

The Impact of Initial Start Distribution Mismatch on Policy Evaluation in Behavior-agnostic Reinforcement Learning

Tiberiu Sabău
T.Sabau@student.tudelft.nl

Responsible Professor: Frans Oliehoek
Supervisor: Stephan Bongers



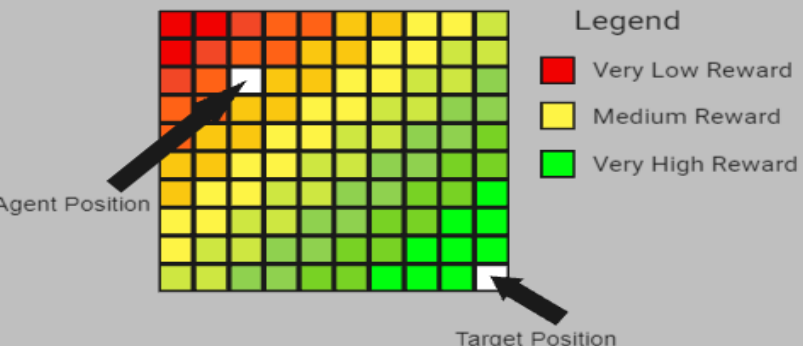
Introduction

- Off-policy learning → capacity to learn from past experiences [1].
- Behavior-agnostic RL → developing algorithms capable of learning effective policies (target policies) without explicit knowledge of the environment or specific behavior policies [3].
- Distribution Correction Estimation (DICE) → correcting the mismatch between the state-action distributions of the behavior and target policy [2].
- **Research Question:** *How does the mismatch in initial start distribution affect the performance of the DICE estimators in off-policy evaluation?*

Background

- Infinite-horizon Markov Decision Process (MDP) → (S,A,R,T, μ_0 , γ): state space, action space, reward function, transition probability function, initial start distribution, and discount factor [3].
- μ_0 → probability distribution that the MDP will start in each state, for the target policy (π).
- Let $D = \{ (s_0^{(i)}, s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)}) \}_{i=1}^N$, where the starting states $s_0^{(i)} \sim \mu'_0$ are samples from some initial start distribution μ'_0 , $(s^{(i)}, a^{(i)}) \sim d^{\pi^b}$ are samples from some distribution d^{π^b} , $r^{(i)} = R(s^{(i)}, a^{(i)})$, and $s'^{(i)} \sim T(s^{(i)}, a^{(i)})$, N is the number of episodes [3].
- Off-policy evaluation aims to evaluate π by sampling experiences from D , which is created using behavior policy π^b .
- Start distribution mismatch → disparity between the starting states observed by an agent in a dataset created using the behavior policy, and the starting states it would encounter in a dataset created using the target policy.

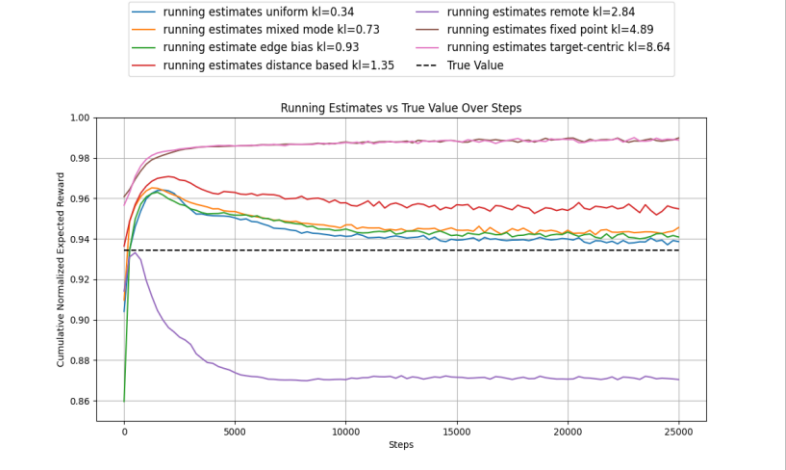
Methodology

- Environment: 10 x 10 grid.
 - The agent is being rewarded for moving closer to the target.
- 
- 7 different initial start distributions were systematically generated and fixed for π^b , and 1 for π .

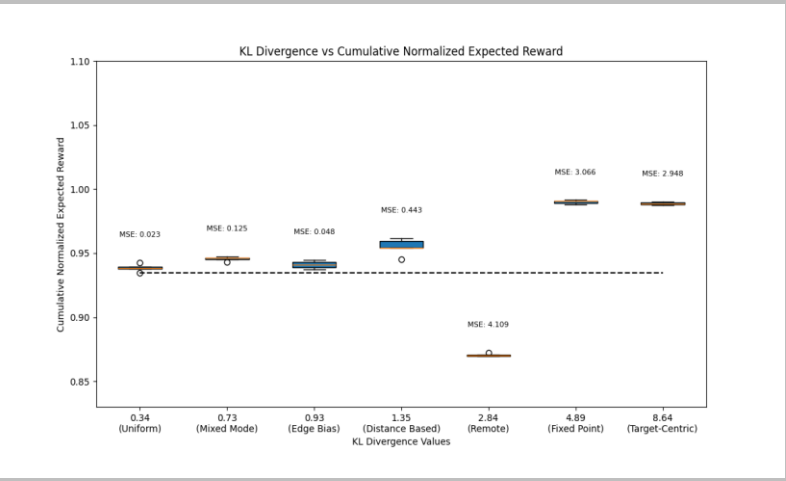
Distribution	Description
Uniform	Equal probability for each state.
Edge Bias	Higher probability at the edges.
Distance Based	Inversely proportional to distance to goal.
Target-Centric	(x, y) always from [8, 9].
Remote	(x, y) always from [0, 2].
Fixed Point	(x, y) always (8, 9).
Mixed	Average of first 3 distributions.
Uniform (for π)	Equal probability for each state.

- The initial start distribution mismatch was computed using KL divergence.
- The cumulative normalized expected reward was computed using the DICE estimators.
- The performance of the DICE estimators is assessed using MSE, based on 5 dataset samples.

Results



- For all but 1 KL value, the estimation is higher than the true value. While for lower KL divergence values the estimation seems to converge closer to true value, for high KL divergence values no trend can be identified.



- While smaller KL divergence values seem to lead to more accurate estimations and lower MSE, higher KL divergence values do not follow a consistent pattern. The highest MSE corresponds to the third-highest KL divergence value and is also the only box plot below the ground truth. The second and third highest MSE values correspond to the two highest KL divergence values. Overall, the results do not indicate any trend that would suggest the KL divergence value has a significant influence on the MSE.

Conclusion

- Based on the results, the initial start distribution mismatch does not appear to significantly impact the performance of DICE estimators. Consequently, further research is needed to expand the experiments and explore this relation further.
- Future research can enhance the study by experimenting on multiple discrete and continuous environments, using different measures for state visitation mismatch and performance, as well as using more dataset samples. Another interesting direction would be to analyze the relation between the start distribution and the policies.

Limitations

- Choice of environment.
- Number of dataset samples.
- Policies.
- Choice of measures.
- Length of each dataset.

References

[1] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience, 2019.

[2] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[3] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian, 2020.