Multi-Objective Reinforcement Learning

A. Nowé K. Van Moffaert M. M. Drugan

Artificial Intelligence Lab

Vrije Universiteit Brussel

Pleinlaan 2, 1050 Brussels, Belgium

In multi-objective reinforcement learning (MORL) the agent is provided with multiple feedback signals when performing an action. These signals can be independent, complementary or conflicting. Hence, MORL is the process of learning policies that optimize multiple criteria simultaneously. In this abstract, we briefly describe our extensions to single-objective multi-armed bandits and reinforcement learning (RL) algorithms to make them applicable in multi-objective environments.

Bellow, we first summarize the main parts of our *multi-objective multi-armed bandits* (MOMAB) [2] framework that incorporates techniques from multi-objective optimization.

A first straightforward approach is to use a scalarization function which transforms the multi-objective environment into a single-objective environment, and therefore they can be implemented into the standard MAB framework. Since single-objective environments, in general, results in a single optimum, we need to vary the scalarization functions to generate a variety of elements belonging to the Pareto optimal set. The efficiency of these MOMAB algorithms depends very much on the scalarization type. The *linear scalarization* is the most popular scalarization function due to its simplicity. Moreover, a known problem with linear scalarization is its incapacity to potentially find all the points in a non-convex Pareto set. *The Chebyshev scalarization* has the advantage that it can find all the points in a non-convex Pareto set at the cost of introducing another parameter i.e. the utopian point, that needs to be optimized.

To measure the performance of a single scalarized MOMAB we introduce the scalarized regret that is defined by the difference between the maximum scalarized value and the scalarization of an suboptimal arm. The upper bound of a multi-objective UCB1 that randomly alternates between a number of scalarized UCB1 is then the sum of upper bounds for each single UCB1. While this definition of regret seems natural, it only partially serves our goal because it gathers a collection of independent regrets instead of minimizing the regret of a multi-objective strategy

in all objectives. As an improvement, we introduce the concept of *fairness* that is the *variance* of a set of the optimal arms. This concept of fairness allowed us to develop a MOMAB which starts with a large set of scalarized MAB and progressively deletes the irrelevant scalarizations. We are currently proving the efficiency of this algorithm based on [1, 4].

Our Pareto MAB uses the Pareto dominance order relationship to explore the multi-objective environment directly. As a result, there could be a large set of optimal arms and therefore a larger time interval to explore them. An adequate regret definition for the Pareto MAB algorithm measures the distance between the *set* of optimal reward vectors and a suboptimal reward vector. Our measure is inspired by ϵ-dominance as proposed in [3]. In terms of upper regret bounds, the worst case occurs when all the arms are optimal, and the best case results in a similar behaviour as the standard UCB1 when there are only few optimal arms.

Inspired by the results on the MOMABs, we also developed into a multi-objective Q-learning framework [5] that can incorporate any scalarization function. The principle extension compared to single-objective Q-learning are the $Q(s, a, o)$-values that allow a separate Q-value for each state-action-objective tuple. The particular scalarization function guides the action selection process based on the weighted preference. For example, in the greedy case, the action with the largest scalarized $Q(s, a, o)$-value for its objectives or $SQ(s, a)$-value is selected. The update rule adjusts the $Q(s, a, o)$-values into the direction of the largest scalarized Q-value of the next state s' for objective o, i.e. $\max_{SQ(s',a')} Q(s', a', o)$.

We noted that in environments that consist of a convex shape of the Pareto front, the linear scalarization function is a powerful technique to guide the search towards every Pareto optimal policy, given the proper weight parameter. However, in non-convex environments, the limitations of the linear scalarization function become clear as it is unable to capture concave parts of the Pareto front. The Chebyshev scalarization function however attempts to minimize the weighted distance from each Q-value to a particular reference point whereby it is able to cover non-convex Pareto fronts.

Currently we are developing an approach to learn directly the set of Pareto incomparable solutions. Therefore, we are using the Pareto dominance relation directly in the Q-learning update rule. Hence, the update rule performs backups with sets instead of (weighted) scalar values. This idea was previously explored in dynamic programming where [6] and [7] examined this approach in deterministic environments. Their approach is however infeasible in environments with stochastic reward schemes or cyclic environments, as they simply append Pareto dominating Q-vectors. We propose to store the average immediate reward and the Pareto dominating future discounted reward vector separately. Hence, these two entities can converge separately but can easily be combined with a vector-sum operator when the actual Q-vectors are requested.

References

[1] P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

[2] M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits: an analysis. In *Proc of International Joint Conference of Neural Networks (IJCNN)*, 2013.

[3] C. Horoba and F. Neumann. Benefits and drawbacks for the use of ϵ-dominance in evolutionary multi-objective optimization. In *Proc of Genetic and Evolutionary Computation Conference (GECCO'08)*, 2008.

[4] A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *NIPS*, pages 3284–3292, 2012.

[5] K. Van Moffaert, M. M. Drugan, and A. Nowé. Scalarized Multi-Objective Reinforcement Learning: Novel Design Techniques. In *2013 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. IEEE, 2013.

[6] D. J. White. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 89(2), 1982.

[7] M. A. Wiering and E. D. de Jong. Computing Optimal Stationary Policies for Multi-Objective Markov Decision Processes. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 158–165. IEEE, Apr. 2007.