Potential-Based Reward Shaping Preserves Pareto Optimal Policies

Patrick Mannion
Galway-Mayo Institute of Technology
patrick.mannion@gmit.ie

Sam Devlin
University of York
sam.devlin@york.ac.uk

Karl Mason
National University of Ireland Galway
k.mason2@nuigalway.ie

Jim Duggan
National University of Ireland Galway
jim.duggan@nuigalway.ie

Enda Howley
National University of Ireland Galway
enda.howley@nuigalway.ie

ABSTRACT

Reward shaping is a well-established family of techniques that have been successfully used to improve the performance and learning speed of Reinforcement Learning agents in single-objective problems. Here we extend the guarantees of Potential-Based Reward Shaping ($PBRS$) by providing theoretical proof that $PBRS$ does not alter the true Pareto front in MORL domains. We also contribute the first empirical studies of the effect of $PBRS$ in MORL problems.

Keywords

Multi-Objective, Reinforcement Learning, Reward Shaping

1. INTRODUCTION

In Reinforcement Learning (RL), an agent learns to improve its performance with experience by maximizing the return from a reward function. The majority of RL research focuses on optimising systems with respect to a single objective, despite the fact that many real world problems are inherently multi-objective in nature. Single-objective approaches seek to find a single solution to a problem, whereas in reality a system may have multiple conflicting objectives that could be optimised. Examples of multi-objective problems include water resource management [5], traffic signal control [3] and electricity generator scheduling [4].

Compromises between competing objectives can be defined using the concept of Pareto dominance [7]. The Pareto optimal or non-dominated set consists of solutions that are incomparable, where each solution in the set is not dominated by any of the others on every objective. In multi-objective Reinforcement Learning (MORL) the reward signal is a vector, where each component represents the performance on a different objective.

Reward shaping augments the reward function with additional knowledge provided by the system designer, with the goal of improving learning speed. Potential-Based Reward Shaping [6] ($PBRS$) is a specific form of reward shaping that provides theoretical guarantees including policy invariance in single-objective single-agent domains [6], and consistent Nash equilibria in single-objective multi-agent domains [1].

Our work [2] has extended the previous guarantees of $PBRS$ with theoretical proof that the set of Pareto optimal solutions remains consistent when $PBRS$ is used in multi-objective domains, regardless of the quality of the heuristic used. This means that the increased learning speed that is a characteristic of $PBRS$ can be leveraged in multi-objective problem domains, without any risk of altering the intended goals of the problem. The remainder of this paper provides an empirical demonstration of the effect of $PBRS$ in a single-agent MORL domain, and concludes with a discussion of our findings.

2. DEEP SEA TREASURE RESULTS

The Convex Deep Sea Treasure (CDST) environment, shown in Fig. 1, consists of 10 rows and 11 columns, and is a modified version of the Deep Sea Treasure environment [8]. An agent controls a submarine, which searches for undersea treasures. There are 10 treasure locations in all, and the agent begins each episode in the top left state. An episode ends after 1000 actions, or when the agent reaches a treasure location. The agent's state is defined as its current position on the grid, and the actions available correspond to moving one square in one of the four cardinal directions.

There are two objectives in this domain: to minimise the time taken to reach a treasure, and to maximise the reward received when a treasure is reached. After each action selection, the agent receives a reward vector with two elements. The first element is the time reward, which is -1 for all turns. The second element is the treasure reward, which is the value for the corresponding cell in Fig. 1 if a treasure is reached, and zero for all other turns. The Pareto front for this problem (Fig. 2) consists of 10 elements, with a non-dominated policy corresponding to each of the 10 treasure locations.

We test three different Q-learning agents in the CDST: an agent without reward shaping, an agent with a good $PBRS$ heuristic, and an agent with a poorly designed $PBRS$ heuristic. The good heuristic is intended to demonstrate the effect of $PBRS$ when useful domain knowledge is available, and is expected to improve learning speed. Conversely, the poor $PBRS$ heuristic has been purposely designed to mislead the agent receiving it, and is expected to reduce learning speed. However, our formal proof of consistent Pareto fronts states that all agents should learn the same set of policies, regardless of the quality of the $PBRS$ heuristic used. The parameters used are $num\_episodes = 3000$, $\alpha = 0.1$, $\gamma = 1.0$, and $\epsilon = 0.998^{episode}$. Action values are optimistically initialised to [0,125] for all non-terminal states, and to [0,0] for terminal states. In order to sample all policies on the Pareto front, we test each agent with 100 different objective weights uniformly distributed in the continuous range [0.0,1.0]. The non-dominated policies learned are then used to compute

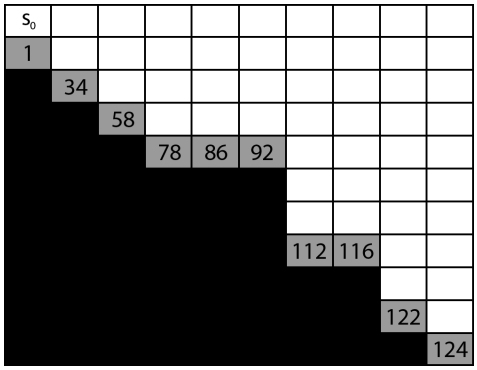

Figure 1: The CDST domain

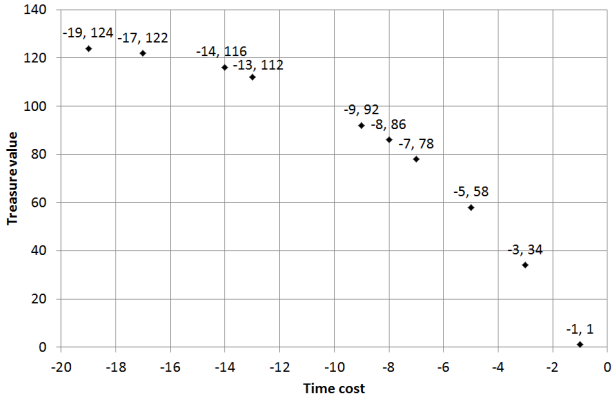

Figure 2: CDST Pareto front

the hypervolume of the agents' policies during learning. Experiments are repeated 30 times, and Fig. 3 shows the average hypervolume during learning. The hypervolume of the Pareto front for the CDST is 2166, computed using a reference point of [-25,0]. The hypervolume measures the quality of the policies learned, and values close to the maximum of 2166 indicate good learning performance.

From the learning curves in Fig. 3, it is evident that all approaches have reached the maximum hypervolume of 2166 after 1200 episodes, and therefore have learned all 10 policies on the true Pareto front of the problem. When a good $PBRS$ heuristic is added, there is a substantial improvement in learning speed, and the maximum hypervolume of 2166 is reached more quickly when compared to an agent learning without $PBRS$. Here $PBRS$ has improved the learning speed, without altering the set of Pareto optimal policies for the problem. When using a poor $PBRS$ heuristic, the learning speed is reduced compared to an agent learning without $PBRS$, but the agent learning with a poor heuristic eventually converges to the maximum hypervolume, and successfully learns all 10 Pareto optimal policies. Thus, $PBRS$ has not altered the set of Pareto optimal policies, regardless of the quality of the heuristic used, as per our theoretical proof.

Acknowledgements

Patrick Mannion's PhD work at the National University of Ireland Galway was funded by the Irish Research Council.

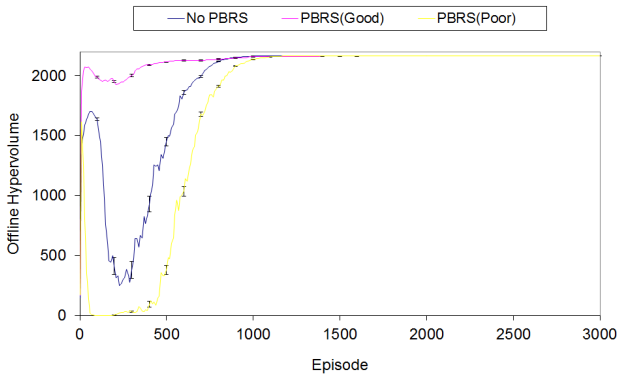

Figure 3: Learning curves for the CDST domain

REFERENCES

[1] S. Devlin and D. Kudenko. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 225–232, 2011.

[2] P. Mannion, S. Devlin, K. Mason, J. Duggan, and E. Howley. Policy invariance under reward transformations for multi-objective reinforcement learning. *Neurocomputing*, 2017 (in press).

[3] P. Mannion, J. Duggan, and E. Howley. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. In L. T. McCluskey, A. Kotsialos, P. J. Müller, F. Klügl, O. Rana, and R. Schumann, editors, *Autonomic Road Transport Support Systems*, pages 47–66. Springer International Publishing, 2016.

[4] P. Mannion, K. Mason, S. Devlin, J. Duggan, and E. Howley. Dynamic economic emissions dispatch optimisation using multi-agent reinforcement learning. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2016)*, May 2016.

[5] K. Mason, P. Mannion, J. Duggan, and E. Howley. Applying multi-agent reinforcement learning to watershed management. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2016)*, May 2016.

[6] A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[7] V. Pareto. *Manual of political economy*. Macmillan, 1971.

[8] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, 84(1):51–80, 2010.