REINFORCEMENT LEARNING
(CS3316)

Assignment Report

Assignment 2

SMALL GRIDWORLD
MONTE-CARLO AND TEMPORAL DIFFERENTIAL

Name: Kezhi Li ID: 520021911013
Date: 23 March 2023

Contents

1 Introduction

The goal of this assignment is to apply Monte-Carlo Learning and Temporal Differential Learning (TD) on a Gridworld to find the shortest way from any state to the terminal state in the grid. As shown in Fig. 1, the Gridworld has 6×6 grids and hence the state space can be denoted as $S = \{s_t | t \in 0, \ldots, 35\}$, with S_1 and S_{35} the terminal states. The action space is $A = \{n, e, s, w\}$, which represent moving north, east, south and west respectively. Moving out of the Gridworld will not leave the grid. We suppose that we do not know the whole backups in advance. Each movement get a reward of -1 until the terminal state is reached.

0	1	2	3	4	5
6	7	8	9	10	11
12	13	14	15	16	17
18	19	20	21	22	23
24	25	26	27	28	29
30	31	32	33	34	35

Figure 1: Gridworld

2 Experiment

For a given uniform random policy $\pi(n|\cdot) = \pi(e|\cdot) = \pi(s|\cdot) = \pi(w|\cdot) = 0.25$, the Monte-Carlo Learning and TD should be used to evaluate the policy in the Gridworld with incomplete backups.

3 Results

The state value function for the uniform random policy with complete backups is shown in Fig. 2 to compare the evaluation results in this assignment.

3.1 Parameters Chosen

λ of the whole assignment is chosen to be 1. α of TD is chosen to be 0.005.

3.2 Monte-Carlo Learning

After applying first-visit MC method on the uniform random policy with 10000 iterations, the state value function resulted in Fig. 3.

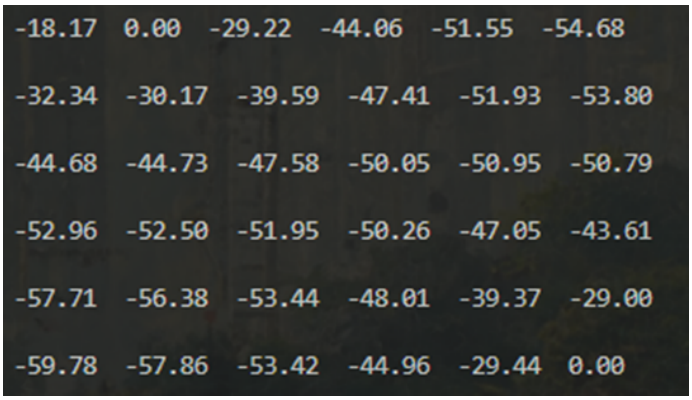

Figure 2: State function of uniform random policy

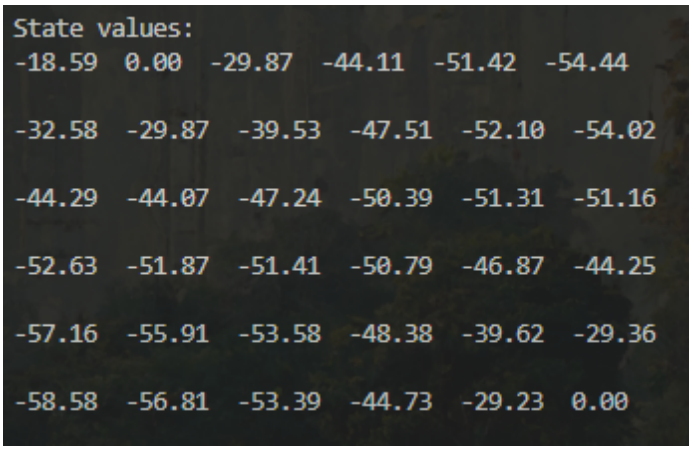

Figure 3: State value function evaluated by first-visit Monte-Carlo method.

Namely, the state value function of every-visit MC method is shown in Fig. 4.

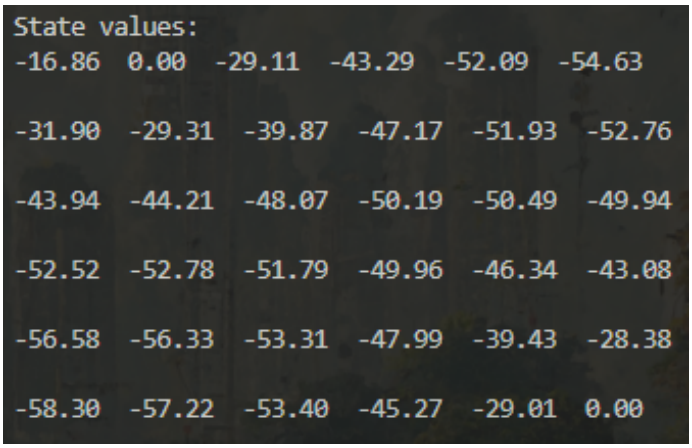

Figure 4: State value function evaluated by every-visit Monte-Carlo method.

3.3 Temporal Differential Learning

Fig. 5 shows the state value function evaluated by TD method after 1000000 samples.

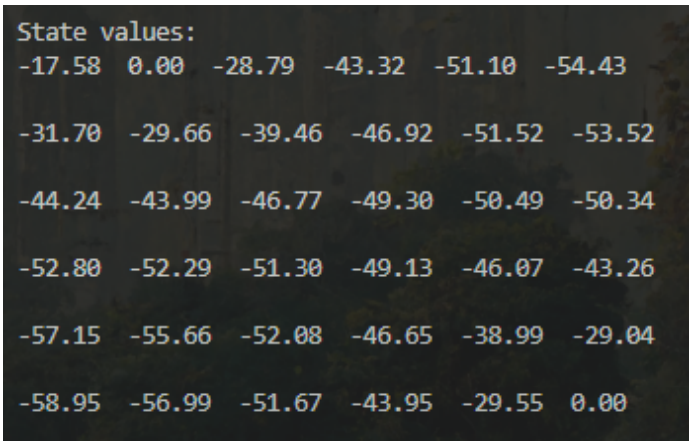

Figure 5: The state value function evaluated by TD method.

4 Conclusion

As shown in the above three figures about the result of the three methods, the state value functions are all close to the principle one. In my process of experiment, the TD method need to be fed with more samples to converge, which accords to the fact that the TD only take one step

to get refresh the value function. Besides, the TD method seems to converge more stably than the two MC methods, where the MC methods seem to have higher variance. More experiments can be conducted by testing different α and λ.