THE UNIVERSITY OF BRITISH COLUMBIA

Institut für Data Science im Maschinenbau

RWTH AACHEN UNIVERSITY

Self-tuning Q-ensembles (STQE): Using Monte-Carlo Estimates for Adaptive Overestimation Bias Reduction

Lily Xie - Supervised by Bernd Frauenknecht

Reinforcement Learning

Reinforcement learning is a subfield of machine learning that learns the **optimal behavior** from environment data to obtain the **maximum reward**. It mainly involves an **agent** and an **environment**: initially the environment is in a certain state s_t; at time t, the agent performs an action a_t, which results in a new state in the environment s_{t+1}, as well as a next reward value r_{t+1}.

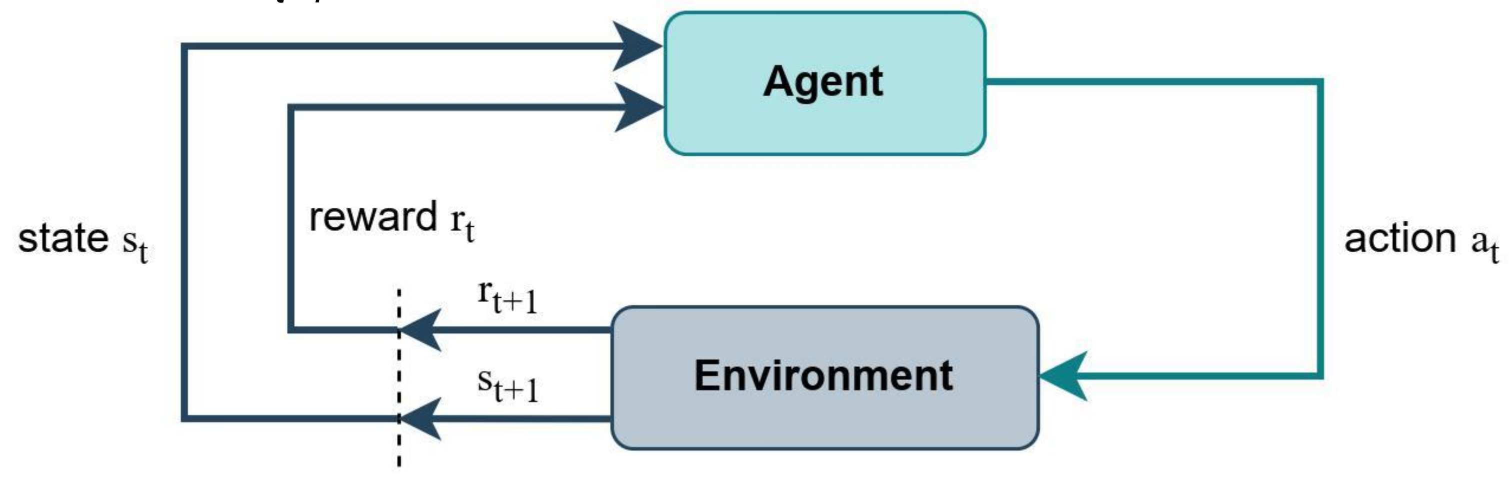

Figure 1: The agent–environment interaction in a Markov decision process.

This process repeats and the objective of reinforcement learning is to **maximize the cumulative reward**, which is referred to as **return**.

$$J(\pi) = argmax_{\pi}\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}]$$

Background

Q-function

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}\left[\sum_{k=t}^{\infty}\gamma^k r_{k+1} \,|\, s_t, a_t\right]$$

The Q-function gives the expected return under policy π, given a particular state and action. Thus, we can define a **greedy policy**:

$$\pi(s) = \arg\max_{a} Q^{\pi}(s_t, a_t)$$

Q-update

The Q-function is updated with **temporal difference (TD) learning**. The TD target is computed with respect to the greedy policy.

$$\underbrace{Q(s_t,a_t)}_{\text{New Q-value estimation}} \leftarrow \underbrace{Q(s_t,a_t)}_{\text{Former Q-value estimation}} + \underbrace{\alpha}_{\text{Learning Rate}} \underbrace{[\underbrace{\underbrace{r_{t+1}}_{\text{Immediate Reward}} + \underbrace{\gamma \max_a Q(s_{t+1},a)}_{\text{Discounted Estimate optimal Q-value of next state}}}_{\text{TD Target}} - \underbrace{Q(s_t,a_t)}_{\text{Former Q-value estimation}}]}_{\text{TD Error}}$$

Overestimation Bias

Q-functions are typically represented by **function approximators** like artificial neural networks. However, function approximators introduce noise and cause the Q-values to vary, yet the max operator always selects the largest Q-value. Over time, **a systematic overestimation bias** is created, which severely impairs the learning performance.

Methodology

Q-ensembles

- **Ensemble methods** are a way to mitigate overestimation bias
- Approaches like [1] propose to train N estimates of the Q-function

To calculate the **TD-target**:

- the different Q-predictions are **ordered by value**
- **K** denotes the number of most optimistic, discarded Q-estimates
- The average over the **most pessimistic N-K estimates** represents the TD-target

Objective

We propose **Self-tuning Q-ensembles (STQE)** that auto-tunes K. This is achieved by combining the previous method with **Adaptively Calibrated Critic (ACC)** [2].

Tuning Mechanism

By comparing the **biased TD** estimates obtained from the Q-ensembles (on the right) with the **unbiased Monte-Carlo estimates** on the most recent observed data, we can obtain the overestimation bias, and then use it to tune the number of dropped Q-functions.

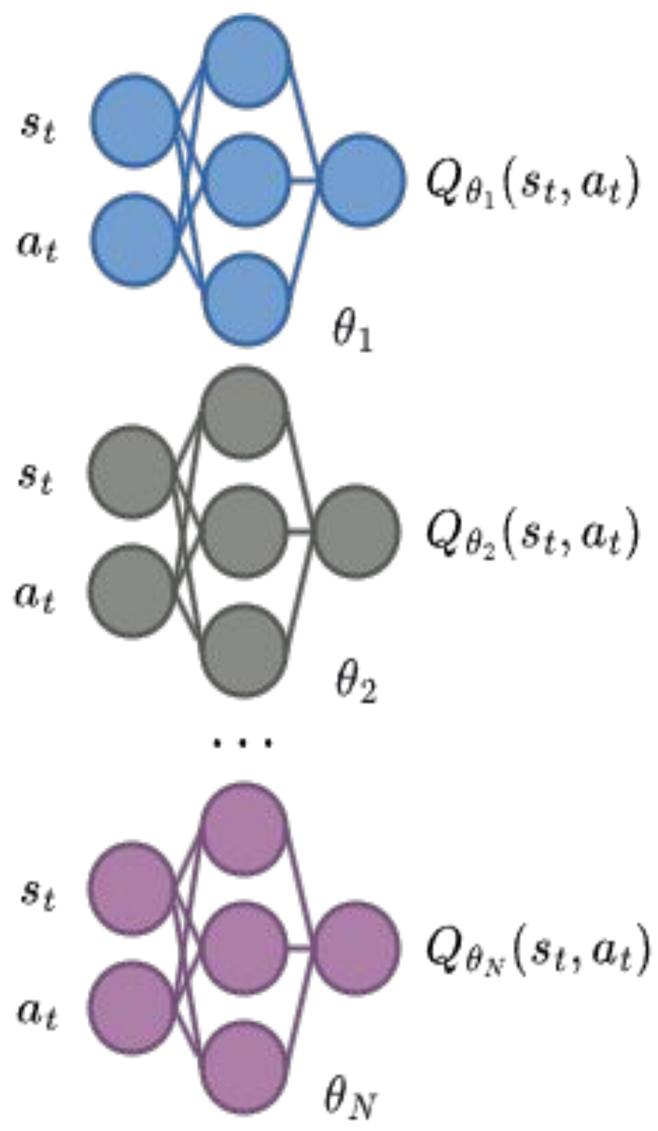

Figure 2: A Q-ensemble with N networks and Q-function outputs.

Result

The resulting Self-tuning Q-ensembles algorithm is implemented and thoroughly tested. It is capable of **adjusting K automatically** and **instantaneously** with response to the normalized bias (shaded in green).

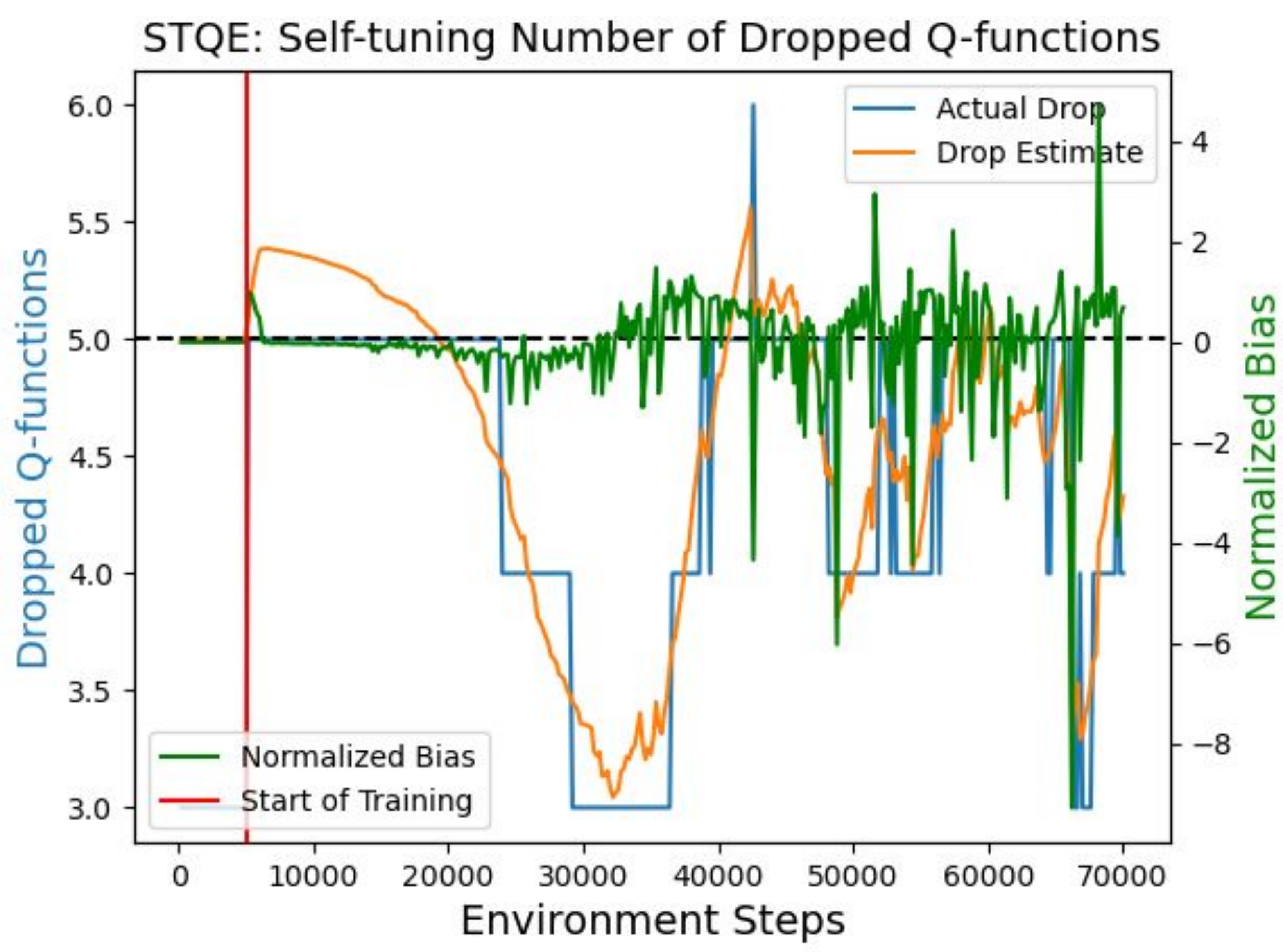

Figure 3: Adjusted dropped number estimates (orange), actual dropped Q-functions (blue) and normalized bias (green) against environment steps, tested on the OpenAI Gymnasium Pendulum-v0 environment.

Conclusion

The orange/blue dropped numbers change simultaneously with the bias in green, demonstrating the self-tuning bias reduction mechanism has been successfully implemented in the Q-ensembles method. Further testing on STQE with more challenging environments and repeated trials are strongly encouraged, and we envision STQE to contribute to the advancement of reinforcement learning in the near future.

References

[1] Y. Wu, X. Chen, C. Wang, Y. Zhang, Z. Zhou, and K. W. Ross, “Aggressive Q-Learning with Ensembles: Achieving Both High Sample Efficiency and High Asymptotic Performance,” International Conference on Information Processing Systems, 2021

[2] N. Dorka, T. Welschehold, J. Bodecker, and W. Burgard, “Adaptively Calibrated Critic Estimates for Deep Reinforcement Learning,” IEEE Robotics and Automation Letters, 2022