Deflating Q-values : investigating overestimation bias

Team#3:Subin Kim(20233143), Kyungwook Name(20170203), Doojin Baek(20190289)

1. Motivation & Objectives

Overestimation bias is a property of Q-learning in which the maximization of a noisy value estimate induces a consistent overestimation. Rolling a dice is an intuitive example of this, as taking the maximum value across 100 rolls is likely to be higher than the true expected value of 3.5. This can lead to poor performance in reinforcement learning since incorrect value estimation can lead to a suboptimal policy.

We aim to investigate the **overestimation bias of Q-value**[Figure 1] in various model-free RL algorithms in diverse settings. Our specific objective is as follows:

- Compare different approaches derived from DQN[1] such as Averaged DQN[2], Double DQN[3], and Maxmin DQN[4], or actor-critic methods like SAC[5] and TQC[6] and analyze the result.
- Investigate the condition of environments where overestimation of Q-value can occur in varying degrees, such as stochasticity of state transitions or exploration rate of epsilon-greedy policy.
- Propose a reasonable and novel methodology to reduce the overestimation problem of Q-values.

2. Experiments

Algorithms including Averaged DQN, Double DQN, Maxmin DQN are considered in this work. Further approach we suggest is replacing max operation in Q target as below:

$$Q^i(s,a) = Q^i(s,a) + \alpha[r + \gamma\, max_n Q^n(s',a') + (1-\gamma)\, min_m Q^m(s',a') - Q^i(s,a)],\ i,m,n \in \{1,2,\dots,N\}$$

Metrics The most important metric in this project is the differences between estimation and the true value, representing the magnitude of overestimation/underestimation. More specifically, we plan to calculate the ground truth value through policy evaluation in descrete states. In continuous environment, we plan to compare baseline Q-learning with other algorithms. Also, performance of the task and variance of values will be considered as metrics, too.

Stochasticity of Environment If Q-value of a state deviates in wide range, the maximum value does not properly represent the true Q-value. Thus, we can hypothesize that overestimation bias would be bigger in stochastic states. To experiment whether the algorithms properly handle the issue in a randomness, we are planning to investigate several randomness settings for each environment by extending codes from OpenAI Gym.

Exploration rate Exploration allows the agent to sample a wider range of actions in different states, which can help it better estimate the expected rewards of each action. In contrast, if an agent does not explore enough, it may rely too heavily on its current estimates of Q-values, leading to overestimation bias.[7]. Therefore, we plan to test differing epsilon values and analyzed for each algorithm and environment.

Appendix

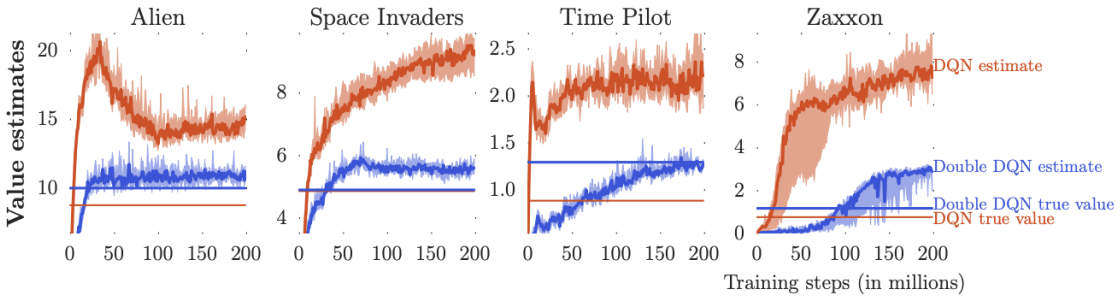

Figure 1. Value estimates by DQN (orange) and Double DQN (blue) on Atari games.

Reference

[1] Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." *arXiv preprint arXiv:1312.5602* (2013).

[2] Anschel, Oron, Nir Baram, and Nahum Shimkin. "Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning." *International conference on machine learning*. PMLR, 2017.

[3] Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep reinforcement learning with double q-learning." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. No. 1. 2016.

[4] Lan, Qingfeng, et al. "Maxmin q-learning: Controlling the estimation bias of q-learning." *arXiv preprint arXiv:2002.06487* (2020).

[5] Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." *International conference on machine learning*. PMLR, 2018.

[6] Kuznetsov, Arsenii, et al. "Controlling overestimation bias with truncated mixture of continuous distributional quantile critics." *International Conference on Machine Learning*. PMLR, 2020.

[7] Wagenbach, Julius, and Matthia Sabatelli. "Factors of Influence of the Overestimation Bias of Q-Learning." *arXiv preprint arXiv:2210.05262* (2022).