Why Multi-objective Reinforcement Learning?

Diederik M. Roijers, Shimon Whiteson {D.M.ROIJERS, S.A.WHITESON}@UVA.NL
Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Peter Vamplew, Richard Dazeley {P.VAMPLEW, R.DAZELEY}@FEDERATION.EDU.AU
Federation Learning Agents Group, Federation University Australia, Ballarat, Victoria, Australia

Editor: Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos

Abstract

We argue that multi-objective methods are underrepresented in RL research, and present three scenarios to justify the need for explicitly multi-objective approaches. Key to these scenarios is that although the utility the user derives from a policy — which is what we ultimately aim to optimize — is scalar, it is sometimes impossible, undesirable or infeasible to formulate the problem as single-objective at the moment *when the policies need to be learned.* We also present the case for a utility-based view of multi-objective RL, i.e., that the appropriate multi-objective solution concept should be derived from what we know about the user's utility function, rather than axiomatically assumed to be the Pareto front.

Many tasks have multiple, possibly conflicting, objectives. However, while interest in multi-objective RL (MORL) has grown in recent years, the majority of RL research and applications still assume only a single objective. Some researchers argue that explicit multi-objective modeling is not necessary, and that a scalar reward function is adequate. For example Sutton's *reward hypothesis*, states "that all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)". The implication is that a multi-objective MDP can always be converted into a single-objective one. Such a conversion process would involve two steps. The first is to specify a *scalarization function* f, that projects the multi-objective value $\mathbf{V}^\pi$ to a scalar utility $V^\pi_{\mathbf{w}}(s) = f(\mathbf{V}^\pi(s), \mathbf{w})$, where $\mathbf{w}$ is a vector parameterizing f. For example, if f is a linear combination of the values, $\mathbf{w}$ specifies the relative weight of the objectives. The second step is to define a single-objective MDP such that, for all π and s, the expected return equals the scalarized value $V^\pi_{\mathbf{w}}(s)$.

However, there are three scenarios for which one or both of these conversion steps is impossible, infeasible, or undesirable (see Figure 1) (Roijers et al., 2013). The *unknown weights scenario* occurs when $\mathbf{w}$ is unknown at the moment of learning (e.g., when the weights correspond to prices that may vary over time). In this scenario, it can be preferable to use a multi-objective method to compute in advance a set of policies and then use the current weights to select the optimal policy at any point in time.

In the *decision support scenario*, scalarization is infeasible throughout the entire decision-making process because of the difficulty of specifying $\mathbf{w}$, or even f (for example, the user may have subjective preferences that defy meaningful quantification). In such a system, MORL can be used to calculate an optimal solution set with respect to any known constraints about f and $\mathbf{w}$ and then the user selects a policy according to their arbitrary preferences, rather than via explicit scalarization.

In the *known weights scenario*, **w** is known and thus scalarization is possible. However, it may be undesirable because of the difficulty of the second step in the conversion. If f is nonlinear, the resulting MDP may not have *additive returns* and the optimal policy may be non-stationary or stochastic. Consequently, many standard single-objective methods are not applicable and so it may be preferable to use methods specially designed for MOMDPs.

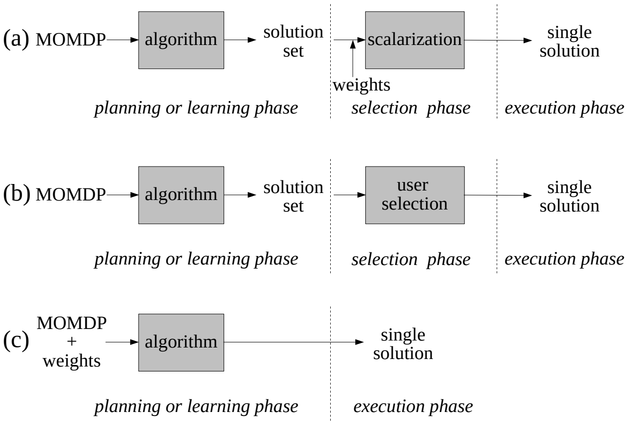

Figure 1: MOMDP motivating scenarios: (a) unknown weights, (b) decision support, (c) known weights.

Hence, scalar RL is preferable to MORL only if there genuinely is only a single objective, or if f is known and linear, and **w** is known in advance and fixed. We argue this occurs less frequently than indicated by existing practice and applying single-objective methods to multi-objective tasks may not fully meet the user's needs.

The decision to adopt a multi-objective approach to RL is often seen as requiring the agent to identify all policies belonging to the *Pareto front*, which we call the *axiomatic approach* to MORL. However, in many cases the full Pareto front is not required. Therefor, we advocate a *utility-based* approach where the algorithm is tailored to produce specifically the solution concept required by the user, i.e., it deduces the solution concept from the nature of the scalarization function, the types of policies that are admissible, and the scenario to which the task belongs.

The most obvious difference between the axiomatic and utility-based approaches is for the known weights scenario, in which only a single policy is required. Even when multiple policies are required, as in the unknown weights or decision support scenarios, the user's needs are often satisfied by a subset of the Pareto front. For example, given a linear scalarization function, it suffices to find only those policies which lie on the *convex hull.* as there are no values for **w** for which policies outside of the convex hull will be optimal. Similarly, if the problem allows for the use of nonstationary or stochastic policies, then the convex hull is a sufficient solution set even for nonlinear scalarization functions as an optimal solution can always be formed from a mixture policy of two or more members of the convex hull. In fact, the only context in which the full Pareto front is required is when the scalarization function is either unknown or known to be nonlinear, and a strictly deterministic stationary policy is required. Therefore, a utility-based approach will often be both simpler and less computationally intensive than deriving the entire Pareto front, and more appropriate to the needs of the user.

Acknowledgments

This work is supported by NWO (project #612.001.109).

References

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