Evaluating POMDP Rewards for Active Perception (Extended Abstract)

Adam Eck and Leen-Kiat Soh
Department of Computer Science and Engineering
University of Nebraska-Lincoln
256 Avery Hall, Lincoln, NE, 68588, USA
+1-402-472-4257
{aeck, lksoh}@cse.unl.edu

ABSTRACT

One popular approach to active perception is using POMDPs to maximize rewards received for sensing actions towards task accomplishment and/or continually refining the agent's knowledge. Multiple types of reward functions have been proposed to achieve these goals: (1) state-based rewards which minimize sensing costs and maximize task rewards, (2) belief-based rewards which maximize belief state improvement, and (3) hybrid rewards combining the other two types. However, little attention has been paid to understanding the differences between these function types and their impact on agent sensing and task performance. In this paper, we begin to address this deficiency by providing (1) an intuitive comparison of the strengths and weaknesses of the various function types, and (2) an empirical evaluation of our comparison in a simulated active perception environment.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence – *intelligent agents, multiagent systems*

General Terms

Performance, Design, Experimentation

Keywords

Active Perception, POMDP, Reward Functions, Sensing

1. INTRODUCTION

Recently, one application of intelligent agents growing in popularity is intelligent information gathering. Here, developers commonly model the agent's reasoning about sensing as an **active perception** problem (e.g., [9]), where the agent makes explicit decisions about sensing to maximize the quality and/or quantity of its gathered information. One popular approach to active perception is to make sequential decisions using the partially observable Markov decision process (POMDP) [5], e.g., in user preference elicitation [2, 3] and agent-based classification [4].

To illustrate, we consider a robotic mining simulation called *MineralMiner*, a testbed for sensing research similar to RockSample [8]. Here, an intelligent agent completes frequent mineral collection tasks with firm deadlines. To discover minerals (gold, silver, uranium), the agent performs sensing on various mines in the environment. The agent models its sensing at each mine with an active perception POMDP, where the states and observations represent the possible mineral types in the mine and the actions represent various sensing activities with different cost and accuracy, as well as drilling actions (which stop sensing) for each type of mineral. Of note, drilling for an incorrect mineral type destroys a mine. Thus, quality sensing is necessary for completing tasks.

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.

2. REWARD FUNCTION COMPARISON

Several types of reward functions for active perception POMDPs have been proposed in the literature. First, **state-based rewards** $R(s, a)$ (e.g., [3, 4]) follow the traditional design of reward functions in the POMDP literature [5], where rewards are the benefit or cost of actions in different states with respect to the accomplishment of tasks and environment impact. An agent handles its uncertainty about the hidden state of the environment by marginalizing expected rewards over beliefs about each state:

$$\sum_{s \in S} b(s) R(s, a) \qquad (1)$$

We present $R(s, a)$ values for two state-based reward functions for MineralMiner in Table 1 (similar to [3]), where (1) Cost Sensing encodes the actual costs incurred by the agent for each action, and (2) Zero Cost Sensing focuses only on task rewards.

Alternatively, recently proposed **belief-based rewards** [1] break from tradition and consider only measures on the entire belief state of the agent, independent of individual states. For example, if the primary goal of sensing is to reduce the uncertainty in the agent's beliefs, the agent can use the entropy in its belief state as a measure of uncertainty, then maximize the negative of its expected entropy to minimize uncertainty:

$$-E[H(b^{a,o})] = E\left[\sum_{s \in S} b^{a,o}(s) \log_{|S|} b^{a,o}(s)\right] \qquad (2)$$

Other belief-based reward functions accomplish similar goals including maximizing information gain or the expected top belief:

$$E[\max_{s \in S} b^{a,o}(s)] \qquad (3)$$

The intuitive strengths and weaknesses of these functions include:

- State-based rewards *directly encode the costs of sensing activities*, allowing the agent to minimize sensing costs, whereas belief-based rewards ignore such information.
- Belief-based rewards *directly encode the benefits of sensing* (i.e., belief state improvement), whereas state-based rewards only implicitly consider this information through finding policies of actions that reach task accomplishment the fastest.
- State-based rewards *provide a natural stopping condition for sensing*: when the expected reward of using information exceeds further sensing costs, and thus are appropriate for task-based environments. Belief-based rewards, instead, *require*

Table 1: State-based Rewards for MineralMiner

Action	Cost Sensing		Zero Cost Sensing	
	Correct State	Incorrect State	Correct State	Incorrect State
Advanced (80% Accuracy)	-5	-5	0	0
Basic (50% Accuracy)	-2	-2	0	0
Wait (do nothing)	0	0	0	0
Drilling	100	-500	100	-500

an external stopping condition (e.g., stop when a confidence threshold is reached for the top belief: $b(s) \geq 0.85$).

- Belief-based rewards *optimize beliefs for continual sensing when it is unknown when information will be used*, whereas state-based rewards might be inappropriate for such environments [1] due to lacking task rewards to guide sensing.

Finally, **hybrid rewards** consider both of the other types simultaneously in the form of a weighted function [1, 6], e.g.,

$$w \sum_{s\in S} b(s)R(s,a) - (1-w)\, E[H(b^{a,o})] \qquad (4)$$

where w weights the impact of the two reward types. Below, we use a hybrid of Cost Sensing (c.f., Table 1) with negative expected Entropy (Eq. 3) using three weights: $w = 0.25, 0.50, 0.75$.

Hybrid functions have the potential to merge the strengths of state- and belief-based rewards while mitigating their weaknesses:

- Hybrid rewards *add cost information for sensing activities to belief-based reward functions* to improve sensing.
- Hybrid rewards *incorporate belief state revision into state-based rewards* to speed up belief state convergence and promote faster task accomplishment.
- However, the weight between the two types of rewards must be properly tuned, which can be difficult to set *a priori.*

3. INVESTIGATION

We now provide results from an empirical investigation using MineralMiner to evaluate our intuitive comparison of the various active perception POMDP reward function types. To maximize rewards, we choose actions from limited depth policy trees created online [7] from the current belief state for the current mine. This approach (1) finds exact solutions with low computational cost due to the small POMDP size, and (2) allows us to compare how performance depends on policy depth $(1, 2, 3, 4, 5)$ due to the different properties of the functions. We used 600 mines/tasks to provide many opportunities for sensing and ran our experiments 30 times with different random seeds to minimize variance.

Figures 1-2 present: (1) the number of mines correctly identified/drilled, measuring sensing *effectiveness*, and (2) the task-based rewards earned by the agent (Cost Sensing, c.f., Table 1), measuring sensing *efficiency*. Due to space constraints, we highlight the key analyses from our results, confirming our intuitions:

- State-based functions almost always improved as policy depth increased. This is because myopically, state-based functions only minimize *immediate* costs, whereas non-myopically they minimize *total* costs through less sensing.
- Belief-based functions performed consistently for all policy depths due to always looking ahead to expected belief improvement, but achieved lower task rewards (higher costs) than state-based functions for longer policy depths $(3, 5)$.
- Hybrid rewards $(w = 0.75)$ performed the best due to combining cost-awareness (state-based rewards) and rapid belief improvements (belief-based rewards).

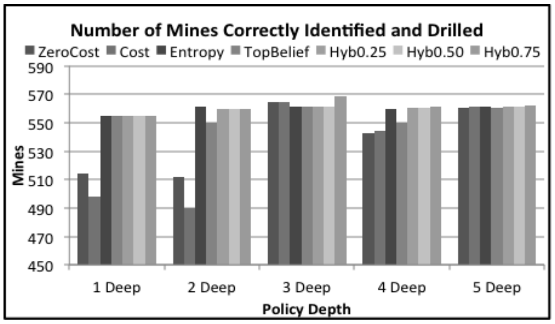

Figure 1: Number of Mines Correctly Identified and Drilled

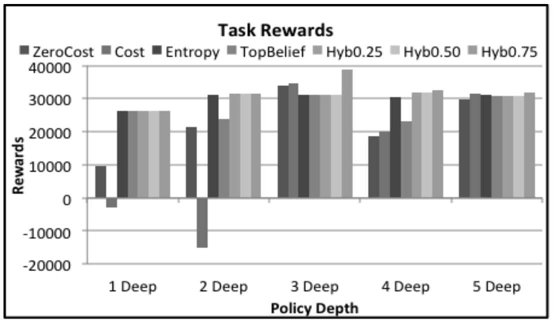

Figure 2: Cumulative Task-based Rewards
Note: Cost has negative task rewards due to very poor sensing

- Increasing policy depth from 3 to 4 worsened performance of all functions but at different rates. Thus, looking farther ahead is not always beneficial, which we intend to further study.

4. ACKNOWLEDGMENTS

This research was supported by a NSF Graduate Research Fellowship (grant DGE-054850) and was completed utilizing the Holland Computing Center of the University of Nebraska.

5. REFERENCES

[1] Araya-Lopez, M., Buffet, O., Thomas, V., and Charpillet, F. 2010. A POMDP extension with belief-dependent rewards. *Proc. of NIPS'10*.

[2] Boutilier, C. 2002. A POMDP formulation of preference elicitation problems. *Proc. of AAAI'02*. 239-246.

[3] Doshi, F. and Roy, N. 2008. The permutable POMDP: fast solutions to POMDPs for preference elicitation. *Proc. of AAMAS'08*. 493-500.

[4] Guo, A. 2003. Decision-theoretic active sensing for autonomous agents. *Proc. of AAMAS'03*. 1002-1003.

[5] Kaelbling, L.P., Littman, M.L., and Cassandra, A.R. 1998. Planning and acting in partially observable stochastic domains. *AI*. 101. 99-134.

[6] Mihaylova, L. et. al. 2002. Active sensing for robotics – a survey. *Proc. of NM&A'02*.

[7] Ross, S., Pineau, J., Paquet, S., and Chaib-draa, B. 2008. Online planning algorithms for POMDPs. *JAIR*. 32. 663-704.

[8] Smith, T. and Simmons, R. 2004. Heuristic search value iteration for POMDPs. *Proc.UAI'04*. 520–527.

[9] Weyns, D., Steegmans, E., and Holvoet, T. 2004. Towards active perception in situated multi-agent systems. *Applied Artificial Intelligence*. 18. 867-883.