Decision-Theoretic Planning for the Expected Scalarised Returns

Extended Abstract

Conor F. Hayes
National University of Ireland Galway (IE)
c.hayes13@nuigalway.ie

Diederik M. Roijers
Vrije Universiteit Brussel (BE)
& HU Univ. of Appl. Sci. Utrecht (NL)

Enda Howley
National University of Ireland Galway (IE)

Patrick Mannion
National University of Ireland Galway (IE)

ABSTRACT

In sequential multi-objective decision making (MODeM) settings, when the utility of a user is derived from a single execution of a policy, policies for the expected scalarised returns (ESR) criterion should be computed. In multi-objective settings, a user's preferences over objectives, or utility function, may be unknown at the time of planning. When the utility function of a user is unknown, multi-policy methods are deployed to compute a set of optimal policies. However, the state-of-the-art sequential MODeM multi-policy algorithms compute a set of optimal policies for the scalarised expected returns (SER) criterion. Algorithms that compute a set of optimal policies for the SER criterion utilise expected value vectors which cannot be used when optimising for the ESR criterion. We propose multi-objective distributional value iteration (MODVI) that replaces value vectors with distributions over the returns and computes a set of optimal policies for the ESR criterion.

KEYWORDS

Multi-objective; decision making; expected scalarised returns; distributional; planning; value iteration

ACM Reference Format:
Conor F. Hayes, Diederik M. Roijers, Enda Howley, and Patrick Mannion. 2022. Decision-Theoretic Planning for the Expected Scalarised Returns: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

When making decisions in the real world, trade-offs between multiple, often conflicting, objectives must be made [16]. In many real-world decision making settings, a policy is only executed once. The current state-of-the-art multi-objective decision making (MODeM) literature focuses almost exclusively on computing polices that are optimal over multiple executions [14]. Therefore, to fully utilise MODeM in the real world, we must develop algorithms to compute a policy, or set of policies, that are optimal given the single-execution nature of the problem.

MODeM distinguishes between two optimality criteria. In scenarios where the utility of a user is derived from multiple executions of a policy, the scalarised expected returns (SER) criterion should be optimised [8]. In scenarios where the utility of a user is derived from a single execution of a policy, the expected scalarised returns (ESR) criterion should be optimised [6, 7, 13].

The majority of multi-policy MODeM algorithms are designed to compute a set of optimal policies for the SER criterion [1, 4, 5, 18]. However, the current state-of-the-art SER methods [11, 17] are fundamentally incompatible with the ESR criterion [9, 10]. When the utility function of a user is unknown, SER methods use expected value vectors to compute a set of optimal policies [17, 18]. To compute policies under the ESR criterion, a distribution over the returns, or return distribution, must be maintained [9].

We propose multi-objective distributional value iteration (MODVI, Algorithm 2) that computes a set of optimal policies for the ESR criterion in scenarios when the utility function of a user is unknown at the time of planning.

2 MULTI-OBJECTIVE DISTRIBUTIONAL VALUE ITERATION

To compute a set of optimal policies for the ESR criterion when the utility function of a user is unknown, we propose multi-objective distributional value iteration (MODVI, Algorithm 2). MODVI maintains sets of return distributions for each state and uses ESR dominance [9] to compute a set of non-dominated return distributions, known as the *ESR set* [9, 10].

To compute a set of optimal polices for the ESR criterion, expected value vectors must be replaced with return distributions [9]. Generally, expected value MODeM algorithms utilise the Bellman operator [3] to compute the expected value vectors for each state. Given our approach is distributional, we adopt the distributional Bellman operator [2], $\mathcal{T}_D^\pi$, to update the return distribution for each state-action pair:

$$\mathcal{T}_D^\pi \mathbf{z}(s,a) \stackrel{D}{=} \mathbf{r}_{s,a} + \gamma\, \mathbf{z}(s',a'). \tag{1}$$

To represent a return distribution in multi-objective settings, we use a multivariate categorical distribution similar to the distributions used by Reymond et al. [12] and Bellemare et al. [2]. The categorical distribution is paramaterised by a number of atoms, $N \in \mathbb{N}$, where the distribution has a dimension per objective, n. The atoms outline the width of each category and are bounded by the minimum returns, $\mathbf{R}_{min}$, and maximum returns, $\mathbf{R}_{max}$.

To update the multivariate categorical distribution, we utilise the state space, action space and reward function of the model. During an update of the multivariate categorical distribution, we iterate over each atom, j, for each objective. To update the return distribution, $\mathbf{z}_s$, for state s, we compute the distributional Bellman update $\hat{\mathcal{T}}\mathbf{z}_{s,j} = \mathbf{r}_{s,a,s'} + \gamma \mathbf{z}_{s',j}$ for each atom j, for a given reward $\mathbf{r}_{s,a,s'}$ and

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

return distribution, $\mathbf{z}_{s'}$, for state s'. We then distribute the probability, p, for the atom, j, of the return distribution, $p_j(\mathbf{z}_{s'})$, in state s', to the corresponding atom of the updated return distribution, z_s, for state s.

At each iteration, k, of MODVI, for each state, s, and action, a, a set of optimal return distributions is backed up once. In Equation 2, the Bellman operator has been replaced with the distributional Bellman operator [2],

$$\mathbf{Q}_{k+1}(s,a) \leftarrow \bigoplus_{s'} T(s'|s,a)[\mathbf{r}_{s,a,s'} + \gamma \mathbf{Z}_k(s')] \tag{2}$$

where $\mathbf{Q}_{k+1}(s,a)$ and $\mathbf{Z}_k(s')$ represent sets of return distributions, $\oplus$ denotes the cross-sum between sets of return distributions, and $T(s'|s,a)$ represents the probability of transitioning to state s' from state s after taking action a.

To compute a set of ESR non-dominated policies for each state, we define an algorithm known as `ESRPrune` (Algorithm 1) which computes a set of ESR non-dominated policies by removing ESR dominated return distributions from a given set.

$$\mathbf{Z}_{k+1}(s) \leftarrow \texttt{ESRPrune}\left(\bigcup_a \mathbf{Q}_{k+1}(s,a)\right) \tag{3}$$

Equation 3 calculates the set of return distributions for a given state, s, by taking the union of each set of return distributions over each action, a. The resulting set of return distributions is then passed to the `ESRPrune` algorithm as input.

`ESRPrune` utilises ESR dominance defined by Hayes et al. [9, 10]. Like Pareto dominance, ESR dominance is transitive [19], therefore we can apply `ESRPrune` in sequence. To compute ESR dominance, the cumulative distribution function (CDF) of each return distribution in the given set must be calculated. `ESRPrune` iterates over the given set of return distributions and compares the CDFs of the return distributions to determine which are ESR non-dominated. The return distributions that are ESR dominated are removed from the set. A set of non-dominated return distributions is known as the *ESR set* [9].

Algorithm 1: `ESRPrune`

Input: $\mathbf{Z}$ $\leftarrow$ A set of return distributions
$\mathbf{Z}^* \leftarrow \emptyset$
while $\mathbf{Z} \neq \emptyset$ **do**
  $\mathbf{z} \leftarrow$ the first element of $\mathbf{Z}$
  for $\mathbf{z}' \in \mathbf{Z}$ **do**
    if $\mathbf{z}' >_{ESR} \mathbf{z}$ **then**
      $\mathbf{z} \leftarrow \mathbf{z}'$
    end
  end
  Remove $\mathbf{z}$ and all return distributions
  ESR-dominated by $\mathbf{z}$ from $\mathbf{Z}$. Add $\mathbf{z}$ to $\mathbf{Z}^*$
end
Return $\mathbf{Z}^*$

Algorithm 2: MODVI

Initialise all return distributions and sets
while *not converged* **do**
  for $s \in S$ **do**
    for $a \in A$ **do**
      $\mathbf{Q}_{k+1}(s,a) \leftarrow$
        $\bigoplus_{s'} T(s'|s,a)[\mathbf{R}(s,a,s') + \gamma \mathbf{Z}_k(s')]$
    end
    $\mathbf{Z}_{k+1}(s) \leftarrow$ *ESRPrune* $\left(\bigcup_a \mathbf{Q}_{k+1}(s,a)\right)$
  end
end

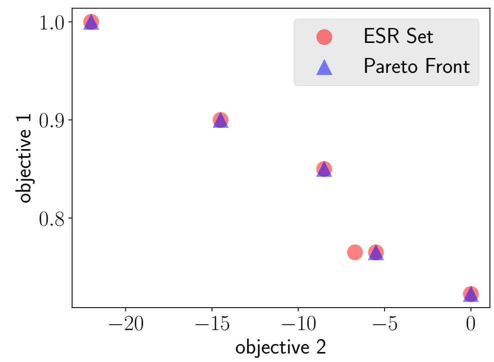

Figure 1: The expected value vectors of the return distributions in the *ESR set* (red) are plotted against the expected value vectors of the Pareto front (blue). The *ESR set* contains one extra policy. Under the SER criterion, the extra policy is Pareto dominated.

3 EXPERIMENTS

We evaluated MODVI using three multi-objective benchmark problem domains. In this paper, we present the results of MODVI evaluated using Space Traders [15]. Space Traders is a problem with nine policies and a small number of returns per policy. Therefore, it is possible to visualise each policy in the *ESR set.*

Figure 1 plots the expected value vectors of each return distribution in the *ESR set* and also plots the expected value vectors for the Pareto front [15]. It is important to note, the *ESR set* for Space Traders contains a policy that is not present on the Pareto front. The Pareto front is a set of optimal policies for the SER criterion. Therefore, certain policies that are optimal under the ESR criterion are not optimal under the SER criterion. In real-world decision making, incorrectly selecting an optimality criterion can lead to sub-optimal performance, given some optimal policies may not be returned to the user.

ACKNOWLEDGEMENTS

Conor F. Hayes is funded by the National University of Ireland Hardiman Scholarship. This research was supported by funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program.

REFERENCES

[1] Leon Barrett and Srini Narayanan. 2008. Learning all optimal policies with multiple criteria. In *Proceedings of the 25th international conference on Machine learning*. 41–47.

[2] Marc G Bellemare, Will Dabney, and Rémi Munos. 2017. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 449–458.

[3] Richard Bellman. 1957. *Dynamic programming*. Courier Corporation.

[4] Daniel Bryce, William Cushing, and Subbarao Kambhampati. 2007. Probabilistic planning is multi-objective. *Arizona State University, Tech. Rep. ASU-CSE-07-006* (2007).

[5] Peichen Gong. 1992. Multiobjective dynamic programming for forest resource management. *Forest Ecology and Management* 48, 1 (1992), 43–54. https://doi.org/10.1016/0378-1127(92)90120-X

[6] Conor F. Hayes, Mathieu Reymond, Diederik M. Roijers, Enda Howley, and Patrick Mannion. 2021. Risk-Aware and Multi-Objective Decision Making with Distributional Monte Carlo Tree Search. *In: Proceedings of the Adaptive and Learning Agents workshop at AAMAS 2021)* (2021).

[7] Conor F. Hayes, Mathieu Reymond, Diederik M. Roijers, Enda Howley, and Patrick Mannion. 2021 In Press. 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.

[8] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2021. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *arXiv preprint arXiv:2103.09568* (2021).

[9] Conor F. Hayes, Timothy Verstraeten, Diederik M. Roijers, Enda Howley, and Patrick Mannion. 2021. Dominance Criteria and Solution Sets for the Expected Scalarised Returns. In *Proceedings of the Adaptive and Learning Agents workshop at AAMAS 2021*.

[10] Conor F. Hayes, Timothy Verstraeten, Diederik M. Roijers, Enda Howley, and Patrick Mannion. 2021. Expected Scalarised Returns Dominance: A New Solution Concept for Multi-Objective Decision Making. *arXiv preprint arXiv:2106.01048* (2021).

[11] Michael Painter, Bruno Lacerda, and Nick Hawes. 2020. Convex Hull Monte-Carlo Tree-Search. In *Proceedings of the Thirtieth International Conference on Automated Planning and Scheduling, Nancy, France, October 26-30, 2020*. AAAI Press, 217–225.

[12] Mathieu Reymond, Conor F. Hayes, Diederik M. Roijers, Denis Steckelmacher, and Ann Nowé. 2021. Actor-Critic Multi-Objective Reinforcement Learning for Non-Linear Utility Functions. *Multi-Objective Decision Making Workshop (MODeM 2021)* (2021).

[13] Diederik M. Roijers, Denis Steckelmacher, and Ann Nowé. 2018. Multi-objective Reinforcement Learning for the Expected Utility of the Return. In *Proceedings of the Adaptive and Learning Agents workshop at FAIM 2018*.

[14] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48 (2013), 67–113.

[15] Peter Vamplew, Cameron Foale, and Richard Dazeley. 2021. The impact of environmental stochasticity on value-based multiobjective reinforcement learning. In *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-05859-1

[16] Peter Vamplew, Benjamin J Smith, Johan Kallstrom, Gabriel Ramos, Roxana Radulescu, Diederik M Roijers, Conor F Hayes, Fredrik Heintz, Patrick Mannion, Pieter JK Libin, et al. 2021. Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). *arXiv preprint arXiv:2112.15422* (2021).

[17] Weijia Wang and Michèle Sebag. 2012. Multi-objective Monte-Carlo Tree Search *(Proceedings of Machine Learning Research, Vol. 25)*, Steven C. H. Hoi and Wray Buntine (Eds.). PMLR, Singapore Management University, Singapore, 507–522.

[18] DJ White. 1982. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of mathematical analysis and applications* 89, 2 (1982), 639–647.

[19] Elmar Wolfstetter. 1999. *Topics in Microeconomics: Industrial Organization, Auctions, and Incentives*. Cambridge University Press. https://doi.org/10.1017/CBO9780511625787