The Effect of State-visitation Mismatch on Off-policy Evaluation in Behaviour-agnostic Reinforcement Learning

Author: Kevin Yi Chen (kychen@student.tudelft.nl), **Supervisor:** Stephan Bongers, **Responsible Professor:** Frans Oliehoek

TU Delft

1. Introduction

- Reinforcement learning (RL) has achieved significant successes in various domains but many real-life applications are too costly/risky to directly interact with the environment to generate training data.
- Behaviour-agnostic RL addresses this challenge by separating the behaviour policy used for training from the target policy used for performance evaluation.
- This is called off-policy evaluation and introduces differences between the probability distributions of states visited by the policies (state-visitation mismatch).
- A method was developed to correct for these mismatches even for infinite horizons [1].
- This method was then used to create the DICE estimators that reduce variance and bias to provide more accurate estimations [2].

2. Research Question

How does the degree of state-visitation mismatch impact the performance of target policies in behaviour-agnostic off-policy evaluation?

3. Methodology

The following **metrics** were used to quantify key aspects:

- **State-visitation mismatch:** The **KL divergence** measures the difference between 2 probability distributions.
- **Target policy performance:** The **cumulative reward** is the estimate of the target policy performance given by the DICE estimator.
- **Effect of state-visitation mismatch on target policy performance:** The **mean squared error (MSE)** calculates the difference between the empirical and estimated cumulative reward.

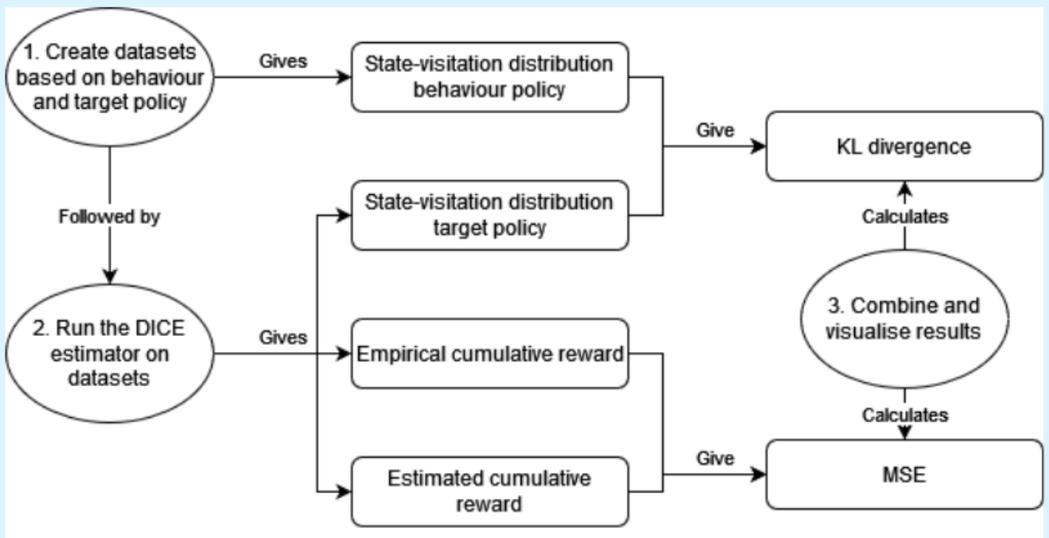

Figure 1: A flowchart representing the procedure to obtain the results to answer the research question.

4. Results

The following **variables** were used to generate datasets and run the DICE estimator on them:

- **Environment:** Frozen Lake
- **Environment size:** The dimensions of the Frozen Lake environment
- **Alpha (α):** How close the behaviour policy is to the target policy
- **Number of datasets:** The number of datasets to generate per alpha (α) value

Figure 2: The default Frozen Lake environment

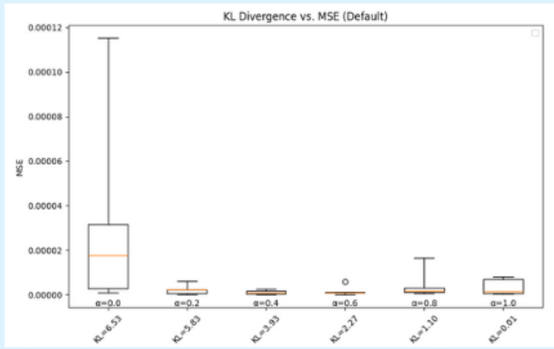

Figure 3: The KL divergence plotted against the MSE for various alpha (α) values with the **default** environment.

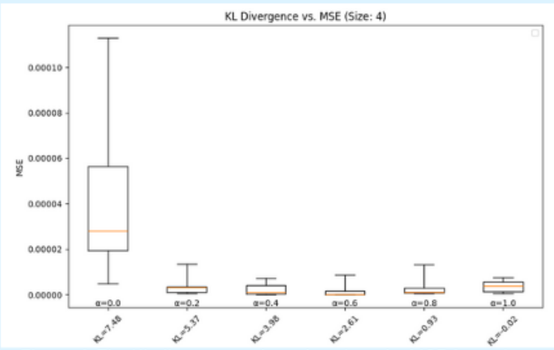

Figure 4: The KL divergence plotted against the MSE for various alpha (α) values with a random **4x4** environment.

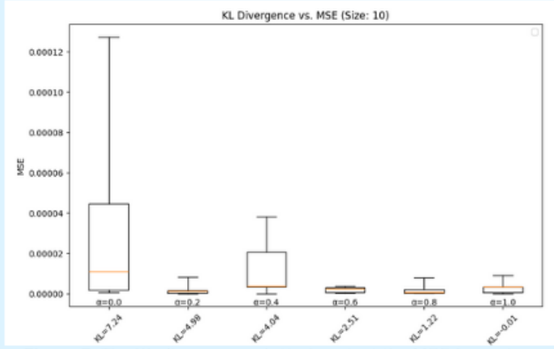

Figure 5: The KL divergence plotted against the MSE for various alpha (α) values with a random **10x10** environment.

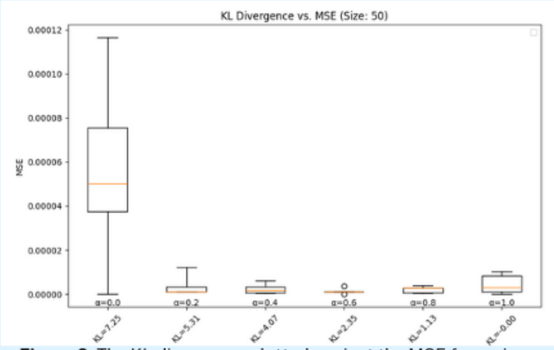

Figure 6: The KL divergence plotted against the MSE for various alpha (α) values with a random **50x50** environment.

5. Conclusion & Future Work

Limitations:

- The KL divergence only works on discrete environments and not continuous ones.
- The experiment only uses 1 environment and 1 DICE estimator which could influence the results.

Conclusion:

- The results suggest that the state-visitation mismatch may influence the target policy performance.
- However, the research is inconclusive due to the limitation regarding the variety of datasets.

Future Work:

- Run the experiment with different estimators and on multiple environments that are bigger or more complex.
- Consider other metrics for the state-visitation mismatch that work on continuous environments.

[1] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation, 2018.
[2] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian, 2020.