Potential-Based Reward Shaping for POMDPs (Extended Abstract)

Adam Eck and Leen-Kiat Soh
Department of Computer Science and Engineering
University of Nebraska-Lincoln
{aeck, lksoh}@cse.unl.edu

Sam Devlin and Daniel Kudenko
Department of Computer Science
University of York, UK
{devlin, kudenko}@cs.york.ac.uk

ABSTRACT
We address the problem of suboptimal behavior caused by short horizons during online POMDP planning. Our solution extends potential-based reward shaping from the related field of reinforcement learning to online POMDP planning in order to improve planning without increasing the planning horizon. In our extension, information about the quality of belief states is added to the function optimized by the agent during planning. This information provides hints of where the agent might find high future rewards, and thus achieve greater cumulative rewards.

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

General Terms
Performance, Design, Experimentation

Keywords
POMDP; Potential-Based Reward Shaping; Online Planning

1. INTRODUCTION
Partially observable Markov decision processes (POMDPs) [5] are a popular approach to agent reasoning and planning. In complex environments (e.g., with large belief state spaces) where the agent needs to adapt its behavior to its unpredictable experiences, online planning approaches [7] are advantageous. In online planning, an agent interleaves planning and execution while it operates, focusing only on belief states the agent actually holds (as opposed to pre-computed belief states in offline approaches).

To plan fast enough to operate in real-time, online planning is generally restricted to limited horizon depths. However, this can result in myopic, suboptimal behavior since the agent might not consider important decisions that could lead to large rewards a little farther in the future. That is, short horizon planning can result in *underestimating* action sequences that earn greater future rewards and *overestimating* sequences that earn large *immediate* rewards but lead to smaller *cumulative* rewards in the future.

Previous solutions [7] to this limited horizon problem for online planning include (1) Monte Carlo methods (e.g., [3]), which focus planning on the most likely beliefs based on environment dynamics modeled by the POMDP, and (2) heuristic search methods (e.g., [10]), which focus planning on the "best" belief states, such as those returning the highest guaranteed reward. Compared to planning for all possible belief states, both of these solutions enable deeper planning in the same amount of time spent since they limit the belief states considered, but they unfortunately neglect possible belief states and can lead to the same under- or over-estimation problems of planning with limited horizons.

To improve online planning, we propose a new solution enabling the agent to *implicitly* look farther forward (1) without increasing the short horizon, and (2) without sacrificing coverage by neglecting possible belief states. Specifically, we propose the first extension of **potential-based reward shaping** (PBRS) [2, 4, 6] from the related field of reinforcement learning (RL) to online POMDP planning (also providing provable performance guarantees for including belief-based POMDP rewards [1]). Recently, PBRS was demonstrated [9] to be useful in planning with the simpler, fully observable MDP. However, extending PBRS to POMDPs is non-trivial (and possibly *richer*) due to partial observability.

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013),* Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

2. POTENTIAL-BASED PLANNING
In RL, PBRS [2, 4, 6] is a method of modifying the agent's reward function in order to address the exploration-exploitation problem: determining how to better maximize long term rewards, given uncertain knowledge of future rewards. PBRS accomplishes this by embedding a priori information in a *potential function* mapping states to potential and providing the agent with additional rewards based on the difference in potential of the initial and final state in a transition. This encourages exploration of high potential states and, through the structure of shaping (c.f., Eqs. 3-4), is guaranteed [2, 4, 6] to still optimize the agent's original reward function.

To extend PBRS to online POMDP planning, we first note that in this setting, the agent makes decisions based on its uncertain belief state b (instead of individual state s), which represents the agent's probabilistic beliefs over which possible environment state is the correct one. Therefore, we define our potential function $\phi(b)$ over belief states. Given this change, such a potential function can measure several different types of information about the agent's beliefs. First, the function could retain information about individual states (as in RL) and simply measure the expected potential given its uncertainty:

$$\phi(b) = \sum_{s \in S} b(s)\phi(s) \quad (1)$$

Second, the potential function could provide measures *independent* from the potential of any particular state. For example, the potential function could measure the certainty in the agent's belief, indicating how well the agent has handled partial observability before acting on its beliefs [1]:

$$\phi(b) = \log|S| + \sum_{s \in S} b(s) \log b(s) \quad (2)$$

Finally, the potential function can also represent a preference ordering over beliefs, encouraging the agent to reach some belief states before others (e.g., for long-term goal directed behavior).

To include ϕ in POMDP planning, we shape the agent's rewards:

$$r_t = R(b_t, a) + F(b_t, a, b_{t+1}) \quad (3)$$

where $F(b_t, a, b_{t+1}) = \gamma\Phi(b_{t+1}) - \phi(b_t)$ (4)

Here, the reward r_t earned at time t is shaped by adding the difference in potential in changing the agent's belief from b_t to b_{t+1}.

The forward-looking term $\gamma\Phi(b_{t+1})$ enables the agent to consider potentially high rewards beyond the reward at time t (and thus eventually beyond the planning horizon). Thus, PBRS can lead to policies that better maximize the agent's long-term cumulative rewards given a fixed, small planning horizon without sacrificing belief states during *planning* (only modifying which are reached during *execution*). Furthermore, based on the inclusion of the discount factor γ in Eq. 4 (which is the same γ used to calculate cumulative rewards during POMDP planning [5]), it can be shown (similar to the proofs for PBRS in RL, e.g. [2, 4]) that optimizing the shaped reward also optimizes the agent's original reward function. Finally, it can also be shown (using Theorem 3.1 in [1]) that as long as the potential function is convex, the agent's shaped rewards remain convex and thus can still be solved by a wide range of POMDP planners relying on convexity for optimization.

3. EMPIRICAL RESULTS

To demonstrate the benefits of using PBRS with online planning, we present an empirical study with a classic POMDP benchmark problem: RockSample [8]. Here, an agent is placed in a $g \times g$ grid from which it must check k rocks of unknown quality and sample only good rocks, then exit the grid. The agent is given a reward of +10 (-10) for sampling a good (bad) rock and +10 for exiting. We adopt the commonly used $g = 7$ and $k = 8$, (e.g., [7, 10]), and perform online planning for the complete plan tree up to policy horizons $n = 1,2,3,4$ both without shaping (**Original**) and with different potential functions:

(1) **TopBelief**, a measure of belief certainty similar to Eq. 2 but exploiting the factored state space:

$$\phi(b) = \max_{s \in S} b(s) \quad (5)$$

(2) **ClosestDistance**, measuring domain dependent knowledge that belief states closer to uncertain rocks achieve greater accuracy and thus most immediate belief improvement:

$$\phi(b) = \begin{cases} -\frac{1}{2g}(d+1) & \text{if some rocks are uncertain} \\ 0 & \text{else} \end{cases} \quad (6)$$

where d is the Euclidian distance to the closest rock whose quality the agent is uncertain about, and

(3) **NoExit,** prioritizing more certain beliefs about rocks before exiting to avoid neglected sampling due to myopic planning:

$$\phi(b) = \begin{cases} -1000 & \text{if exiting with uncertain rocks} \\ 0 & \text{else} \end{cases} \quad (7)$$

Fig. 1 presents the cumulative (unshaped) rewards earned by the agent both with and without PBRS. First, using PBRS (*regardless* of potential function) better maximized the original reward function than planning without PBRS (Original) for the shortest horizons ($n < 4$) and never performed significantly worse. Furthermore, one potential function (NoExit) achieved much greater performance than without PBRS. Specifically, this potential function prioritized belief states in order to avoid myopically exiting the grid before forming certain beliefs about all rocks. Without belief prioritization, the short horizons caused the agent to otherwise take the known, certain +10 reward for exiting the grid whenever possible, achieving less *cumulative* reward than instead sampling additional rocks. Overall, we conclude that PBRS is beneficial to improving online POMDP planning with short horizons, especially given an appropriate potential function.

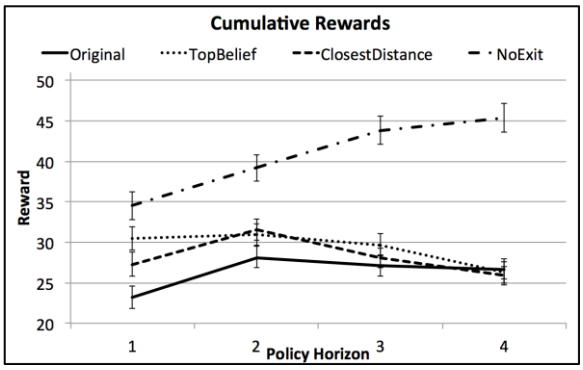

Figure 1: Cumulative Rewards in RockSample

4. CONCLUSIONS

In conclusion, we described the first extension of PBRS to online POMDP planning in order to improve short horizon planning. We suggested several types of information that can be embedded in potential functions hinting of high future rewards and empirically demonstrated that PBRS improved cumulative (unshaped) rewards. In the future, we intend to (1) publish our theoretical results guaranteeing PBRS provides an opportunity to better maximize cumulative rewards, whilst still optimizing the original reward function, (2) establish how to select potential functions based on environment characteristics (including dynamic functions [4]), and (3) apply our results to RL (e.g., partially observable settings).

5. 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.

6. 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] Asmuth, J., Littman, M.L., and Zinkov, R. 2008. Potential-based shaping in model-based reinforcement learning. *Proc. of AAAI'08*. 604-609.

[3] Bertsekas, D.P. and Castanon, D.A. 1999. Rollout algorithms for stochastic scheduling problems. *Journal of Heuristics*. 5. 89-108.

[4] Devlin, S. and Kudenko, D. 2012. Dynamic potential-based reward shaping. *Proc. of AAMAS'12*.

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

[6] Ng, A.Y., Harada, D., and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. *Proc. of ICML'99*. 278-287.

[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. of UAI'04*. 520–527.

[9] Sorg, J., Singh, S., and Lewis, R.L. 2011. Optimal rewards versus leaf-evaluation heuristics in planning agents. *Proc. of AAAI'11*. 465-470.

[10] Zhang, Z. and Chen, X. 2012. FHHOP: A factored heuristic online planning algorithm for POMDPs. *Proc. of UAI'12*.