A Decision-Theoretic Approach to Evaluating Posterior Probabilities of Mental Models

Jonathan Y. Ito and **David V. Pynadath** and **Stacy C. Marsella**
Information Sciences Institute, University of Southern California
4676 Admiralty Way, Marina del Rey CA 90292 USA

Abstract

Agents face the problem of maintaining and updating their beliefs over the possible mental models (whether goals, plans, activities, intentions, etc.) of other agents in many multiagent domains. Decision-theoretic agents typically model their uncertainty in these beliefs as a probability distribution over their possible mental models of others. They then update their beliefs by computing a posterior probability over mental models conditioned on their observations. We present a novel algorithm for performing this belief update over mental models that are in the form of Partially Observable Markov Decision Problems (POMDPs). POMDPs form a common model for decision-theoretic agents, but there is no existing method for translating a POMDP, which generates deterministic behavior, into a probability distribution over actions that is appropriate for abductive reasoning. In this work, we explore alternate methods to generate a more suitable probability distribution. We use a sample multiagent scenario to demonstrate the different behaviors of the approaches and to draw some conclusions about the conditions under which each is successful.

Introduction

Agents face the problem of maintaining and updating their beliefs over the possible mental models (whether goals, plans, activities, intentions, etc.) of other agents in many multiagent domains. Decision-theoretic agents typically model their uncertainty in these beliefs as a probability distribution over their possible mental models of others. They then update their beliefs by computing a posterior probability over mental models conditioned on their observations. For example, many existing methods for plan recognition represent the agent being observed in a probabilistic representation of its plan hierarchy (Goldman, Geib, & Miller 1999; Charniak & Goldman 1993; Pynadath & Wellman 2000; Bui, Venkatesh, & West 2002; Kaminka, Pynadath, & Tambe 2001). These methods provide algorithms that invert the representation to provide an abductive reasoning mechanism that compute the desired posterior probability over the possible plans.

In this work, we are interested in modeling agents who perform decision-theoretic planning, but not over a hierarchy of possible plans. In particular, we consider agent models in the form of Partially Observable Markov Decision Problems (POMDPs) (Smallwood & Sondik 1973). The recognizing agent maintains a probability distribution over possible POMDP models, representing the likelihood that the observed agent is following a given POMDP. Thus, in the language of Interactive POMDPs (Gmytrasiewicz & Doshi 2005), all of the agents' *frames* are POMDPs.

For a recognizing agent to update its belief over the POMDP model of the observed, it must determine the probability that a candidate POMDP model would generate each observed action of that agent. An agent following a given POMDP model will execute one of a set of policies that have the highest expected reward, as determined by whichever POMDP solution algorithm the agent uses. Thus, the recognizing agent could assume that the observed agent's behavior will always conform to one of these optimal policies. Then, it could assign a probability of zero to any action that does not appear in any optimal policies for the given situation.

While such an assumption would lead to the straightforward development of a POMDP recognition algorithm, we could safely do so only if we are sure that the space of possible POMDPs under consideration is exhaustive. If the observed agent happens to be following some other, unknown POMDP, then it could perform an action that does not occur in any of the known optimal policies, leading to all of the existing POMDP models to have a zero posterior probability. To avoid such a degenerate case, the recognizing agent must allow for the possibility of error in its computation of the possible POMDP policies.

In this work, we observe that we can use the POMDP model's *value function* as a measure of likelihood over all actions, rather than only the optimal one. In other words, the larger the gap in expected reward between an observed action and the optimal one under a given POMDP model, the worse candidate that model is. Similarly, the more actions that have a higher expected reward (in a given POMDP model) than the observed action, the worse candidate that model is.

There is no one universal method for modeling such likelihoods, so we examine a variety of methods and their impact on recognition accuracy. In particular, we explore three methods for calculating the probability of an observed action

conditioned on a particular POMDP model: (1) the ratio of expected rewards, (2) a linear ranking method, and (3) an exponential ranking method. We explore and evaluate these different methods in a sample domain within the PsychSim (Marsella, Pynadath, & Read 2005; Pynadath & Marsella 2005) multiagent framework.

PsychSim Simulation Framework

PsychSim is a multiagent framework for running social simulations. It operationalizes existing psychological theories as boundedly rational computations to generate more plausibly human behavior. A central aspect of the PsychSim design is that agents have fully qualified decision-theoretic models of others. Such quantitative recursive models give PsychSim a powerful mechanism to model a range of factors in a principled way. Each agent within PsychSim maintains independent beliefs about the world, has its own goals and its own policies for achieving those goals. The PsychSim framework is an extension to the Com-MTDP model (Pynadath & Tambe 2002) of agent teamwork.

Example Domain

We have taken our example domain from a scenario in childhood aggression modeled within PsychSim. There are agents for three students: a bully, his victim (i.e., the student he focuses his aggression on), and an onlooking student to whom the bully looks for affirmation. There is also a teacher who wishes to prevent any incidents of aggression. The teacher can deter the bully from picking on his victim by doling out punishment. We focus on the problem facing the bully agent, whose decision on whether or not to pick on his victim must consider the possible punishment policy of the teacher.

State

Each agent model includes features representing its "true" state. This state consists of objective facts about the world, some of which may be hidden from the agent itself. For our example bully domain, each agent has a $power$ state feature representing the strength of the agent. Thus $power_{bully}$ represents the power of the bully agent. We represent the state as a vector, $\overrightarrow{s}^t$, where each component corresponds to one of these state features and has a value in the range $[-1, 1]$.

Reward

PsychSim uses a decision-theoretic model of preferences, so the bully agent decides whether or not to pick on his victim through a finite-horizon evaluation of expected reward. There are three components of the bully's reward function: (1) a desire to increase his power, which decreases when he is punished; (2) a desire for affirmation from the onlooking student, which increases when the onlooker laughs along as the bully picks on the victim; and (3) a desire to decrease the victim's power, which decreases when the bully picks on him (as well as when the onlooker laughs at him). We define the reward function as a linear combination of these three components, so that we can reduce the specification of the bully's type as a triple of coefficients, with each coefficient in $[0, 1]$ and with the additional constraint that the three coefficients sum to 1. It is only the relative value of the coefficients that determines behavior. Thus, to simulate the behavior of a bully whose aggression is intended to gain the approval of his peers, we would use an agent with a higher weight for the second of the reward components. On the other hand, to simulate a bully of a more sadistic orientation, we would use a higher weight for the third.

The teacher also has three components to her reward function, corresponding to her desire to increase the power of each of the three students: bully, victim, and onlooker. This reward function gives the teacher a disincentive for punishing anyone (as the punishee's power would then decrease) unless doing so will deter acts of aggression that would reduce the victim's power even more. A perfectly fair teacher would give equal weight to the three students' power. A bully feeling persecuted by the teacher may be inclined to think that she favors the victim's power over his own. On the other hand, a bully may feel that the teacher shares his dislike of the victim, in which case he may model here as having a lower weight for increasing the victim's power.

We can represent the overall preferences of an agent, as well as the relative priority among them, as a vector of weights, $\overrightarrow{g}$, so that the product, $\overrightarrow{g} \cdot \overrightarrow{s}^t$, quantifies the degree of satisfaction that the agent receives from the world, as represented by the state vector, $\overrightarrow{s}^t$.

Actions

The teacher has 7 options in her action set, A_t. She can do nothing; she can scold the bully, onlooker, or the entire class; or she can punish the bully, onlooker, or the entire class. Punishing a student causes a more sever decrease in a student's power than simply scolding. The onlooking student has 2 option in his action set, A_o: laugh at the victim, or do nothing. The bully has 2 actions in his action set, A_b: pick on the victim or do nothing.

The state of the world changes in response to the actions performed by the agents. We model these dynamics using a transition probability function, $T\left(\overrightarrow{s}, \overrightarrow{a}, \overrightarrow{s}'\right)$, to capture the possibly uncertain effects of these actions on the subsequent state:

$$P\left(\overrightarrow{s}^{t+1} = \overrightarrow{s}' | \overrightarrow{s}^t = \overrightarrow{s}, \overrightarrow{a}^t = \overrightarrow{a}\right) = T\left(\overrightarrow{s}, \overrightarrow{a}, \overrightarrow{s}'\right) \quad (1)$$

For example, the bully's attack on the victim impacts the power of the bully, the power of the victim, etc. The distribution over the bully's and victim's changes in power is a function of the relative powers of the two - e.g., the larger the power gap that the bully enjoys over the victim, the more likely the victim is to suffer a big loss in power.

Policies

The policy of the teacher depends on not only her mental models of the students, but also on the prior actions of the students. In other words, the teacher may perform a different action when the bully picks on the victim than when he does not. Thus, the policy, $\pi_T : M_{TB} \times M_{TO} \times A_B \times A_O \rightarrow A_T$. Over the infinite space of possible reward functions for

the teacher in this classroom scenario, there are only eight policies that are ever optimal. However, there are $7^4 = 2401$ possible policies in the overall space, so a bully considering only optimal policies is ignoring a large set of possibilities.

We model each agent's real policy table as including a bounded lookahead policy rule that seeks to best achieve the agent's goals given its beliefs. To do so, the policy considers all of the possible actions/messages it has to choose from and measures the results by simulating the behavior of the other agents and the dynamics of the world in response to the selected action/message. They compute a quantitative value, $V_a\left(\overrightarrow{b}^t\right)$, of each possible action, a, given their beliefs, $\overrightarrow{b}^t$.

$$V_a\left(\overrightarrow{b}^t\right) = \overrightarrow{g} \cdot T\left(\overrightarrow{b}^t, a\right) + V\left(T\left(\overrightarrow{b}^t, a\right)\right) \quad (2)$$

$$V\left(\overrightarrow{b}^{t+1}\right) = \sum_{\tau=1}^{N} \overrightarrow{g} \cdot \overrightarrow{b}^{t+\tau} \quad (3)$$

Thus, an agent first uses the transition function, T, to project the immediate effect of the action, a, and then projects another N steps into the future, weighing each state along the path against its goals, $\overrightarrow{g}$. Thus, the agent is seeking to maximize the expected value of its behavior, along the lines of decision policies in decision theory. However, PsychSim's agents are only boundedly rational, given that they are constrained, both by the finite horizon, N, of their lookahead and the possible error in their belief state, $\overrightarrow{b}$.

Distribution Update Procedure

$$P\left(m_i|A_{obs}\right) = \frac{P\left(m_i\right) \times P\left(A_{obs}|m_i\right)}{\sum_i P\left(m_i\right) \times P\left(A_{obs}|m_i\right)} \quad (4)$$

Here we describe the process by which an agent can update its initial probability distribution over which mental model is being employed by the another agent. In 4 we use Bayes' Theorem to calculate the probability that mental model i is being employed given that actions A have been observed. The marginal probability of a given mental model $P\left(m_i\right)$ can either be determined beforehand by the modeler or we can assume a uniform distribution. The calculation of $P\left(A_{obs}|m_i\right)$ is more complicated since there is no definitive method of calculating the probability of observing a certain sequence of actions given that a particular mental model is being employed. However, since we are in a decision-theoretic framework in which agents have their own beliefs (albeit sometimes incorrect) regarding other agents, we can use these beliefs of other agents to determine an agent's probable expected reward when executing a particular action assuming they are using a particular mental model. And using one of the several methods described below, we can approximate the probability that a certain sequence of actions may be observed.

Expected Value Ratio

One method for generating a probability distribution is to treat the model's value function as a likelihood weighting. In other words, we model the agent as choosing an action randomly, with the probability of a given action being proportional to its expected reward. More precisely we assign a probability that an agent will perform a given action as equal to the ratio of the expected reward of that particular action to the total summation of expected rewards of all actions the agent can perform at any given time step.

$$P_{evr}\left(a_j|m_i\right) = \frac{E\left(a_j\right)}{\sum_j E\left(a_j\right)} \quad (5)$$

Linear and Exponential Ranking

Another intuitive basis for a probability distribution is the notion that an agent is less likely to pick actions when there are other actions with a higher expected reward. The further down the ranking an action is, the less likely that a rational agent will select it. Unlike the expected value ratio method, the ranking-based probability distributions are insensitive to the actual difference in expected rewards. One possible instantiation of this intuition is to assign each action a probability proportional to its rank according to expected reward:

$$P_{linear}\left(a_j|m_i\right) = \frac{rank\left(a_j\right) + 1}{\sum_j \left(rank\left(a_j\right) + 1\right)} \quad (6)$$

This method would model the "irrationality" of the agent as increasing linearly as the action drops down the ranking. We could alternatively make irrationality even less likely with an exponential relationship:

$$P_{exp}\left(a_j|m_i\right) = \frac{e^{rank(a_j)}}{\sum_j e^{rank(a_j)}} \quad (7)$$

The rank function monotonically orders each action, where a ranking of a particular action is higher than the rank of another action if its expected reward is greater than that of the other action. If two actions have identical expected rewards, then the assigned rank is also identical. An example of some sample expected values and associated ranking values is shown in Table 1. Each expected value in Table 1 corresponds to an agent's evaluation of a particular action using the lookahead procedure described in 2 and 3.

Table 1: Ranking Example

EV	$rank\left(x\right)$	$rank\left(x\right)+1$	$e^{rank(x)}$
0.65	1	2	$exp\left(1\right) = 2.718$
0.49	0	1	$exp\left(0\right) = 1.0$
0.73	2	3	$exp\left(2\right) = 7.389$
0.65	1	2	$exp\left(1\right) = 2.718$
0.83	3	4	$exp\left(3\right) = 20.086$

Experimental Results

In our experimental setup, three distinct teacher mental models are used with a fixed bully and onlooker. We need consider only three models, because, for the given bully, all of the other possible reward functions produce the same value function as one of these three. Based on the differing goals associated with each model (seen in Table 2), each mental

Table 2: Teacher Goals

Mental model	Bully	Onlooker	Victim
A	0	1.0	0
B	0.25	0	0.75
C	0.25	0.75	0

Table 3: Policy A

Bully Act	Onlooker Act	Action
pick-on Victim	laugh-at	punish Bully
pick-on Victim	wait	punish Bully
wait	laugh-at	punish Bully
wait	wait	punish Bully

Table 4: Policy B

Bully Act	Onlooker Act	Action
pick-on Victim	laugh-at	punish Onlooker
pick-on Victim	wait	wait
wait	laugh-at	wait
wait	wait	punish Onlooker

Table 5: Policy C

Bully Act	Onlooker Act	Action
pick-on Victim	laugh-at	wait
pick-on Victim	wait	wait
wait	laugh-at	wait
wait	wait	wait

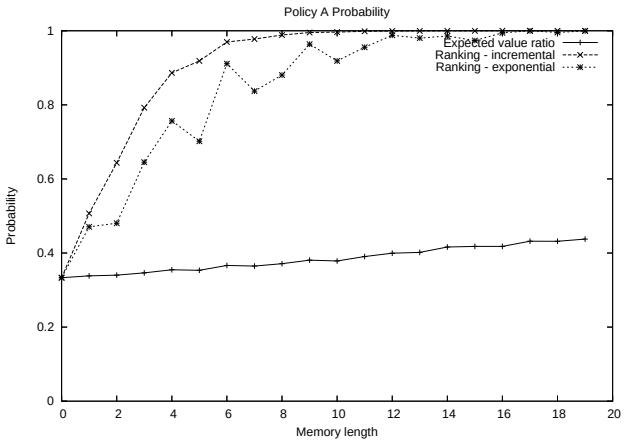

Figure 1: Policy A

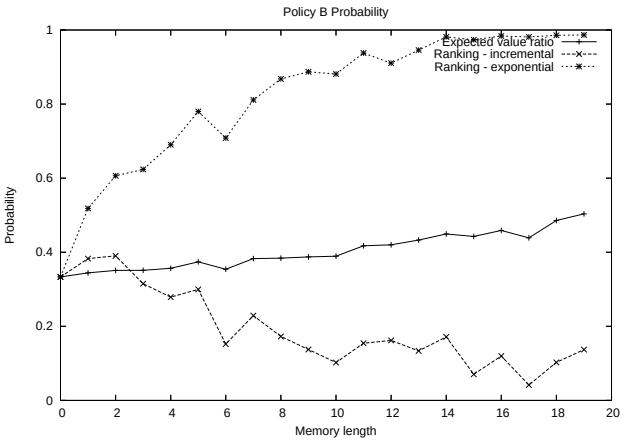

Figure 2: Policy B

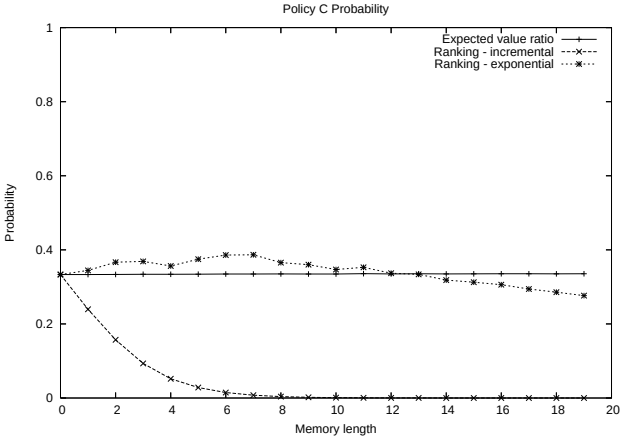

Figure 3: Policy C

model compiled to a unique policy given in Tables 3, 4 and 5. The first and second columns of a given policy table correspond to the various actions that are available to the bully and onlooker respectively. The third column corresponds to the resulting action the teacher will take in response to the observed actions of the bully and onlooker. For example, if the teacher is employing policy B and observes the bully picking on the victim and the onlooker laughing at the victim, then the teacher will punish the onlooker. Graphs 1, 2 and 3 show the results of the simulation experiments. Each graph corresponds to a scenario in which a different initial policy (A, B, or C) was chosen for the teacher. The other observing agents begin the scenario with a uniform marginal probability distribution over the three possible policies. In other words, the agents believe that the teacher is equally likely to employ any of the three given policies. With the initial starting policy, each graph shows the observing agent's belief of the correct policy being employed by the teacher. We vary the memory length from 0 to 20 and plot the different methods (expected value ratio, linear ranking, and exponential ranking) of calculating the conditional probabilities that a particular policy is being used. The ideal method of calculating the probabilities would begin at the initial marginal probability (.33 in our case) at memory length 0 and then quickly converge towards 1 as memory length increases.

We see that when the expected value ratio method shown in 5 is used, it converges very slowly (or sometimes not at all) to the correct policy. In our particular domain the expected values for any given action that an agent can take are generally very similar. So if an agent chooses a sub-optimal action having an expected value close to the optimal action, then the penalty to the probability calculation is extremely small. In other words, as long as actions are chosen that are relatively close in value to the action of the highest expected value there will be very little movement from the initial distribution.

The ranking methods, both linear and exponential, are insensitive to the actual differences in expected values and instead rely on a relative ordering, or ranking, of the expected values. Thus, the ranking methods will assign a high probability that a particular model is being employed if the action with the highest expected value is consistently observed. However, the linear ranking method can sometimes understate the importance of choosing the optimal action while the exponential ranking method can sometimes overstate its importance. Consider the following example: Assume we have two policies X and Y. Each policy has three actions available to it: actions a, b, c. Furthermore, assume the expected values for X as $E(X_a) = E(X_b) = E(X_c)$ and for Y as $E(Y_a) > E(Y_b) > E(Y_c)$. In this case the ranking of the actions will be the following: $Rank(X_a) = Rank(X_b) = Rank(X_c) = 0$ and $Rank(Y_a) = 2$, $Rank(Y_b) = 1$, $Rank(Y_c) = 0$. If optimal action b is chosen under policy X we see that the same action is not optimal for policy Y. But the linear ranking method will evaluate $P_{linear}(b|X) = P_{linear}(b|Y)$ which is clearly not the desired outcome. Employing the exponential ranking method, we can avoid understating the importance of choosing the optimal action, but then become susceptible to overstating its importance. For example, in addition to the aforementioned example consider that after action b is observed, action a is subsequently observed. While these two actions are both optimal for policy X only action a is optimal for policy Y. But using the exponential ranking function described in 7 we find that $P_{exp}(b|Y) \times P_{exp}(a|Y) > P_{exp}(b|X) \times P_{exp}(a|X)$. Figures 2 and 3 both illustrate some limitations to the linear and exponential ranking methods. Policy C has several actions of equal expected value and thus in some cases the linear ranking method understates the importance of choosing the optimal action while the exponential method sometimes overstates its importance.

Discussion

While we have explored three methods of calculating $P(m_i|A_{obs})$ there are still certainly many more probabilistic models that can be investigated. For instance, we could experiment with a multitude of other linear and exponential functions in concert with both he ranking function and the actual expected rewards. Alternatively, we could model the error has being generated by some noise model (e.g., Gaussian), which would be parameterized by the degree of confidence in the space of possible models (e.g., the variance on the Gaussian would decrease with confidence).

Another direction that is currently outside the scope of this paper is to experiment with what to do with the end probability $P(m_i|A_{obs})$. For example, if we use this distribution as a belief state over frames in an interactive POMDP, then we could use this information in the calculation of the recognizing agent's expected reward. We could do so by calculating the expected reward for that agent using each mental model and then weighting those values according to our calculated probability distribution. While such a decision-making procedure would be optimal, we are also interested in using these models to simulate the modeling that people do of each other. Thus, we could instead simply have an agent switch beliefs of a mental model depending on which mental model has the highest current probability given the observed actions. And certainly, agents could have individual preferences or thresholds over which they are willing to change their beliefs over the mental models of other agents.

Conclusion

We have discussed and explored several methods for calculating the posterior probability $P(m_i|A_{obs})$ that a particular mental model is being employed given a sequence of observed actions. In particular we've explored several methods of calculating the conditional probability $P(a_j|m_i)$ that action j is observed given that mental model i is being used. These methods include using an expected reward ratio and linear and exponential ranking functions. We have demonstrated the application of these techniques using an example scenario of childhood aggression within PsychSim, a multiagent framework and show that the expected reward ratio and exponential ranking functions are the most well suited for our domain. However, additional work still needs to be

done to both formalize and explore other methods of calculating the conditional probabilities of actions as well as investigating how to better integrate the newly calculated probability distribution over mental models into our multiagent, decision-theoretic framework.

References

Bui, H. H.; Venkatesh, S.; and West, G. 2002. Policy recognition in the abstract hidden Markov model. *Journal of Artificial Intelligence Research* 17:451–499.

Charniak, E., and Goldman, R. P. 1993. A Bayesian model of plan recognition. *Artificial Intelligence* 64(1):53–79.

Gmytrasiewicz, P., and Doshi, P. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24:49–79.

Goldman, R. P.; Geib, C. W.; and Miller, C. A. 1999. A new model of plan recognition. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 245–254.

Kaminka, G.; Pynadath, D. V.; and Tambe, M. 2001. Monitoring deployed agent teams. In *Proceedings of the International Conference on Autonomous Agents*, 308–315.

Marsella, S.; Pynadath, D.; and Read, S. 2005. Psychsim: Agent-based modeling of social interactions and influence. In *Proceedings of the International Conference on Cognitive Modeling*, 243–248.

Pynadath, D., and Marsella, S. 2005. Psychsim: Modeling theory of mind with decision-theoretic agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1181–1186.

Pynadath, D., and Tambe, M. 2002. Multiagent teamwork: Analyzing the optimality and complexity of key theories and models. In *In Proceedings of the 1st conference on autonomous agents and multiagent systems (AAMAS-2002.*

Pynadath, D. V., and Wellman, M. P. 2000. Probabilistic state-dependent grammars for plan recognition. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 507–514.

Smallwood, R. D., and Sondik, E. J. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* 21:1071–1088.