Landmark Based Reward Shaping in Reinforcement Learning with Hidden States

Extended Abstract

Alper Demir
Middle East Technical University
Ankara, Turkey
ademir@ceng.metu.edu.tr

Erkin Çilden
STM Defense Technologies Engineering and Trade Inc.
Ankara, Turkey
erkin.cilden@stm.com.tr

Faruk Polat
Middle East Technical University
Ankara, Turkey
polat@ceng.metu.edu.tr

ABSTRACT

While most of the work on reward shaping focuses on fully observable problems, there are very few studies that couple reward shaping with partial observability. Moreover, for problems with hidden states, where there is no prior information about the underlying states, reward shaping opportunities are unexplored. In this paper, we show that landmarks can be used to shape the rewards in reinforcement learning with hidden states. Proposed approach is empirically shown to improve the learning performance in terms of speed and quality.

KEYWORDS

reward shaping; landmarks; reinforcement learning; hidden states

ACM Reference Format:
Alper Demir, Erkin Çilden, and Faruk Polat. 2019. Landmark Based Reward Shaping in Reinforcement Learning with Hidden States. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION AND RELATED WORK

In Reinforcement Learning (RL) context, Reward Shaping (RS) methods aim to provide the agent with additional rewards for a problem with sparse or late rewards, so that the agent avoids extensive and probably unnecessary exploration throughout learning. RS is shown to guarantee policy invariance under Markov Decision Process (MDP) model assumption [19].

For most problems, however, the environment is not fully observable as in the MDP. Partially Observable MDP (POMDP) model is a generalization of MDP to fulfill this requirement, where the agent has indirect access to the state space through *observation* space via an observation function of states and actions [12]. An interpretation of POMDP assumes the set of states are entirely hidden, and the model provides a limited set of observations, violating Markov property and giving rise to *perceptual aliasing* [1, 23].

Although perceptual aliasing makes it very difficult, sometimes even impossible, to solve the problem, the agent can benefit from any information that can completely distinguish its state. A *landmark* corresponds to such an information and takes place as a distinctive indicator in different fields of the related literature, such as planning [5, 13] and robotic navigation [6, 21] . Although there is no agreed definition of a landmark, the one used in our work fuses the ideas from the "unique observation" interpretation in [11] and "memory-based" approach used in [14].

Known RL algorithms like Q-Learning [22] lose convergence guarantees when the task is non-Markovian as in POMDP with hidden states [20]. *Eligibility traces* are used to overcome this problem where the agent leaves decaying traces over the previous transitions and employs value updates based on these traces [15, 22]. James et al., proposed an adaptation of the well-known eligibility trace based Sarsa(λ) algorithm for problems containing landmarks, called SarsaLandmark, and showed that this adaptation can further improve the convergence of the algorithm [11].

Various studies adapt RS idea using different approaches, such as potential based RS (PBRS) [19] and plan based RS [9]. The methods were tailored for different settings [7] including multi-agent RL [2, 3], and theoretical analyses were carried out [8, 16]. Automatic learning of the potential function also gained attention, where macro-action oriented abstractions were used [10, 17].

Most of the RS effort, however, assume MDP model. Even the studies for the POMDP case assume belief state formalism [4], which is essentially a continuous MDP. To our best knowledge, this is the first attempt to incorporate RS for problems with hidden state interpretation of POMDP.

2 LANDMARK BASED REWARD SHAPING

Regular RS approach provides a shaping reward for every transition based on the potentials of the states in it. However, if there is perceptual aliasing, finding a unique potential value is impossible for an ambiguous observation which represents multiple states with possibly different potentials. To overcome this challenge, a *state estimate* can be kept and used to find a suitable policy [15]. However, the problem of assigning a potential to an estimated state persists if that estimated state is still aliased.

On the other hand, it is often the case that some estimated states are unique in the problem so that the agent can rely on them. The agent can be informed about its progress in the given task by using those "specific" estimated states.

Definition 2.1. A state estimate x is a *landmark*, if it uniquely represents a state in a partially observable environment.

Following the Definition 2.1, a landmark is free of any form as long as it can completely distinguish a true problem state. Due to this one-to-one mapping, it is now straightforward to assign a potential to a landmark and use it for RS.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

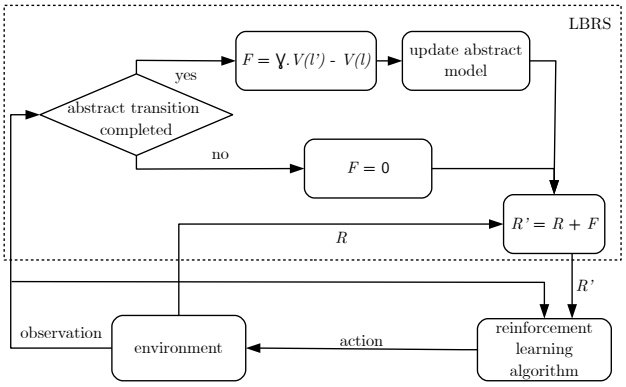

Figure 1: LBRS workflow, combined with the underlying RL algorithm. l and l' represent the previously observed landmark and the recently reached landmark respectively, and $V(.)$ represents the value of a landmark that is used as the potential value. R, F, R' are as defined in [19].

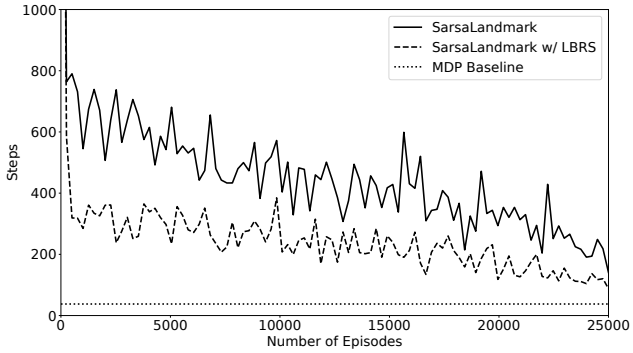

Figure 3: Number of steps taken to reach the goal state in `6 rooms` domain. Results are averaged over 100 experiments with 5000 step limit, $\alpha = 0.01$, $\gamma = 0.9$ and ϵ-greedy action selection linearly decaying from 0.2 downto 0.0001. MDP baseline represents optimal MDP policy performance.

We argue that, a shaping reward can only be reasonable when a transition between two landmarks occurs. Since the potentials of the landmark are consistent, RS, based on these potentials, would also be consistent, providing a reliable information to the agent about its actions. A transition between landmarks may take more than one step, forming an abstract transition within an abstract model of landmarks.

The core idea is to use the landmarks in a problem with hidden states to form an abstract model and apply RS whenever the agent completes an abstract transition between two landmarks. Assuming that the agent knows the landmarks in advance, the remaining question is how to find the potentials of these landmarks. In order to learn the potentials, we follow Grześ's work [10] which makes an abstraction over the set of states and applies value iteration on the abstract states. However, unlike in [10], we only form the abstract model with the landmarks of the problem since it would not be reasonable to make further abstraction over an already aliased abstraction of observations.

Figure 1 summarizes the main RL loop combined with the proposed RS approach, named Landmark Based Reward Shaping (LBRS). Basically, whenever the agent completes an abstract transition between two landmarks, say l and l', it uses their values to calculate a shaping reward. Then, it updates the abstract model composed of landmarks with this new abstract transition and applies value iteration. Finally, the shaping reward is coupled with the regular reward mechanism to be provided to the underlying RL algorithm.

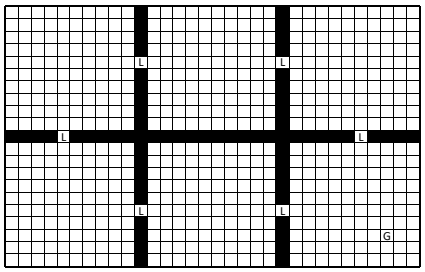

Figure 2: Illustration of the `6 rooms` domain. Landmarks and the goal state are indicated by L and G respectively. The the black grids represent the walls.

3 EXPERIMENT

As an empirical evaluation, we experimented LBRS on `6 rooms` domain [18], by coupling it with SarsaLandmark [11]. In `6 rooms` domain (Figure 2), the landmarks reside in bottleneck regions, except the goal state, which is in the bottom right part of a room. The agent starts from any cell at the top-left room, aiming to reach the goal state with four navigational actions while getting -0.01 punishment for a regular movement and $+1$ reward for ending up in the goal state. The problem is partially observable since the agent's observations are formulated by its distance to the walls in four directions. The distances are enumarated into four categories (one step from the wall, two steps from the wall, closer to the wall in this direction than the other, further from the wall in this direction than the other).

Figure 3 shows SarsaLandmark with LBRS not only learns faster, but helps the agent find a better policy. Since each room provides a similar set of observations, finding a good policy is difficult without keeping a memory. However, it is clear from Figure 3 that LBRS improved the algorithm by leading the agent to the goal state much earlier. Since LBRS provides the agent with reliable feedback from the environment about its progress, the agent reaches the goal state much sooner throughout the learning process.

4 CONCLUSION

This paper proposes that RS can be adapted to problems with hidden states by making use of landmarks. It is shown that LBRS further improves the performance of a landmark based algorithm designed for problems with hidden state, by leading the agent to the goal state much faster. As an immediate future work, we plan to extend our experimentation to more complex domains with sophisticated landmarks.

REFERENCES

[1] Lonnie Chrisman. 1992. Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctions Approach. In *Proc. of the Tenth National Conference on Artificial Intelligence (AAAI'92)*. 183–188.

[2] Sam Devlin and Daniel Kudenko. 2011. Theoretical Considerations of Potential-based Reward Shaping for Multi-agent Systems. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS'11)*. 225–232.

[3] Sam Devlin and Daniel Kudenko. 2016. Plan-based reward shaping for multi-agent reinforcement learning. *The Knowledge Engineering Review* 31, 1 (2016), 44–58.

[4] Adam Eck, Leen-Kiat Soh, Sam Devlin, and Daniel Kudenko. 2016. Potential-based reward shaping for finite horizon online POMDP planning. *Autonomous Agents and Multi-Agent Systems* 30, 3 (2016), 403–445.

[5] Mohamed Elkawkagy, Pascal Bercher, Bernd Schattenberg, and Susanne Biundo. 2012. Improving Hierarchical Planning Performance by the Use of Landmarks. In *Proc. of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. 1763–1769.

[6] Lutz Frommberger. 2008. Representing and Selecting Landmarks in Autonomous Learning of Robot Navigation. In *Intelligent Robotics and Applications (ICIRA 2008, LNCS)*, Vol. 5314. Springer Berlin Heidelberg, 488–497.

[7] Yang Gao and Francesca Toni. 2015. Potential Based Reward Shaping for Hierarchical Reinforcement Learning.. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'15)*. 3504–3510.

[8] Marek Grześ. 2017. Reward Shaping in Episodic Reinforcement Learning. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'17)*. 565–573.

[9] Marek Grześ and Daniel Kudenko. 2008. Plan-based reward shaping for reinforcement learning. In *Proc. of the 4th International IEEE Conference Intelligent Systems*, Vol. 2. 10–22 – 10–29.

[10] Marek Grześ and Daniel Kudenko. 2010. Online learning of shaping rewards in reinforcement learning. *Neural Networks* 23, 4 (2010), 541 – 550. (Special Issue for the 18th International Conference on Artificial Neural Networks, ICANN 2008).

[11] Michael R. James and Satinder Singh. 2009. SarsaLandmark: An Algorithm for Learning in POMDPs with Landmarks. In *Proc. of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS'09)*. 585–591.

[12] Leslie Pack Kaelbling, Michael L. Littman, and Andrew P. Moore. 1996. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research* 4 (1996), 237–285.

[13] Erez Karpas, David Wang, Brian Charles Williams, and Patrik Haslum. 2015. Temporal Landmarks: What Must Happen, and When. In *Proc. of the 25th International Conference on Automated Planning and Scheduling (ICAPS'15)*. 138–146.

[14] Yunlong Liu, Yun Tang, and Yifeng Zeng. 2015. Predictive State Representations with State Space Partitioning. In *Proc. of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15)*. 1259–1266.

[15] John Loch and Satinder P. Singh. 1998. Using Eligibility Traces to Find the Best Memoryless Policy in Partially Observable Markov Decision Processes. In *Proc. of the Fifteenth International Conference on Machine Learning (ICML'98)*. 323–331.

[16] Ofir Marom and Benjamin Rosman. 2018. Belief Reward Shaping in Reinforcement Learning. In *Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI'18)*. 3762–3769.

[17] Bhaskara Marthi. 2007. Automatic Shaping and Decomposition of Reward Functions. In *Proc. of the 24th International Conference on Machine Learning (ICML'07)*. 601–608.

[18] Ishai Menache, Shie Mannor, and Nahum Shimkin. 2002. Q-Cut—Dynamic Discovery of Sub-goals in Reinforcement Learning. In *Proc. of the 13th European Conference on Machine Learning (Mach. Learn.: ECML'02)*. 295–306.

[19] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proc. of the Sixteenth International Conference on Machine Learning (ICML'99)*. 278–287.

[20] Satinder P. Singh, Tommi S. Jaakkola, and Michael I. Jordan. 1994. Learning without State-estimation in Partially Observable Markovian Decision Processes. In *Proc. of the Eleventh International Conference on Machine Learning*. 284–292.

[21] Tuomas Välimäki and Risto Ritala. 2016. Optimizing gaze direction in a visual navigation task. In *Proc. of the 2016 IEEE International Conference on Robotics and Automation (ICRA'16)*. 1427–1432.

[22] Chris Watkins. 1989. *Learning from Delayed Rewards*. Ph.D. thesis. Cambridge University.

[23] Steven D. Whitehead and Dana H. Ballard. 1991. Learning to Perceive and Act by Trial and Error. *Machine Learning* 7, 1 (1991), 45–83.