Final Report

Kangqi Wang
kangqiwa@usc.edu

April 22, 2024

1 Final Problem Description

The basic problem is to train an agent for map navigation. The map includes some irregular walls. The destination is located at the map's bottom-right corner. The agent will begin at a random location. The aim of the agent is to reach the destination while avoiding walls. We need to design reinforcement learning algorithms to achieve the target performance on all maps with minimal training episodes.

2 Final Solution

As the map size increases, the agent faces a significant sparse reward problem. Therefore, my solution introduces a negative bias to the traditional Temporal Difference (TD) algorithm, named NBTD. The new formula of TD target with NBTD is:

$$td\_target_{new} = td\_target_{old} - bias$$

We use SARSA and decaying ϵ-greedy algorithm as our basic algorithm, the new action-value function with NBTD for SARSA:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right) - bias$$

For the bias, the final solution uses a dynamic setting method. The bias reduces after a specific number of episodes N. The formula of dynamic bias is:

$$bias = \begin{cases} base & \text{if } 0 \leq episodes < N \\ base \times scaling\_ratio & \text{if } episodes \geq N \end{cases}$$

In the final solution, there are some hyperparameters we need to set including epsilon_decay, base, scaling_ratio and N.

3 Results

In the final report, I show the results on two hard maps. To show the efficiency of NBTD and the different TD algorithms, we train and test 3 methods for 2 hard maps, including SARSA without NBTD, SARSA with NBTD, QLearning with NBTD. For the setting of NBTD, the epsilon decay is 0.0001, the base as 0.005, scaling_ratio as 0.2 and N as 200. The train processes on the 2 hard maps of 3 methods are shown in Figure 1. Compare to SARSA w/ NBTD and SARSA w/o NBTD(orange

line and blue line), we can see the significant improvement in model training brought by NBTD. Compare to SARSA w/ NBTD and QLearning w/o NBTD(orange line and green line), we can see that after combining these two methods with NBTD, both of them perform well. And the effect of SARSA is slightly better.

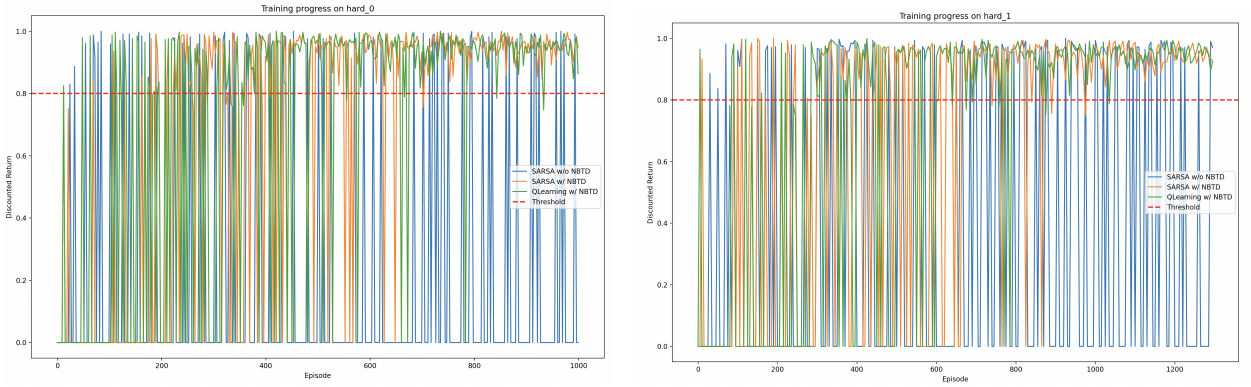

Figure 1: Training progress on two hard maps

The test results on two hard maps are shown in the Figure 2. The results show that both two agent can achieve at least 0.85 discounted return on hard_0 map and at least 0.90 discounted return on hard_1 map, with $\gamma = 0.997$.

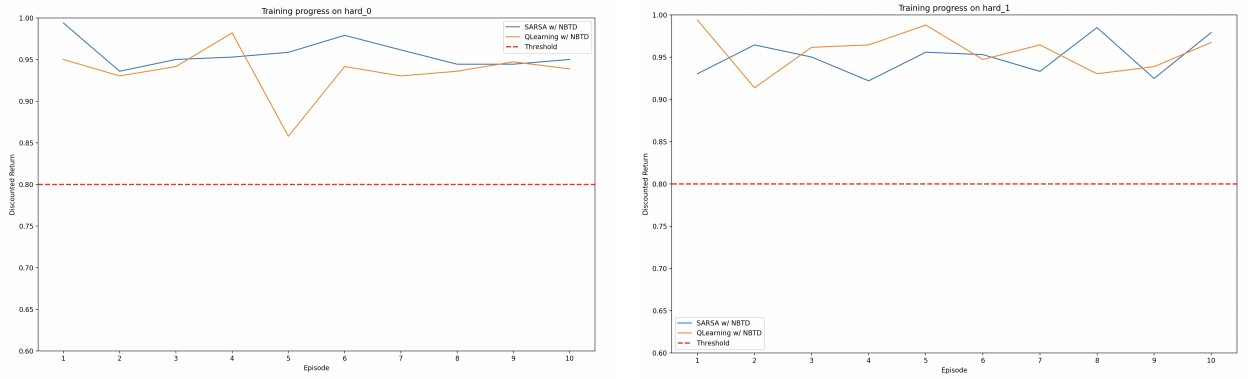

Figure 2: 10 Test episodes on two hard maps

4 Limitations

This solution mainly focus on the sparse reward problem. But there are still some limitations. (1) Adjustment Difficulty: Determining the size and timing of negative bias still requires manual intervention. We can explore a fully automatic bias adjustment strategy in the future. (2) Policy Bias: Negative bias can cause the learned policy to be biased towards avoiding punishment rather than optimizing long-term rewards. This could reduce the efficiency of learning and the effectiveness of the final policy. We can redesign the method to weigh long-term rewards more heavily.