

Appendix: Navigating Trade-offs: Policy Summarization for Multi-Objective Reinforcement Learning

Zuzanna Osika^{a,*}, Jazmin Zatarain-Salazar^a, Frans A. Oliehoek^a and Pradeep K. Murukannaiah^a

^aDelft University of Technology

ORCID (Zuzanna Osika): <https://orcid.org/0000-0002-0602-2812>, ORCID (Jazmin Zatarain-Salazar): <https://orcid.org/0000-0003-2248-2242>, ORCID (Frans A. Oliehoek): <https://orcid.org/0000-0003-4372-5055>, ORCID (Pradeep K. Murukannaiah): <https://orcid.org/0000-0002-1261-6908>

A. Additional Results for MO-Highway Environment

Figure 1 shows the trade-offs offered by the policies in the objective space.

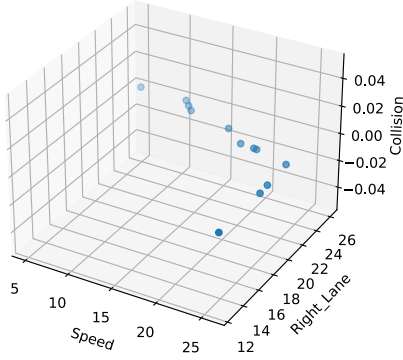


Figure 1: Solution Set consisting of 12 policies represented by the objective vectors in the Objective Space.

Figure 2 displays the normalized distances utilized for k-medoid clustering from Section 5.2. It is evident from the two heatmaps that they appear distinct, indicating the policies are spread differently across the various spaces.

The sankey charts (figure 3) were generated based on the outputs of k-medoid clustering, which can be found in the table 1.

Figure 3 showcases convergence plot of hypervolume over generations for PAN clustering in MO-Highway.

B. PAN parameter configuration

Table 2 showcases parameter configuration for PAN for specific environments.

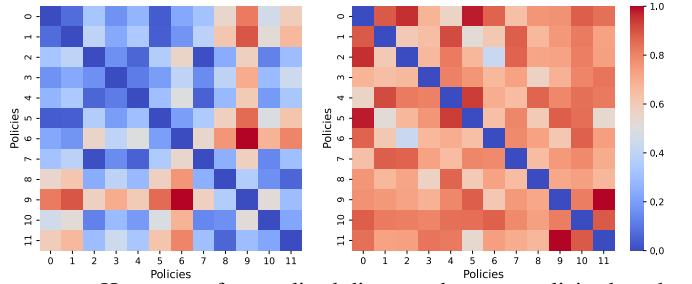


Figure 2: Heatmaps of normalized distances between policies based on different information used (objectives on the left-hand side and highlights on the right-hand side).

k	Clusters: Objective Space	Clusters: behavior Space	ARI
2	[0, 1, 5, 6], [2, 3, 4, 7, 8, 9, 10, 11]	[0, 3, 4, 7, 9], [1, 2, 5, 6, 8, 10, 11]	-0.06
3	[8, 9, 11], [0, 1, 5, 6], [2, 3, 4, 7, 10]	[0, 4, 7], [2, 3, 6, 9], [1, 5, 8, 10, 11]	-0.11
4	[2, 3, 4, 7, 10], [0, 1, 5, 6], [9], [8, 11]	[10], [0, 4, 7], [1, 5, 8, 11], [2, 3, 6, 9]	0.01
5	[6], [2, 3, 4, 7, 10], [0, 1, 5], [8, 11], [9]	[0, 4, 7], [10], [2, 6, 9], [1, 5, 11], [3, 8]	-0.01
6	[2, 4, 7, 10], [8, 11], [6], [0, 1, 5], [3], [9]	[3, 8], [2, 6], [10], [1, 5, 11], [0, 4, 7], [9]	-0.03

Table 1: K-medoids clustering of the policies, with cluster sizes ranging from 2 - 6 done separately for objective space (distance: euclidean distance) and behavior space (distance: Frobenius norm)

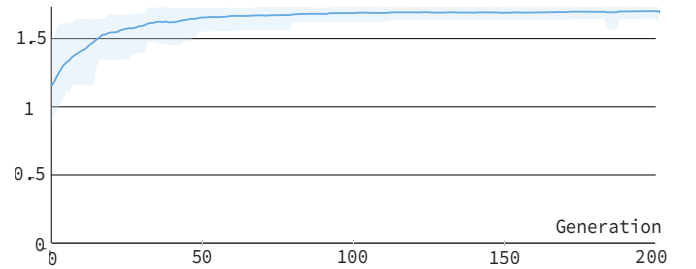


Figure 3: Convergence plot for MO-Highway. Convergence plot showing the mean hypervolume over generations, averaged across multiple seeds (solid line). The shaded area represents the range of performance, illustrating the algorithm's performance variability across different initial conditions.

* Corresponding Author. Email: z.osika@tudelft.nl

environment	g	n	p_r	p_u	p_s	p_m
MO-Highway	200	10	0.7	0.2	0.2	0.6
MO-Reacher	100	10	0.7	0.4	0.4	0.6
MO-Minecart	200	10	0.9	0.6	0.6	0.8
MO-Lunar-lander	500	10	0.9	0.6	0.6	0.8

Table 2: Parameter configuration for PAN for various environments.

C. Additional Environments

All the solution sets were achieved through training a GPI-PD agent with default parameters, as in benchmarks in Felton et al. ([2]).

MO-Minecart In this environment, the agent must collect two types of ores and minimize fuel consumption (3 objectives). The observation is a 7-dimensional vector containing information about the position, speed, and orientation of the cart as well as the contents inside it. The action space is a discrete space with 6 actions, and the reward space is 3-dimensional [1]. The solution set achieved consisted of 302 policies. Figure 4 presents the convergence plot of our clustering approach, where the performance is measured using the hypervolume of the set containing different clustering. The number of generations was chosen based on the tuning of the parameters to achieve the best performance over time. Each point in Figure 5 represents a partitioning of the solution set achieved by PAN and for comparison by iterative k-medoid clustering (as described in section 4). We observe that the majority of the PAN’s clusterings exhibit similar or higher silhouette index values in both spaces. The k-medoids approach demonstrates greater variability, achieving better performance in the behavior space for some clustering. The increased variability in clustering (notably, the clusters being more dispersed) is the reason why, in this specific environment, the k-medoids algorithm outperforms PAN slightly, achieving an improvement of 0.1 in hypervolume.

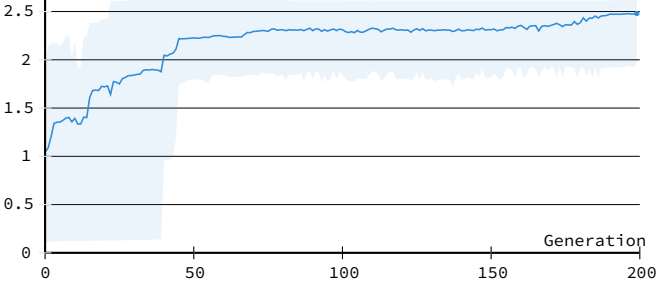


Figure 4: Convergence plot for MO-Minecart. Convergence plot showing the mean hypervolume over generations, averaged across multiple seeds (solid line). The shaded area represents the range of performance, illustrating the algorithm’s performance variability across different initial conditions.

MO-Reacher MO-Reacher is based on the Mujoco’s Reacher, which is a two-jointed robot arm. The goal is to move the robot’s end effector (called fingertip) close to a target that is spawned at a random position. It is a 4-objective problem, where the reward is defined based on the distance of the tip of the arm and the four target locations. The observation is 6-dimensional and contains sin and cos of the angles of the central and elbow joints as well as angular velocity of the central and elbow joints. The solution set achieved consisted of 357 policies. Figure 6 presents the convergence plot of our clustering approach, where the performance is measured using the hypervolume of the set containing different clustering. The number of generations was chosen based on the tuning of the parameters to achieve the best performance over time. Each point in Figure 7

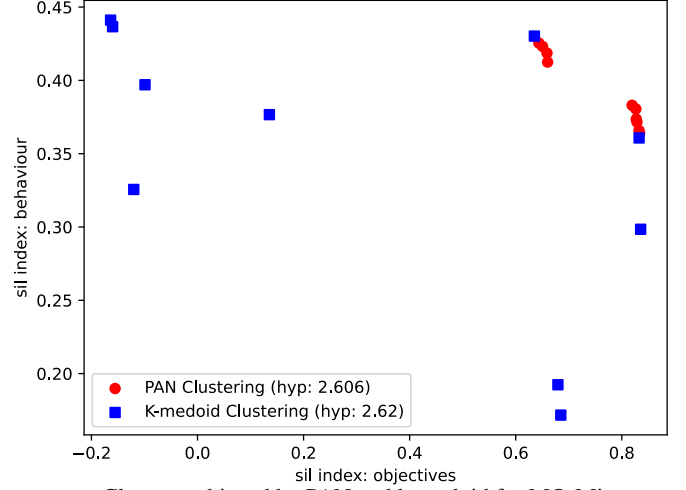


Figure 5: Clusters achieved by PAN and k-medoid for MO-Minecart.

represents a partitioning of the solution set achieved by PAN and for comparison by iterative k-medoid clustering (as described in section 4). We observe that the clusterings of PAN appear to form a Pareto front, presenting trade-offs between silhouette values in the behavior and objective spaces. In contrast, k-medoids clustering performs well either in the objective space or the behavior space but not both. Thus, PAN clustering aligns closely with our desired outcomes.

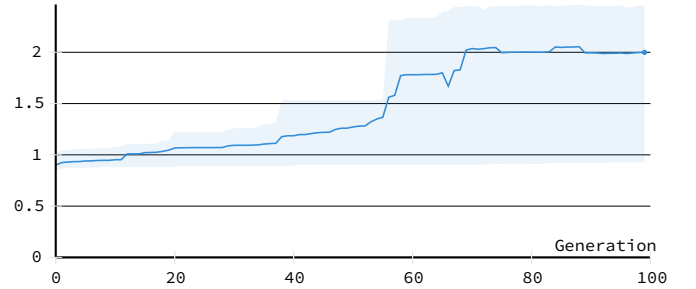


Figure 6: Convergence plot for MO-Reacher. Convergence plot showing the mean hypervolume over generations, averaged across multiple seeds (solid line). The shaded area represents the range of performance, illustrating the algorithm’s performance variability across different initial conditions.

MO-Lunar-lander Multi-objective version of the LunarLander, an environment representing a classic rocket trajectory optimization problem. The reward is 4-dimensional: -100 if crash, +100 if lands successfully, shaping reward, fuel cost (main engine), fuel cost (side engine). The solution set achieved consisted of 157 policies. Figure 6 presents the convergence plot of our clustering approach, where the performance is measured using the hypervolume of the set containing different clustering. The number of generations was chosen based on the tuning of the parameters to achieve the best performance over time. Each point in Figure 7 represents a partitioning of the solution set achieved by PAN and for comparison by iterative k-medoid clustering (as described in section 4). We observe that PAN’s clusterings significantly outperform k-medoids, achieving higher values on both indices for all clusterings and resulting in a much higher hypervolume compared to k-medoids.

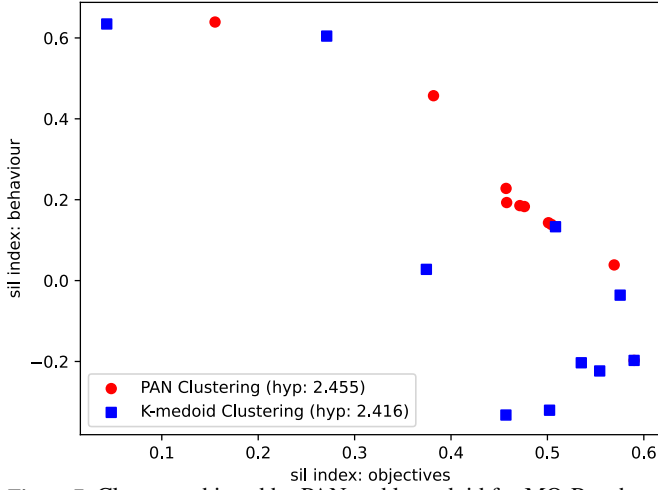


Figure 7: Clusters achieved by PAN and k-medoid for MO-Reacher.

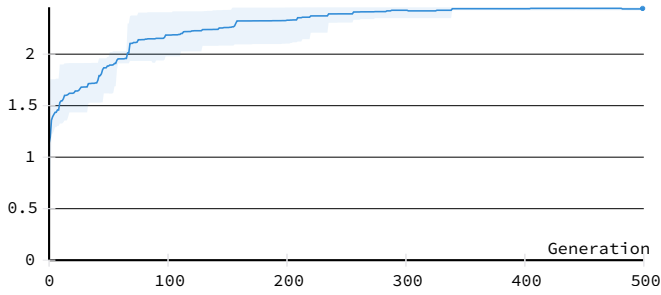


Figure 8: Convergence plot for MO-Lunar-lander. Convergence plot showing the mean hypervolume over generations, averaged across multiple seeds (solid line). The shaded area represents the range of performance, illustrating the algorithm’s performance variability across different initial conditions.

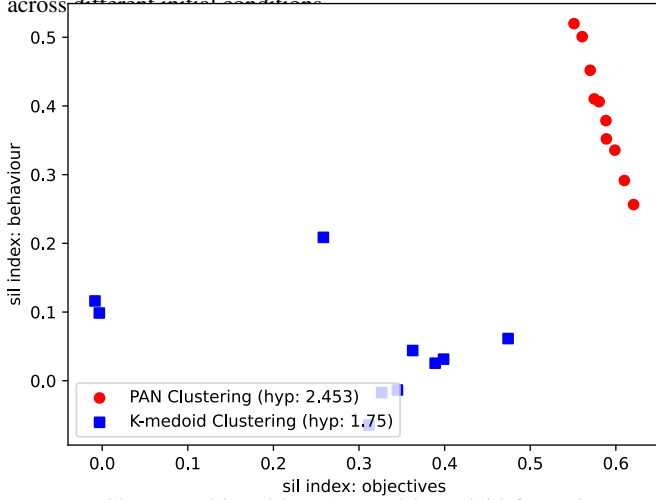


Figure 9: Clusters achieved by PAN and k-medoid for MO-Lunar-lander.

References

- [1] L. N. Alegre, F. Felten, E.-G. Talbi, G. Danoy, A. Nowé, A. L. C. Bazzan, and B. C. da Silva. MO-Gym: A library of multi-objective reinforcement learning environments. In *Proceedings of the 34th Benelux Conference on Artificial Intelligence BNAIC/Benelearn 2022*, 2022.
- [2] F. Felten, L. N. Alegre, A. Nowé, A. L. C. Bazzan, E. G. Talbi, G. Danoy, and B. C. d. Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th*