Potential-Based Reward Shaping for Intrinsic Motivation (Student Abstract)

Grant C. Forbes, David L. Roberts

North Carolina State University
Raleigh, North Carolina 27606 USA
gforbes@ncsu.edu, dlrober4@ncsu.edu

Abstract

Recently there has been a proliferation of intrinsic motivation (IM) reward shaping methods to learn in complex and sparse-reward environments. These methods can often inadvertently change the set of optimal policies in an environment, leading to suboptimal behavior. Previous work on mitigating the risks of reward shaping, particularly through potential-based reward shaping (PBRS), has not been applicable to many IM methods, as they are often complex, trainable functions themselves, and therefore dependent on a wider set of variables than the traditional reward functions that PBRS was developed for. We present an extension to PBRS that we show preserves the set of optimal policies under a more general set of functions than has been previously demonstrated. We also present *Potential-Based Intrinsic Motivation* (PBIM), a method for converting IM rewards into a potential-based form that are useable without altering the set of optimal policies. Testing in the MiniGrid DoorKey environment, we demonstrate that PBIM successfully prevents the agent from converging to a suboptimal policy and can speed up training.

Introduction

An increasing amount of work in reinforcement learning (RL) uses intrinsic reward functions, in addition to environmental rewards, to speed convergence to reasonable policies. This approach is particularly widespread in sparse-reward problems, or those that are exploration-heavy, and has had much success in these domains (Burda et al. 2018b).

However, adding a secondary reward term may lead to changes in the set of optimal policies, with unintended, and potentially adverse, consequences. For example, Burda et al. (2018a) show that an intrinsic reward that incentivizes visiting areas of the state space where the agent is less able to predict what will happen can result in the agent becoming "addicted" to watching a screen with flashing images.

We extend the potential-based, policy-preserving reward shaping term of Ng, Harada, and Russell (1999) to arbitrary rewards, and show this preserves optimality. We:

1. **Extend potential based reward shaping (PBRS) to functions of arbitrary variables in episodic environments**. We derive a boundary condition that is a sufficient condition for preserving optimality, and extend PBRS to reward functions, like intrinsic motivation (IM), which are more general functions than previous methods.

2. **Develop a method to convert arbitrary reward functions into this extended potential form**, maintaining the benefits of that function while mitigating its drawbacks, by guaranteeing that such a shaping reward will not alter the set of optimal policies in the underlying environment, and thus cannot be "hacked" by an optimal agent.

3. **Empirically demonstrate that our method is effective** as both a safety measure to prevent hacking of intrinsic rewards and, in certain cases, at speeding up training.

Theoretical Results

We take the MDP $M = (S, A, T, \gamma, R)$, defined as is standard. Working in an episodic environment with N time steps, and given shaping reward $F_t = \gamma\Phi_{t+1} - \Phi_t$, where Φ_t is an arbitrary potential, we derive the condition

$$\mathbb{E}_{a\sim\pi, s\sim T}\left(\gamma^{N-t}\Phi_N - \Phi_t\right) = \Phi'_t \qquad \forall t \in (0,1,...N-1), \tag{1}$$

where Φ'_t is some function that is constant with respect to action a_t. From here, we can prove Theorem 1:

Theorem 1 (Sufficient Condition For Optimality)**.** *The addition of a shaping reward $F_t = \gamma\Phi_{t+1} - \Phi_t$ leaves the set of optimal policies unchanged if Equation 1 holds.*

This is the most generalized condition for PBRS, compared with prior work (Ng, Harada, and Russell 1999; Wiewiora, Cottrell, and Elkan 2003; Devlin and Kudenko 2012; Grzes 2017). We also present a method for converting most IM rewards to a form that provably satisfies Equation 1, and thus preserves optimality, mitigating issues that can arise from IM. More formally, given the assumption

Assumption 1. *F_t is constant with respect to $a_{t'>t} \forall t, t' \in (0, 1, ...N-1)$*

about a shaping reward F_t (which applies to the vast majority of IM in prior literature), we define a new shaping reward

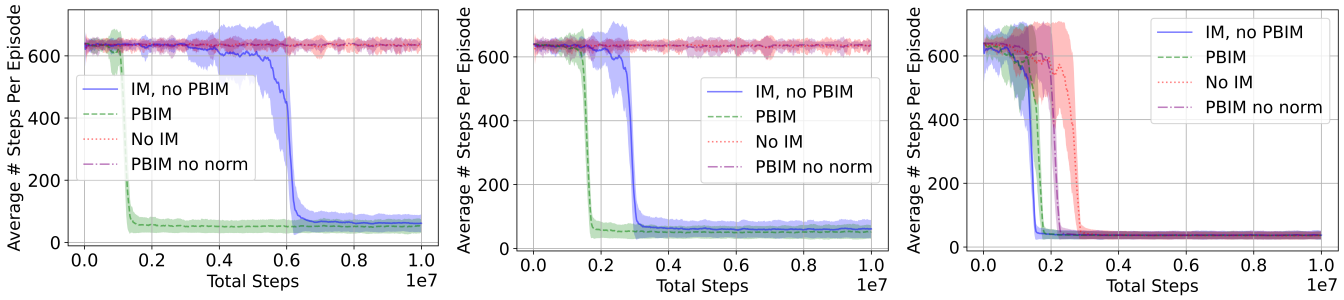
$$F'_t = \begin{cases} \sum_{n=0}^{N-1} -\gamma^{n-N} F_n, & \text{if } t{=}N \\ F_t, & \text{if } t{\neq}N, \end{cases} \tag{2}$$

and, using Theorem 1, derive Theorem 2:

Theorem 2 (PBIM Preserves Optimality)**.** *The addition of a shaping reward F'_t of the form in Equation 2 leaves the set of optimal policies unchanged if Assumption 1 holds.*

	$\alpha = 0.005, \gamma = 0.99$			$\alpha = 0.02, \gamma = 0.995$			$\alpha = 0.25, \gamma = 0.995$		
	T	$\bar{N}$	$\bar{\sigma}$	T	$\bar{N}$	$\bar{\sigma}$	T	$\bar{N}$	$\bar{\sigma}$
PBIM	1.67E6	36.4	12.5	1.67E6	51.8	22.4	1.26E6	51.2	22.7
IM, NO PBIM	1.46E6	37.3	13.1	2.88E6	60.5	27.5	6.07E6	62.0	27.9
PBIM NO NORM	2.29E6	37.1	13.4	N/A	635.9	14.9	N/A	634.7	18.8
NO IM	2.95E6	35.9	12.0	N/A	634.8	18.55	N/A	634.8	18.6

Table 1: Time to convergence (T), mean steps per episode after convergence ($\bar{N}$), and average standard deviation of steps per episode after convergence ($\bar{\sigma}$) for three parameter settings.



(a) $\alpha = .025, \gamma = .995$ (b) $\alpha = .02, \gamma = .995$ (c) $\alpha = .005, \gamma = .99$

Figure 1: Frames per episode 100-point moving average (lower is better). For IM + PBRS, IM no PBRS (a) $T = 36.5, p < 0.01$. For IM + PBRS, IM no PBRS (b) $T = 27.4, p < 0.01$. In (c), No IM converges lower than IM + PBRS, which converges lower than IM + PBRS no norm, which converges lower than IM no PBRS. Respectively, for each of these pairings, $T = 4.3, p < 0.01$, $T = 6.1, p < 0.01$, and $T = 1.8, p = 0.32$. The last isn't significant, but IM + PBRS and IM no PBRS is ($T = 7.9, p < 0.01$).

We also show that optimality is preserved with the addition of a mean-adjusted version of Equation 2, defined as

$$F'_t = \begin{cases} \sum_{n=0}^{N-1} -\gamma^{n-N} F'_n, & \text{if } t{=}N \\ F_t - \bar{F}, & \text{if } t{\neq}N, \end{cases} \tag{3}$$

where $\bar{F}$ is the expectation value of F across prior training.

Empirical Validation

We empirically validate our method in Minigrid Doorkey 8x8 with a tabular exploration reward $F_t = \frac{\alpha}{n(s)}$, where $n(s)$ is the number of times a state has been visited in the episode and α is a hyperparameter. Table 1 contains the time to convergence T (if converged), mean episode length $\bar{N}$, and standard deviation σ after convergence for each set of hyperparameters. Episode lengths are depicted in Figures 1c, 1b, & 1a, respectively. Shaded regions represent standard deviation among 16 trials. All differences were statistically significant, with one exception (Figure 1 caption).

Our method consistently outperformed the baseline IM method. Note also that our modification of Equation 2 into Equation 3 was key in allowing for convergence in the more difficult environments of Figures 1a & 1b, and in statistically significantly outperforming the IM baseline (Figure 1c).

The only experiment in which our (normalized) method did not perform best in both speed of convergence and final policy was with $\alpha = 0.005, \gamma = 0.99$. Note, though, that our method converged to a policy more efficient than that of IM to a significant degree, and converged more quickly than the run with no IM. In this worst-case scenario our method still provides value by facilitating a trade-off between efficiency in training speed and preventing reward hacking.

References

Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; and Efros, A. A. 2018a. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2018b. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

Devlin, S. M.; and Kudenko, D. 2012. Dynamic potential-based reward shaping. In *Proc. of the 11th Int. Conf. on Autonomous Agents and Multiagent Systems*, 433–440.

Grzes, M. 2017. Reward shaping in episodic reinforcement learning. In *Sixteenth International Conference on Autonomous Agents and Multiagent Systems*.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, 278–287.

Wiewiora, E.; Cottrell, G. W.; and Elkan, C. 2003. Principled methods for advising reinforcement learning agents. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 792–799.