Robust Decoupled Q-Learning with Adversarial Attacks

109550119 邵筱庭 109550174 孟祥蓉 109550121 温柏萱

Introduction

Decoupled Q-Learning (DecQN) [4] withholds the advantage of being highly efficient by adapting the critic-only structure and bang-bang discretization. While the study exhibited state-of-the-art performances, there were no experiments conducted on the robustness of such structure. We make the assumption that applying bang bang control makes the model more robust when encountering noisy observations, and noisy actions. We first exhibit that bang-bang discretization demonstrated robustness on both action disturbance and observation disturbance through empirical results, and showed the vulnerability upon encountering intentional action manipulation of bang bang discretization. We aim to prove the robustness on noisy observations, and actions while conducting adversarial action poisoning to further improve robustness and achieve better exploration. To address the problem of extra training time cost when adopting adversarial training, we apply the sampling n batches technique to accelerate training.

Noisy Environment

Motivation

In real-world scenarios, the observations and actions are not as accurate and consistent as in experimental settings due to various factors such as noise, environmental changes, sensor inaccuracies, and unforeseen disturbances or malicious poisoning. These factors can lead to variability and uncertainty in the data and performed action, making it more challenging to develop robust and reliable models that perform well under diverse and dynamic conditions. We aim to mimic these inaccuracies by injecting noise into observations and actions selected by the agent and examine the effect of bang-bang discretization on robustifying DecQN against observation poisoning or action manipulation.

Through empirical results, bang-bang discretization is able to resist minor observation noises, but showed poor results when encountering noisy observations.

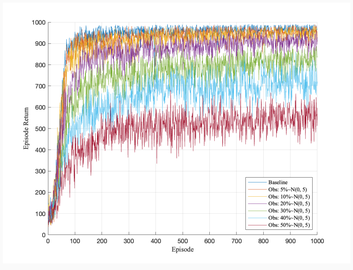


Fig. 1: Pair Comparison Results

Adversaries Formulation

The addition of Gaussian noise on observation induces randomness and foster diverse learning experiences [3]. We adapt the strategy which is formulated as the following.
We construct a random variable $Z \sim \text{Bernoulli}(p), p \in [0, 1]$ and assign $\delta \sim \mathcal{N}(0, \sigma^2)$ if $Z = 1$, otherwise set $\delta = 0$ if $Z = 0$. The setting is characterized by two parameters, the attack probability p and the attack strength σ.
We examine the effect of bang-bang discretization under different attack probability p and its ability to resist such noise or attack and the enhancement in performance when trained and tested under observation perturbance.

Adversarial training

Adversarial training [1, 2], has been recognized as one of the most effective approaches in traditional supervised learning tasks in training time defenses [2]. We adapt the method from [5] and formulate the setting as follows:
Given the target policy value function $Q^\dagger$, the adversarial action manipulation attack aims to minimize the distance between Q^θ and $Q^\dagger$. In our setting, we set the target policy to yield the reward as low as possible.

Given attack probability p and action radius r
$Z \sim$ Bernoulli(p)
if $Z = 1$ **then**
  $a^\dagger \leftarrow$ the action that yields the lowest reward within r
  $(s, a) \leftarrow (s, a^\dagger)$
end if

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[2] Aleksander Madry et al. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083* (2017).

[3] Harsha Putla et al. “A Pilot Study of Observation Poisoning on Selective Reincarnation in Multi-Agent Reinforcement Learning”. In: *Neural Processing Letters* 56.3 (2024), p. 161.

[4] Tim Seyde et al. “Solving continuous control via q-learning”. In: *arXiv preprint arXiv:2210.12566* (2022).

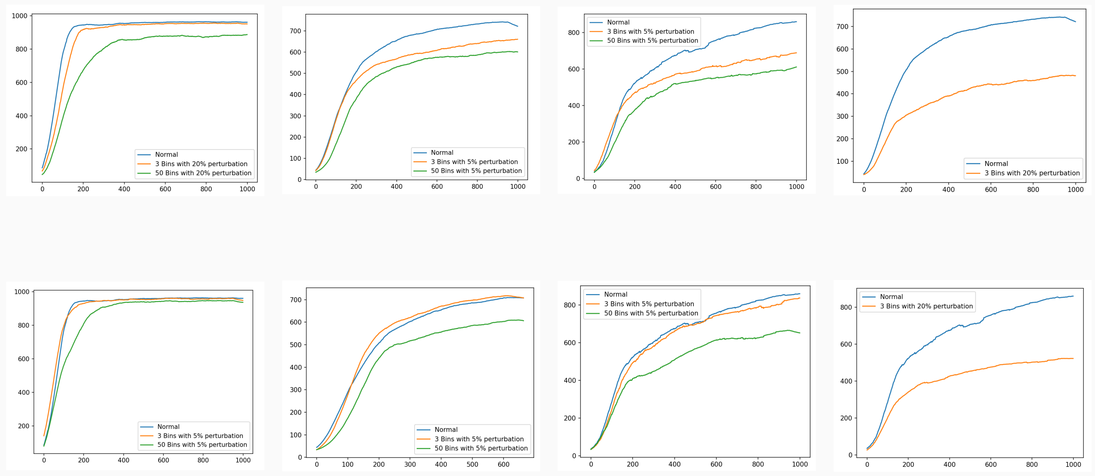
[5] Yinglun Xu and Gagandeep Singh. “Black-box targeted reward poisoning attack against online deep reinforcement learning”. In: *arXiv preprint arXiv:2305.10681* (2023).

[6] Zhe Zhang et al. *M^2DQN: A Robust Method for Accelerating Deep Q-learning Network*. 2022. arXiv: 2209.07809.

Results

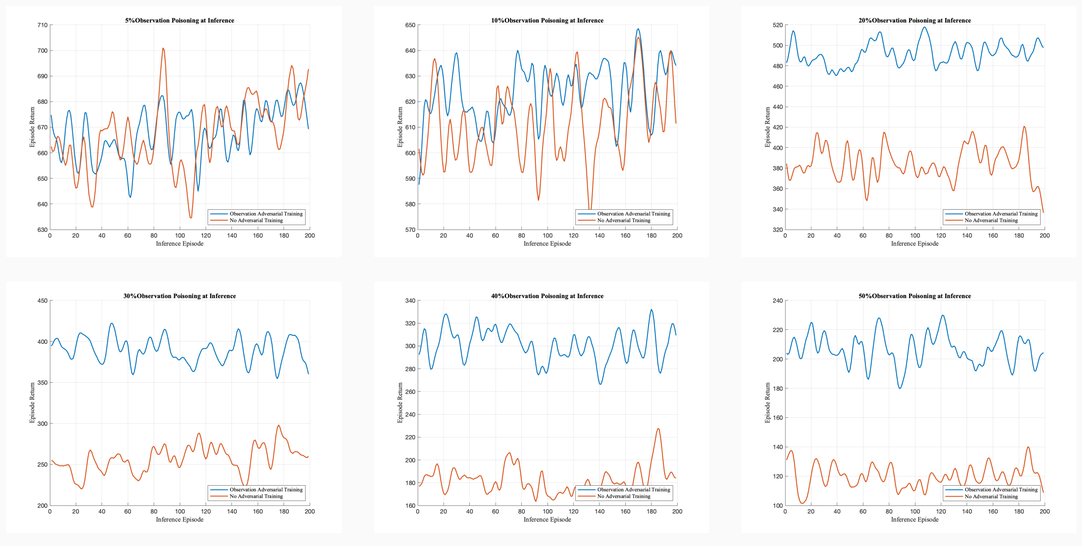
0.1 Noisy Environment

Overall, Decqn with bang bang control has better robustness under all environments and noises. However, large noises would still yield collapses in models.

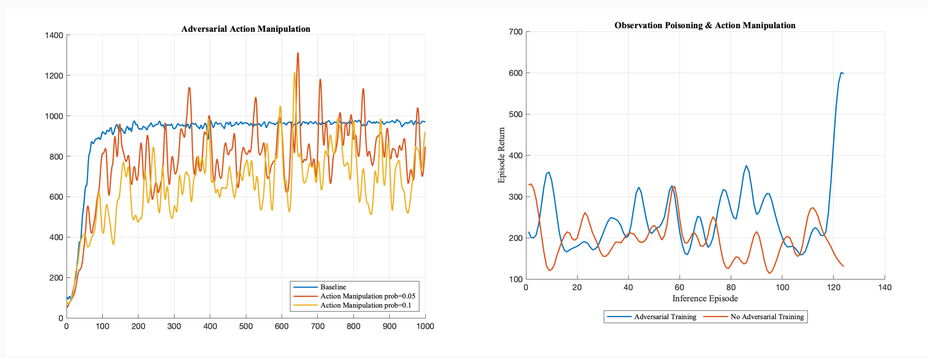


0.2 Training Under Adversarial Observation Attack

We demonstrate the enhanced robustness when trained under adversarial observations.

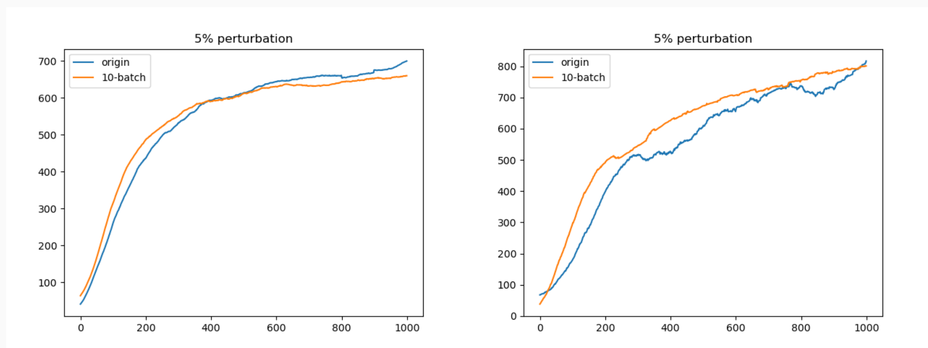


0.3 Adversarial Action Manipulation



Sample N batches

We draw inspiration from the paper ”M^{2}DQN: A Robust Method for Accelerating Deep Q-learning Network”[6] which employs the max-mean algorithm to calculate loss. This algorithm calculates the loss using an n-batch of data and updates the network based on the maximum loss within the batch. The principle is to utilize the worst-case scenario in the replay buffer, thereby facilitating rapid convergence of the model.
In our implementation, we aim to enhance training efficiency by adopting this approach. However, while the original paper updates the network using the maximum loss, we found this method to be suboptimal in our case. Consequently, we have opted to use the mean loss for network updates instead.
We utilize ”Reverb” as our replay buffer framework. Reverb can sort data based on weights related to the Temporal Difference (TD) error while incorporating some randomness. In our implementation, we set the TD-error weight in the replay buffer to 0.6.



Finally, we tested two tasks: ”Walker Run” and ”Cheetah Run”. The results demonstrate that, compared to training in a purely noisy environment, the n-batch method exhibits superior performance in the early stages of training. This indicates that the model converges more rapidly, enabling efficient training of a more robust model.