LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Multi-Objective Reinforcement Learning using Evolutionary Algorithms for Diverse Policy Selection

Master Thesis Defense

Clarissa Kümhof

Table of Contents

1) Motivation
2) Background
   a) (Multi-Objective) Reinforcement Learning (MORL)
   b) Evolutionary Algorithms (EAs)
3) Related Work
4) Methodology
   a) Policy Selection Mechanism
   b) Evolutionary Algorithm
5) Implementation
6) Results and Discussion
7) Future Work

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Motivation

Balancing Trade-offs in a Dynamic World

Motivation

- Real-world problems involve **multiple conflicting objectives**
- Traditional MORL methods have key limitations
  - Output is a single policy based on fixed **predefined preferences** → **less adaptable** to changes
  - No diverse solutions in dynamic environments
- EMOS (Evolutionary Multi-Objective Selector)
  - EAs and policy selection mechanism → **diverse policy population**
  - **Adaptive selection** of policies at each time step

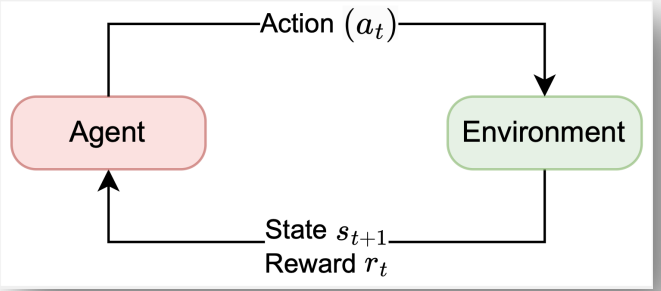
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Background

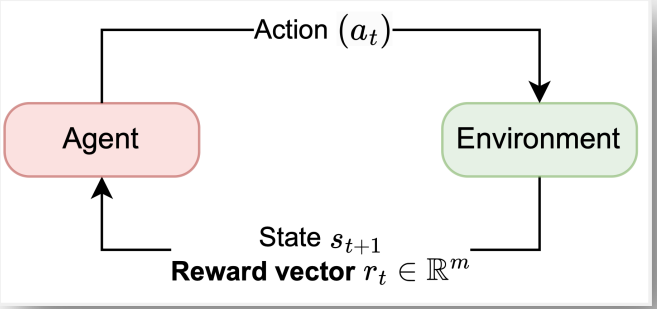
Beyond Reinforcement Learning

Background

- RL trains agents to make **sequential** decisions by optimizing a **reward function** through interactions

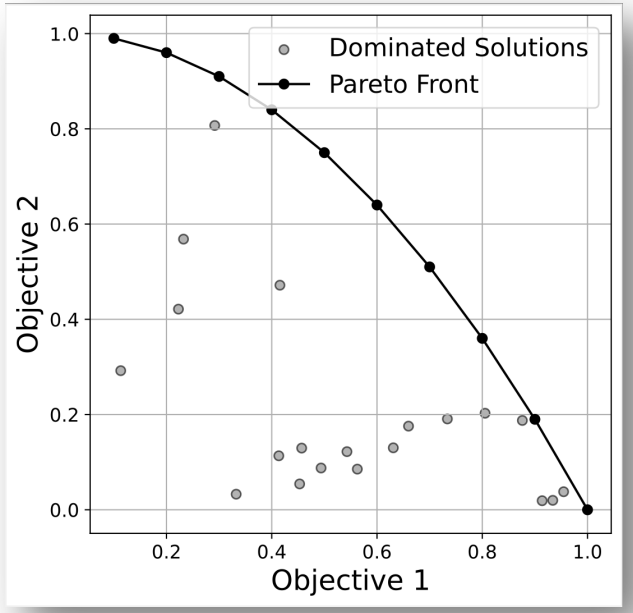


- MORL extends RL to optimize **multiple conflicting** objectives **simultaneously**



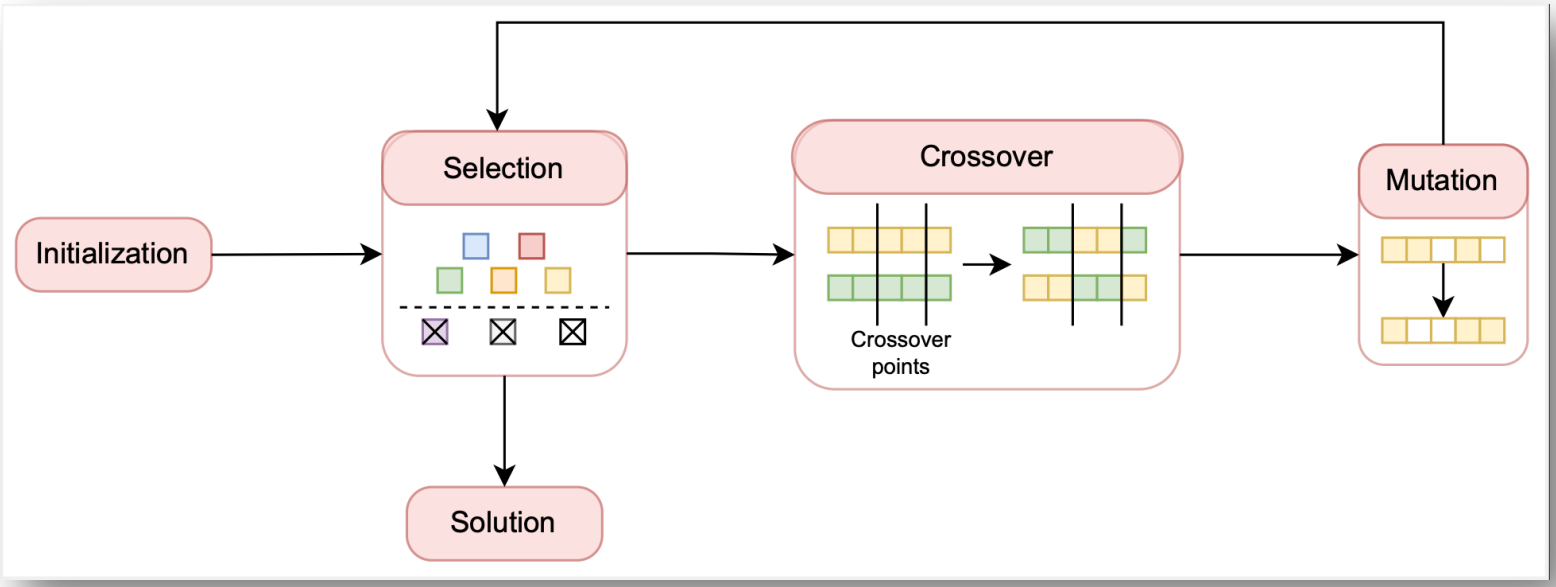
Background

- Pareto Fronts (PFs)
  - Set of dominating policies representing **optimal trade-offs** between multiple objectives
  - Goal is to **approximate** PF with diverse and representative solutions
- Exploration-Exploitation Trade-off



Background

Evolutionary Algorithms (EAs) to **explore** complex solution spaces



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Related Work

Previous Approaches in Multi-Objective Reinforcement Learning

Related Work

- MORL approaches: single- vs. multi-policy
- Evolutionary Reinforcement Learning (ERL)[3]
  - Combines EAs and Deep RL
  - Population evolves via **EAs** while integrating **gradient-based** learning
- Prediction-Guided MORL (PGMORL)[6]
  - Prediction model to guide policy selection
  - Based on **predicted improvements**

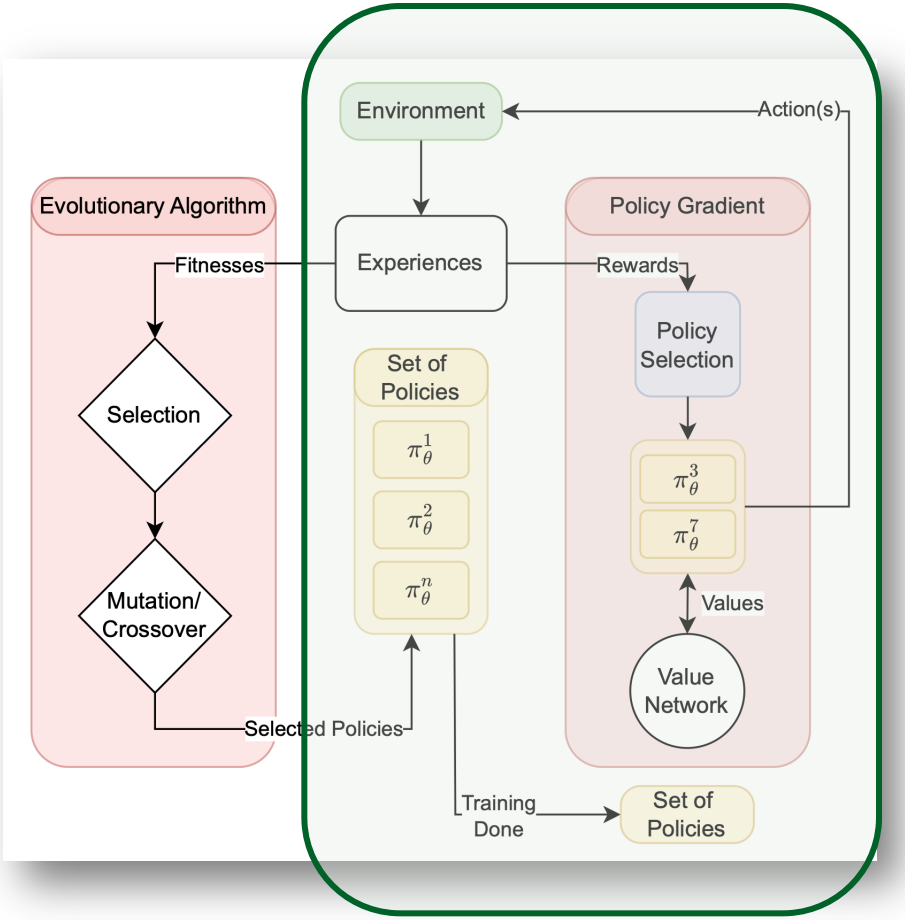
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Methodology

Inside the Framework: *EMOS – Evolutionary Multi-Objective Selector*

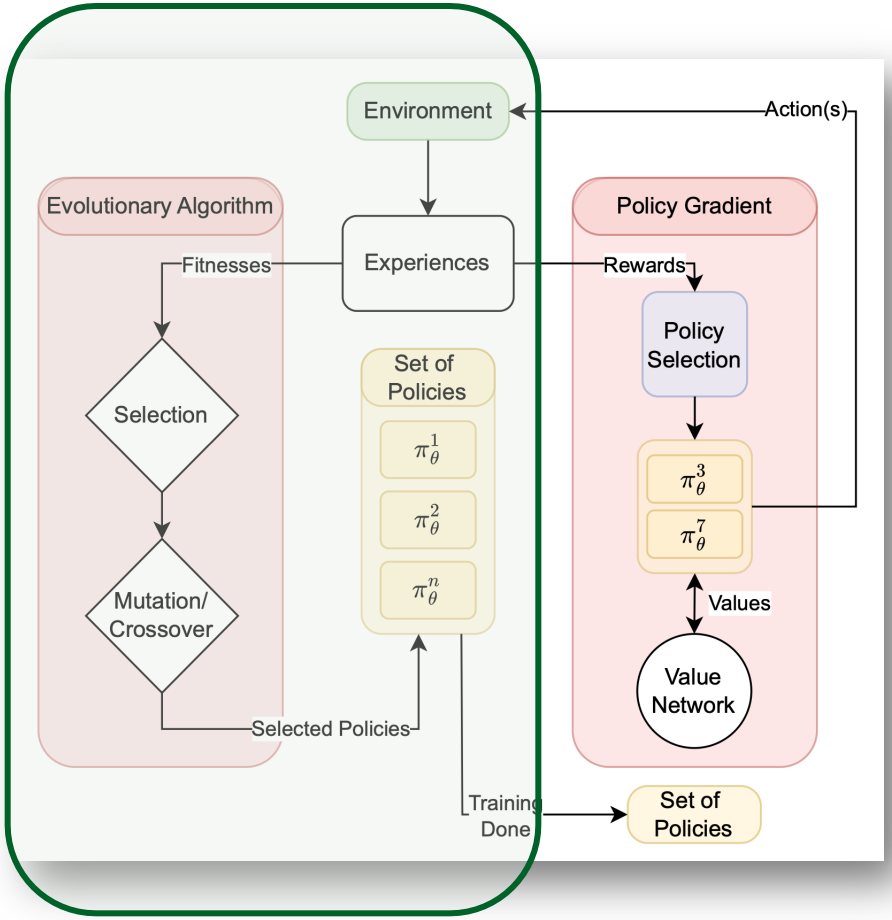
Methodology

- Returning a set of trained policies forming PF
- Using **heuristics** (regret and uncertainty) for policy selection instead of a prediction model (PGMORL[4])
- Dynamically adapting **exploration** and **exploitation** during training



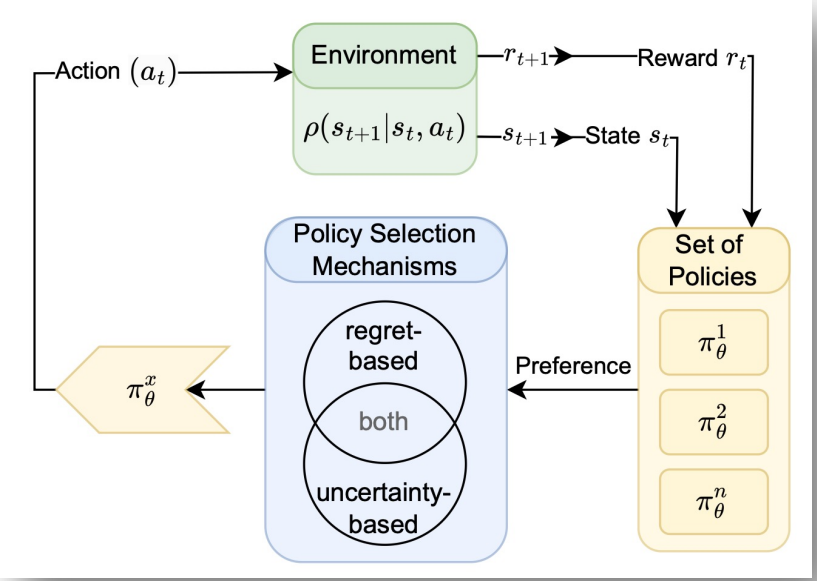
Methodology

- Evolutionary operators to **refine policies** beyond traditional policy gradient methods
  - Fitness = weighted combination of reward and adaptive novelty
  - Multi-point crossover
  - Adaptive mutation rate



Methodology

- Final trained policies enable **dynamic** policy selection based on user preferences at **inference stage**



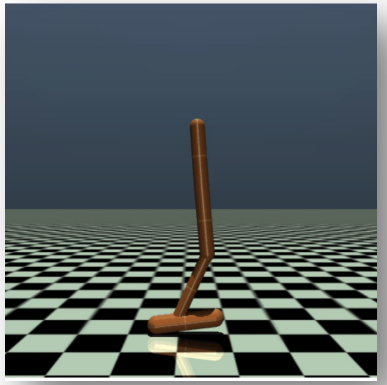
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Implementation

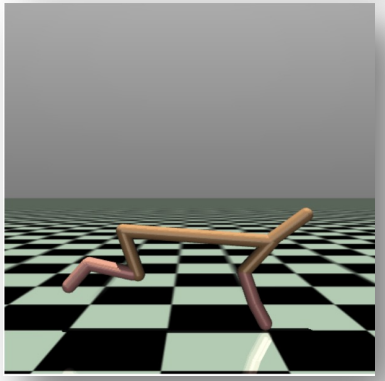
Benchmarking and Performance Evaluation

Implementation

- Evaluation Framework: **MO-Gymnasium API**[1], a library for MORL environments following the Gymnasium API[5] but returning **vectorized rewards**

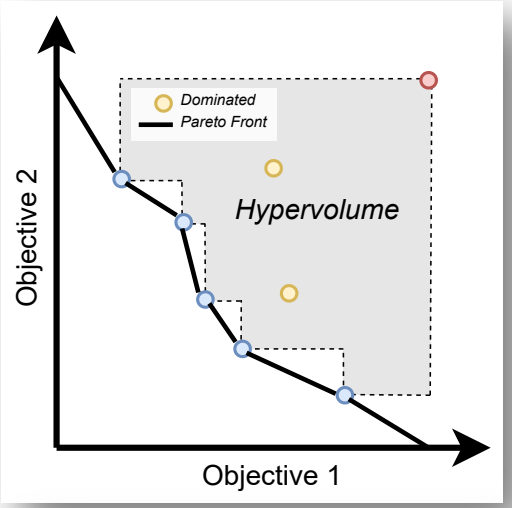


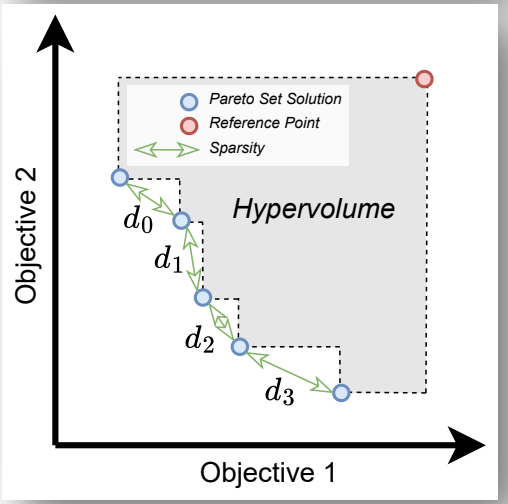
- *MuJoCo*-based environments are ideal for complex, real-world scenarios (*hopper* and *halfcheetah*)



Implementation

- Hypervolume (↑): **volume** covered by PF relative to **reference point** (lower limit of each objective)
- Sparsity (↓): assesses how evenly policies are **distributed** on the PF





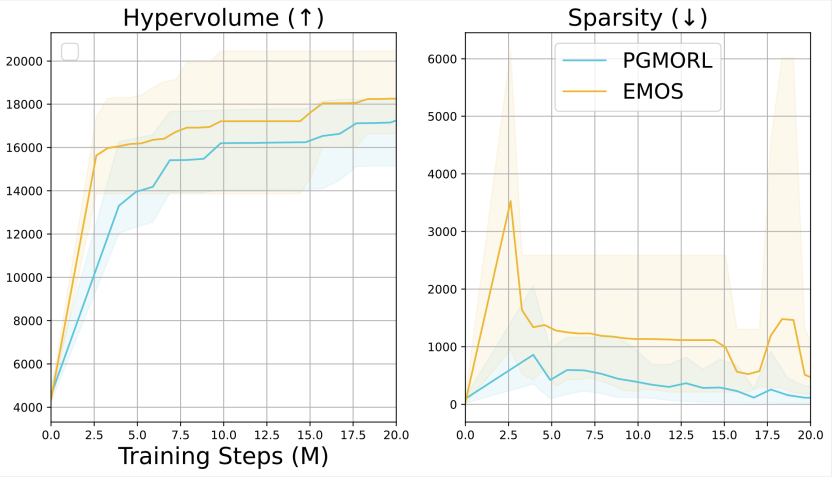
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Results and Discussion

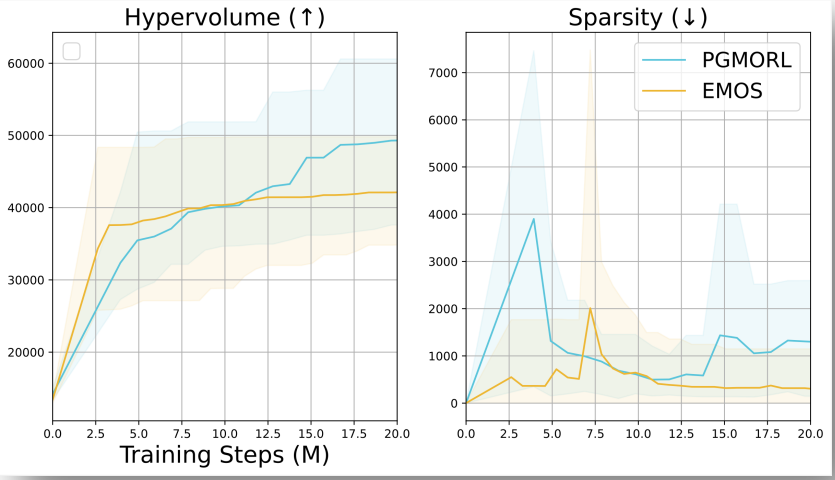
Evaluation: Insights, Hyperparameters, and Comparison with state-of-the-art Algorithms

Results and Discussion

- *Halfcheetah*
  - EMOS consistently outperforms PGMORL due to its EAs enabling **greater diversity**
  - PGMORL lower sparsity, while EMOS exhibits higher sparsity early but **stabilizes**
- *Hopper*
  - EMOS performs better **early in training**, but PGMORL catches up eventually
  - PGMORL's sparsity increases after 10 million steps, indicating a shift from exploration to exploitation → **widely spread solutions**



mo-halfcheetah-v4

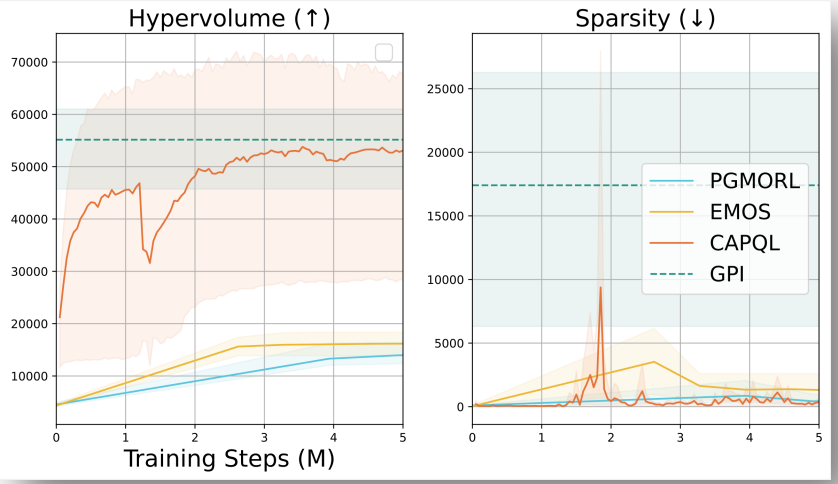


mo-hopper-2d-v4

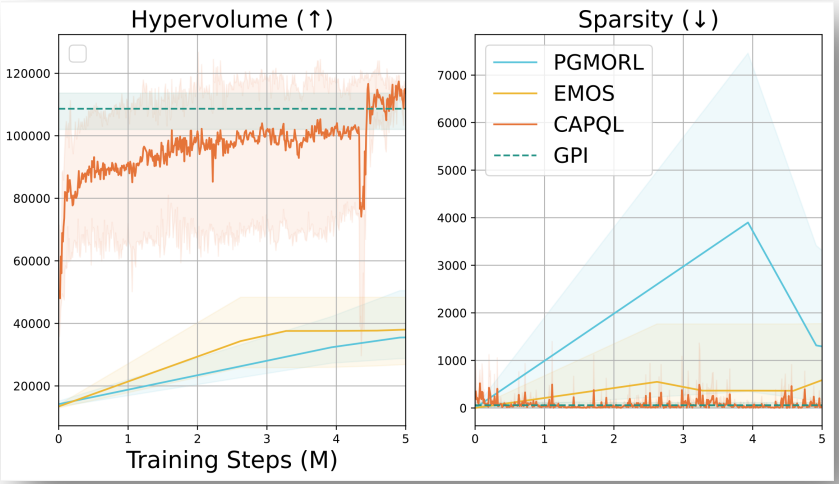
Results and Discussion

- **Concave-Augmented Pareto Q-Learning (CAPQL)**[4] and **Generalized Policy Improvement (GPI)**[2]

- CAPQL and GPI significantly **outperform** both EMOS and PGMORL
- *Halfcheetah*: GPI has extreme sparsity and **high variance**
- *Hopper:* CAPQL briefly drops around 4.4 million steps (**instability**)



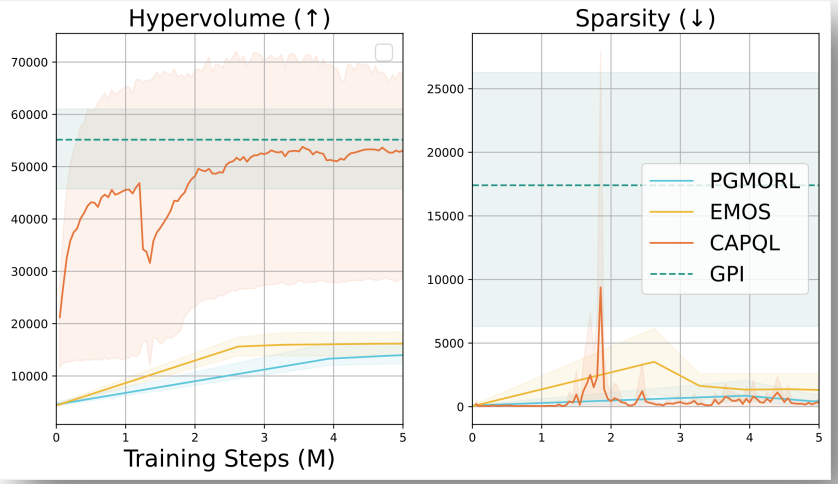
mo-halfcheetah-v4



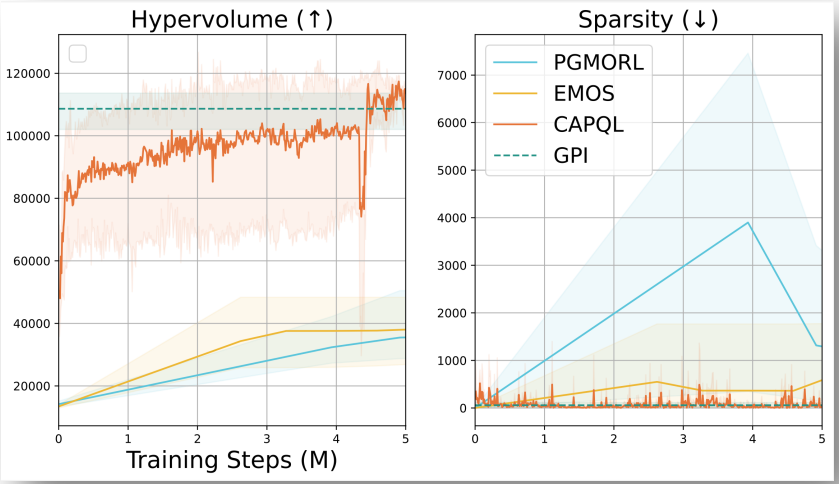
mo-hopper-2d-v4

Results and Discussion

- Superiority of CAPQL and GPI
  - **Off-policy** learning → reusing past experiences, whereas EMOS follows on-policy approach
  - Q-Learning enhances structured learning and **knowledge sharing**, while EMOS relies on EAs (randomness)



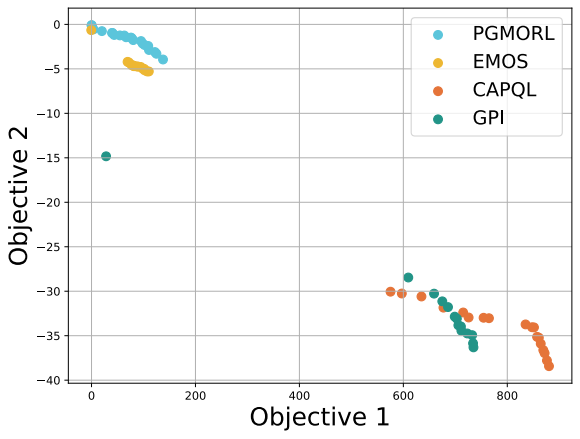
mo-halfcheetah-v4



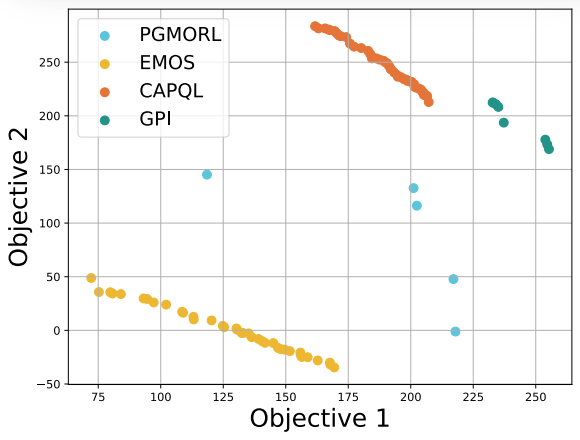
mo-hopper-2d-v4

Results and Discussion

- Best-performing front for each algorithm is selected based on best metric value
- Goal in MORL is not just a well-shaped PF, but one that **maximizes** the evaluation metrics
- Different algorithms explore **different regions** in search space



mo-halfcheetah-v4



mo-hopper-2d-v4

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Key Takeaways

Evaluating EMOS in Multi-Objective RL

Key Takeaways

- EMOS **performs well** compared to PGMORL but struggles against CAPQL and GPI due to its on-policy nature and **evolutionary randomness**
- CAPQL and GPI **dominate** due to their structured learning, knowledge sharing, and sample efficiency
- **Mutation rate** and **novelty weight** significantly impact performance
- Different algorithms explore **different regions** in search space

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Future Work

Expanding MORL’s Potential

Future Work

- Refining policy selection mechanism
  - **Alternative heuristics** beyond regret and uncertainty
- Improved novelty mechanisms
  - Adaptive novelty mechanism that **adjusts based on actual population diversity** rather than global steps
- **Knowledge sharing** across policies (e.g., clustering-based selection)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thank you for your attention!

Master Thesis Defense

Clarissa Kümhof

References

[1] **Alegre**, L. N., Felten, F., Talbi, E. G., Danoy, G., Nowé, A., Bazzan, A. L., & da Silva, B. C. (2022). MO-Gym: A library of multi-objective reinforcement learning environments. In *Proceedings of the 34th Benelux Conference on Artificial Intelligence BNAIC/Benelearn* (Vol. 2022, p. 2).

[2] **Alegre**, L. N., Bazzan, A. L., Roijers, D. M., Nowé, A., & da Silva, B. C. (2023). Sample-efficient multi-objective learning via generalized policy improvement prioritization. *arXiv preprint arXiv:2301.07784*.

[3] **Khadka**, S., & Tumer, K. (2018). Evolution-guided policy gradient in reinforcement learning. *Advances in Neural Information Processing Systems*, *31*.

[4] **Lu**, H., Herman, D., & Yu, Y. (2023). Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*.

[5] **Towers**, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., ... & Younis, O. G. (2024). Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*.

[6] **Xu**, J., Tian, Y., Ma, P., Rus, D., Sueda, S., & Matusik, W. (2020, November). Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning* (pp. 10607-10616). PMLR.