Impact of state visitation mismatch methods on the performance of on-policy reinforcement learning methods

Hongwoo Cho, Stephan Bongers, Frans Oliehoek
h.cho-1@student.tudelft.nl, s.r.bongers@tudelft.nl, f.a.oliehoek@tudelft.nl
Delft University of Technology

1. Background

Reinforcement learning (RL) has shown success in different domains, such as autonomous driving. However, they rely on large amounts of data generated by simulators, which may not be feasible in real-world applications

Instead, we often have access to offline data collected from unknown behavior policies, posing significant challenges for policy evaluation and optimization.

Behavior-agnostic data: In many real-world scenarios, the data available for training RL models comes from unknown behavior policies.

By using off-policy evaluation methods such as the Distribution Correction Estimation (DICE) to initialize Q-values, we can provide a more accurate starting point for on-policy RL methods like Q-learning [1]. This approach aims to enhance convergence speed, effectively bridging the gap between off-policy and on-policy paradigms, and making the most of available offline data to inform online learning processes.

Research Question: How does state visitation mismatch methods impact the performance of on-policy RL methods?

2. Methodology

To answer the RQ, we used the following steps:

- **Dataset Generation:** Created two types of datasets, which are the behavior policy dataset and target policy dataset.
- **Running the DICE Estimator**: Used to correct distribution mismatches by adjusting state-action visitation distributions. Q-values are extracted from the DICE estimator.
- **Q-Learning:** With two sets of Q-values: DICE-initialized and zero-initialized Q values, we measure the average reward per step to assess performance and convergence speed.

3. Experiment

Frozenlake environment was used as shown in Figure 1.

- Navigate the agent from the start (top-left) to the goal (bottom-right) of a grid.
- Consists of two types of tiles: F = Frozen lake (safe to move), H: Hole (falling ends the episode)
- Movement: The ice is slippery, so the agent may not move in the intended direction
- Rewards: +1 for reaching the goal and 0 for falling into a hole or moving to a safe tile.

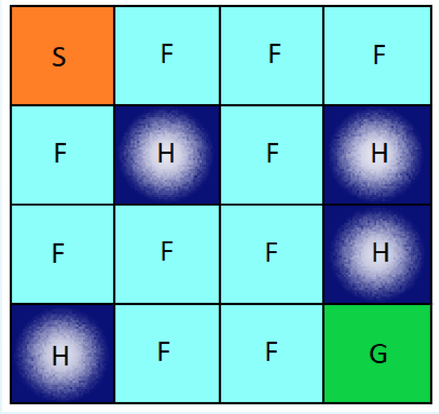

Figure 1: FrozenLake Environment, from [2]

4, Results and Discussion

Extracted Q-Values from the DICE Estimator are shown in Table 1:

Table 1: Q-Values from DICE Estimator

State	Move Left	Move Down	Move Right	Move Up
State 1	0.50368656	0.28417100	0.24744200	0.30284700
State 2	0.00666858	0.00369054	0.00309231	0.00149879
State 3	0.02507504	0.00371495	0.00309155	0.01011084
State 4	0.05221552	0.01277991	0.00067256	0.00430612
State 5	0.26031873	0.00403238	0.00468556	0.00260281
State 6	0.00233800	0.00251356	0.00273770	0.00179974
State 7	0.02266099	0.00031998	0.00077225	0.00000000
State 8	0.00000000	0.00000000	0.00000000	0.01678710
State 9	0.00223199	0.00417178	0.00007266	0.13045901
State 10	0.01144835	0.11829945	0.00535556	0.00362227
State 11	0.00075249	0.01784247	0.00250205	0.00025089
State 12	0.01815625	0.00211173	0.00307075	0.00129442
State 13	0.00279297	0.00271804	0.00299118	0.00082534
State 14	0.00014303	0.00039642	0.02397827	0.00021320
State 15	0.00124678	0.01325846	0.00001431	0.00456260
State 16	0.00000000	0.00000000	0.00000000	0.00000000

With the DICE Q-values and zero Q-values, Q-Learning algorithm was ran to compare the average rewards over time step, shown in Figure 2.

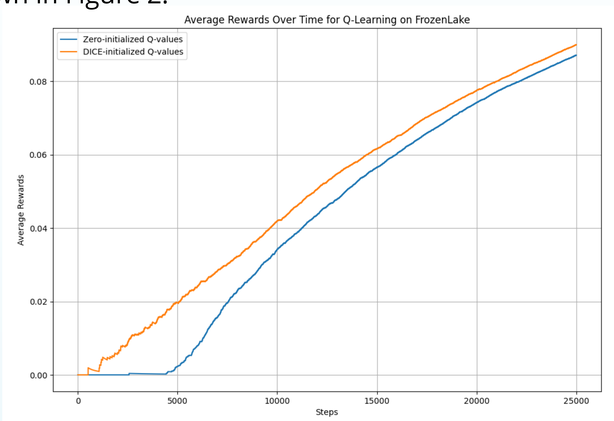

Figure 2: Average Rewards over time for Q-Learning on FrozenLake

- Average reward higher for DICE Q-Values as DICE Q-values provide more accurate initial estimate of the expected rewards, which reduces the amount of exploration

5. Conclusions and Future Work

Main Contributions

- Investigated the impact of state visitation mismatch methods on the performance of on-policy RL methods
- Our study demonstrated that initializing Q-learning with DICE Q-values enhances convergence speed and performance.

Future Work

- Additional experiments on larger and more complex environments, which offer higher complexity and variability.
- Explore different target policies to assess how they affect the Q-values generated by the DICE.

6. References

[1] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian, 2020.
[2] Bethany D. Pena and Daniel T. Banuti. Reinforcement learning for pathfinding with restricted observation space in variable complexity environments, 2021