Detecting Environment Changes via Quantile Spread in Quantile Regression Deep-Q Networks

Paul-Gabriel Stan
Supervised by Mustafa Celikok, Frans Oliehoek | 25-06-2025

1 Background

- Reinforcement learning is the process of learning what to do - how to map situations to actions - to maximize a numerical reward signal [4].
- Deep Q-Networks (DQN) [3] achieve strong performance by approximating value functions using neural networks, but provide a point estimate that fails to capture uncertainty about outcomes.
- Quantile Regression Deep Q-Networks[1] (QR-DQN) address this weakness by modeling the complete distribution of possible returns using quantile regression.
- In DQN, the Q-function is approximated using a neural network, but QR-DQN addresses this by learning the full return distribution $Z^{\pi}(s, a)$ instead of its expectation $Q^{\pi}(s, a)$.
- The quantile outputs are trained using the Quantile Huber loss[2, 1].
- We measure the agent's uncertainty through the inter-quantile range for the selected action at the current time step. IQR $= Q_{0.75} - Q_{0.25}$

2 Motivation

The training environment of the agent may differ from the one in which it is deployed. Even small changes in the dynamics and physics of the environment can impact the agent's performance if the policy has not been exposed to these variations in the training setting.

- Can changes of deterministic environments in reinforcement learning agents be detected by the quantile spread of QR-DQN?

3 Methodology

- Train agents with ideal parameters in deterministic environments, this experiment used Cartpole-v1.
- At test time, modify the environment's parameters to perturb the dynamics of the testing space, pole length for Cartpole-v1.
- In the testing loop, extract the quantile distribution of the chosen action and calculate its quantile spread, characterized by the inter-quantile range.
- Use the trained agents to run 100 episodes on the modified environment and calculate the average quantile spread over the episodes.
- Early into the episode, most of the dynamics are not shifted, therefore we only take into account the spread after the midpoint of the time steps.
- Plot the averaged quantile spread against the testing time steps and inspect for trends.
- If the average quantile spread difference from the baseline training environment is bigger than a threshold, categorize the environment as "DIFFERENT", else as "SAME".
- Make the experiments reproducible by using seeds.

4 Results

Pole Length	Δ Spread (vs. 0.50)	Shift Detected
0.10	0.4137	**DIFFERENT**
0.45	0.0002	SAME
0.55	-0.0000	SAME
1.00	1.9948	**DIFFERENT**
2.00	2.4521	**DIFFERENT**
3.00	2.0093	**DIFFERENT**
10.00	1.2640	**DIFFERENT**
20.00	0.8486	**DIFFERENT**

Table: Environment shift detection based on change in quantile spread relative to pole length 0.50

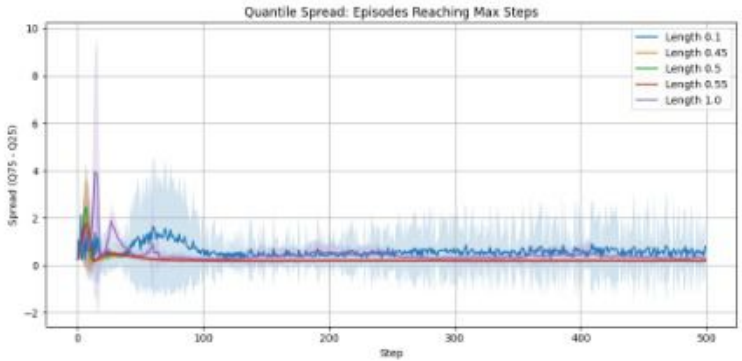

Figure: Quantile spread over time for environments where the agent reaches the maximum episode length

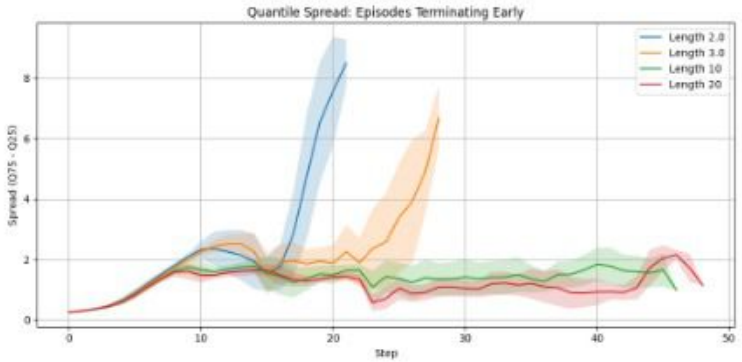

Figure: Quantile spread over time for environments where episodes terminate early due to failure

Pole Length	Mean Spread
0.10	0.5991
0.45	0.1855
0.50	0.1853
0.55	0.1853
1.00	2.1801
2.00	2.6374
3.00	2.1947
10.00	1.4493
20.00	1.0339

Table: Mean quantile spread for different pole lengths

5 Conclusion

- The quantile spread behaved as hypothesized: as the environment's pole length deviated to a greater extent from the training length of 0.5, the spread increased, reflecting rising uncertainty in the learned value estimates.
- Figure 1 shows environments where the agent performs reliably, reaching the episode length limit of 500 steps. In these environments, the spread remains small, with negligible differences across neighboring lengths like 0.45 and 0.55.
- Figure 3 visualizes environments in which the agent fails to maintain balance consistently. These unknown conditions (pole lengths 2.0 and above) favor early termination and are associated with a higher spread.
- Interestingly, the pole length 0.1 stands out. Even though the agent consistently com- pletes the full 500 steps, the average quantile spread is higher than in other environments that reach the maximum steps.

6 Future Work

- We expect that the use of quantile spread to capture environmental shifts to extend to non-deterministic settings.
- Future work will need to explore how to calibrate and adapt the thresholds in the presence of noise in the environment.
- One direction is to use the IQR increases to fall back on a more conservative policy that prioritizes safety while the agent is retrained or adapts in the background.
- On top of this, the quantile spread increase could be used as a way to trigger human intervention for the systems.

References

[1] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. "Distributional reinforcement learning with quantile regression". In: *AAAI Conference on Artificial Intelligence* (2018).
[2] P. J. Huber. "Robust Estimation of a Location Parameter". In: *Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101.
[3] V. Mnih, K. Kavukcuoglu, D. Silver, et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (2015), pp. 529–533.
[4] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. 2nd. MIT press, 2018.

TU Delft Delft University of Technology