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

Tiberiu Sabău
Supervisor(s): Frans Oliehoek, Stephan Bongers
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Tiberiu Sabău
Final project course: CSE3000 Research Project
Thesis committee: Frans Oliehoek, Stephan Bongers, Catholijn Jonker

Abstract

Behavior-agnostic reinforcement learning is a rapidly expanding research area focusing on developing algorithms capable of learning effective policies without explicit knowledge of the environment's dynamics or specific behavior policies. It proposes robust techniques to perform off-policy evaluation, namely Distribution Correction Estimation (DICE) methods, in the context of infinite horizon Markov Decision Processes (MDPs). This research paper investigates the impact of the initial start distribution mismatch on the accuracy of DICE estimators in behavior-agnostic reinforcement learning. To achieve this, seven systematic initial start distributions were created and utilized to calculate the initial start distribution mismatch via Kullback–Leibler (KL) divergence. Furthermore, off-policy evaluation performance was assessed using DICE estimators, with Mean Squared Error (MSE) comparisons against ground truth values. The study reveals that, based on the conducted experiments, the initial start distribution mismatch does not have a clear influence on the performance of the DICE estimators. Therefore, future research is required to increase the scope of the experiments and address some of the limitations of this study to accurately assess the impact of the initial start distribution mismatch on off-policy evaluation using DICE methods. This paper underscores the complexity of the initial start distribution choice in behavior-agnostic reinforcement learning, calling for further research to effectively evaluate its impact across diverse environments and measures. Additionally, exploring the relation between the initial start distribution and policies could provide deeper insights and further refine the understanding of their influence on DICE estimators.

1 Introduction

Reinforcement learning (RL) algorithms have received significant attention recently for demonstrating important successes in various real-life domains such as robotics [1], and games [2], among others. Off-policy learning represents one of the diverse approaches within RL, distinguished by its capacity to learn from past experiences, even when they were generated by potentially unknown behavior policies [3]. By separating the policy used for exploration from the one being optimized, these algorithms can leverage valuable experience from diverse sources. This has inspired significant interest in behavior-agnostic RL, a rapidly expanding area of research aimed at developing algorithms capable of learning effective policies without explicit knowledge of, or reliance on, the underlying dynamics of the environment or specific behavior policies [4].

Despite the challenges of behavior-agnostic off-policy reinforcement learning, several estimators have been developed over time for this scenario. Those estimators are part of Distribution Correction Estimation (DICE) and are designed to address these challenges by quantifying and correcting the mismatch between the state-action distributions of the behavior policy and the target policy. These estimators aim to improve the efficiency and effectiveness of learning algorithms in behavior-agnostic RL scenarios [5]. These methods operate within an infinite horizon Markov Decision Process (MDP), which consists of a state space, action space, reward function, transition probability function, initial state distribution, and discount factor [6].

The impact of the initial state distribution on the learning objective has been investigated in prior studies [7]. Leveraging experience memory to facilitate exploration by adapting the restart distribution, these approaches have shown promise in improving learning performance. Additionally, earlier works have explored the use of reset capacity and initial start distributions to enhance learning outcomes [8]. However, the extent to which the initial start distribution mismatch impacts the policy evaluation and state-action visitation mismatch, particularly in the context of DICE methods is yet to be assessed.

Within this research paper, the main question that is being answered is *How does the mismatch in initial start distribution affect the performance of the DICE estimators in off-policy evaluation?*. The main findings of this research can be used to understand the correlation between the state-action visitation distribution and the initial start distribution and, what is the impact of the initial start distribution on the accuracy of estimating the target policy. Specifically, the methods considered for this research are the DICE estimators, as showcased in [4].

This research paper is structured as follows. Section 2 provides a background into this field of study, with an emphasis on behavior-agnostic off-policy reinforcement learning, while reiterating the key concepts of existing research. Section 3 outlines the research methodology, detailing the selection of measures used in the experiment. Section 4 describes the experiments used to answer the (sub)questions. This involves showcasing the environment, different approaches to generate different initial start distributions, and providing context on how the measures described in the methodology were used. Section 4 also outlines the results following the experimentation. Section 5 places the results in a broader context and includes a reflection on the performed research, including limitations and future work. Section 6 delves into the application of responsible research throughout the experimenting and documentation process. Section 7 brings the report to a close by summarizing the main findings of the study.

2 Background

2.1 Problem Setting

An infinite horizon Markov Decision Process (MDP) can formally be described as a tuple $M = (S, A, R, T, \mu_0, \gamma)$ [6]. The main components are the state space S, representing all possible states in the environment, an action space A, representing all possible actions the agent can take, a reward function $R(s_i, a_i)$, which provides a reward for each action taken in a given state, a transition probability function $T(s_{i+1} \mid s_i, a_i)$, which specifies the probability of transitioning to state s_{i+1} from state s_i after taking action a_i, an initial state distribution μ_0, which defines the probability distribution over initial states, and a discount factor $\gamma \in [0, 1]$, which represents the importance of future rewards compared to immediate rewards. A policy π controls the agent's behavior, by defining the probability of taking each action in each state. The interaction with the environment begins by sampling an initial state s_0 from the initial state distribution μ_0. At each step $i \geq 0$, the agent selects an action according to the policy π, receives a corresponding reward r_i, and transitions to a new state based on the transition probability function [4]. This process continues for a fixed number of steps, generating a sequence of states, actions, and rewards. Such a process is called an episode or trajectory.

2.2 Policy Evaluation

The value of a policy π is defined as the expected average reward it generates per step [4], also known as normalized expected cumulative reward [5], which can be expressed as follows:

$$\rho(\pi) := (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 \sim \mu_0, \forall t, a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t)\right]. \tag{1}$$

In the context of policy evaluation, the policy under assessment is known as the target policy. The value of a policy can be expressed in two equivalent ways, according to [4]:

$$\rho(\pi) = (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0),\, s_0 \sim \mu_0}[Q^\pi(s_0, a_0)] = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)], \tag{2}$$

where $Q^\pi(s,a)$ is the state-action value function, which represents the expected return when starting from state s_0, taking action a_0, and then following policy π thereafter. It satisfies the equation:

$$Q^\pi(s,a) = R(s,a) + \gamma P^\pi Q^\pi(s,a). \tag{3}$$

In this case, P^π is the policy transition operator [4], that is:

$$P^\pi Q(s,a) = \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[Q(s', a')] = \sum_{s' \in S, a' \in A} \pi(a'|s') T(s'|s,a) Q(s',a'). \tag{4}$$

The function d^π corresponds to the distribution of state-action pairs (s,a) within policy π. It measures the probability that, given a policy π and a state s_i and action a_i, the agent will take action a_i, when in state s_i. It satisfies the equation:

$$d^\pi(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma P_*^\pi d^\pi(s,a), \tag{5}$$

where P_*^π is the transpose policy transition operator [4], given by:

$$P_*^\pi d(s,a) := \pi(a \mid s) \sum_{\tilde{s},\tilde{a}} T(s \mid \tilde{s}, \tilde{a}) d(\tilde{s}, \tilde{a}). \tag{6}$$

2.3 Off-policy Evaluation Using DICE Estimators

Off-policy evaluation (OPE) aims to estimate $\rho(\pi)$ using only a fixed dataset of experiences. For instance, such a dataset would be $D = \{(s_0^{(i)}, s^{(i)}, a^{(i)}, r^{(i)}, s'^{(i)})\}_{i=1}^N$, 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 $s'^{(i)} \sim T(s^{(i)}, a^{(i)})$ [4]. In other words, OPE focuses on estimating the performance of a target policy using data collected under a different policy, known as the behavior policy [9]. The behavior policy used to create dataset D is denoted as π^b.

In the context of DICE methods, the following expression is used for the value of the target policy:

$$\rho(\pi) = \mathbb{E}_{(s,a,r)\sim d^{\pi^b}}[\zeta^*(s,a) \cdot r], \tag{7}$$

where $\zeta^*(s,a)$ can be written as:

$$\zeta^*(s,a) = \frac{d^\pi(s,a)}{d^{\pi^b}(s,a)}, \tag{8}$$

which is the distribution correction ratio [4]. The ratio can be formally described as the state-action visitation distribution under the target policy over the state-action visitation distribution under the behavior policy. The DICE estimators work towards approximating this ratio, without knowledge of d^{π} or d^{π^b}, then estimate $\rho(\pi)$ by applying (7) [4].

Considering (5), the function d^{π^b}, which represents the state-action visitation distribution by following the behavior policy can be written as:

$$d^{\pi^b}(s,a) = (1-\gamma)\mu_0'(s)\pi^b(a|s) + \gamma P_*^{\pi^b} d^{\pi^b}(s,a), \tag{9}$$

In the research conducted by [4], the initial start distribution of D is denoted as μ_0, indicating that both d^{π} and d^{π^b}, are generated under the same initial start distribution. However, since this study focuses on the mismatch between initial start distributions, the initial start distribution of D is denoted as μ_0' instead. This notation highlights that, in this research, these two initial start distributions, μ_0 and μ_0' are not necessarily the same.

To observe the impact of the initial start distributions (μ_0 and μ_0') on the distribution correction ratio, equations (5) and (9) can be used in equation (8), by substituting $d^{\pi}(s,a)$ and $d^{\pi^b}(s,a)$ as follows:

$$\zeta^*(s,a) = \frac{(1-\gamma)\mu_0(s)\pi(a \mid s) + \gamma P_*^{\pi} d^{\pi}(s,a)}{(1-\gamma)\mu_0'(s)\pi^b(a \mid s) + \gamma P_*^{\pi^b} d^{\pi^b}(s,a)}. \tag{10}$$

2.4 Initial Start Distribution Mismatch

In the context of initial start distributions, the concept of state visitation mismatch highlights the 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. In other words, it quantifies how the agent's initial state experiences diverge between these two datasets.

Given a finite number of reinforcement learning episodes, the initial start distribution of the dataset created under the target policy, μ_0, can be defined empirically as follows:

$$\hat{\mu}_0(s) = \frac{1}{N}\sum_{i=1}^{N} f(i,s,\pi,\mu_0), \tag{11}$$

where N is the total number of episodes or trajectories and f is a function defined as:

$$f(i,s,\pi,\mu_0) = \begin{cases} 1, & \text{if episode i of the dataset created under } \pi \text{ using } \mu_0 \text{ starts in state s} \\ 0, & \text{otherwise} \end{cases}$$

Similarly, for the initial start distribution of the dataset created under the behavior policy, μ_0', the empirical start distribution $\hat{\mu}_0'$ can be defined as follows:

$$\hat{\mu}_0'(s) = \frac{1}{N}\sum_{i=1}^{N} f(i,s,\pi^b,\mu_0'), \tag{12}$$

The initial start distribution mismatch revolves around the divergence between $\hat{\mu}_0$ and $\hat{\mu}_0'$. This study is concerned with the impact of this mismatch on the accuracy of the DICE methods when estimating the value of the target policy.

3 Methodology

3.1 Choice of Measure for Initial Start Distribution Mismatch

To verify whether the initial start distribution mismatch has an impact on the accuracy of the DICE estimators in estimating the target policy, it is necessary to compute the disparity in state visitation between $\hat{\mu}_0'$ and $\hat{\mu}_0$. For this, several measures are available that can be used to estimate the distance between the state visitation distributions.

For computing the similarity between two probability distributions, a widely used measure is the Kullback-Leibler (KL) divergence. It aims to quantify how similar a probability distribution is to a candidate or model distribution [10]. It is not a distance metric due to its asymmetry [11]. The KL divergence between two probability distributions P and Q is given by:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{13}$$

The value of the KL divergence ranges from 0 to infinity, where the value 0 is obtained when the two probability distributions P and Q are identical.

Using equation (13), the state visitation mismatch between $\hat{\mu}_0'$ and $\hat{\mu}_0$ can be expressed as follows:

$$D_{\mathrm{KL}}(\hat{\mu}_0' \parallel \hat{\mu}_0) = \sum_{s \in \mathcal{S}} \hat{\mu}_0'(s) \log \frac{\hat{\mu}_0'(s)}{\hat{\mu}_0(s)}, \tag{14}$$

where s represents the states from the state space S of an MDP.

3.2 Choice of Measure for Target Policy Estimation

This research aims to assess the impact of the initial start distribution mismatch on off-policy evaluation. Specifically, the policy evaluation was conducted using behavior-agnostic methods, particularly the DICE methods used in [4]. As previously stated, the goal of these methods is to approximate the distribution correction ratio, defined in equation (8), and then use it to estimate the value of the target policy, also known as the normalized expected cumulative reward by applying equation (7). This estimator, which is called the dual estimator, is proposed by [4], and can be expressed as follows:

$$\hat{\rho}_\zeta(\pi) := \mathbb{E}_{(s,a,r)\sim d^{\pi^b}} \left[\hat{\zeta}(s,a) \cdot r \right]. \tag{15}$$

The dual estimator is highly effective based on the findings of [4]. Therefore it was the chosen estimator for this research paper to estimate the normalized expected cumulative reward, for evaluating the target policy. This estimator is unbiased across multiple configurations and consistently outperforms other estimators. Furthermore, it demonstrates robustness to both scaling and shifting of MDP rewards, making it a reliable choice for various applications [4].

3.3 Choice of Measure for Assessing Performance

After determining the measures for evaluating the state visitation mismatch and target policy estimation, the final important step was to find a measure that can assess the quality of the policy evaluation estimation ($\hat{\rho}_\zeta(\pi)$), by measuring how closely it aligns with the ground truth. A widely used metric for assessing performance is Mean Squared Error (MSE), which aims to determine the average squared distance between the actual and predicted values [12]. Generally, MSE can be expressed as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{16}$$

where n represents the total number of dataset samples.

In the context of this research, to assess the accuracy of the DICE methods in estimating the normalized expected cumulative reward of the target policy, the following equation is used for MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\rho(\pi)_i - \hat{\rho}_\zeta(\pi)_i)^2, \tag{17}$$

where $\rho(\pi)_i$ and $\hat{\rho}_\zeta(\pi)_i$ represent the ground truth and estimated normalized cumulative reward, respectively, for dataset i.

The MSE is calculated over multiple datasets to obtain a more robust and reliable error value. This approach ensures that the MSE provides a comprehensive assessment of the influence that initial start distribution mismatch has on the accuracy of DICE methods.

3.4 Integration of Steps

To systematically assess the impact of initial start distribution mismatch on the accuracy of DICE methods, a structured approach was followed involving the generation and comparison of various initial start distributions.

Initially, multiple different initial start distributions $\mu_0^1, \mu_0^2, \mu_0^3, \ldots, \mu_0^k$ were generated. These distributions were carefully selected to ensure a reasonable variance in KL divergence values relative to the target initial start distribution μ_0. For each initial start distribution μ_0^i, a fixed number of datasets were created using the behavior policy π^b. Let n denote the number of datasets for each μ_0^i. For each dataset, the empirical initial start distribution $\hat{\mu}_0^i$ was recorded. This empirical distribution remained constant for all datasets generated under μ_0^i.

Similarly, n datasets were created using the target policy π with the initial start distribution μ_0. The empirical initial start distribution $\hat{\mu}_0$ was also recorded and remained constant across all target policy datasets. By fixing the empirical initial start distributions ($\hat{\mu}_0^i, \hat{\mu}_0$), the KL divergence values between each pair ($\hat{\mu}_0^i, \hat{\mu}_0$) would remain constant between datasets. Then, the KL divergence values were computed, between each pair ($\hat{\mu}_0^i, \hat{\mu}_0$), using equation (14). Each initial start distribution, $\hat{\mu}_0^i$ is associated with a specific KL divergence value, relative to $\hat{\mu}_0$.

Despite fixed policies and initial start distributions, the actions taken in each dataset varied due to the stochastic nature of the policies. This variability ensured that each dataset produced a unique ground truth value and normalized cumulative expected reward. Creating multiple datasets helps mitigate biases that may arise from the interaction between specific initial start distributions and the policies.

At fixed intervals, the normalized expected cumulative reward was computed using equation (15), and averaged across the n datasets for each start distribution. This average was then compared to the ground truth value to monitor the convergence of the estimated values. For each initial start distribution μ_0^i, the MSE was calculated across the n datasets using (17). This calculation incorporated the ground truth value and the normalized cumulative expected reward, computed using equation (15), for each dataset, created under π^b, using μ_0^i.

After calculating the MSE for all k initial start distributions, k MSE values were obtained, each computed from n samples. In other words, each initial start distribution $\hat{\mu}_0^i$ is associated with an MSE value and a KL divergence value. These MSE values were then compared for each KL divergence value to assess the impact of initial start distribution mismatch on the accuracy of the DICE methods.

This structured approach ensured that the analysis covered a range of initial start distributions with varying KL divergence values, providing a comprehensive assessment of how initial start distribution mismatch affects DICE method performance.

4 Experiments

4.1 Environment

The experiments were conducted using a 10x10 grid-world environment, similar to the one used for the experiments of [4], where the agent can move up/down/left/right. The states/observations are given by the x, y coordinates of the agent's location. The reward function for this environment is the following:

$$\exp\left(-2\frac{|x - t_x| + |y - t_y|}{\text{length}}\right), \tag{18}$$

where t_x and t_y are the coordinates of the goal or target. In these experiments, the target was fixed at the bottom-right corner of the grid, specifically at coordinates $(t_x, t_y) = (9, 9)$ on a 0-indexed grid. The *length* value corresponds to the length of the grid which is 10. This reward function implies that the reward decreases exponentially as the distance between the agent and the target increases. When the agent is very close to the target, the distance between the agent's current location and the target is small, leading to a high reward close to 1. As the distance grows, the reward rapidly decreases towards 0. This environment is depicted in Figure 1.

The optimal policy for this task involves moving all the way to the right, followed by moving all the way down. The target policy π is taken to be the optimal policy plus 0.1 weight on uniform exploration, allowing for some degree of random actions to encourage exploration of the state space. The behavior policy π^b is taken to be the optimal policy plus 0.3 weight on uniform exploration, allowing for a greater degree of random action compared to π. The datasets generated using policies π and π^b, each consist of 400 trajectories. Every trajectory contains 100 time steps.

It is important to note that discrete environments were required for the experiments due to the chosen metric for computing the state visitation distribution mismatch, as well as for reduced complexity.

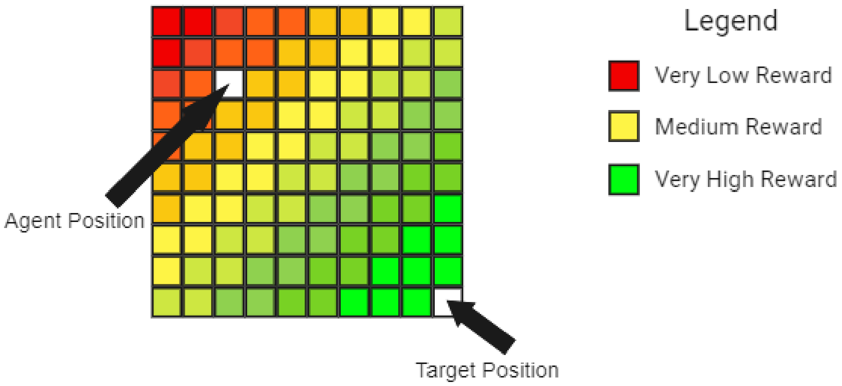

Figure 1: 10x10 Grid Environment

4.2 Initial Distributions

For the experiments, 7 different initial start distributions have been used. For each of those distributions, 5 datasets were created using π^b. Each different start distribution was fixed, to ensure the same distribution is used across the 5 datasets. These distributions were selected systematically, to be able to generate a reasonable variance in the KL-divergence values, without having a large sample size. The initial start distributions that were used were the following:

- Uniform Distribution: Each state in the grid has an equal probability.
- Edge Bias Distribution: The states around the edges of the grid were assigned a higher probability than others.
- Distance-Based Distribution: The states have a probability inversely proportional to their distance to the goal/target state.
- Target-Centric Distribution: This distribution prioritizes points closest to the target. Specifically, it focuses on points with (x, y) coordinates where both x and y are sampled from the interval $[8, 9]$, with equal probability.
- Remote Distribution: This distribution prioritizes points that are in the proximity of the furthest possible point from the target. Specifically, it focuses on points with (x, y) coordinates where both x and y are sampled from the interval $[0, 2]$, with equal probability.
- Fixed Point Distribution: This assigns probability 1 to the grid cell $(8, 9)$, thus always starting one cell away from the target state.
- Mixed Mode Distribution: This distribution averages the probabilities from the first three distributions with equal weights and then normalizes the resulting values.

Furthermore, 5 datasets were created using a fixed uniform initial start distribution under the target policy π, to generate 5 different ground truth values. After every dataset was created, the empirical initial start distributions were calculated using equations (11) and (12), respectively. Then, the KL divergence was computed between each of the 7 empirical initial start distributions (${\hat{\mu}_0}^1$, ${\hat{\mu}_0}^2$, ..., ${\hat{\mu}_0}^7$) used in datasets created under π^b, and the empirical initial start distribution ($\hat{\mu}_0$) used in datasets created under π. As such, 7 different KL divergence values were created, by applying equation (14). It is important to note that for some states $s \in S$, the empirical initial start distribution ($\hat{\mu}_0(s)$) can be 0. In such cases, a very small arbitrary constant value is used instead, 10^{-10}, to avoid issues in calculating the KL divergence.

The next step in the experiment involves using the DICE estimators. The estimator is trained using a feed-forward neural network, as described in [4]. The network has two hidden layers, each with 64 neurons and ReLU activation functions. The learning rate is set to 0.00003, and the total number of training steps is 25,000, with a batch size of 512.

The estimator takes as input a dataset created under π^b and a dataset created under π. Every 250 training steps, the normalized cumulative expected reward is computed using equation (15), which is compared to the ground truth, to observe the convergence of the DICE estimator. The reward value at the final training step is then used as the estimate of the target policy, $\hat{\rho}_\zeta(\pi)$, for the dataset created under π^b. This process is repeated for each of the 5 datasets created under π^b, using $\hat{\mu}_0^i$ for i from 1 to 7, corresponding to each of the 7 initial start distributions. Subsequently, the MSE is computed for the 5 dataset samples for each $\hat{\mu}_0^i$, with i from 1 to 7, following equation (17). Here, $\hat{\rho}_\zeta(\pi)_k$ represents the normalized expected cumulative reward or the estimation value of π from dataset k and $\rho(\pi)_k$ is the ground truth value from dataset k for π, with k from 1 to 5. As such, 7 MSE values are computed, which correspond to the 7 KL divergence values.

4.3 Results

After conducting the experiment outlined in the previous subsections, a set of results was obtained. Initially, for each of the 7 initial start distributions utilized in datasets created under π^b, the normalized cumulative reward values were averaged across the 5 datasets. These averages were then plotted alongside the ground truth value, which was averaged over the 5 datasets created under π, spanning a total of 25,000 steps, to observe the convergence of the DICE estimators. Additionally, the legend of the plot includes the KL divergence values for each initial start distribution. These values represent the state visitation mismatch between each distribution compared to μ_0.

As depicted in Figure 2, systematically creating the 7 initial start distributions leads to a reasonable variance in KL divergence values. The normalized expected cumulative reward is higher than the ground truth value for all distributions, except the remote distribution, which converges around 0.87. The two highest values are obtained by the fixed and target-centric distributions, both converging around 0.99, which also have the highest KL divergence values of 4.89 and 8.64, respectively. The values for the remaining distributions are converging closer to the true value, around 0.93, with the closest being the uniform distribution, which also has the lowest KL divergence value of 0.34.

Overall, no clear trend can be observed from this graph regarding the impact of the initial start distribution mismatch on the performance of the DICE estimators. In some cases, a lower KL divergence value leads to a more accurate estimation, while in others, specifically for higher KL divergence values, there is no clear trend. For instance, for the remote initial start distribution, which is furthest from the goal, the estimation is lower compared to the ground truth value. Conversely, the two closest start distributions to the goal, namely target-centric and fixed point, with higher KL divergence values, yield estimations higher than the ground truth value.

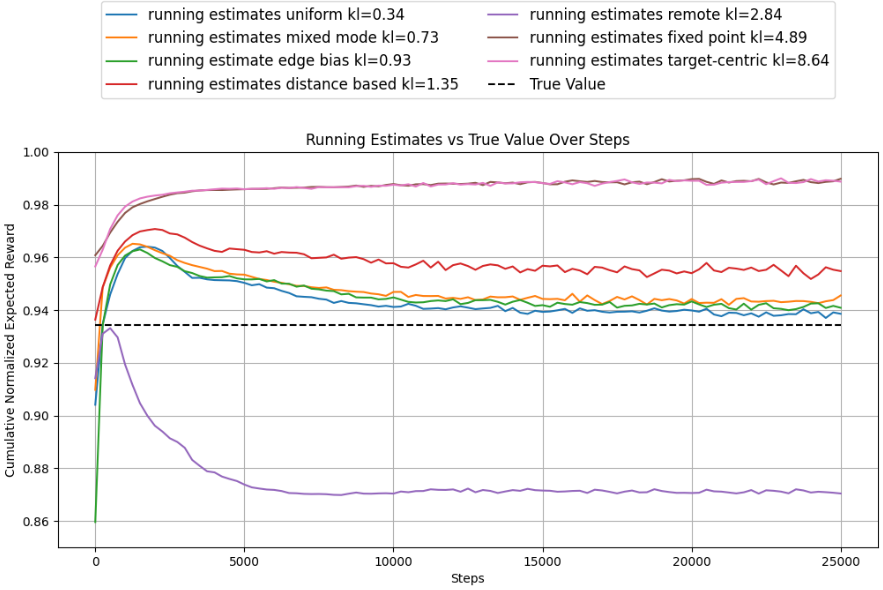

Figure 2: Normalized Cumulative Reward Over Time Steps

The MSE for each KL divergence value is depicted in Figure 3. The x-axis shows the ascending sorted KL divergence values corresponding to each initial start distribution. Each box plot displays the cumulative expected reward values for each of the five dataset samples after convergence. Above each box plot, the MSE values are indicated. To enhance observability, the MSE values were multiplied by 1000. The dashed line represents the average ground truth value across the five datasets.

From the graph, no clear trend is identifiable. 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 particular trend that would suggest the KL divergence value has a significant influence on the MSE.

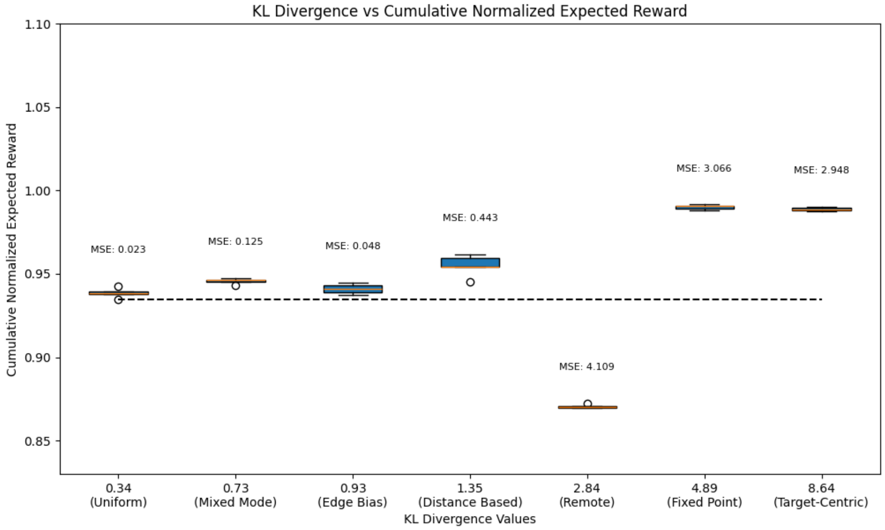

Figure 3: Box Plot of Mean Squared Error for Each KL Divergence Value

5 Discussion

5.1 Limitations

Based on the experiments conducted in this research, the initial start distribution mismatch does not have a clear influence on the performance of the DICE estimators. There are several limitations to consider, which influenced the results of the experiments. Firstly, it is important to note that the experiment was run on relatively few datasets. For each initial start distribution, of the 7 selected, only 5 datasets were generated. That is because, creating each dataset using a reasonable amount of trajectories, which is 400 in this case, each being of length 100, takes a significant amount of time to run. This is partly a limitation of Python itself not being regarded as a particularly fast programming language relative to others, but also due to the specifications of the machine used to conduct the experiments, as highlighted in section 6.

As such, the box plots showcased in Figure 3, would have benefited greatly from additional samples to provide additional support for the results since the MSE would have been more accurate. Furthermore, the same observation can be made about the number of initial start distributions and the respective KL divergence values, as well as the number of training steps. Due to time constraints, generating a large number of start distributions and KL divergence values respectively, along with a longer training process, was not feasible. To combat this, the distributions were chosen systematically, such that a reasonable variance between the KL divergence was present.

Moreover, another aspect that has to be considered is the type of environment in which the experiment is conducted. Due to the chosen metric, the KL divergence, the experiments had to be run in discrete environments, to be able to calculate the initial start distribution mismatch. As such, only grid-world-like environments were used to conduct the experiments. Analyzing the impact of the initial start distribution mismatch on continuous environments is beyond the scope of this research. It is also important to note that the experiments were conducted in a single environment. Other discrete environments could have been explored to enhance the study. Other similar limitations include the choice of metrics. Though some research was conducted to motivate the choice for each measure, it is hard to say whether results would hold under different measures for state visitation mismatch, or performance assessment, without a thorough comparison of those measures on the performed experiments.

Furthermore, a possible limitation of the experiment is the relation between the initial start distribution and the policies. As described in subsection 4.2, the policies utilized weights for uniform exploration to enhance overall state space exploration and mitigate biases between the initial start distributions and the optimal policy. Given the nature of the environment, with the goal always positioned at coordinates $(9, 9)$, and the nature of the optimal policy, certain initial start distributions had an advantage by reaching the reward more consistently. The added weights for uniform exploration helped mitigate these biases.

However, the risk of using excessively high exploration weights is that it could lead to extensive state space exploration, regardless of the start distribution. This would diminish the impact of the start distributions. To address this, the target policy used a weight of 0.1, and the behavior policy used a weight of 0.3, aiming for a balance that mitigates the bias between the optimal policy and start distributions while avoiding over-exploration. The exploration weights used in this research are the same as those employed in the experiments conducted by [4]. Although fine-tuning these parameters could further optimize the balance, it is beyond the scope of this research.

Lastly, one more limitation to consider is the reproducibility of the results. Due to time constraints, it was not possible to perform multiple runs of the DICE estimators with the same seed. Consequently, running the DICE estimators on the same dataset would yield slightly different results in terms of the normalized cumulative expected reward. This implies that the plots highlighted in subsection 4.3 cannot be exactly replicated. However, it is important to note that the datasets used in the research can be replicated, as they have been saved and seeded successfully. Therefore, the initial start distributions and Kl divergence values are reproducible. Given the relatively large number of training steps and the fact that each estimation averages results from 5 fixed datasets, the differences between multiple separate runs on the same dataset would be minimal. Thus, while the exact values may differ slightly, the overall convergence of the plots and the normalized cumulative expected reward values will be very close to those presented in the paper. Therefore, the main conclusions drawn from the plots would remain unchanged.

5.2 Future Work

Looking ahead, several adjustments can be made to the experiments to further validate and enhance the results. With fewer time constraints and access to higher-performing machines, a greater number of experiments could be conducted to ensure the research is as robust as possible. This would include exploring a wider range of initial distributions, which means a more accurate representation of the variance in KL divergence. Then, each initial distribution can be run on a much larger number of datasets to increase the sample size of each initial start distribution. After that, the training process can be adjusted, by increasing the number of steps to improve convergence, to get a more accurate representation of the reward values for each initial start distribution.

Furthermore, another way of validating the experiments and enhancing the research includes, using different performance metrics, as well as different measures for computing the initial start distribution mismatch, to assess whether the results hold under different configurations. Lastly, the experiments can be conducted on multiple discrete environments, or even in more complex continuous environments, to be able to accurately assess to what extent the initial start distribution mismatch impacts off-policy evaluation using DICE methods.

Lastly, as mentioned previously in subsection 5.1, certain initial start distributions benefit more from specific policies depending on the environment. This is shown in equation (10), as the distribution correction ratio depends on both the policy and the initial start distribution. Future research could explore this relation further by employing various policies for each initial start distribution and assessing their interactions across multiple environments. This approach would provide a clearer understanding of how the initial start distribution mismatch influences policy performance and vice versa, in the context of DICE estimators.

6 Responsible Research

It is very important to understand that conducting research requires adhering to certain ethical considerations. It is part of the researcher's responsibility to ensure that the research was conducted ethically and that results are reproducible. Throughout this study, ethical considerations were of utmost importance and upheld at every stage of the process.

To ensure transparency, the results of the performed experiments are publicly available in a GitHub repository[1]. This repository is an extension of the original repository[2], used in [4], which can be reproduced and modified free of charge under the Apache-2.0 license. That repository serves as a foundation for this research, as it contains the implementation for the environments, dataset creation for behavior and target policies respectively, and the training process of the DICE estimators. Any modifications made to files from the original repository are documented in the header of those files.

Finally, the experiments should be reproducible by following the described methodology and experiment setup. Running the experiments from the repository will yield similar results to those described in subsection 4.3, though not identical, as explained in subsection 5.1. It is important to note that the main takeaways from the results will remain unchanged. All experiments were conducted locally on an HP Laptop 15s-fq1xxx with an Intel(R) Core(TM) i5-1035G1 CPU and Windows 11 PRO.

[1] `https://github.com/tibisabau/dice_rl`

[2] `https://github.com/google-research/dice_rl`

7 Conclusion

To conclude, this research paper aims to assess the impact of the initial start distribution mismatch on the accuracy of DICE estimators. The DICE estimators presented in [4] served as the foundation for this study. To address the research question, seven initial start distributions were systematically created and assigned to the behavior policy. The initial start distribution mismatch between each behavior policy start distribution and the target policy start distribution was then computed using the Kullback–Leibler (KL) divergence. This approach ensured a reasonable variance in KL divergence values, allowing for an analysis of whether it influences the performance of DICE estimators.

For each start distribution, the performance of off-policy evaluation was assessed by running the DICE estimators for 5 dataset samples. The normalized expected cumulative reward values obtained from these estimators were compared to the ground truth value of the target policy using Mean Squared Error (MSE). The initial start distribution mismatch value, given by the KL divergence, was then compared to the MSE for each distribution to determine if the initial start distribution mismatch impacted the MSE values. The results from the conducted experiments did not reveal a clear trend, indicating that further research is required. Expanding the scope of the experiments is necessary to accurately assess whether the initial start distribution mismatch can influence the performance of DICE estimators.

Overall, this research aimed to test the significance of the choice of initial start distribution in behavior-agnostic reinforcement learning by evaluating its impact on the performance of DICE methods. However, further research is needed to accurately assess the overall impact of the start distribution on off-policy evaluation. Future studies should include experiments on multiple types of environments, a larger number of datasets and start distribution samples, and different measures to compute the initial start distribution mismatch and assess performance. Additionally, exploring the relation between the initial start distribution and policies could be an interesting direction for future research. This could provide deeper insights into how different start distributions interact with various policies and further refine the understanding of their influence on DICE estimators.

References

[1] OpenAI, M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, “Learning dexterous in-hand manipulation,” 2019.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” 2013.

[3] O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans, “Algaedice: Policy gradient from arbitrary experience,” 2019.

[4] M. Yang, O. Nachum, B. Dai, L. Li, and D. Schuurmans, “Off-policy evaluation via the regularized lagrangian,” 2020.

[5] O. Nachum, Y. Chow, B. Dai, and L. Li, “Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections,” in *Advances in Neural*

Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/cf9a242b70f45317ffd281241fa66502-Paper.pdf

[6] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. USA: John Wiley & Sons, Inc., 1994.

[7] A. Tavakoli, V. Levdik, R. Islam, C. M. Smith, and P. Kormushev, "Exploring restart distributions," 2020.

[8] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," Proceedings of the Nineteenth International Conference on Machine Learning, 2002.

[9] Q. Liu, L. Li, Z. Tang, and D. Zhou, "Breaking the curse of horizon: Infinite-horizon off-policy estimation," 2018.

[10] J. Shlens, "Notes on kullback-leibler divergence and likelihood," 2014.

[11] Y. Zhang, W. Liu, Z. Chen, J. Wang, and K. Li, "On the properties of kullback-leibler divergence between multivariate gaussian distributions," 2023.

[12] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, p. 045–076, 2019. [Online]. Available: http://dx.doi.org/10.28945/4184