A Framework for Optimal Sequential Planning in Multiagent Settings

Prashant J. Doshi*
Dept. of Computer Science
Univ of Illinois, Chicago, IL 60607
pdoshi@cs.uic.edu

Introduction

Research in autonomous agent planning is gradually moving from single-agent environments to those populated by multiple agents. In single-agent sequential environments, partially observable Markov decision processes (POMDPs) provide a principled approach for planning under uncertainty. They improve on classical planning by not only modeling the inherent non-determinism of the problem domain, but also by producing "universal" plans or policies which represent complete control mechanisms. We are motivated by these reasons to generalize POMDPs from their traditional single-agent application setting to an environment populated by several interacting autonomous agents.

The formalism of Markov decision processes has been extended to multiple agents previously, giving rise to stochastic games or Markov games. Other extensions of POMDPs to multiple agent environments have also appeared and are called DEC-POMDPs (Bernstein *et al.* 2002) in the literature. Both these formalisms employ the solution concept of Nash equilibria. Specifically, solutions are plans (policies) that are in mutual equilibrium with each other. However, while Nash equilibria are useful for describing a multiagent system when, and if, it has reached a stable state, this solution concept is not sufficient as a general control paradigm. The main reasons are that there may be multiple equilibria with no clear way to choose among them (non-uniqueness), and the fact that equilibria do not specify actions in cases in which agents believe that other agents may not act according to their equilibrium strategies (incompleteness). Furthermore, at present, researchers have inadequate understanding of the intermediate stages before Nash equilibrium is reached.

In this thesis, we present a new framework called Interactive POMDPS (I-POMDPs) for optimal planning by an agent interacting with other autonomous agents in a sequential environment and maximizing its reward that depends on joint actions of all agents. As expected, the generalization of POMDPs from a single-agent setting to multiple agents is not trivial. In addition to maintaining beliefs about the physical environment, each agent must also maintain beliefs about the other agents: their sensing capabilities, beliefs, preferences, and intentions. Analogously to POMDPs, each agent will locally compute its actions that optimize its preferences given what it believes in. The resulting control paradigm complements and generalizes the traditional equilibrium approach in that, if the agent believes that other agents will act according to an equilibrium, then it will also act out its part of the equilibrium. However, if it believes that other agents will diverge from equilibrium, then it will choose the appropriate optimal response. The unique aspect of I-POMDPs is that, by prescribing actions based on the agent's belief about other agents' beliefs and parameters, an agent maintains a possibly infinitely nested interactive belief system.

*Joint work with my advisor Piotr Gmytrasiewicz

Interactive POMDPs

The proposed framework attempts to bring together game-theoretic solution concepts of equilibrium and decision-theoretic control paradigms such as policies as put forward by frameworks like POMDPs. The conceptual pieces of our framework are similar to those of POMDPs, thereby facilitating its easy adoption by the research community for multiagent settings. In the next few paragraphs, we briefly discuss our framework, its properties, and preliminary results.

For simplicity of presentation let us consider an agent, i, that is interacting with one other agent, j. An *interactive POMDP* of agent i, *I-POMDP*$_i$, is:

$$\textit{I-POMDP}_i = \langle IS_i, A, T_i, \Omega_i, O_i, R_i \rangle$$

where:

• IS_i is a set of **interactive** states defined as $IS_i = S \times \Theta_j$, where S is the set of states of the physical environment, and Θ_j is the set of possible intentional models of agent j. An intentional model of j is, $\theta_j = \langle b_j, A, \Omega_j, T_j, O_j, R_j, OC_j \rangle$ where OC_j is agent j's optimality criterion, together with the assumption that agent j is Bayesian rational. For the sake of simplicity, let us rewrite θ_j as, $\theta_j = \langle b_j, \widehat{\theta}_j \rangle$ where $\widehat{\theta}_j$ is the agent j's *frame*. Agent j's belief is a probability distribution over the states of the world and the models of the agent i, $b_j \in \Delta(S \times \Theta_i)$. Each interactive state of agent i therefore contains possibly infinitely nested beliefs over others' types and their beliefs about others. This nesting may be terminated by assuming that at some arbitrary level, an agent (say i) models the other agent (say j) using

a *no-information* model – i considers all of j's actions to be equally likely.

• $A = A_i \times A_j$ is the set of joint moves of all agents.

• T_i is a transition function $T_i : IS_i \times A \times IS_i \rightarrow [0, 1]$ which describes results of agents' actions. Actions can change the physical state, as well as the frames of other agents, for example by changing their observation function.

• Ω_i is the set of agent i's observations.

• O_i is an observation function $O_i : IS_i \times A \times \Omega_i \rightarrow [0, 1]$.

• R_i is defined as $R_i : IS_i \times A \rightarrow \mathbf{R}$. We allow the agent to have preferences over physical states and models of other agents, but usually only the physical state will matter.

In a manner similar to POMDPs, we can show that agent's beliefs over their interactive states are *sufficient statistics* i.e. they fully summarize the agent's observable histories. Furthermore, we propose the following equation which captures the belief update process of an agent modeled as an I-POMDP.

$$\begin{aligned} b_i^t(is^t) = \quad & \beta \int_{IS^{t-1}} \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(is^t, a^{t-1}, o_i^t) \\ & \times \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) T_i(is^{t-1}, a^{t-1}, is^t) \\ & \times O_j(is_j^t, a^{t-1}, o_j^t) db_i^{t-1}(is^{t-1}) \end{aligned}$$

Here $is = (s, \theta_j)$, $is_j = (s, \theta_i)$, b_j^{t-1} and b_j^t are the belief elements of θ_j^{t-1} and θ_j^t, β is a normalizing constant, O_j is the observation function in θ_j^t, and $Pr(a_j^{t-1}|\theta_j^{t-1})$ is the probability of other agent's action given its model.

Though our proposed belief update has a lot in common with that of POMDPs, two important differences manifest due to the multiagent application setting. First, since the predicted state of the environment depends on the actions performed by both agents, a probability measure on the other agent's actions ($Pr(a_j^{t-1}|\theta_j^{t-1})$) must be obtained. Second, changes in the models of the other agent must be included in the update. Specifically, update of the other agent's beliefs due to its observations ($\tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t)$) is included. A formal derivation of our belief update appears in (Gmytrasiewicz & Doshi 2004). We note that this update *recurses* through the entire belief nesting, with the recursion bottoming out when a no-information model is encountered.

The local policy of each agent is computed by solving the associated I-POMDP using value iteration. Agent i's optimal action, a^*, for the case of infinite horizon criterion with discounting, is an element of the set of optimal actions for the belief state, $OPT(\theta_i)$, which is defined as:

$$\begin{aligned} OPT(\theta_i) = \underset{a_i \in A_i}{argmax} \Bigg\{ & \int_{IS} ER_i^{a_j}(is, a_i) db_i(is) + \\ & \gamma \sum_{o_i \in \Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i), \widehat{\theta}_i \rangle) \Bigg\} \end{aligned}$$

where, $ER_i^{a_j}(is, a_i) = \sum_{a_j} R_i(is, a_i, a_j) Pr(a_j|\theta_j)$

It turns out that, as in POMDPs, the value function is piecewise linear and convex (PWLC) w.r.t. the belief. Additionally, the sequence of value functions converges as the horizon approaches infinity. However, both these properties hold in the case of a finite belief nesting only. By establishing these critical properties, we can apply the wide array of POMDP solution techniques to I-POMDPs as well.

We have applied the I-POMDP framework to the multiagent tiger game (Tambe *et al.* Aug 2002). Our preliminary results have shown that an agent's policy changes as its belief about the other agent's beliefs changes. For example, if the agents are operating as a team (coordination is rewarded), and if agent i believes that agent j likely believes that the tiger is behind the left door, then i gives preference to opening the right door. However, if i believes that j likely believes that the tiger is behind the right door, then i prefers opening the left door. In addition to modeling multiple agents as a team, we have also modeled them as enemies, and as being neutral towards each other.

Future Work

We have adopted a two-stage approach towards my thesis research. The first stage involved developing the framework, identifying and formalizing its properties and solution techniques, and conducting preliminary experiments on interesting problem domains. High computational complexity of solving I-POMDPs forces us to search for efficient approximation techniques. The second stage of this thesis will concentrate on developing these approximation techniques.

A promising approximation technique seems to be stochastic sampling methods such as particle filters. Particle filters utilize Monte Carlo sampling to approximate a belief state, and propagate it forwards in time. As frequently is the case when it comes to Bayesian update methods, here over the space of possible models of agents, the choice of the "right" prior arises. To address this issue we have turned to algorithmic probability and Kolmogorov complexity (Li & Vitanyi 1997).

Conclusion

We have proposed a new framework, called Interactive POMDPs, for optimal sequential decision-making in multiagent settings. Our framework is applicable to autonomous agents operating in partially observable environments, who locally compute actions they should execute to optimize their preferences given what they believe. I-POMDPs infuse decision-theoretic planning with notions of game theory thereby blending strategic and long term planning into a single framework. We have established its definition, its properties, and gathered some preliminary empirical data. Future work will revolve around developing approximate solution techniques that will tradeoff complexity with the quality of the solution, and performing more experimentation.

References

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*.

Gmytrasiewicz, P., and Doshi, P. 2004. A framework for sequential optimality in multiagent settings. Technical Report UIC-AIL-TR05, University of Illinois at Chicago.

Li, M., and Vitanyi, P. 1997. *An Introduction to Kolmogorov Complexity and its Applications*. Springer.

Tambe, M.; Nair, R.; Pynadath, D.; and Marsella, S. Aug 2002. Towards computing optimal policies for dec-pomdps. In *AAAI Workshop on Game and Decision Theoretic Agents*.