Action-Dependent Optimality-Preserving Reward Shaping

Extended Abstract

Grant C. Forbes
North Carolina State University
Raleigh, United States
gforbes@ncsu.edu

Jianxun Wang
North Carolina State University
Raleigh, United States
jwang75@ncsu.edu

Leonardo Villalobos-Arias
North Carolina State University
Raleigh, United States
lvillal@ncsu.edu

Arnav Jhala
North Carolina State University
Raleigh, United States
ahjhala@ncsu.edu

David L. Roberts
North Carolina State University
Raleigh, United States
dlrober4@ncsu.edu

ABSTRACT

Recent RL research has utilized reward shaping–particularly complex shaping rewards such as intrinsic motivation (IM)–to encourage agent exploration in sparse-reward environments. While often effective, "reward hacking" can lead to the shaping reward being optimized at the expense of the extrinsic reward. Prior techniques have mitigated this, allowing for implementing IM without altering optimal policies, but have only thus far been tested in simple environments. In this work we show that they are effectively unsuitable for complex, exploration-heavy environments with long episodes. To remedy this, we introduce Action-Dependent Optimality Preserving Shaping (ADOPS), a method of converting arbitrary intrinsic rewards to an optimality-preserving form that allows agents to utilize them more effectively in the extremely sparse environment of Montezuma's Revenge. We demonstrate significant improvement over prior SOTA optimality-preserving IM-conversion methods, and argue that these improvements come from ADOPS's ability to preserve 'action-dependent' IM terms.

KEYWORDS

Reinforcement Learning; Reward Shaping; Intrinsic Motivation; Game-Playing Agents

ACM Reference Format:
Grant C. Forbes, Jianxun Wang, Leonardo Villalobos-Arias, Arnav Jhala, and David L. Roberts. 2025. Action-Dependent Optimality-Preserving Reward Shaping: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

There is growing interest in the Reinforcement Learning (RL) literature in using reward shaping to train agents in sparse-rewards environments that would otherwise be intractable [2, 10, 14]; specifically, interest in Intrinsic Motivation (IM): complex, non-Markovian reward functions used to encourage general exploration [4, 13].

It has been noted, however, that both traditional reward shaping [14] and IM [6] can be "hacked," with the agent learning to optimize the shaping reward at the expense of the actual reward. Prior methods of mitigating this reward-hacking exist—most notably Potential-Based Intrinsic Motivation (PBIM) [7, 8], Generalized Reward Matching (GRM) [9], and Policy-Invariant Explicit Shaping (PIES) [1], but each of these methods has only been tested in simple environments, and not in any environments wherein training with IM is itself essential in consistently obtaining any extrinsic rewards. Testing in Montezuma's Revenge, a benchmark environment that has been widely acknowledged [6, 11] to require IM or similar exploration-encouraging incentives, we find that all of these prior methods, while preserving the optimal policy set, detract from the agent's ability to learn to the extent that none of them can outperform an agent training on RND alone.

Motivated by this, we develop Action-Dependent Optimality-Preserving Shaping (ADOPS), a method for converting any arbitrary shaping reward (including IM) to a form that preserves optimality. ADOPS drops several key assumptions required for prior methods to ensure optimality, encompasses a provably-wider set of optimality-preserving functions, and empirically outperforms SOTA methods in a complex, sparse-reward environment.

2 THEORETICAL RESULTS

To preserve optimality, we want to ensure the optimal policy set is the same for both the original MDP M and the shaped MDP M', that is:

$$\operatorname*{argmax}_a Q_E^* = \operatorname*{argmax}_a (Q_E^* + Q_I^*) \quad \forall s, t, \pi^*, \tag{1}$$

where all variables are defined as in a standard MDP, and the subscripted E and I denote extrinsic and (possibly non-Markovian) intrinsic rewards, respectively.[1] If we now define $\bar{a}$ as any action not in $\operatorname{argmax}_a Q_E^*$, then Equation 1 becomes equivalent to

$$V_{IE}^{\pi_1^*}(s,t) = V_{IE}^{\pi_2^*}(s,t) \quad \forall s, t, \pi_1^*, \pi_2^* \tag{2}$$

$$Q_{IE}^*(s,\bar{a},t) < V_{IE}^*(s,t) \quad \forall s, \bar{a}, t, \pi^*. \tag{3}$$

Intuitively, the first of these conditions says that every action that would be optimal without IM must remain optimal after the

[1] Note the enumeration over all π^*: this is because, while every optimal policy by definition has identical Q_E^* and V_E^* to every other optimal policy, they may differ from each other intrinsically, resulting in different Q_I^* and V_I^*. Note also the direct t-dependence, as we're generally dealing with non-Markovian reward functions.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

addition of IM, while the second says that any suboptimal action must remain suboptimal with the addition of IM.[2]

In prior work in optimality-preserving reward shaping, particularly in Potential-Based Reward Shaping (PBRS), the optimal policy set is provably unchanged due to mathematical guarantees that Q_I^* is *independent* of the agent's actions at a given time step, and thus drops out of the argmax_a: see for example the $\Phi(s)$ of [12], or the Φ'_t of [7]. While this is a sufficient condition to ensure that optimality is preserved (as this term then drops out of the argmax_a), it is not a necessary one. It leaves out a theoretically interesting and empirically useful subset of optimality-preserving reward shaping functions: those whose cumulative intrinsic returns are allowed to depend on the agent's actions.

Inspired by the conditions in Equations 2 and 3, we first introduce an "ideal" reward-shaping conversion method that actively checks whether these conditions are satisfied, and if not, modifies the initial shaping reward just enough to ensure that they are. We prove the optimal policy remains unchanged by the addition of a reward $F' = F + F^2$, where F is some arbitrary initial IM, and F^2 is defined according to as

$$F^2 = \begin{cases} \min(0, V_E^* - Q_E^* + V_I^* - \gamma Q_{I,t+1}^* - F - \epsilon) & \text{if} \quad Q_E^* < V_E^* \\ \max(0, V_E^* - Q_E^* + V_I^* - \gamma Q_{I,t+1}^* - F) & \text{if} \quad Q_E^* \geq V_E^*. \end{cases} \tag{4}$$

Here, ϵ is an arbitrarily small positive constant, and V_I^* is defined as the *maximum* IM achievable while following an extrinsically optimal policy.

The first case of this equation can be intuitively thought of as checking to see if Equation 3 is violated, and if so adjusting the intrinsic reward downwards until it is not. Conversely, the second case checks to see if Equation 2 is violated, and adjusts the IM upwards if so until it is not.

While it would be ideal, it is usually not practically feasible to implement Equation (4), as it requires an accurate estimate of the optimal value function. Let's assume instead that we have access to some critic function that allows us to make approximations of V and Q of a given state-action pair, under the agent's current policy π (rather than a strictly optimal policy). Let us also assume that this critic handles the extrinsic and intrinsic rewards separately, such that we can deal with them independently (this is already a common practice, for example in [5], whose example we follow in Section 3). We then prove that an intrinsic reward of the form $F' = F + F^2$ with F^2 defined according to Equation 5

Given some initial shaping reward F, about which we make no assumptions, we define the ADOPS reward to be

$$F^2 = \begin{cases} \min(0, V_E^\pi - Q_E^\pi + V_I^\pi - \gamma Q_{I,t+1}^\pi - F - \epsilon) & \text{if} \quad Q_E^\pi < V_E^\pi \\ \max(0, V_E^\pi - Q_E^\pi + V_I^\pi - \gamma Q_{I,t+1}^\pi - F) & \text{if} \quad Q_E^\pi \geq V_E^\pi \end{cases} \tag{5}$$

will leave the set of optimal policies unchanged.

3 EMPIRICAL RESULTS

We test ADOPS, as well as prior optimality-preserving reward shaping methods in Montezuma's Revenge [3] with RND IM [6]. We find that all prior methods fail to converge to a policy that outperforms the policy trained on the baseline IM, while all versions of ADOPS achieve higher performance than the baseline policy. We plot our results in Figure 1.

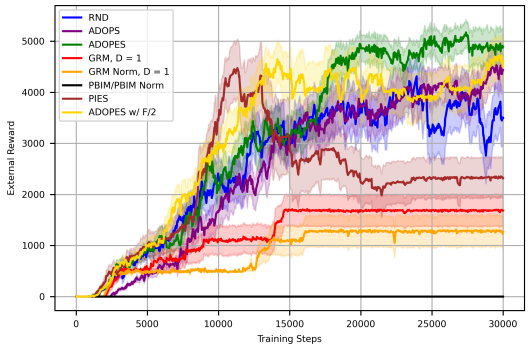

Figure 1: Comparison of all methods. Error bars are errors on the mean. Differences between RND and GRM are statistically significant, with $p = 0.009$ and $p = 0.031$ for norm and no-norm. ADOPES statistically outperforms RND with $p = 0.038$. ADOPS, ADOPES, and ADOPES w/ $F/2$ all statistically outperform PIES, with $p = 4.4e-5$, $p = 2.4e-6$, and $p = 6.4e-5$, respectively. They similarly improve over GRM and PBIM. $N = 10$ for GRM runs, $N = 1$ for PBIM runs, and $N = 20$ for all others.

We find that PBIM with RND, whether normalized or not, fails to ever obtain nonzero extrinsic rewards in Montezuma's Revenge. We find that this is due to the exponential nature of the denominator in the final reward for an episode under PBIM, combined with this environments' long episode lengths, exploding the reward and thus saturating the agent's action probability. Other GRM methods fared somewhat better,[3] but failed to reach the same average cumulative extrinsic reward as RND. We tested PIES as described in [1] with a decay rate such that it begins theoretically conserving the optimal policy at exactly the halfway point in training, as this is the point wherein it returns an IM of zero. While PIES performs well initially, it decreases in performance rapidly upon approaching this halfway point, and never recovers: in other words, its performance worsens as soon as PIES begins to approach conserving the optimal policy. PIES thus can either conserve the optimal policy or benefit from IM, but not both simultaneously.

We test three versions of our method. The first simply implements Equation (5) using the preexisting network critics' estimations of the relevant quantities. Noting that these critics' estimations are much better later on in training than earlier, we also implement Action Dependent Optimality Preserving Explicit Shaping (ADOPES), which uses a PIES-like t-dependence to phase in F^2 as the critics become gradually more accurate. We also test a version of ADOPES with the starting IM coefficient halved, in order to approximately harness the early benefits of PIES's incidental scan to a better hyperparameter for this coefficient, while preserving these gains later in training. All three of our methods consistently outperform prior optimality-preserving work.

[2] Implicit in the step from Equation 1 to Equation 2 is the fact that $Q_E^*(a^*) = V_E^* \forall a^* \pi^*$.

[3] PBIM is equivalent to a long-time-horizon version of GRM, as noted in [9].

REFERENCES

[1] Paniz Behboudian, Yash Satsangi, Matthew E Taylor, Anna Harutyunyan, and Michael Bowling. 2022. Policy invariant explicit shaping: an efficient alternative to reward shaping. *Neural Computing and Applications* (2022), 1–14.

[2] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems* 29 (2016).

[3] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47 (2013), 253–279.

[4] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355* (2018).

[5] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* (2018).

[6] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by random network distillation. In *7th International Conference on Learning Representations (ICLR 2019)*. 1–17. https://iclr.cc/ Seventh International Conference on Learning Representations, ICLR 2019 ; Conference date: 06-05-2019 Through 09-05-2019.

[7] Grant C Forbes, Nitish Gupta, Leonardo Villalobos-Arias, Colin M Potts, Arnav Jhala, and David L Roberts. 2024. Potential-Based Reward Shaping for Intrinsic Motivation. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 589–597.

[8] Grant C. Forbes and David L. Roberts. 2024. Potential-Based Reward Shaping For Intrinsic Motivation (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

[9] Grant C Forbes, Leonardo Villalobos-Arias, Jianxun Wang, Arnav Jhala, and David L Roberts. 2024. Potential-Based Intrinsic Motivation: Preserving Optimality With Complex, Non-Markovian Shaping Rewards. *arxiv preprint* (2024).

[10] Maja J Mataric. 1994. Reward functions for accelerated learning. In *Machine learning proceedings 1994*. Elsevier, 181–189.

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[12] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, Vol. 99. 278–287.

[13] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR, 2778–2787.

[14] Jette Randløv and Preben Alstrøm. 1998. Learning to Drive a Bicycle Using Reinforcement Learning and Shaping.. In *ICML*, Vol. 98. Citeseer, 463–471.