC-POMDP planners, they have also assumed model correctness. Xuan and Lesser [12] studied the trade-offs between centralized and decentralized policies in terms of communication requirements, which differs from our own given its focus on distributed MDPs rather than DEC-POMDPs, and its assumption of model correctness. ACE-PJB-COMM (APC) [7] and MAOP-COMM (MAOP) [11] rely on a single-agent POMDP planner at plan-time, and agents execute the plan in a decentralized fashion, communicating to avoid miscoordination at execution time. APC and MAOP respectively use *GrowTree* and *JointHistoryPool*, the set of possible belief nodes to reason about the entire team's belief space, which are different from our work. Williamson et al. [10] also handle online policy computation that incorporates communication and reward shaping. Although their reward shaping is similar to the markup function, MODERN differs from this research since we use the markup function motivated by model uncertainty to encourage communication in order to reduce uncertainty.

While BDI is unable to quantitatively reason about costs and uncertainties, prior BDI works [5, 9] are related to our work in a sense of execution-centric framework and emphasizing communication at execution time, which will be explained more in Section 4.

Cite as: Teamwork in Distributed POMDPs: Execution-time Coordination Under Model Uncertainty (Extended Abstract), J. Kwak, R. Yang, Z. Yin, M. E. Taylor and M. Tambe, *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. XXX-XXX.

3. PROBLEM STATEMENT

DEC-POMDPs have been used to tackle real-world multi-agent collaborative planning problems under transition and observation uncertainty, which are described by a tuple $\langle I, S, \{A_i\}, \{\Omega_i\}, T, R, O, \mathbf{b}^0 \rangle$, where $I = \{1, ..., n\}$ is a finite set of agents, and $S = \{s_1, ..., s_k\}$ is a finite set of joint states. A_i is the finite set of actions of agent i, $A = \prod_{i \in I} A_i$ is the set of joint actions, where $\mathbf{a} = \langle a_1, ..., a_n \rangle$ is a particular joint action (one individual action per agent). Ω_i is the set of observations of agent i, $\Omega = \prod_{i \in I} \Omega_i$ is the set of joint observations, where $\mathbf{o} = \langle o_1, ..., o_n \rangle$ is a joint observation. $T : S \times A \times S \mapsto \mathbb{R}$ is the transition function, where $T(s'|s, \mathbf{a})$ is the transition probability from s to s' if joint action $\mathbf{a}$ is executed. $O : S \times A \times \Omega \mapsto \mathbb{R}$ is the observation function, where $O(\mathbf{o}|s', \mathbf{a})$ is the probability of receiving the joint observation $\mathbf{o}$ if the end state is s' after $\mathbf{a}$ is taken. $R(s, \mathbf{a}, s')$ is the reward that agents get by taking $\mathbf{a}$ from s and reaching s', and $\mathbf{b}^0$ is the initial joint belief state.

Here, we assume the presence of model uncertainty, which is modeled with a Dirichlet distribution [3]. A separate Dirichlet distribution for the observation and transition function is used for each joint state, action, and observation. An L-dimensional Dirichlet distribution is a multinomial distribution parameterized by positive hyper-parameters $\boldsymbol{\beta} = \langle \beta_1, \ldots, \beta_L \rangle$ that represents the degree of model uncertainty. The probability density function is

$$f(x_1, ..., x_L; \boldsymbol{\beta}) = \frac{\prod_{i=1}^{L} x_i^{\beta_i - 1}}{B(\boldsymbol{\beta})}, B(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{L} \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{L} \beta_i)},$$

and $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ is the standard gamma function. The maximum likelihood point can be easily computed: $x_i^* = \frac{\beta_i}{\sum_{j=1}^{L} \beta_j}$, for $i = 1, ..., L$. Let $\mathbf{T}_{s,\mathbf{a}}$ be the vector of transition probabilities from s to other states when $\mathbf{a}$ is taken and $\mathbf{O}_{s',\mathbf{a}}$ be the vector of observation probabilities when $\mathbf{a}$ is taken and s' is reached. Then $\mathbf{T}_{s,\mathbf{a}} \sim Dir(\boldsymbol{\beta})$ and $\mathbf{O}_{s',\mathbf{a}} \sim Dir(\boldsymbol{\beta}')$, where $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ are two different hyper-parameters.

We assume that the planner is not provided the precise amount of model uncertainty (i.e., the precise amount of uncertainty over transition or observation uncertainty). Our goal is effective teamwork, i.e., achieving high reward in practice, at execution time.

4. SUMMARY OF DESIGN DECISIONS

MODERN's design is explicitly driven by model uncertainty, leading to three major key ideas. First, MODERN maintains an exponentially smaller model of other agents' beliefs and actions than the entire set of joint beliefs as done in previous work via *Individual estimate of joint Beliefs (IB)*; then it further reduces the computation-time and space expense of this model via *Bounded Pruning*. IB is a concept used in MODERN to decide whether or not communication would be beneficial and to choose a joint action when not communicating. IB can be conceptualized as a subset of team beliefs that depends on an agent's local history, leading to an exponential reduction in belief space compared to *GrowTree* mentioned earlier. However, the number of possible beliefs in IB still grows rapidly, particularly when agents choose not to communicate for long time periods. Hence, we propose a new pruning algorithm that provides further savings. In particular, it keeps a fixed number of most likely beliefs per time step in IB.

Second, MODERN reduces execution-time computation by: (i) engaging in decision-theoretic reasoning about communication only at *Trigger Points* — instead of every agent reasoning about communication at every step, only agents encountering trigger points perform such reasoning; and (ii) utilizing a pre-planned policy for actions that do not involve interactions, avoiding on-line planning at every step. Note that trigger points include any situation involving ambiguity in mapping an agent's observation to its action in the joint policy. The key idea is that in sparse interaction domains, agents will not have to reason about coordination at every time step and only infrequently encounter trigger points, thus significantly reducing the burden of execution-time reasoning.

Lastly, MODERN's reasoning relies on two novelties — how it computes the expected utility gain and how it uses the *Markup Function*. In particular, MODERN's reasoning about communication is governed by the following formula: $f(\kappa, t) \cdot (U_{\mathrm{C}}(i) - U_{\mathrm{NC}}(i)) > \sigma$, where κ is a markup rate, t is a time step, $U_{\mathrm{C}}(i)$ is the expected utility of agent i if agents were to communicate, $U_{\mathrm{NC}}(i)$ is the expected utility of agent i when it does not communicate, and σ is a given communication cost. $U_{\mathrm{C}}(i)$ is calculated by considering two-way synchronization, which emphasizes the benefits from communication. $U_{\mathrm{NC}}(i)$ is computed based on the individual evaluation of heuristically estimated actions of other agents. The markup function, $f(\kappa, t)$, helps agents to reduce uncertainty among team members by marking up the expected utility gain from communication rather than perform precise local computation over erroneous models.

5. ACKNOWLEDGMENTS

We thank Perceptronics Solutions, Inc. for their support of this research, and Maayan Roth for providing us with the source code for ACE-PJB-COMM.

6. REFERENCES

[1] D. S. Bernstein, S. Zilberstein, and N. Immerman. The complexity of decentralized control of markov decision processes. In *UAI*, 2000.

[2] C. V. Goldman and S. Zilberstein. Optimizing information exchange in cooperative multi-agent systems. In *AAMAS*, 2003.

[3] R. Jaulmes, J. Pineau, and D. Precup. A formal framework for robot learning and control under model uncertainty. In *ICRA*, 2007.

[4] L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[5] H. J. Levesque, P. R. Cohen, and J. H. T. Nunes. On acting together. In *AAAI*, 1990.

[6] D. V. Pynadath and M. Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *JAIR*, 16:389–423, 2002.

[7] M. Roth, R. Simmons, and M. Veloso. Reasoning about joint beliefs for execution-time communication decisions. In *AAMAS*, 2005.

[8] S. Seuken and S. Zilberstein. Formal models and algorithms for decentralized decision making under uncertainty. *JAAMAS*, 17:190–250, 2008.

[9] M. Tambe. Towards flexible teamwork. *JAIR*, 7:83–124, 1997.

[10] S. A. Williamson, E. H. Gerding, and N. R. Jennings. Reward shaping for valuing communications during multi-agent coordination. In *AAMAS*, 2009.

[11] F. Wu, S. Zilberstein, and X. Chen. Multi-agent online planning with communication. In *ICAPS*, 2009.

[12] P. Xuan and V. Lesser. Multi-agent policies: from centralized ones to decentralized ones. In *AAMAS*, 2002.