Autonomous Acquisition of Arbitrarily Complex Skills for Continuous Reinforcement Learning Domains

Zeynep Ferah Akkurt
zferah@marun.edu.tr

Bahadır Alacan
bahadiralacan@marun.edu.tr

Merve Rana Kızıl
ranakizil@marun.edu.tr

Advisor: Prof. Dr. Borahan Tümer

Abstract

Real-world problems mostly have a continuous state/action space. Skill Coupling (SC) [1] is a method proposed only for discrete environments. SC solves the oversegmentation problem in Dynamic Community Detection (DCD) algorithms. The motivation of this project is to create a setup for continuous domains before the SC method can be used.

Problem: Continuous State Space
Solution: Since states are Non-Markov, there must be a function approximation or state aggregation process. Fourier Basis is used.

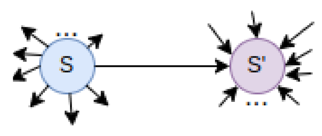


Fig 1. Non-Markov state example

Problem: Community Detection and Skill Construction
Solution: A DCD algorithm named Dynamo [4] can detect communities.

Problem Skill Learning and Learning the Sub-policies of Skills
Solution: With SARSA and Intra-Option Learning, primitive actions and skills can be learned.

Problem: Skill coupling is applicable or not
Solution: After proper setup, SC algorithm can be implemented and examined.

Reinforcement Learning

Reinforcement learning (RL) is a type of machine learning technique where an agent takes an action in an environment, moves to the next state, and receives the environment's feedback (reward or punishment) regarding that action.

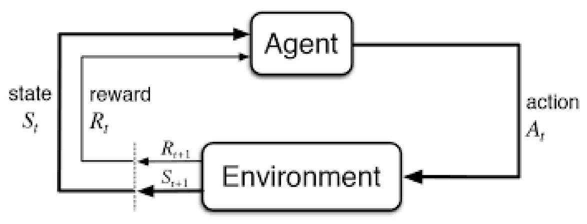


Fig 2. RL Diagram

Hierarchical Reinforcement Learning (HRL)

As the environment grows too large, converging to a satisfactory policy for regular RL algorithms such as flat Q–learning becomes quickly infeasible.

In HRL [3]:

- Environment is split into sub-regions (communities).
- A sub-policy (sequence of primitive actions) is learned for each sub-region.
- The sequence of primitive actions is called skill/option.

An option consists of three components: a policy π: S×A → [0,1], an initiation set IS and a termination condition β: S+ → [0,1]

Continuous Environment: Pinball Domain

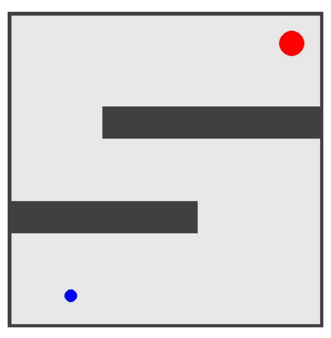


Fig 3. Env 1

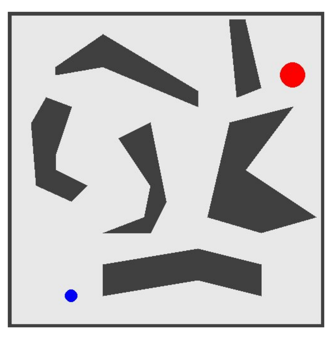


Fig 4. Hard env 2

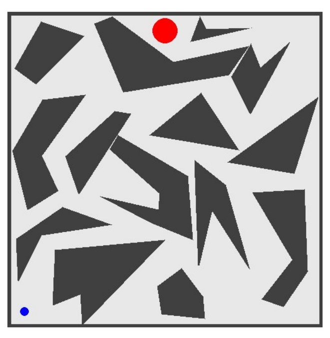


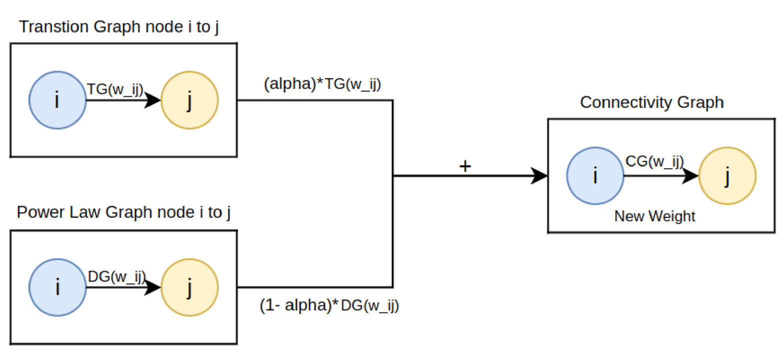
Fig 5. Very hard env 3

Pinball is one of the most challenging environments for RL algorithms because of its dynamic aspect, sharp discontinuities, and extended dynamics control characteristics.

Actions: There are five primitive actions: adding or subtracting a small force to x velocity or y velocity, or leaving them unchanged
Representation of a state (4D) = (x_coordinate, y_coordinate, x_velocity, y_velocity)
Goal: Manoeuvre the blue ball into the red hole

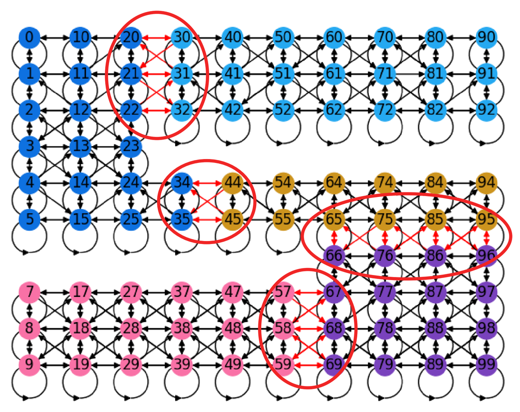
Connectivity Graph (CG) for Community Detection

The combination of a transition Graph (TG) and power-law distance graph (DG) is called a connectivity graph [2]. This graph will be the input to the our DCD algorithm that finds the communities. This method is called Graph-Based Skill Learning (GSL) [2].



$$CG(w_{ij}) = \alpha_w * TG(w_{ij}) + (1 - \alpha_w) * DG(w_{ij})$$

Fig 6. Connectivity graph generation with its formula



State coordinates: (0.x, 0.y) → label = xy

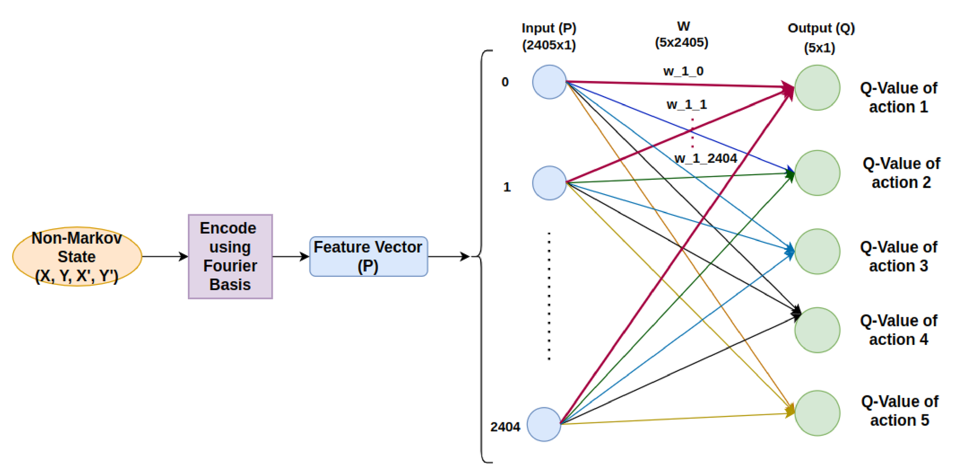
Fig 7. Partial Graph

Function Approximation for Skill Learning: Fourier Basis

Table 1. Qtable for an environment that has a finite number of states

	Action 1	Action 2	Action 3
State 1	0	3.1	-2.5
State 2	12.5	53.2	0
State 3	-8.1	7.6	5

Since the states cannot be represented with a Q-table like in Table 1, in continuous domains, function approximation methods that return the Q-value of a given encoded state for actions are required.

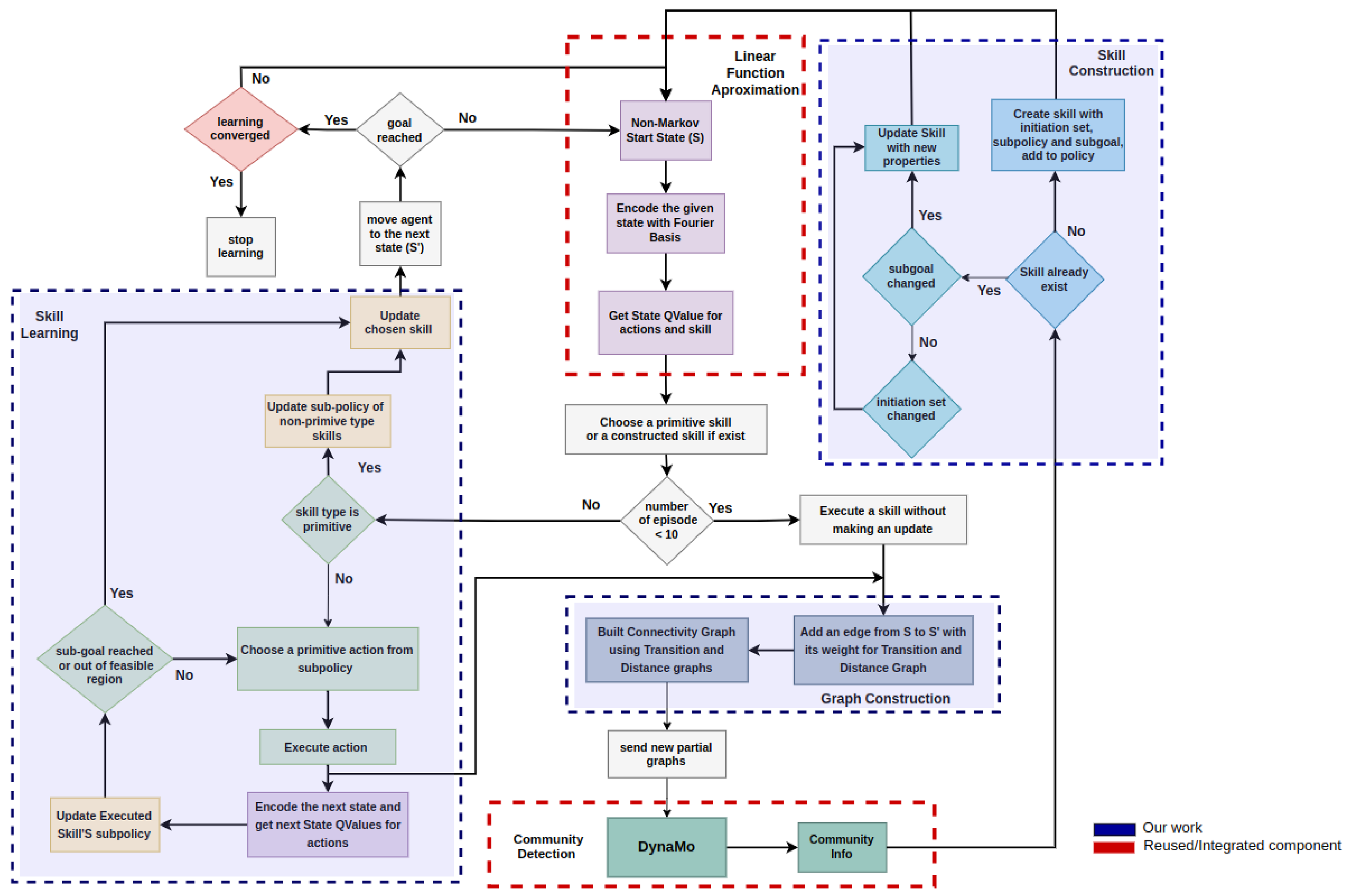


$$Q = W \cdot P$$

$$W = \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \end{bmatrix} \text{ where } W_i = [w_{i0}, w_{i1}, \dots w_{i2404}] \text{ and } Q = \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \\ Q_5 \end{bmatrix}, \; Q_i\text{: QValue of action } i$$

Fig 8. Q-value approximation using Fourier basis

Flow Chart



Experimental Results

Community Detection and Skill Learning

Primitive action · Skill

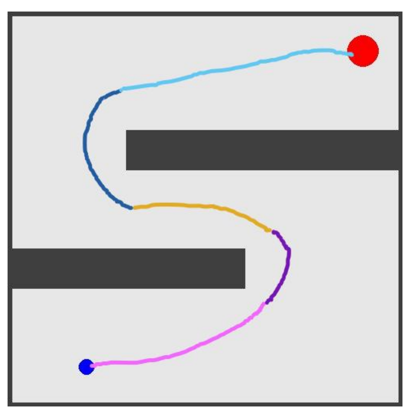


Fig 9. Optimal path

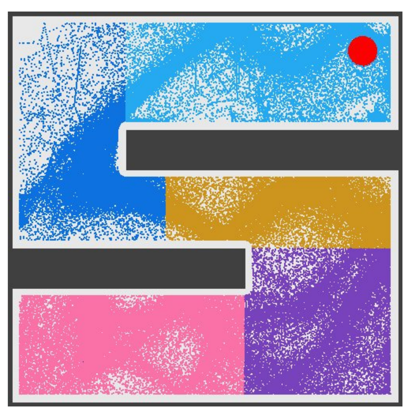


Fig 10. Detected communities

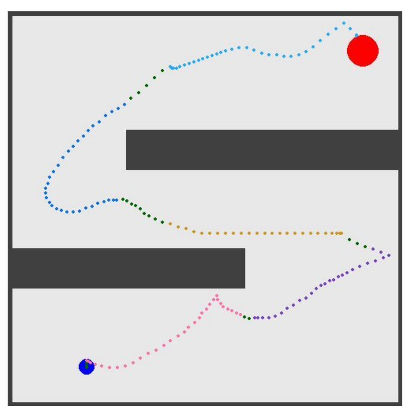


Fig 11. Sub-optimal converged path

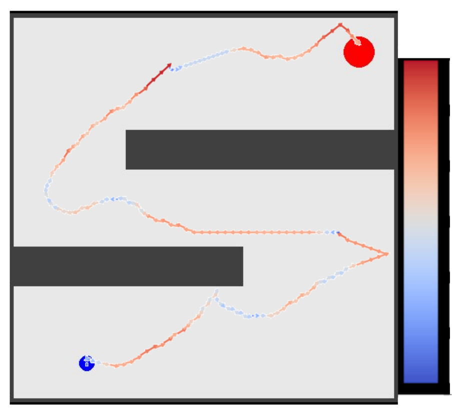


Fig 12. Path with velocities

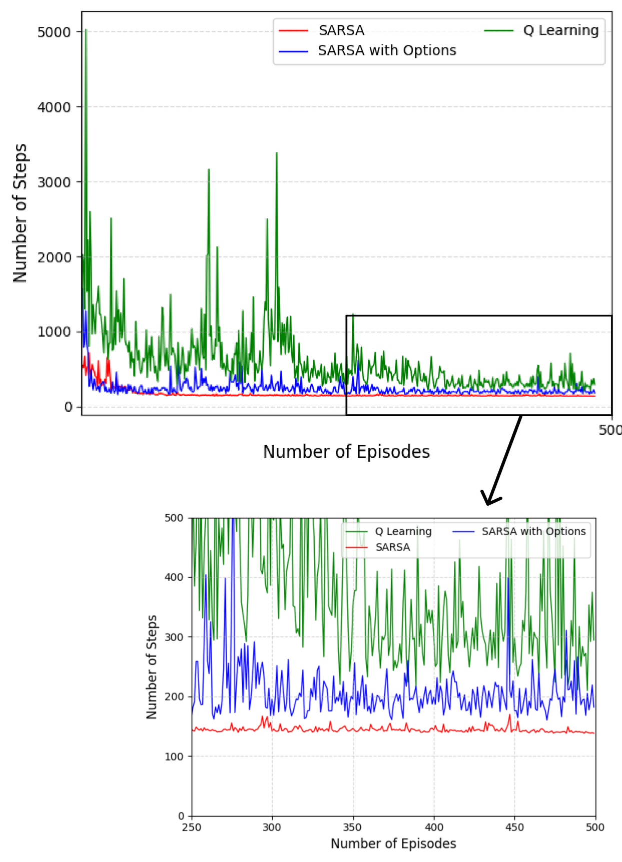


Fig 14. Number of steps graph of 10 experiments

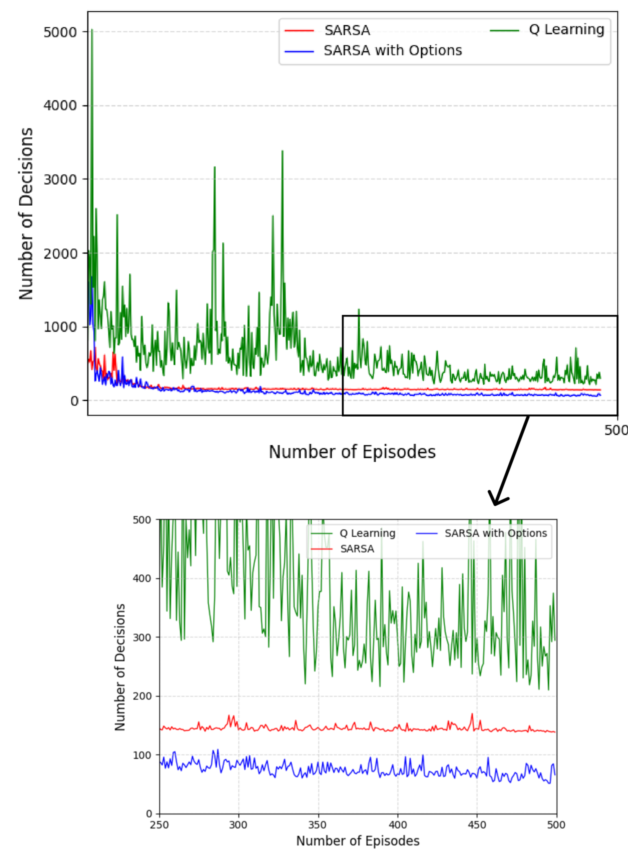


Fig 15. Number of decisions graph of 10 experiments

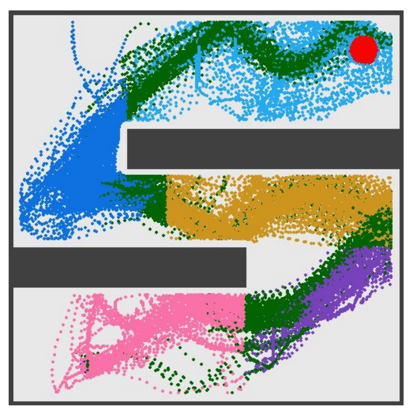


Fig 13. Band of the converged path

Table 2. Comparison of algorithms in env1 in terms of, number of steps to the goal state, and number of decisions.

Agent	# steps	# decision
GSL(TG+DGPower-Law)	205	6.1
SARSA with Options	200.15	71.63
SARSA	143.90	143.90
Q-Learning	347.99	347.99

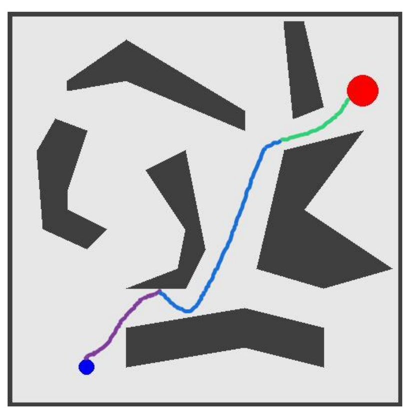


Fig 16. Optimal path

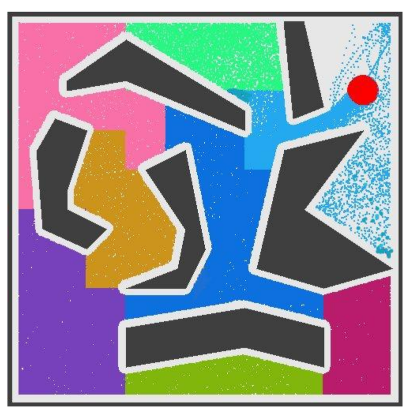


Fig 17. Detected communities

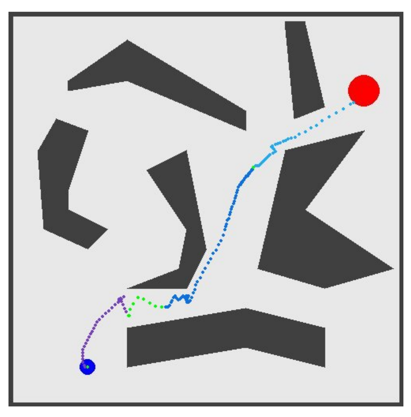


Fig 18. Sub-optimal converged path

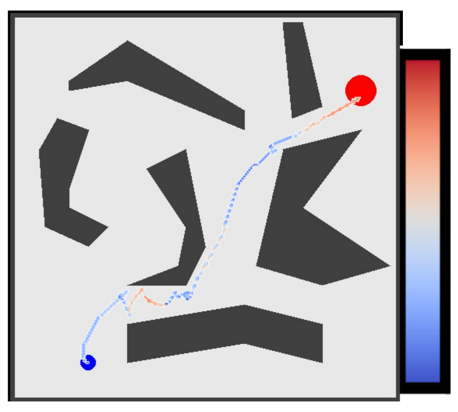


Fig 19. Path with velocities

Conclusion

- Community detection using connectivity graph with Dynamo does not have the problem of oversegmentation. So, skill coupling is not needed for env 1.
- SARSA with skills algorithm for pinball domain converges to a sub-optimal policy.
- CG, DynaMo and an algorithm like Intra-Option Learning work well together.

Future Work

- Skill coupling will be enabled for hard environments.
- Experiments with different environment settings
- Improvements for sub-policies of skills

Technologies Used

Acknowledgement

We would like to thank Kutalmış Coşkun, Zeynep Kumralbaş and Hazel Çavuş for their precious contributions to the project.

References

[1] Z. Kumralbaş , S. H. Çavus , K. Coşkun, B. T ümer, Autonomous acquisition of arbitrarily complex skills using locality based graph theoretic features: a syntactic approach to hierarchical reinforcement learning, Evolving Systems (2023) 1–24.
[2] F. Shoeleh, M. Asadpour, Graph based skill acquisition and transfer learning for continuous reinforcement learning domains, Pattern Recognition Letters 87 (2017)104–116.
[3] R. S. Sutton, D. Precup, S. Singh, Between mdps and semi-mdps: A frameworkfor temporal abstraction in reinforcement learning, Artificial intelligence 112 (1999)181–211.
[4] D. Zhuang, J. M. Chang, M. Li, Dynamo: Dynamic community detection by incre-mentally maximizing modularity, IEEE Transactions on Knowledge and Data Engi-neering 33 (2019) 1934–1945.