

# Multi-Scale Reward Shaping via an Off-Policy Ensemble

## (Extended Abstract)

Anna Harutyunyan  
Vrije Universiteit Brussel  
aharutyu@vub.ac.be

Tim Brys  
Vrije Universiteit Brussel  
timbrys@vub.ac.be

Peter Vrancx  
Vrije Universiteit Brussel  
pvrancx@vub.ac.be

Ann Nowé  
Vrije Universiteit Brussel  
anowe@vub.ac.be

### ABSTRACT

We propose a potential-based reward shaping architecture that is able to reduce learning speed, with no prior tuning and extra environment samples required, via considering an off-policy ensemble of value functions learning on a variety of heuristics with a variety of scales.

### Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning

### Keywords

reinforcement learning, potential-based reward shaping, horde

## 1. INTRODUCTION

We are interested in methods that are capable of aiding reinforcement learning (RL) [9] with as little extra maintenance as possible. Potential-based reward shaping (PBRS) is a simple framework for integrating domain knowledge into RL, particularly attractive for its policy invariance guarantees [8]. The efficacy of PBRS in reducing learning speed, while repeatedly demonstrated in practice [3], is conditioned on precise knowledge of both quality *heuristics* and their *magnitudes*, which together define the *potential function*. Recent literature in both active [1, 2] and latent [4] settings has argued and demonstrated the benefits of maintaining *ensembles* of policies shaped with simple-heuristic-based potentials, rather than limiting to a single (but complex) one. In this work we take this intuition further, to remove the second requirement of knowing correct value magnitude for the potentials,<sup>1</sup> which is typically found via behind-the-scenes tuning. The assumption of an ability to do so is unrealistic, and defeats the purpose of a method intended to reduce learning speed. By removing this assumption, we achieve a PBRS architecture, that reduces learning speed at no extra sample cost. Together with previous work [1, 4,

<sup>1</sup>Brys et al. [2] address the issue of *relative* scalings within an ensemble, while our focus is the unknown absolute scale for each heuristic.

**Appears in:** *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AA-MAS 2015)*, Bordini, Elkind, Weiss, Yolum (eds.), May, 4–8, 2015, Istanbul, Turkey.

Copyright © 2015, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

2], this allows the designer to benefit from a handful of simple heuristics, with no requirements on their quality, and no additional tuning steps introduced, making the architecture practical to use out of the box.

## 2. APPROACH

We assume the usual RL framework [9]. PBRS [8] augments the reward function  $R$  with an additional reward  $F = \gamma\Phi' - \Phi$ , where  $\Phi$  is the potential function over the state(-action) space. We assume an off-policy *latent* learning setup, and maintain our Horde [10] of shapings as a set  $\mathcal{D}$  of Greedy-GQ( $\lambda$ )-learners [6]. Given a set of potential functions  $\Phi = \{\Phi_1, \dots, \Phi_\ell\}$ , a range of scaling factors  $\mathbf{c}^i = \langle c_1^i, \dots, c_{k_i}^i \rangle$  for each  $\Phi_i$ , and the base reward function  $R$ , the ensemble reward function is a vector:

$$\mathbf{R} = R + \langle F_{c_1^1}, F_{c_2^1}, \dots, F_{c_{k_\ell}^\ell} \rangle \quad (1)$$

where  $F_{c_j^i}^{\Phi_i}$  (or simply  $F_j^i$ ) is the potential-based shaping reward w.r.t. the potential function  $\Phi_i$ , scaled with the factor  $c_j^i$ . Adopting the terminology of Sutton et al. [10], we refer to individual agents within Horde as *demons*. Each demon  $d_j^i$  learns a greedy policy  $\pi_j^i$  w.r.t. its reward  $R + F_j^i$ . Our latent setting implies a fixed behavior policy  $\pi_b$ , with all  $\pi_j^i$  learning in parallel from the experience generated by  $\pi_b$ . Because each policy  $\pi_j^i$  is available separately at each step, an *ensemble* policy  $\pi_E$  can be devised by collecting votes on action preferences from the demons  $d_j^i$ , or any other suitable ensemble technique [2].

## 3. EXPERIMENTS

We evaluate<sup>2</sup> our approach in two common benchmark problems: mountain car [9] and cart-pole [7]. We empirically show that an indiscriminate ensemble of simple heuristics on general scaling ranges performs as well as one with cherry-picked components. The behavior  $\pi_b$  is a uniform distribution over all actions at each time step. Evaluation is done by interrupting learning every  $z$  episodes and executing the queried greedy policy  $\pi_j^i$  or ensemble policy  $\pi_E$  once. We report our results w.r.t. *rank* voting [11].

<sup>2</sup>For experiment details, see the full version of this paper [5].

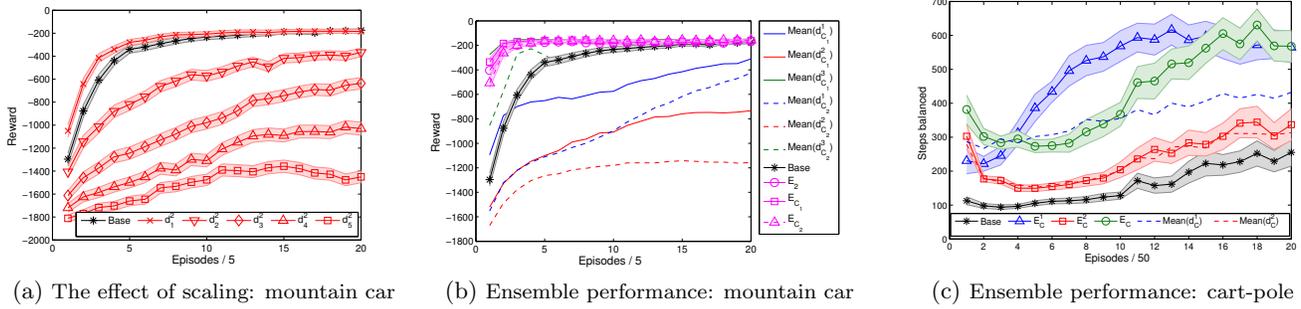


Figure 1: (a) Each curve corresponds to the performance of a demon shaped with  $\Phi_2$ , with a scaling factor from the range  $C_1$ . (b),(c) The solid and dashed lines denote the *mean* performance of the demons w.r.t. a single shaping on a scale range, serving as reference for the performance of the ensemble components. Note that there is no single demon with this performance.

### Mountain Car

We define 3 shaping potentials, corresponding to the position ( $\Phi_1$ ), height ( $\Phi_2$ ), and speed ( $\Phi_3$ ) of the car. We consider two scaling ranges  $C_1 = \langle 20, 40, 60, 80, 100 \rangle$  and  $C_2 = \langle 1, 10, 10^2, 10^3, 10^4 \rangle$ , with the first being a reasonably close range to the optimal scales  $c_1, c_2, c_3$ , and the second being a general sweep, with no intuition or knowledge of the optimal scale. Fig. 1(a) presents a comparison of the performance of  $\Phi_2$  over the (reasonable) scaling range  $C_1$ , illustrating the dramatic effect small differences in scale can have on a shaping’s performance. Now let  $E_{C_1}$  and  $E_{C_2}$  be the ensembles w.r.t. all three shapings on  $C_1$  and  $C_2$ , resp., each totaling in 16 demons (including the base learner), and let  $E$  be the ensemble w.r.t. the three shapings on tuned scalings  $c_1, c_2, c_3$ .  $E_{C_1}$  and  $E_{C_2}$  are both statistically the same ( $p > 0.05$ ) as the *tuned* ensemble  $E$ , despite their components having a much wider range of performance (Fig. 1(b)).

### Cart-Pole

We define 2 shaping potentials, corresponding to the angle ( $\Phi_1$ ) and angular speed ( $\Phi_2$ ) of the pole. We consider a general scaling range  $C = \langle 1, 10, 10^2, 10^3, 10^4 \rangle$ , and three ensembles:  $E_C^1$  resp.  $E_C^2$  only comprised of the demons shaped w.r.t.  $\Phi_1$  resp.  $\Phi_2$  across  $C$  (5 demons each), and  $E_C$  containing all 11 demons (including the base learner). All ensembles improve over the base learner (Fig. 1(c)). The performance of  $E_C^2$  matches that of its average, as all of its components perform similarly, while  $E_C^1$  does much better than the corresponding average. The global ensemble  $E_C$  correctly identifies both *which shaping* to follow: its performance lies between the average of  $\Phi_1$  across  $C$  and  $E_C^1$ , always outperforming  $\Phi_2$ , and on *what scales*: its final performance matches that of  $E_C^1$ , significantly improving over the average of  $\Phi_1$  across  $C$ .

## 4. CLOSING REMARKS

We described a PBRs architecture that, through the use of an ensemble, can speed up learning by leveraging information from just a handful of imperfect heuristics, with no prior tuning required. In realistic settings, where little information is available a priori and environment samples are costly, this is the first practical reward shaping method, readily usable off-the-shelf. Note that the added computational expense is only linear in the number of non-zero

features: Horde has been demonstrated to be able to learn thousands of policies in real time [10].

## 5. ACKNOWLEDGMENTS

Anna Harutyunyan is supported by the IWT-SBO project MIRAD (grant nr. 120057). Tim Brys is funded by a Ph.D grant of the Research Foundation-Flanders (FWO).

## 6. REFERENCES

- [1] T. Brys, A. Harutyunyan, P. Vrancx, M. E. Taylor, D. Kudenko, and A. Nowé. Multi-objectivization of reinforcement learning problems by reward shaping. In *Proc. of IEEE IJCNN*, 2014.
- [2] T. Brys, A. Nowé, D. Kudenko, and M. E. Taylor. Combining multiple correlated reward and shaping signals by measuring confidence. In *Proc. of AAAI*, 2014.
- [3] S. Devlin, D. Kudenko, and M. Grzes. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems (ACS)*, 14(02):251–278, 2011.
- [4] A. Harutyunyan, T. Brys, P. Vrancx, and A. Nowé. Off-policy shaping ensembles in reinforcement learning. In *Proc. of ECAI*, pages 1021–1022, 2014.
- [5] A. Harutyunyan, T. Brys, P. Vrancx, and A. Nowé. Off-policy reward shaping with ensembles. Technical report, arXiv:1502.03248, 2015.
- [6] H. Maei and R. Sutton. GQ( $\lambda$ ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proc. of AGI*, 2010.
- [7] D. Michie and R. A. Chambers. Boxes: An experiment in adaptive control. In *Machine Intelligence*. 1968.
- [8] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *In Proc. of ICML*, 1999.
- [9] R. Sutton and A. Barto. *Reinforcement learning: An introduction*, volume 116. Cambridge Univ Press, 1998.
- [10] R. Sutton, J. Modayil, M. Delp, T. Degris, P. Pilarski, A. White, and D. Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proc. of AAMAS*, 2011.
- [11] M. Wiering and H. van Hasselt. Ensemble algorithms in reinforcement learning. *Systems, Man, and Cybernetics, Part B: Cybernetics*, 38(4):930–936, 2008.