Reinforcement Learning on Multiple Correlated Signals

Tim Brys and **Ann Nowé**
Vrije Universiteit Brussel
{timbrys, anowe}@vub.ac.be

Abstract

This extended abstract provides a brief overview of my PhD research on multi-objectivization and ensemble techniques in reinforcement learning.

Problem

Multi-objective problems (MOP) require the simultaneous optimization of multiple feedback signals or objectives. As conflicts may exist between objectives, there is in general a need to identify (a set of) trade-off solutions. The set of optimal, i.e. non-dominated, incomparable solutions is called the Pareto-front. We identify multi-objective problems with correlated objectives (CMOP) as a specific subclass of multi-objective problems, defined to contain those MOPs whose Pareto-front is so limited that one can barely speak of trade-offs (Brys et al. 2014b). By consequence, the system designer does not care about which of the very similar optimal solutions is found, but rather how fast it is found (and perhaps how well it is approximated). Specifically, we investigate such reinforcement learning problems formulated as *Correlated Multi-Objective Markov Decision Processes* (CMOMDP). (Single-objective) *MDPs* describe a system as a set of potential observations of that system's state S, a set of possible actions A, transition probabilities T for state-action-state triplets, and a reward function R that probabilistically maps these transitions to a scalar reward indicating the utility of that transition. The goal of a reinforcement learning agent operating in an MDP is to maximize the expected, discounted return of the reward function. Popular temporal-difference learners such as SARSA (Rummery and Niranjan 1994) attempt this by estimating the Q-function, which represents the utility of each state-action pair. *MOMDPs* (Roijers et al. 2013) extend this framework to multiple objectives, with the reward function returning a vector of rewards to be maximized, and the added difficulty of finding trade-off solutions. Finally, *CMOMDPs* remove the need to find trade-offs, as (near-) optimal solutions for any objective are (near-) optimal for every objective.

The relevance of this class of problems may not immediately be obvious, but we show in our work that (1) there exist problems that naturally fall into this problem class, and more importantly, (2) that any MDP can be framed as a CMOMDP. For the first type, we identify the traffic light control problem as being a natural CMOMDP (Brys, Pham, and Taylor 2014), as policies improving the average delay for cars also improve the throughput of the system, showing that the objectives are strongly correlated. For the second type, we show that any single-objective MDP can be *multi-objectivized*, i.e. turned into a CMOMDP, by using several potential-based reward shaping functions (heuristic signals guiding exploration) (Brys et al. 2014a). We prove that this modification preserves the total order, and thus also optimality, of policies, mainly relying on the results by Ng, Harada, and Russell (1999). This insight – that any MDP can be framed as a CMOMDP – significantly increases the importance of this problem class, as well as techniques developed for it, as these could be used to solve regular single-objective MDPs *faster* and *better*, provided several meaningful shapings can be devised.

Solution techniques

CMOMDPs can in principle be solved with a single-objective solution method using feedback from only one of the objectives. But, as the different objectives are correlated and basically provide multiple sources of information for a single-objective optimization problem, combining these objectives intelligently could allow an agent to better solve these problems. Calculating a scalarization of the objectives is the most naive approach, and has been done by Devlin, Grześ, and Kudenko (2011), who defined two shapings for Keepaway Soccer, and combined them using a linear scalarization (implicit multi-objectivization).

We developed a novel technique inspired by work in evolutionary computation, where it is proposed to make every optimization decision based on feedback from only a single of the correlated objectives. Before every decision, they select one objective and use that to measure solution quality and accept/reject candidate solutions. Jensen (2005) makes this objective selection decision uniformly at random, while in (Buzdalova and Buzdalov 2012), the authors treat this selection as a dynamically changing (as optimization progresses) multi-armed bandit problem, solving it using Q-learning.

Our approach works similarly for temporal-difference learners in CMOMDPs. We call it *adaptive objective selec-*

tion (Brys et al. 2014c; 2014b). The learner estimates the Q-function for every objective o in parallel, and decides before every action selection decision which objective's estimates to use. To make this objective selection decision, we introduce the concept of *confidence* in learned estimates, defining confidence as an estimation of the likelihood that the estimates are correct. Higher-variance reward distributions will make any estimate of the average reward less confident, and always selecting the objective whose estimates are most likely to be correct will maximize the likelihood of correctly ranking the action set.

This approach has several interesting properties. It makes its decisions a function of the state-space, which can account for different objectives being more or less reliable in different parts of the state space. Furthermore, it uses the objectives in a scale-invariant way. That is, its workings do not depend on the relative scalings of the objectives, since all confidence metrics proposed are scale-invariant, and thus no parameters are introduced. This is a significant improvement over scalarization techniques (the most common approach to multi-objective problems), which usually require weight tuning, if only to align the magnitudes of the different correlated objectives in CMOPs. Adaptive objective selection can be said to do this implicitly and automatically.

The most interesting variant we introduced exploits the inherent Q-value decomposition of tile-coding function approximation to measure confidence in estimates. The technique was shown to improve performance on a traffic light control problem, a natural CMOMDP, as well as on the Pursuit domain, an example of a single-objective MDP framed as a multi-objective problem using multiple shaping functions. Furthermore, the technique's objective selection decisions yield intuitive insights into the nature of the problems being solved, e.g. indicating where in the state space each shaping function correlates best with the value function.

Work in Progress

Better characterization of problem class Our current mathematical definition for the class of problems we investigate encapsulates the intuition that (near-) optimal solutions for any objective are (near-) optimal for every objective. I am investigating whether we could provide a better characterization of the type of problems considered, (1) mathematically, through a definition that reflects on the shape of the whole set of solutions in objective space, not just (near-) optimal solutions, and (2) intuitively, using simple abstract optimization problems that fall in this class.

Validation of current techniques To further demonstrate the usefulness of this class of problems, we are working on several other domains to demonstrate the potential of framing regular single-objective MDPs as CMOMDP. These domains include Mountain Car, KeepAway and StarCraft, each multi-objectivized (Brys et al. 2014a) using several shaping functions. Initial results with adaptive objective selection in the first two domains are promising.

Other solution techniques We believe there exist (in potentia) many more techniques to solve CMOMDPs. One example of a set of techniques that holds much promise is ensemble techniques for reinforcement learning (Wiering and van Hasselt 2008). These combine different (weak) predictors (of unknown performance) for the same signal in order to create a predictor that is better than any of the constituting parts. 'The same signal' can be relaxed to the condition necessary for CMOMDPs, allowing their application to this type of problem. Initial experiments in Mountain Car and the Pursuit domain also show promising results for ensemble techniques.

References

Brys, T.; Harutyunyan, A.; Vrancx, P.; Taylor, M. E.; Kudenko, D.; and Nowé, A. 2014a. Multi-objectivization of reinforcement learning problems by reward shaping. In *Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN 2014)*.

Brys, T.; Kudenko, D.; Nowé, A.; and Taylor, M. E. 2014b. Selecting reward and shaping signals based on confidence in estimates. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*.

Brys, T.; Van Moffaert, K.; Nowé, A.; and Taylor, M. E. 2014c. Adaptive objective selection for correlated objectives in multi-objective reinforcement learning. In *Proceedings of the 13th Internation Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Brys, T.; Pham, T. T.; and Taylor, M. E. 2014. Distributed learning and multi-objectivity in traffic light control. *Connection Science* (DOI:10.1080/09540091.2014.885282).

Buzdalova, A., and Buzdalov, M. 2012. Increasing efficiency of evolutionary algorithms by choosing between auxiliary fitness functions with reinforcement learning. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, 150–155. IEEE.

Devlin, S.; Grześ, M.; and Kudenko, D. 2011. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems* 14(02):251–278.

Jensen, M. T. 2005. Helper-objectives: Using multi-objective evolutionary algorithms for single-objective optimisation. *Journal of Mathematical Modelling and Algorithms* 3(4):323–347.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, volume 99, 278–287.

Roijers, D. M.; Vamplew, P.; Whiteson, S.; and Dazeley, R. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48:67–113.

Rummery, G. A., and Niranjan, M. 1994. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering.

Wiering, M. A., and van Hasselt, H. 2008. Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 38(4):930–936.