Dominance Criteria and Solution Sets for the Expected Scalarised Returns

Conor F. Hayes*[,1], Timothy Verstraeten [2], Diederik M. Roijers[2,3], Enda Howley[1] and Patrick Mannion[1]

[1]School of Computer Science, National University of Ireland Galway, Ireland
[2]AI Lab, Vrije Universiteit Brussel, Belgium
[3]HU University of Applied Sciences Utrecht, the Netherlands
*email: c.hayes13@nuigalway.ie

Introduction

- When a user's preferences over objectives (utility function) are unknown, MORL methods learn a set of optimal solutions
- For the ESR criterion a set of optimal solutions has yet to be defined
- Under the ESR criterion we must consider a distribution over the returns when learning policies
- We use first-order stochastic dominance to define a partial ordering over policies under the ESR criterion and define set of optimal policies

Background

- Multi-Objective Reinforcement Learning
- Utility Functions
- SER and ESR
- Stochastic Dominance

Multi-Objective Reinforcement Learning

- Decision problems with multiple objectives
- Multi-Objective Markov Decision Process
  - M = (S, A, T, γ $\mathcal{R}$)
  - $\mathcal{R}$ is an n-dimensional vector, where n is the number of objectives
- MORL methods can be single policy or multi-policy
- Various MORL scenarios exist depending on the availability of the user's utility function e.g. unknown utility function scenario [4, 3]
- Utility function represents the user's preferences over objectives
- Two optimality criteria exit: SER and ESR

Utility Functions

- Linear utility functions are widely used to represent a user's preferences, $u = \sum_{i=1}^{n} w_i r_i$, where w_i is the preference weight and r_i is the value at position i of the return vector
- In this paper, we consider monotonically increasing utility functions [4], i.e.,

$$(\forall i, V_i^{\pi} \geq V_i^{\pi'} \wedge \exists i, V_i^{\pi} > V_i^{\pi'}) \implies (\forall u, u(\mathbf{V}^{\pi}) > u(\mathbf{V}^{\pi'}))$$

  where $\mathbf{V}^{\pi}$ and $\mathbf{V}^{\pi'}$ are the values of executing policies π and π' respectively
- A monotonically increasing utility function includes both linear and non-linear utility functions

SER and ESR

- Scalarised Expected Returns (SER) is the most commonly used optimality criterion in MORL
- Utility of a user is derived from multiple executions of a policy
- First the expectation is computed, then utility function is applied

$$V_u^\pi = u\left(\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t \mid \pi, \mu_0\right]\right)$$

SER and ESR

- Expected Scalarised Returns (ESR) has largely been ignored by the MORL community
- Utility of a user is derived from the single execution of a policy
- Applies the utility function to the return vector first, and then computes the expectation

$$V_u^\pi = \mathbb{E}\left[u\left(\sum_{t=0}^{\infty}\gamma^t \mathbf{r}_t\right) \mid \pi, \mu_0\right]$$

SER and ESR

- For linear utility functions the policies learned under the ESR and the SER criterion are the same
- For non-linear utility functions the policies learned under the ESR and the SER criterion are different
- In the real world utility functions can be non-linear!

Stochastic Dominance

- Stochastic dominance gives a partial ordering over distributions
  - Can be used to make decisions under uncertainty
- Useful when a distribution must be taken into consideration instead of the expected value
- We focus on first-order stochastic dominance (FSD)
  - $X \geq_{FSD} Y : P(X > z) \geq P(Y > z), \forall z$
  - $X \geq_{FSD} Y : F_X(z) \leq F_Y(z), \forall z$

Expected Scalarised Returns

- ESR and SER criteria learn different policies under non-linear utility functions
- Let's consider the following example:

L_1

$P(L_1 = \mathbf{R})$	$\mathbf{R}$
0.5	(4, 3)
0.5	(2, 3)

L_2

$P(L_2 = \mathbf{R})$	$\mathbf{R}$
0.9	(1, 3)
0.1	(10, 2)

- Utility function : $u(\mathbf{x}) = x_1^2 + x_2^2$

Expected Scalarised Returns

- Under the SER criterion the agent will prefer L_1

$$L_1 : E(L_1) = 0.5(4,3) + 0.5(2,3) = (2,1.5) + (1,1.5)$$

$$L_1 : u(E(L_1)) = (2^2 + 1.5^2) + (1^2 + 1.5^2) = 6.25 + 3.25 = 9.5$$

$$L_2 : E(L_2) = 0.9(1,3) + 0.1(10,2) = (0.9,2.7) + (1,0.2)$$

$$L_2 : u(E(L_2)) = (0.9^2 + 2.7^2) + (1^2 + 0.2^2) = 8.1 + 1.04 = 9.14$$

- Under the ESR criterion the agent will prefer L_2

$$L_1 : u(L_1) = u(4,3) + u(2,3) = (4^2 + 3^2) + (2^2 + 3^2) = (25) + (13)$$

$$L_1 : \mathbb{E}(u(L_1)) = 0.5(25) + 0.5(13) = 12.5 + 6.5 = 19$$

$$L_2 : u(L_2) = u(1,3) + u(10,2) = (1^2 + 3^2) + (10^2 + 2^2) = (10) + (104)$$

$$L_2 : \mathbb{E}(u(L_2)) = 0.9(10) + 0.1(104) = 9 + 10.4 = 19.4$$

Expected Scalarised Returns

- Hayes et al. [2] show that a distribution over the returns is required to learn optimal policies under the ESR criterion
- To understand why a distribution over the returns is needed, let's consider the following example:

L_3

$P(L_3 = \mathbf{R})$	$\mathbf{R}$
0.5	(-20, 1)
0.5	(20, 3)

L_4

$P(L_4 = \mathbf{R})$	$\mathbf{R}$
0.9	(0, 2)
0.1	(10, 2)

- Utility function : $u(\mathbf{x}) = x_1 + x_2^2$
- Both lotteries have the same expected utility of 5. However, both have significantly different outcomes...

Expected Scalarised Returns

- The current MORL literature on the ESR criterion assumes a scalar expected utility
- This is not sufficient to exploit positive outcomes and avoid negative outcomes
- We define a multi-objective version of the value distribution [1]

$$\mathbb{E}\,\mathbf{Z}^{\pi} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} \mathbf{r}_{t} \,\middle|\, \pi, \mu_{0}\right]$$

- $\mathbf{Z}^{\pi}$ is the distribution over returns of a random vector when a policy π is executed
- $\mathbf{Z}^{\pi}$ can be used to learn an optimal set of policies for the ESR criterion

Stochastic Dominance for the Expected Scalarised Returns

- We show that first-order stochastic dominance can be used to give a partial ordering over random vectors

$$\mathbf{X} \geq_{FSD} \mathbf{Y} \Leftrightarrow \forall u : (\forall \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) \geq P(u(\mathbf{Y}) > u(\mathbf{v}))$$

- For MORL we prove the following:

$$\mathbf{X} \geq_{FSD} \mathbf{Y} \implies \mathbb{E}(u(\mathbf{X})) \geq \mathbb{E}(u(\mathbf{Y}))$$

- First-order stochastic dominance can be used to give a partial ordering over policies under the ESR criterion for MORL

$$\mathbf{Z}^{\pi} \geq_{FSD} \mathbf{Z}^{\pi'}$$

Solution Sets for the Expected Scalarised Returns

- The undominated set is a sub-set of all possible policies for where there exists some utility function, u, where a policy's value distribution is FSD dominant

$$U(\Pi) = \left\{ \pi \in \Pi \;\middle|\; \exists u, \forall \pi' \in \Pi : \mathbf{Z}^{\pi} \geq_{FSD} \mathbf{Z}^{\pi'} \right\}$$

- The coverage set is a subset of the undominated set, $U(\Pi)$, where, for every utility function, u, the set contains a policy that has a FSD dominant value distribution

$$CS(\Pi) \subseteq U(\Pi) \wedge \left(\forall u, \exists \pi \in CS(\Pi), \forall \pi' \in \Pi : \mathbf{Z}^{\pi} \geq_{FSD} \mathbf{Z}^{\pi'} \right)$$

- Using FSD to determine an optimal set of policies is difficult since FSD relies on having the utility function of a user available
- To overcome this limitation we define ESR dominance

$$\mathbf{X} >_{ESR} \mathbf{Y} \Leftrightarrow$$

$$\forall u : (\forall \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) \geq P(u(\mathbf{Y}) > u(\mathbf{v}))$$

$$\wedge \exists \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) > P(u(\mathbf{Y}) > u(\mathbf{v})))$$

Solution Sets for the Expected Scalarised Returns

- By introducing a Pareto dominance into ESR dominance it is possible to remove the need to compute the utility

$$\mathbf{X} >_{ESR} \mathbf{Y} \Leftrightarrow$$

$$\forall \mathbf{v} : P(\mathbf{X} >_P \mathbf{v}) \geq P(\mathbf{Y} >_P \mathbf{v}) \wedge \exists \mathbf{v} : P(\mathbf{X} >_P \mathbf{v}) > P(\mathbf{Y} >_P \mathbf{v})$$

- It is possible to extend ESR dominance to value distributions and therefore policies

$$\mathbf{Z}^{\pi} >_{ESR} \mathbf{Z}^{\pi'}$$

- Using ESR dominance, it is possible to define a set of optimal policies, known as the ESR set

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

Conclusion

- This work defined the nessesary notation for the value distribution for the ESR criterion
- We also defined multiple solution sets which prior to this work did not exist
- The solution sets define can now be used to learn an optimal set of policies in MORL when the user's utility function is unknown

Future Work

- We aim to define a MORL algorithm that can learn the ESR set in a bandit setting and MOMDP
- We also hope this work inspires more research into the area of ESR
- The ESR criterion has been ignored by the MORL community and lacks benchmarks, metrics and algorithms
- *In order to fully deploy MORL algorithms in the real world the MORL community needs to consider the ESR criterion!*

References i

- Bellemare, M.G., Dabney, W., Munos, R.: A distributional perspective on reinforcement learning.
  In: International Conference on Machine Learning, pp. 449–458. PMLR, Sydney (2017)
- 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)

References ii

- 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 (2021)
- 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)