A Theoretical and Empirical Analysis of Reward Transformations in Multi-Objective Stochastic Games*

Patrick Mannion
Galway-Mayo Institute of Technology
patrick.mannion@gmit.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 has been proposed as a means to address the credit assignment problem in Multi-Agent Systems (MAS). Two popular shaping methods are Potential-Based Reward Shaping and difference rewards, and both have been shown to improve learning speed and the quality of joint policies learned by agents in single-objective MAS. In this work we discuss the theoretical implications of applying these approaches to multi-objective MAS, and evaluate their efficacy using a new multi-objective benchmark domain where the true set of Pareto optimal system utilities is known.

Keywords

Multi-Objective, Stochastic Game, Reinforcement Learning, Reward Shaping, Multi-Agent Systems, Credit Assignment

1. INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) has proven to be successful in developing suitable joint policies in numerous complex single-objective problems (e.g. [2, 5, 10, 17]), but research into multi-objective applications is still at a very early stage. MARL problems may be formalised using the Stochastic Game (SG) framework [3]. A SG is defined as a tuple $< S, A_{1...N}, T, R_{1...N} >$, where N is the number of agents, S is the set of states, A_i is the set of actions for agent i (and A is the joint action set), T is the transition function, and R_i is the reward function for agent i. In MARL, agents learn to maximise the return from R, and thus the design of R directly affects the joint policies learned. R may be augmented by an additional shaping reward F to provide additional feedback to the agents, thus improving learning speed and/or the final joint policy learned. Two typical MARL reward functions exist: **local rewards** (L_i) based on the utility of the part of a system that agent i can observe directly, and **global rewards** (G) based on the utility of the entire system. **Potential-Based Reward Shaping** ($PBRS$) is a form of reward shaping which has been proven to preserve the Nash equilibria of a SG [6]. In $PBRS$ the shaping term is $F(s, s') = \gamma\Phi(s') - \Phi(s)$, where $\Phi(s)$ is a potential function representing preferences for agents to reach certain system states. A **difference reward** (D_i) is a shaped reward signal that aims to quantify each agent's individual contribution to the system performance [20]:

$$D_i(s_i, a_i) = G(s, a) - G(s_{-i} \cup s_i^c, a_{-i} \cup a_i^c) \quad (1)$$

*An extended version of this paper is available [12].

where $G(s_{-i} \cup s_i^c, a_{-i} \cup a_i^c)$ is the counterfactual which represents the global utility for a theoretical system without the contribution of agent i. The terms s_{-i} and a_{-i} refer to all the states and actions not involving agent i, while s_i^c and a_i^c are fixed states and actions not dependent on agent i. Difference rewards have featured in many successful MARL applications (e.g. [7, 8, 11, 13, 14, 17, 19, 21]).

2. REWARD SHAPING THEORY

In a Multi-Objective Stochastic Game (MOSG) the reward function $\mathbf{R}$ returns a vector $\mathbf{r}$ consisting of the rewards for each individual objective $c \in C$. Empirical results have shown that both D [13, 14, 21] and $PBRS$ [9] can outperform agents learning using unshaped G in MOSGs, in terms of learning speed, average performance on system objectives, and quality of the non-dominated solutions found. It has been proven that applying $PBRS$ in multi-objective domains does not alter the true Pareto optimal policies [9]. $PBRS$ does affect the agents' exploration and may alter the Nash equilibrium reached by the agents; therefore different policies could be learned compared to agents learning without $PBRS$. Colby and Tumer [4] proved that the relative ordering of expected returns is not altered when agents are rewarded using D instead of G in a two-player single objective matrix game. We generalise this result to the case of a co-operative MOSG with $|C| \geq 1$ objectives and N agents.

THEOREM 1. *For any state $s \in S$ in a co-operative MOSG, the relative ordering of rewards is not altered when D is used in place of the system evaluation function.*

PROOF. For any state $s \in S$ in a co-operative MOSG, agents select a joint action a according to their joint policy π, and are rewarded for this state transition using the system evaluation function $\mathbf{G}$. If all agents except agent i follow a joint policy $\pi_{-i}^{\dagger} \in \Pi_{-i}$, and agent i follows a policy $\pi_i \in \Pi_i$, the resulting joint policy is $\pi_{-i}^{\dagger} \cup \pi_i$. Suppose that the reward for an objective $c \in C$ is greater if agent i follows policy $\pi_i^1 \in \Pi_i$ rather than $\pi_i^2 \in \Pi_i$ in state s when all other agents follow their respective policies from π_{-i}. Formally:

$$\mathbf{G}_c(s, a_{-i}^{\dagger} \cup a_i^1) > \mathbf{G}_c(s, a_{-i}^{\dagger} \cup a_i^2) \quad (2)$$

where $\mathbf{G}_c(s, a)$ is the return from the system evaluation function for objective c when joint action a is selected in system state s, $a_{-i}^{\dagger}$ are the actions selected in state s by all agents except agent i when following their policies from $\pi_{-i}^{\dagger}$, and a_i^1 and a_i^2 are the actions selected by agent i when following policy π_i^1 or π_i^2 respectively. If objectives are shaped

independently when using D, a counterfactual term must be calcuated for each objective c in order to apply Eqn. 1 to the global reward vector. As the counterfactual term $\mathbf{G}_c(s_{-i} \cup s_i^c, a^\dagger_{-i} \cup a_i^c)$ for any objective c does not depend on the policy being followed by agent i, for each possible system state s we can infer that the counterfactuals for agent i must be a fixed quantity. Thus we can add $-\mathbf{G}_c(s_{-i} \cup s_i^c, a_{-i} \cup a_i^c)$ to each side of Eqn. 2 while preserving the inequality:

$$\begin{aligned}\mathbf{G}_c(s, a^\dagger_{-i} \cup a_i^1) - \mathbf{G}_c(s_{-i} \cup s_i^c, a^\dagger_{-i} \cup a_i^c) > \\ \mathbf{G}_c(s, a^\dagger_{-i} \cup a_i^2) - \mathbf{G}_c(s_{-i} \cup s_i^c, a^\dagger_{-i} \cup a_i^c)\end{aligned} \tag{3}$$

And noting that the difference evaluation for objective c for agent i is: $\mathbf{D}_{c,i}(s_i, a_i) = \mathbf{G}_c(s, a) - \mathbf{G}_c(s_{-i} \cup s_i^c, a_{-i} \cup a_i^c)$

$$\begin{aligned}\forall c \in C, s \in S, i \in \{1, ..., N\} [\mathbf{D}_{c,i}(s_i, a_i^1) > \mathbf{D}_{c,i}(s_i, a_i^2) \\ \iff \mathbf{G}_c(s, a^\dagger_{-i} \cup a_i^1) > \mathbf{G}_c(s, a^\dagger_{-i} \cup a_i^2)]\end{aligned} \tag{4}$$

Therefore D does not alter the relative order of rewards for actions in any system state s, although it does alter the absolute values. Any property that relies on the ordering of rewards, and not the absolute value is therefore unaffected for each system state s. For example, if an action a_i in state s leads to a Nash equilibrium reward with respect to $\mathbf{G}$, it also leads to a Nash equilibrium reward with respect to $\mathbf{D}_i$. And if an action a_i in state s is Pareto optimal with respect to $\mathbf{G}$, it is also Pareto optimal with respect to $\mathbf{D}_i$. □

3. RESULTS & DISCUSSION

The Multi-Objective Beach Problem Domain (MOBPD) [12] is the first MOSG where the true Pareto optimal system utilities are known. It extends an earlier single-objective version [7], in a similar manner to Yliniemi and Tumer's multi-objective extension [21] to the El-Farol bar problem [1]. Each agent begins at a beach section $s \in S$, and then decides at which section they will spend their day. At each timestep an agent knows which beach section it is currently attending, and can choose to move to an adjacent section or to stay still. Agents must coordinate their actions to maximise two conflicting objectives: "capacity" and "mixture".

Each section has a capacity ψ, and the highest capacity reward for a section is received when the number of agents present is equal to ψ. Sections which are too crowded or too empty receive lower rewards as they are less desirable. Agents in the MOBPD are assigned one of two static types: m or f. The maximum mixture reward for a section is received when the number of m agents in attendance is equal to the number of f agents, while sections with an unequal mixture of agents receive a lower reward as they are less desirable. The local, global and difference rewards for capacity and mixture are calculated as per previous work [21]. Rewards for each objective are first normalised [15] before linear scalarisation [16] is applied with an even weighting.

We test multiple individual Q-learning [18] agents using credit assignment structures L, G, $G + PBRS$ and D in the MOBPD, as well as agents that randomly select actions. The $PBRS$ heuristics used are adapted from the work of Devlin et al. [7]. We set $\psi = 5$, $num\_agents\_M = 70$, $num\_agents\_F = 30$, $|S| = 5$, $num\_episodes = 10000$, $num\_timesteps = 1$, $\alpha = 0.1$, $\epsilon = 0.05$ with decay rate 0.9999 and $\gamma = 0.9$. Half of the m and f agents begin each episode at section 1, while the rest begin at section 3.

Table 1 lists the number of true Pareto optimal solutions found across all 50 statistical runs (PO Solns.), the average hypervolume of the non-dominated solutions found (Avg. HV), and the hypervolume of the best non-dominated solutions found across all runs (Best HV). Best HV gives an indication of how close an approach can get to finding the true Pareto front of the problem, while Avg. HV shows how consistent the performance of an approach is. D offered the best overall performance, sampling 16 of 19 Pareto optimal solutions and achieving hypervolumes very close to that of the true Pareto front. Fig. 1 shows that the best non-dominated solutions found by D and $G + PBRS(Mid)$ match closely with those of the true Pareto front. All the solutions found by L and G are dominated by those found by D and $G + PBRS(Mid)$; these typical MARL credit assignment structures are not informative enough to guide agents towards good solutions in the MOBPD.

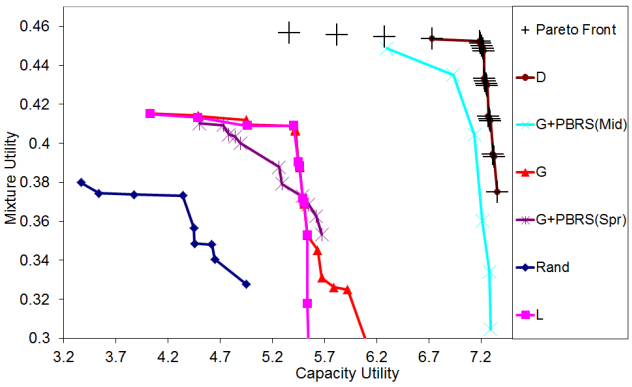

Figure 1: Best non-dominated episodes over all runs

Table 1: Experimental Results

	PO Solns.	Avg. HV	Best HV
True Pareto Front	19		3.347111
D	16	3.322784	3.329418
$G + PBRS(Mid)$	0	2.852388	3.238757
G	0	2.158390	2.474170
$G + PBRS(Spr)$	0	1.966866	2.300028
L	0	1.939231	2.338821
$Random$	0	1.609211	1.849685

The performances of D and $G+PBRS(Mid)$ demonstrate that well designed shaping techniques can guide agents towards true Pareto optimal solutions in MOSGs by making G more informative. Appropriate credit assignment is just as important in MOSGs as it is in traditional single-objective SGs, and this work demonstrated for the first time that agents learning using D can sample true Pareto optimal solutions in MOSGs. More sophisticated scalarisation approaches combined with D may allow further improvements in coverage along the true Pareto front in complex MOSGs.

Acknowledgements

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

REFERENCES

[1] W. B. Arthur. Inductive reasoning and bounded rationality. *The American economic review*, pages 406–411, 1994.

[2] T. Brys, T. T. Pham, and M. E. Taylor. Distributed learning and multi-objectivity in traffic light control. *Connection Science*, 26(1):65–83, 2014.

[3] L. Buşoniu, R. Babuška, and B. Schutter. Multi-agent reinforcement learning: An overview. In D. Srinivasan and L. Jain, editors, *Innovations in Multi-Agent Systems and Applications - 1*, volume 310 of *Studies in Computational Intelligence*, pages 183–221. Springer Berlin Heidelberg, 2010.

[4] M. Colby and K. Tumer. An evolutionary game theoretic analysis of difference evaluation functions. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 1391–1398. ACM, 2015.

[5] S. Devlin, M. Grzes, and D. Kudenko. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems*, 14(2):251–278, 2011.

[6] 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.

[7] S. Devlin, L. Yliniemi, D. Kudenko, and K. Tumer. Potential-based difference rewards for multiagent reinforcement learning. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 165–172, 2014.

[8] P. Mannion, S. Devlin, J. Duggan, and E. Howley. Avoiding the tragedy of the commons using reward shaping. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2016)*, May 2016.

[9] 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).

[10] 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.

[11] P. Mannion, J. Duggan, and E. Howley. Generating multi-agent potential functions using counterfactual estimates. In *Proceedings of Learning, Inference and Control of Multi-Agent Systems (at NIPS 2016)*, December 2016.

[12] P. Mannion, J. Duggan, and E. Howley. Analysing the effects of reward shaping in multi-objective stochastic games. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2017)*, May 2017 (in press).

[13] 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.

[14] P. Mannion, K. Mason, S. Devlin, J. Duggan, and E. Howley. Multi-objective dynamic dispatch optimisation using multi-agent reinforcement learning. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1345–1346, May 2016.

[15] R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369–395, 2004.

[16] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[17] K. Tumer and A. Agogino. Distributed agent-based air traffic flow management. In *Proceedings of the 6th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 330–337, Honolulu, HI, May 2007.

[18] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, 1989.

[19] D. H. Wolpert and K. Tumer. Collective intelligence, data routing and braess' paradox. *Journal of Artificial Intelligence Research*, pages 359–387, 2002.

[20] D. H. Wolpert, K. R. Wheeler, and K. Tumer. Collective intelligence for control of distributed dynamical systems. *EPL (Europhysics Letters)*, 49(6):708, 2000.

[21] L. Yliniemi and K. Tumer. Multi-objective multiagent credit assignment in reinforcement learning and nsga-ii. *Soft Computing*, pages 1–19, 2016.