Abstract: Expected Scalarised Returns Dominance: A New Solution Concept for Multi-Objective Decision Making*

Conor F. Hayes, Timothy Verstraeten, Diederik M. Roijers, Enda Howley, and Patrick Mannion

NUI Galway (IE), Vrije Universiteit Brussel (BE), and HU University of Applied Sciences Utrecht (NL)
Corresponding: c.hayes13@nuigalway.ie

Abstract. In many real-world scenarios, the utility of a user is derived from a single execution of a policy. In this case, to apply multi-objective reinforcement learning, the expected utility of the returns must be optimised. Various scenarios exist where a user's preferences over objectives (also known as the utility function) are unknown or difficult to specify. In such scenarios, a set of optimal policies must be learned. However, settings where the expected utility must be maximised have been largely overlooked by the multi-objective reinforcement learning community and, as a consequence, a set of optimal solutions has yet to be defined. In this work we define a new dominance criterion, known as expected scalarised returns (ESR) dominance, that extends first-order stochastic dominance to allow a set of optimal policies to be learned in practice. Additionally, we define a new solution concept called the ESR set, which is a set of policies that are ESR dominant.

1 Introduction

Many real-world sequential decision making problems have multiple, often conflicting, objectives [15]. A modern approach to solving such problems is to apply multi-objective reinforcement learning (MORL) by taking a utility-based perspective [2]. For MORL a utility function is used to model the preferences over objectives of a user (human decision maker). However, in certain scenarios a user may be uncertain about their preferences, and therefore their utility function may be unknown [11].

The majority of MORL literature focuses on two optimality criteria: the scalarised expected returns (SER) and the expected scalarised returns (ESR) criterion. When a user has multiple opportunities to execute a policy, the SER criterion must be optimised. Under the SER criterion, the expected value vector is calculated, the utility function is applied, and the utility of the expectation is computed. The SER criterion is the most commonly used optimality criterion in the MORL literature [7,13,16,12]. In scenarios where a user may only have

* Original article published in Neural Computing & Applications, April 19, 2022 [4].

a single opportunity to execute a policy, the ESR criterion must be optimised. Under the ESR criterion, the utility function is applied to the vector returns, then the expected utility is calculated. The ESR criterion has largely been overlooked with some exceptions [10,3,14,6,9].

Under the SER criterion, when the utility function is unknown, expected value vectors can be utilised to determine a partial ordering over policies (e.g Pareto dominance [8]) and a solution set can be computed (e.g. Pareto front) [16,7]. However, expected value vectors are fundamentally incompatible with the ESR criterion, given the utility function must first be applied to the vector returns before the expectation can be computed. Furthermore, to date no solution concept has been defined to determine a partial ordering over policies under the ESR criterion. Therefore, new solution concepts must be derived in order to compute sets of optimal policies under the ESR criterion.

In this work we define a new solution concept known as ESR dominance, which takes a distributional perspective to MORL to determine a partial ordering over policies under the ESR criterion. By utilising ESR dominance it is possible to compute a set of policies under the ESR criterion. Moreover, we define the resulting set of optimal policies as the *ESR set*. To compute ESR dominance, we first define a return distribution, $\mathbf{Z}^{\pi}$, as the multivariate distribution over the vector returns received from executing a policy π. Therefore, we can define ESR dominance as follows:

Definition 1. *For return distributions $\boldsymbol{Z}^{\pi}$ and $\boldsymbol{Z}^{\pi'}$, $\boldsymbol{Z}^{\pi} \succ_{ESR} \boldsymbol{Z}^{\pi'}$ for all monotonically increasing utility functions if, and only if, the following is true:*

$$\boldsymbol{Z}^{\pi} \succ_{ESR} \boldsymbol{Z}^{\pi'} \Leftrightarrow$$

$$\forall \mathbf{v} : F_{\boldsymbol{Z}^{\pi}}(\mathbf{v}) \leq F_{\boldsymbol{Z}^{\pi'}}(\mathbf{v}) \wedge \exists \mathbf{v} : F_{\boldsymbol{Z}^{\pi'}}(\mathbf{v}) < F_{\boldsymbol{Z}^{\pi'}}(\mathbf{v}),$$

where $F_{\mathbf{Z}^{\pi}}$ is the cumulative distribution function (CDF) of the return distribution Z^{π}. ESR dominance extends first-order stochastic dominance [5,17,1] for MORL settings. Using ESR dominance, it is possible to define a set of optimal policies for the ESR criterion, known as the *ESR set*.

Definition 2. *The* ESR set*, $ESR(\Pi)$, is a sub-set of all policies where each policy in the* ESR set *is ESR dominant,*

$$ESR(\Pi) = \{\pi \in \Pi \mid \nexists \pi' \in \Pi : \mathbf{Z}^{\pi'} \succ_{ESR} \mathbf{Z}^{\pi}\}.$$

By utilising ESR dominance to compute the *ESR set* it is now possible to compute sets of optimal solutions under the ESR criterion in MORL settings. Given expected value vectors cannot be used to compute sets of optimal policies under the ESR criterion, new distributional MORL methods must be developed for the ESR criterion.

Acknowledgements This research was supported by the following organisations and grants: FWO #1SA2820N, the Flemish Government "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen", and the the National University of Ireland Galway Hardiman Scholarship.

References

1. Atkinson, A.B., Bourguignon, F.: The Comparison of Multi-Dimensioned Distributions of Economic Status. The Review of Economic Studies **49**(2), 183–201 (04 1982). https://doi.org/10.2307/2297269, https://doi.org/10.2307/2297269
2. Hayes, C.F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L.M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A.A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., Roijers, D.M.: A practical guide to multi-objective reinforcement learning and planning. Autonomous Agents and Multi-Agent Systems **36**(1), 1–59 (2022)
3. Hayes, C.F., Reymond, M., Roijers, D.M., Howley, E., Mannion, P.: Distributional monte carlo tree search for risk-aware and multi-objective reinforcement learning. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. vol. 2021. IFAAMAS (2021 In Press)
4. Hayes, C.F., Verstraeten, T., Roijers, D.M., Howley, E., Mannion, P.: Expected scalarised returns dominance: a new solution concept for multi-objective decision making. Neural Computing and Applications pp. 1–21 (2022)
5. Levy, H.: Stochastic dominance and expected utility: Survey and analysis. Management Science **38**(4), 555–593 (1992), http://www.jstor.org/stable/2632436
6. Malerba, F., Mannion, P.: Evaluating tunable agents with non-linear utility functions under expected scalarised returns. In: Multi-Objective Decision Making Workshop (MODeM 2021) (2021)
7. Moffaert, K.V., Nowé, A.: Multi-objective reinforcement learning using sets of pareto dominating policies. Journal of Machine Learning Research **15**(107), 3663–3692 (2014), http://jmlr.org/papers/v15/vanmoffaert14a.html
8. Pareto, V.: Manuel d'Economie Politique, vol. 1. Giard, Paris (1896)
9. Reymond, M., Hayes, C., Roijers, D.M., Steckelmacher, D., Nowé, A.: Actor-critic multi-objective reinforcement learning for non-linear utility functions. In: Multi-Objective Decision Making Workshop (MODeM 2021) (2021)
10. Roijers, D.M., Steckelmacher, D., Nowé, A.: Multi-objective reinforcement learning for the expected utility of the return. In: Proceedings of the Adaptive and Learning Agents workshop at FAIM 2018 (2018)
11. Roijers, D.M., Vamplew, P., Whiteson, S., Dazeley, R.: A survey of multi-objective sequential decision-making. Journal of Artificial Intelligence Research **48**, 67–113 (2013)
12. Roijers, D.M., Whiteson, S., Oliehoek, F.A.: Linear support for multi-objective coordination graphs. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems. p. 1297–1304. AAMAS '14, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2014)
13. Vamplew, P., Dazeley, R., Berry, A., Issabekov, R., Dekker, E.: Empirical evaluation methods for multiobjective reinforcement learning algorithms. Machine Learning **84**, 51–80 (07 2011). https://doi.org/10.1007/s10994-010-5232-5
14. Vamplew, P., Foale, C., Dazeley, R.: The impact of environmental stochasticity on value-based multiobjective reinforcement learning. In: Neural Computing and Applications (2021). https://doi.org/https://doi.org/10.1007/s00521-021-05859-1
15. Vamplew, P., Smith, B.J., Kallstrom, J., Ramos, G., Radulescu, R., Roijers, D.M., Hayes, C.F., Heintz, F., Mannion, P., Libin, P.J., et al.: Scalar reward is not enough: A response to silver, singh, precup and sutton (2021). arXiv preprint arXiv:2112.15422 (2021)

16. Wang, W., Sebag, M.: Multi-objective Monte-Carlo tree search. In: Hoi, S.C.H., Buntine, W. (eds.) Proceedings of Machine Learning Research. vol. 25, pp. 507–522. PMLR, Singapore (Nov 2012)
17. Wolfstetter, E.: Topics in Microeconomics: Industrial Organization, Auctions, and Incentives. Cambridge University Press (1999). https://doi.org/10.1017/CBO9780511625787