Intention-Aware Planning under Uncertainty for Interacting with Self-Interested, Boundedly Rational Agents

(Extended Abstract)

Trong Nghia Hoang and Kian Hsiang Low
Department of Computer Science, National University of Singapore, Republic of Singapore
{nghiaht, lowkh}@comp.nus.edu.sg

ABSTRACT

A key challenge in non-cooperative multi-agent systems is that of developing efficient planning algorithms for intelligent agents to perform effectively among boundedly rational[1], self-interested (i.e., non-cooperative) agents (e.g., humans). To address this challenge, we investigate how intention prediction can be efficiently exploited and made practical in planning, thereby leading to efficient intention-aware planning frameworks capable of predicting the intentions of other agents and acting optimally with respect to their predicted intentions.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Experimentation, Theory

Keywords

Planning (single and multi-agent), Modeling other agents and self

1. INTRODUCTION

To date, existing planning frameworks for non-cooperative multi-agent systems (MAS) can be generally classified into: (a) *game-theoretic* frameworks rely on the well-established solution concepts of classical game theory to characterize interactions among self-interested agents; (b) *decision-theoretic* frameworks extend single-agent decision-theoretic planning framework (e.g., MDP, POMDP) by considering other agents as a stochastic part of the environment. However, such frameworks suffer from the following drawbacks: (a) the restrictive assumptions on other agents' behaviors, as implied by the solution concepts [3, 4]; (b) the failure in accounting for agents' deliberative and boundedly rational behaviors that cannot be sufficiently modeled as stochastic noise in the transition model. Alternatively, the *interactive POMDP* (I-POMDP) framework [2] attempts to explicitly account for the bounded rationality of self-interested agents by maintaining an agent's interactive beliefs over both the physical states and the other agents' beliefs. As a result, solving I-POMDP requires solving an exponential number of POMDPs [2], which are prohibitively expensive. To resolve the above issues, we propose practical and efficient formal, principled intention-aware planning frameworks for interacting with boundedly rational, self-interested agents:

- *Nested MDP* framework for interacting in fully observable environments (Section 2): inspired by [1], it constitutes a recursive reasoning formalism to predict the other agents' intention efficiently and such predictive information is then exploited to plan our agent's optimal interaction policy. The cost of solving nested MDP is linear in the length of time horizon and the depth of reasoning.
- *Intention-aware POMDP* (IA-POMDP) framework for interacting in partially observable environments (Section 3): it extends nested MDP by integrating it into POMDP for tracking our agent's belief. By exploiting problem structure in terms of the other agents' full observability, IA-POMDP can be solved efficiently in polynomial time.

2. NESTED MDP

Nested MDP constitutes a recursive reasoning process comprising k reasoning levels: at level 0, our agent's best response is computed by considering the other agent's actions as stochastic noise in an MDP's transition model. At level $k \geq 1$, our agent plans its optimal strategy by assuming that the other agent's strategy is based only on lower reasoning levels $0, \ldots, k-1$. Formally, nested MDP at level $k \geq 1$ for agent 1 is a tuple $M_1^k \triangleq (S, U, V, T, R, \{\pi_2^i\}_{i=0}^{k-1}, \phi)$ where

- S is a set of all possible states of the environment;
- U and V are sets of all possible actions available to agents 1 and 2, respectively;
- $T : S \times U \times V \times S \rightarrow [0, 1]$ denotes the transition probability of going from state $s \in S$ to state $s' \in S$ using agent 1's action $u \in U$ and agent 2's action $v \in V$;
- $R : S \times U \times V \rightarrow \mathbb{R}$ is the reward function of agent 1;
- $\pi_2^i : S \times V \rightarrow [0, 1]$ is the reasoning model at level $i < k$ predicting the mixed strategy of agent 2 for each state;
- $\phi \in (0, 1)$ is a discount factor.

The optimal value function of nested MDP M_1^k at level $k \geq 1$ for agent 1 satisfies the following Bellman equation:

$$U_1^k(s) \triangleq \max_{u \in U} \sum_{v \in V} \widehat{\pi}_2^{k-1}(s, v)\ Q_1^k(s, u, v)$$
$$Q_1^k(s, u, v) \triangleq R(s, u, v) + \phi \sum_{s' \in S} T(s, u, v, s')\ U_1^k(s') \quad (1)$$

[1]Boundedly rational agents are subject to limited information, cognition, and time while making decisions.

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.

where the mixed strategy $\widehat{\pi}_2^{k-1}$ of the other agent 2 is predicted by averaging uniformly over all its reasoning models $\{\pi_2^i\}_{i=0}^{k-1}$ at levels $0, 1, \ldots, k-1$ because its actual level of reasoning is not known to our agent 1:

$$\widehat{\pi}_2^{k-1}(s,v) \triangleq \beta \sum_{i=0}^{k-1} \pi_2^i(s,v) \ . \tag{2}$$

Agent 2's reasoning model π_2^0 at level 0 is induced by solving a conventional MDP that represents agent 1's actions as stochastic noise in its transition model. To obtain agent 2's reasoning models $\{\pi_2^i\}_{i=1}^{k-1}$ at levels $i = 1, \ldots, k-1$, let $Opt_2^i(s)$ be the set of agent 2's optimal actions for state s induced by solving its nested MDP M_2^i, which recursively involves building agent 1's reasoning models $\{\pi_1^l\}_{l=0}^{i-1}$ at levels $l = 0, 1, \ldots, i-1$, by definition. Then,

$$\pi_2^i(s,v) \triangleq \begin{cases} |Opt_2^i(s)|^{-1} & \text{if } v \in Opt_2^i(s), \\ 0 & \text{otherwise.} \end{cases}$$

After predicting agent 2's strategy $\widehat{\pi}_2^{k-1}$ (2), agent 1's optimal policy (i.e., reasoning model) π_1^k at level k can be induced by solving its corresponding nested MDP M_1^k (1).

3. INTENTION-AWARE POMDP

To tackle partial observability, it seems obvious to first consider generalizing the recursive reasoning formalism of nested MDP. This approach yields two practical complications: (a) our agent's belief over both the physical states and the other agent's belief has to be modeled, and (b) the other agent's mixed strategy has to be predicted for each of its infinitely many possible beliefs. The I-POMDP framework faces both difficulties and consequently incurs a prohibitively expensive processing cost that involves solving exponential number of POMDPs [2]. In practice where we are subject to limited information, cognition, and time, we hardly recall performing such sophisticated modeling of our human counterpart during interaction. Instead, we often make satisficing decisions by limiting our predictions of counterpart's strategy to some specific states and considering how likely each state is based on our belief over these states.

To realize this intuition, we propose an alternative *intention-aware POMDP* (IA-POMDP) framework by exploiting the following structural assumption: the environment is fully observable to the other agent. Such an assumption is practical to make when the other agent's sensing capability is superior (e.g., human) or we do not know nor want to underestimate the other agent's sensing capability, especially in competitive scenarios. Surprisingly, this simple assumption alleviates both difficulties faced by I-POMDP, thus making IA-POMDP computationally efficient. Compared to existing game-theoretic frameworks [3, 4], our assumption is far less restrictive. More importantly, though it makes IA-POMDP less expressive than I-POMDP, it significantly boosts the practicality of decision-theoretic planning frameworks for non-cooperative MAS. Formally, IA-POMDP for agent 1 is defined as a tuple $(S, U, V, O, T, Z, R, \widehat{\pi}_2^k, \phi, b_0)$ where

- S is a set of all possible states of the environment;
- U and V are sets of all possible actions available to our agent 1 and the other agent 2, respectively;
- O is a set of all possible observations of our agent 1;
- $T : S \times U \times V \times S \to [0,1]$ is a transition function that depends on the joint actions of both agents;
- $Z : S \times U \times O \to [0,1]$ denotes the probability $Pr(o|s', u)$ of making observation $o \in O$ in state $s' \in S$ using our agent 1's action $u \in U$;
- $R : S \times U \times V \to \mathbb{R}$ is the reward function of agent 1;
- $\widehat{\pi}_2^k : S \times V \to [0,1]$ denotes the predictive probability $Pr(v|s)$ (i.e., predicted mixed strategy) of agent 2 selecting action v in state s and is derived using (2) by solving its nested MDPs at levels $0, \ldots, k$;
- $\phi \in (0,1)$ is a discount factor; and
- $b_0 \in \Delta(S)$ is a prior belief over the states of environment.

Solving IA-POMDP involves choosing the policy that maximizes the expected total reward with respect to the prediction of agent 2's mixed strategy using nested MDP. The optimal value function of IA-POMDP for our agent 1 satisfies the following Bellman equation:

$$V_{n+1}(b) = \max_u Q_{n+1}(b,u)$$
$$Q_{n+1}(b,u) = R(b,u) + \phi \sum_{v,o} Pr(v,o|b,u) \ V_n(b')$$

where our agent 1's expected immediate payoff is

$$R(b,u) = \sum_{s,v} R(s,u,v) \ Pr(v|s) \ b(s)$$

and the belief update is

$$b'(s') = \beta \ Z(s',u,o) \sum_s T(s,u,v,s') \ Pr(v|s) \ b(s) \ .$$

Like POMDP, the optimal value function $V_n(b)$ of IA-POMDP can be approximated arbitrarily closely (for infinite horizon) by a piecewise-linear and convex function that takes the form of a set V_n[2] of α vectors: $V_n(b) = \max_{\alpha \in V_n}(\alpha \cdot b)$. Thus, solving IA-POMDP is equivalent to computing the corresponding set of α vectors, which grows exponentially with the time horizon: $|V_{n+1}| = |U||V_n|^{|V||O|}$. To avoid this exponential blow-up, IA-POMDP inherits essential properties from POMDP that make it amenable to be solved by existing sampling-based algorithms, such as [5], of POMDP in polynomial time. For interested readers, a further technical discussion of IA-POMDP as well as an empirical evaluation of our proposed frameworks can be found in the extended version of this paper[3].

4. REFERENCES

[1] C. F. Camerer, T. H. Ho, and J. K. Chong. A cognitive hierarchy model of games. *Quarterly J. Economics*, 119(3):861–898, 2004.

[2] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *JAIR*, 24:49–79, 2005.

[3] J. Hu and M. P. Wellman. Multi-agent reinforcement learning: Theoretical framework and an algorithm. In *Proc. ICML*, pages 242–250, 1998.

[4] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. ICML*, pages 157–163, 1994.

[5] J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proc. IJCAI*, pages 1025–1032, 2003.

[2]With slight abuse of notation, the value function is also used to denote the set of corresponding α vectors.

[3]http://www.comp.nus.edu.sg/∼lowkh/pubs/aamas2012e.pdf