# Using Ensemble Techniques and Multi-Objectivization to Solve Reinforcement Learning Problems

**Tim Brys**[1] and **Matthew E. Taylor**[2] and **Ann Nowé**[1]

**Abstract.** Recent work on multi-objectivization has shown how a single-objective reinforcement learning problem can be turned into a multi-objective problem with correlated objectives, by providing multiple reward shaping functions. The information contained in these correlated objectives can be exploited to solve the base, single-objective problem faster and better, given techniques specifically aimed at handling such correlated objectives. In this paper, we identify ensemble techniques as a set of methods that is suitable to solve multi-objectivized reinforcement learning problems. We empirically demonstrate their use on the Pursuit domain.

## 1 Introduction

Reinforcement learning [6] is a framework that allows an autonomous agent to adapt its behaviour in order to maximize the cumulative return of a given reward signal. These agents often learn from scratch, meaning that in more complex environments, they may require an impractical amount of exploration before a satisfactory level of behaviour is obtained; in other words: learning may be very slow. A significant amount of research goes into developing techniques that speed up or improve the learning [3, 7]. One of those techniques is reward shaping, which adds a feedback signal on top of the base reward signal, typically providing the agent with heuristic knowledge about the problem it is trying to solve. If this signal is properly formulated, using a potential function over the state space [3], reward shaping preserves the optimality of solutions, and its sole effect is to guide the exploration behaviour of the agent. This guidance can help the agent avoid spending time gathering experience on actions in situations that the domain expert knows to be suboptimal. Typically only a single shaping function is applied, but Devlin et al. [2] propose the use of multiple shaping functions, allowing the inclusion of different pieces of heuristic knowledge. They evaluate the use of a linear combination of two shaping functions in KeepAway Soccer [4], and show improvements compared to using a single shaping function.

Recent work on *multi-objectivization* abstracts and formalizes this use of multiple shapings as turning a single-objective problem into a multi-objective problem [1]. Applying different shaping functions to several copies of the base reward signal turns the problem into a multi-objective problem that preserves the total ordering of the solutions, while each objective provides different heuristic knowledge. More formally, a Markov Decision Process (MDP) with scalar reward function $R$ is multi-objectivized using potential-based shaping functions $F_1$ through $F_m$, by constructing a Multi-Objective MDP with vector reward function $\mathbf{R} = [R + F_1, \ldots, R + F_m]$. This formulation ensures that no conflicts are introduced between the objectives, and that by consequence no trade-offs need to be identified, which is the main concern in multi-objective optimization. On the contrary, these correlated objectives all have the same target (optimal solution), but each provides some different piece of heuristic knowledge, that, when strategically combined, may speed up learning even more than using a single of the shaping functions. To date, only a linear scalarization (weighted sum) was considered to combine these objectives. In this paper, we explore ensemble techniques as a promising set of methods to solve multi-objectivized reinforcement learning problems.

[1] Vrije Universiteit Brussel, Belgium, email: {timbrys, anowe}@vub.ac.be
[2] Washington State University, WA, email: taylorm@eecs.wsu.edu

## 2 Ensemble Techniques for Reinforcement Learning

Ensemble techniques were developed to combine multiple algorithms operating on the same problem class, in order to improve performance on those problems. In reinforcement learning, these techniques have been used to combine different learning algorithms ($Q$-learning, SARSA, ACLA, etc.) learning on the same task [8]. The algorithms are combined at the action selection stage, where each algorithm scores every action, these scores are aggregated, and an action is selected according to the combined scores. We investigate two such strategies proposed by Wiering et al. [8]. With *majority voting*, every algorithm scores their estimated best action with a 1, and every other action 0. *Rank voting* extends this to a full ranking of the $n$ actions, with each algorithm scoring its estimated best action $n-1$, and its estimated worst action 0. These scores are then aggregated over all algorithms through summation, yielding a combined score for every action. Action selection strategies such as $\epsilon$-greedy or Boltzmann exploration can then use these preference values to select an action.

We apply these techniques in a fundamentally new way, not combining different algorithms learning on the same signal, but by constructing different correlated versions of the same reward signal (multi-objectivization), and combining different algorithms learning on those signals. In fact, we can even use the same algorithm on every signal, as in this case, the diversity required for ensembles to provide a benefit lies in the signals themselves.

## 3 Pursuit Domain

The Pursuit domain was proposed to investigate coordination mechanisms in a multi-agent system. The basic idea of pursuit is that a number of predators must capture a (number of) prey(s) by moving through a simple gridworld. In [5], Stone and Veloso identify many variants of the problem, and our implementation is as follows. There

are two predators and one prey, and these can move in the four cardinal directions as well as choose to stay in place. The prey is caught when a predator moves onto the same gridworld cell as the prey; predators are not allowed to share the same cell. The prey takes a random action $20\%$ of the time, with the rest of the time devoted to moving away from the predators. To do that, it takes the action that maximizes the summed distance from both predators, making the problem harder than with a fully random prey. The predators are controlled by $Q(\lambda)$-learning agents, and both receive a reward of $1$ when the prey is caught by either one of them, and a reward of $0$ the rest of the time. The predators observe the relative $x$ and $y$ coordinates of the other predator and the prey. Tile-coding is used to discretize the state-space, with 32 tilings, and tile-width 10, hashed down to $4096$ weights. Action selection is $\epsilon$-greedy, with $\epsilon = 0.1$. Further parameters are $\gamma = 0.9$, $\lambda = 0.9$ and $\alpha = \frac{1}{10 \times 32}$.

We multi-objectivize the problem using three potential-based shaping functions:

**Proximity** encourages a predator to move closer to the prey. Its potential function is defined as $\Phi_P(s) = -d(pred, prey)$, with $d$ as the Manhattan distance.

**Angle** encourages the predators to move to different sides of the prey, trapping it. It is defined to maximize the angle between them and the prey to $\pi$: $\Phi_A(s) = \arccos(\frac{x \cdot y}{|x||y|})$, with $x$ and $y$ vectors pointing from the prey to the two predators respectively.

**Separation** encourages the predators to move away from each other. Its potential function is defined as $\Phi_S(s) = d(pred_1, pred_2)$ where $d$ is again the Manhattan distance.

We will investigate both normalized and non-normalized shaping functions, as the magnitude of a shaping relative to the basic reward can have a significant impact on learning. Proximity and Separation are normalized by dividing by $2 \times size$, with $size = 20$ both the width and height of the world; Angle is normalized by dividing by $\pi$. Furthermore, Proximity is implemented as $2 \times size - d(pred, prey)$, so that all shaping functions are positive, and thus optimistic.

## 4 Results and Discussion

Tables 1 and 2 summarize the results obtained in $1000$ runs of $1000$ episodes each, with a maximum number of steps per episode of $5000$, for normalized and non-normalized shapings respectively. We compare solving the problem using only the base reward (no shaping), the base reward plus one of the shapings ($x$ shaping), and a multi-objectivized version of the problem (with all three shapings), solving it using a linear scalarization[3] or ensemble techniques. The goal is to minimize the number of steps it takes to catch the prey, and we measure both final and cumulative performance.

In the case of normalized shapings, using the proximity shaping alone yields best performance, but the ensemble techniques are able to match it in cumulative performance, yielding slightly worse final performance. While tuning the linear scalarization weights yields performance similar to the proximity shaping alone,[4] the ensemble techniques are able to automatically approximate the best possible behaviour without parameter tuning.

In the non-normalized case, the difference in magnitude of the shapings is shown to have a significant impact on performance, with the proximity and separation shapings drowning the base reward, and resulting in very bad performance. A linear scalarization with weights compensating for this difference in magnitude can improve performance, but again the ensemble techniques achieve best performance without parameter tuning. They suffer less from this change in relative magnitude between the signals, because they combine the signals in a scale-invariant way.

These results support the hypothesis that the combination of multi-objectivization and ensemble techniques can improve learning in regular, single-objective reinforcement learning problems.

[3] Using uniform weights and weights that align the domains of the shapings for normalized and non-normalized shapings respectively.

[4] Results not included because of space constraints.

| Variant | Cumulative | Final |
|---|---|---|
| No shaping | $215794 \pm 2128$ | $116 \pm 2.4$ |
| Proximity shaping | $\mathbf{129555 \pm 1640}$ | $\mathbf{88 \pm 2.1}$ |
| Angle shaping | $209962 \pm 2031$ | $109 \pm 2.2$ |
| Separation shaping | $244513 \pm 3461$ | $101 \pm 2.5$ |
| Linear scalarization | $152670 \pm 1665$ | $96 \pm 2.2$ |
| Majority Voting Ensemble | $\mathbf{134899 \pm 7106}$ | $96 \pm 6.3$ |
| Rank Voting Ensemble | $\mathbf{130822 \pm 776}$ | $92 \pm 1.6$ |

**Table 1.** Cumulative and final performance for normalized shapings. The best results and those not significantly different from the best (Student's t-test, $p > 0.05$) are indicated in bold.

| Variant | Cumulative | Final |
|---|---|---|
| No shaping | $217554 \pm 2089$ | $116 \pm 2.4$ |
| Proximity shaping | $470809 \pm 5905$ | $438 \pm 10.1$ |
| Angle shaping | $235667 \pm 2625$ | $112 \pm 2.6$ |
| Separation shaping | $1216167 \pm 21970$ | $1142 \pm 33.5$ |
| Linear scalarization | $203131 \pm 3449$ | $\mathbf{104 \pm 3.3}$ |
| Majority Voting Ensemble | $159610 \pm 13156$ | $127 \pm 14.3$ |
| Rank Voting Ensemble | $\mathbf{142131 \pm 11267}$ | $\mathbf{99 \pm 11.2}$ |

**Table 2.** Cumulative and final performance for non-normalized shapings. The best results and those not significantly different from the best (Student's t-test, $p > 0.05$) are indicated in bold.

## ACKNOWLEDGEMENTS

Tim Brys is funded by a Ph.D grant of the FWO. This work was supported in part by NSF IIS-1149917.

## REFERENCES

[1] Tim Brys, Anna Harutyunyan, Peter Vrancx, Matthew E. Taylor, Daniel Kudenko, and Ann Nowé, 'Multi-objectivization of reinforcement learning problems by reward shaping', in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, (2014).

[2] Sam Devlin, Marek Grześ, and Daniel Kudenko, 'An empirical study of potential-based reward shaping and advice in complex, multi-agent systems', *Advances in Complex Systems*, **14**(02), 251–278, (2011).

[3] Andrew Y. Ng, Daishi Harada, and Stuart Russell, 'Policy invariance under reward transformations: Theory and application to reward shaping', in *Proceedings of ICML*, volume 99, pp. 278–287, (1999).

[4] Peter Stone, Gregory Kuhlmann, Matthew E. Taylor, and Yaxin Liu, 'Keepaway soccer: From machine learning testbed to benchmark', in *RoboCup-2005: Robot Soccer World Cup IX*, volume 4020, 93–105, Springer-Verlag, Berlin, (2006).

[5] Peter Stone and Manuela Veloso, 'Multiagent systems: A survey from a machine learning perspective', *Autonomous Robots*, **8**(3), 345–383, (2000).

[6] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: An introduction*, volume 1, Cambridge Univ Press, 1998.

[7] Matthew E. Taylor and Peter Stone, 'Transfer learning for reinforcement learning domains: A survey', *JMLR*, **10**, 1633–1685, (2009).

[8] Marco A. Wiering and Hado van Hasselt, 'Ensemble algorithms in reinforcement learning', *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **38**(4), 930–936, (2008).