# Off-Policy Shaping Ensembles in Reinforcement Learning

**Anna Harutyunyan** and **Tim Brys** and **Peter Vrancx** and **Ann Nowé**[1]

**Abstract.** In this work we propose learning an ensemble of policies related through potential-based shaping rewards via the off-policy Horde framework.

## 1 Introduction and background

Ensemble techniques are widespread in supervised learning, but their use in reinforcement learning (RL) [9] has been extremely sparse thus far. We seek to formulate a RL ensemble that is effective at improving *learning speed*, the bottleneck of RL. In the context of this goal, the ensemble needs to learn in parallel efficiently. Recently proposed Horde architecture [10] fills this bill, and is the first to also possess convergence guarantees in realistic setups. In contrast to the previous uses of Horde, we exploit its power for learning a *single* task faster. The policies in our ensemble are obtained through potential-based reward shaping (PBRS), each expressing different pieces of heuristic domain knowledge, and are then combined via a voting rule. Maintaining multiple shapings allows leveraging the strengths of different heuristics without having to design a complex shaping reward [2].

The scenario we consider is that of off-policy learning under fixed behavior, i.e. *latent learning*. Such is often the setup in applications where the environment samples are costly and a failure is highly penalized. To our knowledge, this is the first validation of PBRS effective in such a latent setting, where it does not actively guide exploration.

For omitted details in discussion and experiment setup, please see the full version of this paper [5]. For standard background on reinforcement learning, see Sutton and Barto [9]. We briefly give the ingredients of the described approach.

**Reward shaping** augments the true reward signal with an additional heuristic *shaping reward*, provided by the designer. Ng et al. [8] show that grounding the shaping rewards in *state potentials* is both necessary and sufficient for ensuring preservation of the (optimal) policies of the original task. PBRS maintains a potential function $\Phi : S \to \mathbb{R}$, and defines the auxiliary reward function $F$ as $F(s, a, s') = \gamma\Phi(s') - \Phi(s)$, where $\gamma$ is the usual discounting factor.

**Horde** Learning about a (*target*) policy that is different from the (*behavior*) one currently followed, is referred to as learning *off-policy*. Despite its versatility, off-policy learning suffers from convergence issues, when combined with function approximation (FA) [1]. This problem was recently addressed by the family of *gradient temporal-difference* methods, which became the cornerstone for Horde: a convergent scalable architecture for learning multiple value functions off-policy from a shared stream of experience [10].

## 2 Ensembles of Shapings

Most previous uses of ensembles of policies involved independent runs for each policy, with the combination happening post-factum [3]. This is limited in practical utility, since it requires a large computational and sample overhead, assumes a repeatable setup, and does not improve learning speed. Others, in general, lack convergence guarantees,[2] by either using mixed on- and off-policy learners [12], or Q-learners under FA [2]. In general, when considering policy ensembles, an off-policy learning setup seems inevitable; it is only useful if the policies reflect information different from the behavior, since the strength of ensemble learning lies in the diversity of information its components contribute [7]. Horde is the first framework that allows to soundly and efficiently learn multiple value functions off-policy in parallel in a realistic setup. For this reason, we believe it to be well-suited for ensemble learning in RL.

Diversity in the RL context can be expressed through several aspects, related to dimensions of the learning process: (1) diversity of *experience*, (2) diversity of *algorithms* and (3) diversity of *reward signals*. Diversity of experience naturally implies high sample complexity, and assumes either a multi-agent setup, or learning in stages. Diversity of algorithms is computationally costly, as it requires separate representations. In the context of our aim of improving learning speed, we focus on the latter aspect of diversity: diversity of reward signals. Recall that shaping rewards encode heuristics about the desirability of states. Prior to solving a task, it is typically much easier to think of many simple, albeit imperfect heuristics, than a complex but accurate cure-all. Combining these simple potentials beforehand naïvely (e.g. with linear scalarization) is uninformative, since they may counterweigh each other in some parts of the space, and "cancel out". Learning and maintaining *all* of them simultaneously, on the other hand, has been infeasible prior to Horde, given a desire to maintain general convergence guarantees and stay efficient. Having access to all shapings at each step opens up new opportunities for autonomous combination, e.g. with ensemble methods.

**Shaping off-policy** The effects of PBRS on the learning process are usually considered to lie in the guidance of exploration during learning [4, 8], while in our setting we assume no control over the agent's behavior. The performance benefits then can be explained by

[1] AI Lab, Vrije Universiteit Brussel, Belgium, email: {anna.harutyunyan, timbrys, pvrancx, anowe}@vub.ac.be

[2] See the discussion on convergence in Section 6.1.2 of van Hasselt's dissertation [11].

faster *knowledge propagation* through the temporal-difference updates, which we now observe decoupled from guidance of exploration. These effects of off-policy reward shaping may be of independent interest.

## 3 Experiments

We focus our attention to a classical benchmark domain of mountain car [9]. The task is to drive an underpowered car up a hill. The base reward is $-1$ for each step, except the final one. We define three intuitive shaping potentials:

**Position** Encourage progress to the right (the goal): $\Phi_1(\mathbf{x}) = c_r \times x$.
**Height** Encourage higher positions: $\Phi_2(\mathbf{x}) = c_h \times h$.
**Speed** Encourage higher speeds: $\Phi_3(\mathbf{x}) = c_s \times |\dot{x}|^2$.[3]

We let the behavior be uniform over all actions (discrete throttle of $-1$, 0, or 1), and deploy Horde to concurrently learn[4] the base task, and the three policies shaped w.r.t. the above potentials. We devise the *ensemble policy* via *rank voting* [12]. The evaluation is done by interrupting the learning every 5 episodes, and executing each greedy policy once. No learning was allowed during evaluation.

The speed shaping turns out to be the most helpful universally. If this is the case, one would prefer to just use that shaping on its own, but we assume no access to such information a priori. To make our experiment more interesting, we consider two distinct scenarios: with and without this shaping. We would like our ensemble to outperform both comparable shapings in the former, and at least match the performance of the best one, in the latter case.

**Table 1**: Average of 1000 independent runs of 100 episodes each. Initial/final labels refer to the first/last 20% of a run. The results that are not significantly different from the best ($p > 0.05$) are in bold.

| Variant | Cumulative | Initial | Final |
|---|---|---|---|
| | Without best shaping | | |
| No shaping | -336.3 ± 279.5 | -784.7 ± 385.9 | -185.1 ± 9.9 |
| Right shaping | -310.4 ± 96.9 | **-378.5 ± 217.4** | -290.3 ± 19.3 |
| Height shaping | -283.2 ± 205.2 | -594.2 ± 317.0 | **-182.3 ± 7.5** |
| Ensemble | **-211.2 ± 94.2** | **-330.6 ± 179.5** | **-180.2 ± 1.5** |
| | With best shaping | | |
| No shaping | -349.7 ± 285.2 | -818.6 ± 373.7 | -193.2 ± 10.9 |
| Right shaping | -303.4 ± 81.4 | -346.7 ± 181.2 | -295.1 ± 16.7 |
| Height shaping | -292.4 ± 213.8 | -619.8 ± 328.3 | -190.1 ± 5.3 |
| Speed shaping | **-158.6 ± 23.7** | **-182.1 ± 50.6** | **-150.2 ± 2.9** |
| Ensemble | **-168.7 ± 44.7** | **-214.8 ± 94.8** | -161.7 ± 4.0 |

The results in Table 1 show that individual shapings alone aid learning speed significantly, and the ensemble policy meets our desiderata: it either statistically matches or is better than the best shaping at any stage. The exception is the final performance in the second scenario, but the difference in the collected reward is still rather small.

## 4 Conclusions and future work

We believe the Horde architecture to be well-suited for ensemble learning in general, and, as it provides the necessary tools to learn many PBRS policies simultaneously at no added cost, a convenient framework for leveraging diverse heuristic knowledge. We demonstrated our method to be effective even on a simple task, and with an ad-hoc combination method. Larger problems with many locally good shapings are the target benchmark, and we expect them to yield larger benefits.

There are many directions for future work. Latent parallel learning of diverse value functions suggests exploring ways to *learn* good combination strategies, or the potential functions themselves. Naturally, such meta-learning has to happen at a much faster pace in order to be useful in speeding up the main learning process. The scalability of Horde allows for learning thousands of value functions efficiently. While it is rarely sensible to define thousands of distinct shapings, one could imagine maintaining many different *scaling factors* for the existing shaping potentials. This would not only mitigate the scaling problem, but make the representation more flexible by having non-static scaling factors throughout the state space.

The primary limitation of Horde is the requirement to keep the behavior policy fixed (or change it slowly). Relaxing this constraint is a topic of ongoing research. Horde tackles convergence, which is one of the two main theoretical challenges with off-policy learning under FA. The other has to do with the *quality* of solutions under off-policy sampling, which may, in general, fall far from optimum. Kolter gives a way of constraining the solution space, achieving stronger guarantees [6], but his algorithm is quadratic in complexity and not scalable.

[3] Here $\mathbf{x} = \langle x, \dot{x} \rangle$ is the state (position and velocity), $h$ is the height of the hill, and $\mathbf{c} = \langle c_r, c_h, c_s \rangle$ is a vector of tuned scaling constants.
[4] We used $\gamma = 0.99$, $\lambda = 0.4$, $\alpha = 0.1$, $\beta = 0.0001$, and approximated the state space via 10 tilings of $10 \times 10$.

## ACKNOWLEDGEMENTS

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). The authors thank the anonymous reviewers who helped to improve the paper.

## REFERENCES

[1] L. Baird, 'Residual algorithms: Reinforcement learning with function approximation', in *In Proc. 12th ICML*, pp. 30–37, (1995).
[2] T. Brys, A. Harutyunyan, P. Vrancx, M.E. Taylor, D. Kudenko, and A. Nowé, 'Multi-objectivization of reinforcement learning problems by reward shaping', in *IJCNN*, (2014).
[3] S. Fausser and F. Schwenker, 'Ensemble methods for reinforcement learning with function approximation.', in *MCS*, volume 6713 of *LNCS*, pp. 56–65. Springer, (2011).
[4] M. Grzes, *Improving Exploration in Reinforcement Learning through Domain Knowledge and Parameter Analysis*, Ph.D. dissertation, University of York, 2010.
[5] A. Harutyunyan, T. Brys, P. Vrancx, and A. Nowé, 'Off-policy shaping ensembles in reinforcement learning', Technical report, arXiv:1405.5358, (2014).
[6] J. Zico Kolter, 'The fixed points of off-policy td.', in *NIPS*, pp. 2169–2177, (2011).
[7] A. Krogh and J. Vedelsby, 'Neural network ensembles, cross validation, and active learning', in *NIPS*, pp. 231–238. MIT Press, (1995).
[8] A. Y. Ng, D. Harada, and S. Russell, 'Policy invariance under reward transformations: Theory and application to reward shaping', in *In Proc. 16th ICML*, pp. 278–287, (1999).
[9] R.S. Sutton and A.G. Barto, *Reinforcement learning: An introduction*, volume 116, Cambridge Univ Press, 1998.
[10] R.S. Sutton, J. Modayil, M. Delp, T. Degris, P.M. Pilarski, A. White, and D. Precup, 'Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction', in *In Proc. 10th AAMAS*, pp. 761–768, (2011).
[11] H. van Hasselt, *Insights in reinforcement learning : formal analysis and empirical evaluation of temporal-difference learning algorithms*, Ph.D. dissertation, Utrecht University, 2011.
[12] M.A. Wiering and H. van Hasselt, 'Ensemble algorithms in reinforcement learning', *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **38**(4), 930–936, (Aug 2008).