# Improving the Effectiveness of Potential-Based Reward Shaping in Reinforcement Learning

Extended Abstract

Henrik Müller
L3S Research Center
Hannover, Germany
hmueller@l3s.de

Daniel Kudenko
L3S Research Center
Hannover, Germany
kudenko@l3s.de

## ABSTRACT

Potential-based reward shaping is often used to incorporate prior knowledge of how to solve the task into reinforcement learning because it can formally guarantee policy invariance. In this work, we highlight the dependence of effective potential-based reward shaping on the initial Q-values and external rewards, which determine the agent's ability to exploit the shaping rewards to guide its exploration and achieve increased sample efficiency. We formally derive how a simple linear shift of the potential function can be used to improve the effectiveness of reward shaping without changing the structure of the potential function and thus its implicitly encoded preferences, and without having to adjust the initial Q-values. We verify our theoretical findings on tabular Q-learning and demonstrate the application of our findings in deep reinforcement learning.

## KEYWORDS

Reinforcement Learning; Reward Shaping; Potential-Based Reward Shaping

**ACM Reference Format:**
Henrik Müller and Daniel Kudenko. 2025. Improving the Effectiveness of Potential-Based Reward Shaping in Reinforcement Learning: 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

Reward shaping is a common approach to accelerate the convergence of reinforcement learning agents by incorporating external guidance into the reward function, thereby improving the exploration of the environment.

In this work, we focus on potential-based reward shaping [9]. The primary appeal of potential-based reward shaping is the guarantee of policy invariance. Despite the change in rewards following the reward shaping, the optimal policy given the shaped reward function remains identical to that of the original MDP. Potential-based reward shaping utilizes a potential function to assign a heuristic value of *goodness* (or potential) to each state, with the reward shaping subsequently derived from the difference between the potential of the states before and after the execution of an action.



*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).

Previous theoretical evaluations of potential-based reward shaping have given pointers of how to structure an effective potential function [2, 3], but have not included the intrinsic link between the reward and the Q-value initialization. While potential-based reward shaping has been shown to be equivalent to shifting the Q-value initialization by adding the potential function [10], previous work has not addressed the impact of the Q-value initialization on the sample efficiency in potential-based reward shaping and how to optimize a potential function for a given Q-value initialization to improve sample efficiency.

Notably, potential-based reward shaping does not alter the optimal policy, and our method does not change the preferences encoded in the potential function. Consequently, our approach is applicable in any situation where additional (possibly approximate) knowledge of the MDP can be exploited for more sample-efficient reinforcement learning.

In summary, the primary contributions of this paper are:

(1) We introduce a generalized framework of requirements for an effective potential-based reward shaping.
(2) We explore how to choose the scale and offset for potential functions to adapt the potential function to a given Q-value initialization and reward function, thus improving the effectiveness of reward shaping.
(3) We verify our findings empirically first on tabular RL, and then extend our experiments to the deep RL setting.

## 2 EFFECTIVE POTENTIAL-BASED REWARD SHAPING

Potential-based reward shaping (PBRS) is defined by its potential function $\Phi(s)$ mapping each state to a heuristic scalar value. Given the potential function, the reward shaping function $F$ is defined as:

$$F(s, a, s') = \gamma\Phi(s') - \Phi(s) \tag{1}$$

where $s'$ is the state reached after executing action $a$ in $s$ and $\gamma$ is the discount factor. The shaped reward $R'$ can then be defined as:

$$R'(s, a, s') = R(s, a, s') + F(s, a, s') \tag{2}$$

We focus on sparse reward functions that offer little (intermediate) feedback to the exploration strategy of the agent, thus being inherently difficult to solve efficiently. We adopt the reward formulation of Matignon et al. [4] for goal-directed reward functions with a goal state $s_g$:

$$R(s, a, s') = \begin{cases} r_g & \text{if } s' = s_g \\ r_\infty & \text{otherwise} \end{cases} \tag{3}$$

The goal for an effective application of PBRS is for the agent to initially exploit the potential function to guide its exploration at the start of the training. As such, our theoretical results will focus on the first updates of the initial estimates.

The following set of requirements for the relation of the shaped rewards to the initial Q-values is an extension of the requirements proposed in Grzes and Kudenko [3] to include the relation between TD-updates and initial Q-values focusing on non-terminal transitions between the states $s$ and $s'$ with $\Phi(s') > \Phi(s)$:

$$r_\infty + \gamma\Phi(s') - \Phi(s) > (1-\gamma)Q_{init} \tag{4}$$

$$r_\infty + \gamma\Phi(s) - \Phi(s') < (1-\gamma)Q_{init} \tag{5}$$

$$r_\infty + \gamma\Phi(s) - \Phi(s) \leq (1-\gamma)Q_{init} \tag{6}$$

Only transitions that lead to states with a higher potential value should be incentivized. Other transitions should be disincentivized initially. As a result, any of the commonly used advantage-based action selection schemes would repeat actions that lead to the largest possible next potential value while avoiding to repeatedly explore actions that lead to states with lower potential values. Extended versions of the proofs can be found in Müller and Kudenko [8].

## 2.1 Potential Scale

As a direct result of the requirement of potential values of zero in terminal states [2], we can obtain upper and lower bounds on the scale of the potential function in goal-directed MDPs:

$$r_\infty - (1-\gamma)Q_{init} < \Phi(s) < r_g - (1-\gamma)Q_{init} \tag{7}$$

As a direct result of these bounds, it is not possible to utilize the scale of the potential function to compensate the mismatch between the original rewards and the initial Q-values, and thus satisfying the general requirements outlined in equations 4-6.

Notably, equation 7 implicitly requires that $r_g > r_\infty$. Accordingly, the incentive to terminate in a goal state must originate from the reward of a goal-directed MDP not from the reward shaping.

## 2.2 Potential Shift

The equations 4 to 6 show, that the effectiveness of PBRS depends on the external reward and the initial Q-values. If the initial Q-values and the (constant) rewards are known, one can add a simple bias to the potential function to remove the dependence. We define the shifted potential function for any non-terminal state $s$ as:

$$\Phi_b(s) = \Phi(s) + \frac{b}{\gamma - 1} \tag{8}$$

This constant bias term shifts all rewards (except when moving into terminal states) by $b$. If we set $b = (1-\gamma)Q_{init} - r_\infty$, we are able to remove the dependence on both the external reward and the initial Q-values. This allows us to make direct use of the prior results on how to create an effective PBRS [3, 7].

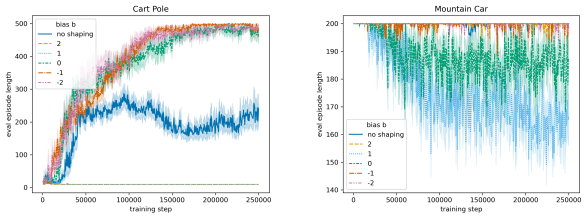
This shift of the potential values has to exclude the potential values for terminal states as the potential of any terminating state has to be zero [2]. For transitions into terminal states the additional term in the shaped reward is $Q_{init} + \frac{r_\infty}{\gamma-1}$. The benefit of removing $r_\infty$ and $Q_{init}$ in the requirements for non-terminal transitions therefore can come at the cost of incorrectly (dis-)incentivizing transitions into terminal states.

# 3 EXPERIMENTS

Our experiments show that our theoretical results hold exactly for tabular Q-learning in a simple Gridworld environment with both a goal-directed reward function (only non-zero reward at goal) and a on-step reward function (same negative reward per step). The optimal bias value leads to rapid convergence while incorrect bias values lead to non-convergence within the training budget.

We further extend our results to deep RL exemplarily using DQN [5]. In the Cart Pole environment [1], experiments tested the impact of bias selection when using function approximation in deep RL using the pole's angle as potential function. Results showed that biases $b \geq 0$ allowed the agent to leverage the potential function effectively, resulting in faster convergence to a better policy compared to training without reward shaping. Bias values $b < 0$ led to deteriorating performance, with evaluation scores remaining near zero. In the Mountain Car environment [6], experiments tested potential-based reward shaping using the car's absolute velocity as the potential function. Results showed that only biases of 0 and 1 reliably led to solving the task, with a bias of 1 producing the best-performing policies by effectively compensating for the constant negative on-step rewards.

Importantly, while the choice of exact bias values mattered less in the function approximation experiments, the results confirm the applicability of the bias-shifting theory in deep RL. In contrast to the experiments on the tabular Gridworld environment, the incorrect reward incentives when transitioning into terminal states for uniform reward functions do not lead to increasing variance between runs as highlighted in Figure 1.




(a) Cart Pole (b) Mountain Car

**Figure 1: Results of Cart Pole and Mountain Car for different values of the bias $b$ plotting the mean episode length and standard error of the mean over ten independent runs.**

# 4 CONCLUSION

We have introduced a framework of requirements for effective potential-based reward shaping and have shown that a constant shift of the potential function can help alleviate problems caused by a mismatch between the original reward and initial Q-values. We empirically verified that this approach also holds for function approximation in deep RL.

## ACKNOWLEDGMENTS

This work was supported by the Lower Saxony Ministry of Science and Culture (MWK), in the zukunft.niedersachsen program of the Volkswagen Foundation (HybrInt).

## REFERENCES


[1] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13, 5 (1983), 834–846. https://doi.org/10.1109/TSMC.1983.6313077

[2] Marek Grześ. 2017. Reward Shaping in Episodic Reinforcement Learning. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (São Paulo, Brazil) *(AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 565–573.

[3] Marek Grzes and Daniel Kudenko. 2009. Theoretical and Empirical Analysis of Reward Shaping in Reinforcement Learning. In *2009 International Conference on Machine Learning and Applications*. 337–344. https://doi.org/10.1109/ICMLA.2009.33

[4] Laëtitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. 2006. Reward Function and Initial Values: Better Choices for Accelerated Goal-Directed Reinforcement Learning. In *International Conference on Artificial Neural Networks*. https://api.semanticscholar.org/CorpusID:3448745

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR* abs/1312.5602 (2013). arXiv:1312.5602 http://arxiv.org/abs/1312.5602

[6] Andrew William Moore. 1990. *Efficient Memory-based Learning for Robot Control*. Technical Report. University of Cambridge.

[7] Henrik Müller, Lukas Berg, and Daniel Kudenko. 2025. Using incomplete and incorrect plans to shape reinforcement learning in long-sequence sparse-reward tasks. *Neural Computing and Applications* (10 Jan 2025). https://doi.org/10.1007/s00521-024-10615-2

[8] Henrik Müller and Daniel Kudenko. 2025. Improving the Effectiveness of Potential-Based Reward Shaping in Reinforcement Learning. arXiv:2502.01307 [cs.LG] https://arxiv.org/abs/2502.01307

[9] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 278–287.

[10] Eric Wiewiora. 2003. Potential-Based Shaping and Q-Value Initialization are Equivalent. *J. Artif. Intell. Res.* 19 (2003), 205–208. https://doi.org/10.1613/JAIR.1190