Planning and Evaluating Multiagent Influences Under Reward Uncertainty

(Extended Abstract)

Stefan Witwicki
GAIPS / INESC-ID
Instituto Superior Técnico
UTL (Porto Salvo, Portugal)
stefan.witwicki@ist.utl.pt

Inn-Tung Chen, Edmund Durfee, and Satinder Singh
Computer Science and Engineering
University of Michigan (Ann Arbor, MI, USA)
{inntung, durfee, baveja}@umich.edu

ABSTRACT

Forming commitments about abstract influences that agents can exert on one another has shown promise in improving the tractability of multiagent coordination under uncertainty. We now extend this approach to domains with meta-level reward-model uncertainty. Intuitively, an agent may actually improve collective performance by forming a weaker commitment that allows more latitude to adapt its policy as it refines its reward model. To account for reward uncertainty as such, we introduce and contrast three new techniques.

Categories and Subject Descriptors

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

General Terms

Algorithms, Theory, Performance

Keywords

Multiagent Planning, Transition-Decoupled POMDP, Model Uncertainty, Bayesian Rewards, Influence Abstraction, Commitments

1. INTRODUCTION

Implicit in the problem of optimal multiagent coordination is the need to balance the local value of one's actions with the nonlocal value gained (or lost) from helping (or hindering) others. This problem is complicated by the presence of transition and observation uncertainty, where agents cannot be certain of the effects of their actions on their peers nor be fully aware of the situations their peers are encountering. Influence abstraction has proven useful in reducing the computational burden of optimal coordination by restricting consideration to an abstracted space of possible probabilistic nonlocal effects [1, 5]. In a running example shown in Figure 1 (top), two military field units G_1 and G_2 (where G_1 can use one of two switches to open a gate for G_2) can successfully coordinate by G_1 committing to a desirable *influence* in the form of a time and probability of opening the gate. By abstracting away local policy details that are superfluous to other agents, influences can enable agents to effectively cope with transition and observation uncertainty.

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

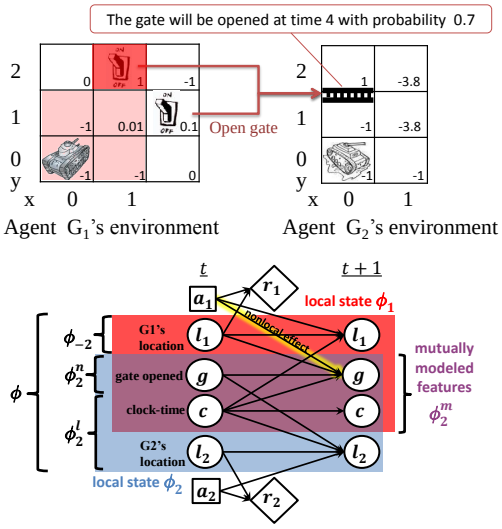

Figure 1: Example Problem and corresponding TD-POMDP

In this paper, we consider a third complicating factor: *dynamic and uncertain rewards*. In the example problem, the rewards G_1 receives in different locations depend upon the presence of an encroaching enemy; as time progresses, the enemy might render some locations more harmful, as manifested by nondeterministically decreasing rewards (with an intensity reflected in the shading in Figure 1). If the agent were alone, it could leverage the reward dynamics to reactively select the best actions depending on how its rewards progress. (For instance, it could navigate away from a switch as it starts to become more harmful.) Committing to a particular influence (e.g., raising the gate at a given time), on the other hand, may constrain the agent's policy in such a way as to preclude taking these actions and saving itself from harm. When planning its influences under reward uncertainty, the agent should account for the latitude that each influence allows in improving its local value. This insight motivates our investigation into the efficacy of influence-based planning under reward uncertainty, which we summarize below.

2. INFLUENCE-BASED PLANNING

There are several decision-theoretic formulations for problems like that portrayed in Figure 1, where agents act largely independently but can sometimes achieve preconditions that affect others [1, 4, 5]. Each of these formulations decomposes the conventional joint

decision model [2] into a set of $\mathcal{N}$ *local models*, one per agent i that includes a *local state* feature vector ϕ_i; similarly, they decompose the *joint reward* function into a summation of local reward functions: $R(\phi(t), a(t)) = \sum_{i=1}^{\mathcal{N}} R_i(\phi_i(t), a_i(t))$. The TD-POMDP of Witwicki and Durfee (W&D) [5], an instance of which is depicted in Figure 1 (bottom), further divides a local state ϕ_i into *nonlocally-affected* features ϕ_i^n (that only other agents' actions immediately affect) and *locally-affected* features ϕ_i^l, and explicitly distinguishes those *mutually-modeled* features ϕ_i^m through which i's interactions occur. In our example, agent G_2 models a single nonlocally-affected feature g (gate-opened) that depends on G_1's action.

As W&D have derived, an agent j can plan optimally using a local belief state, $\boldsymbol{b}_i(t) = \langle \phi_i(t), \phi_i^m(1..t-1) \rangle$), and thereby account for the influence of other agents by modeling a probability distribution over changes to its nonlocal features: $\Gamma_{\rightarrow j} = Pr(\phi_j^n(1..T))$. This distribution, which refer to as agent j's *incoming influence*, in our example encodes the probability that agent G_1 will open the gate for agent G_2 (at each time): $\Gamma_{\rightarrow G_2} = Pr(g(1..T))$. Specifying $\Gamma_{\rightarrow j}$ fully decouples agent j from all other agents, allowing j to compute and evaluate its local policy without having to consider the other agents' policies. Moreover, the optimal joint policy can be computed by searching a finite space of *joint influence* points, which W&D have shown can be significantly smaller than the joint policy space. In evaluating a given point in the *influence space*, agent j should also reason about its *outgoing influence* $\Gamma_{j\rightarrow}$, selecting an *influence-constrained* policy $\pi_j^{*|\Gamma_{j\rightarrow}}$ that achieves $\Gamma_{j\rightarrow}$.

3. THREE ALGORITHMS FOR HANDLING REWARD UNCERTAINTY

Given a fixed, known model of the agents' environment, outgoing influence achievement, incoming influence evaluation, and influence-based planning are all well defined [5]. We now extend them to dynamic or unknown environments wherein agents may be uncertain as to which model is the correct model. In particular, let there be K possible local reward functions $\{R_i^k\}_{k=1}^K$ per agent (independently distributed). Prior to execution, each agent i has only a prior distribution over its reward function, but during execution, i's observations can inform a posterior distribution over the true reward function. In the subsections that follow, we introduce and contrast three different influence-based planning extensions that afford different levels of computational efficiency and approximation.

Extended Belief State (EBS).

First, consider that W&D's approach can be directly applied to a TD-POMDP wherein each agent's belief state has been extended to include a distribution over the true reward function. By branching for every realizable posterior reward distribution after every action, the agent can account for the uncertainty precisely as it plans and evaluates influence points. However, the computation of each such evaluation will depend heavily on the size of the reward distribution, over which the extended-belief-state space grows exponentially.

Mean Reward (MR).

A simple approximation to the EBS algorithm is to completely collapse the uncertainty over each agents' rewards into a single expected or mean reward function, i.e., use the reward distribution to induce a mean-reward TD-POMDP where agent i's local reward function is $\bar{R}_i(\phi_i(t), a_i(t)) = \sum_{k=1}^K Pr(R_i^k) R_i^k(\phi_i(t), a_i(t))$. W&D's influence-based planning method then implicitly accounts for reward uncertainty at no additional computational cost. Although generally an approximation, we have proven that the mean-reward (MR) algorithm is optimal in special cases where agents cannot gain information (informing a new posterior distributions) about their true reward functions as they act and observe.

Influence-Constrained Iterative MR (ICIMR).

Finally, we develop a hybrid approach that builds off of the iterative mean-reward algorithm (IMR) for single-agent Bayesian-MDPs [3]. IMR reapplies the mean-reward technique after each belief update, because changes to the posterior distribution over reward functions can yield a different mean reward function $\bar{R}_i^{t+1} = \mathbb{E}_{R_i^k \sim \boldsymbol{b}_i(t+1)}[R_i^k]$, and hence adopting the policy $\tilde{\pi}_i^{t+1}$ optimal with respect to the updated mean reward may outperform the current policy $\tilde{\pi}_i^t$. Effectively, this involves (perhaps pre-)computation and adoption of a new policy at each time step.

Our ICIMR algorithm's novel departure from IMR comes from our multiagent setting and the role of commitments to influences. An agent who has already committed to probabilistically influencing others cannot iteratively shift from policy to policy without taking its committed outgoing influences into account. A stringent constraint that we could place on this agent is that its policy at the current iteration must *from its current state* satisfy all its outgoing commitments. Unfortunately, this is untenable, because stochastic state transitions could have put the agent into a state from which *no* policy can achieve the requisite commitments. Instead, we require that the agent's adopted policy must have satisfied its commitments, in expectation, *from its initial state*. Formally, agent i should adopt policy $\tilde{\pi}_i^{t+1} = \pi_i^{*|\boldsymbol{b}_i(t+1), \Gamma_{i\rightarrow}}$ that achieves outgoing influences $\Gamma_{i\rightarrow}$ *and* is consistent with its previous action choices $\tilde{\pi}_i^t(0..t)$.

With ICIMR, agent i plans and evaluates outgoing influences by iteratively considering each possible next mean-reward MDP that it could encounter. This resembles the lookahead performed with the EBS algorithm, except that whereas EBS considers every possible action at each successive state, ICIMR only considers the action dictated by the mean-reward policy in the state (given the posterior reward distribution). ICIMR branches for transition and reward uncertainty, but not future action, thereby allowing more efficient planning. And, although ICIMR is an approximation of EBS, we have proven that ICIMR yields solutions whose quality is greater than or equal to those of MR. A preliminary empirical analysis indicates that ICIMR can strike a good compromise between solution quality and computational overhead, making it a useful technique for tackling reward uncertainty in an efficient, yet principled, manner.

4. ACKNOWLEDGEMENTS

This work was supported, in part, by the Fundação para a Ciência e a Tecnologia (FCT) and the CMU-Portugal Program under project CMU-PT/SIA/0023/2009, and by the National Science Foundation under grants IIS-0964512, IIS-1148668, and IIS-0905146.

5. REFERENCES

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

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

[3] P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of ICML '06*, pages 697–704, 2006.

[4] 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,* 313–320, 2009.

[5] S. Witwicki and E. Durfee. Influence-based policy abstraction for weakly-coupled Dec-POMDPs. In *ICAPS,* 185–192, 2010.