Ex 5 Temporal Difference Learning

Thursday, November 3, 2022 6:31 PM

Q.1 Temporal Vs Monte-Carlo

Monte Carlo – Requiring End of the episode to determine state value function

Temporal Difference – Only waits until the next step meaning at time $t+1 \rightarrow$ TD method uses observed reward R_{t+1}

$$\text{TD target} = R_{t+1} + V_{(S_{t+1})}$$

The Scenario where MC would be better where we have the episodic scenarios like **Black Jack**, where reward only affects after the termination of Episode, & updating intermediate state value function will not be of any benefit, It will only take the computational resource

Q.2 Q-Learning Vs SARSA

SARSA

$$Q(s,a) = Q(s,a) + \alpha\left[r + \gamma Q(s',a') - Q(s,a)\right]$$

Q Learning

$$Q(s,a) = Q(s,a) + \alpha\left[r + \gamma \max_a Q(s',a) - Q(s,a)\right]$$

Q Learning ← selects best next action

& SARSA next action depends on the policy it follows

both Converges to true value but at different rates.

2.a Q-Learning is Considered off policy because it follows the greedy approach i.e goes for the best action instead of Sticking to the Current policy therefore Q-Learning is the off policy Control method

2.b If the action is chosen greedily, Still the SARSA & Q-Learning are not Same, as

SARSA will choose Greedy action based on the Q value of the previous iterations. Q Learning on the other hand take the best action based on the Current Q

Hence Q-Learning & SARSA are not Same even after making the action greedy.

Q.3

3a) Which Algorithm is better would be affected if a wider range of α Values are used

With Small α values its more likely that it Converges to the optimal value as we can observe TD methods performing better for the Small α, & large α TD would have more oscillation around true Values.

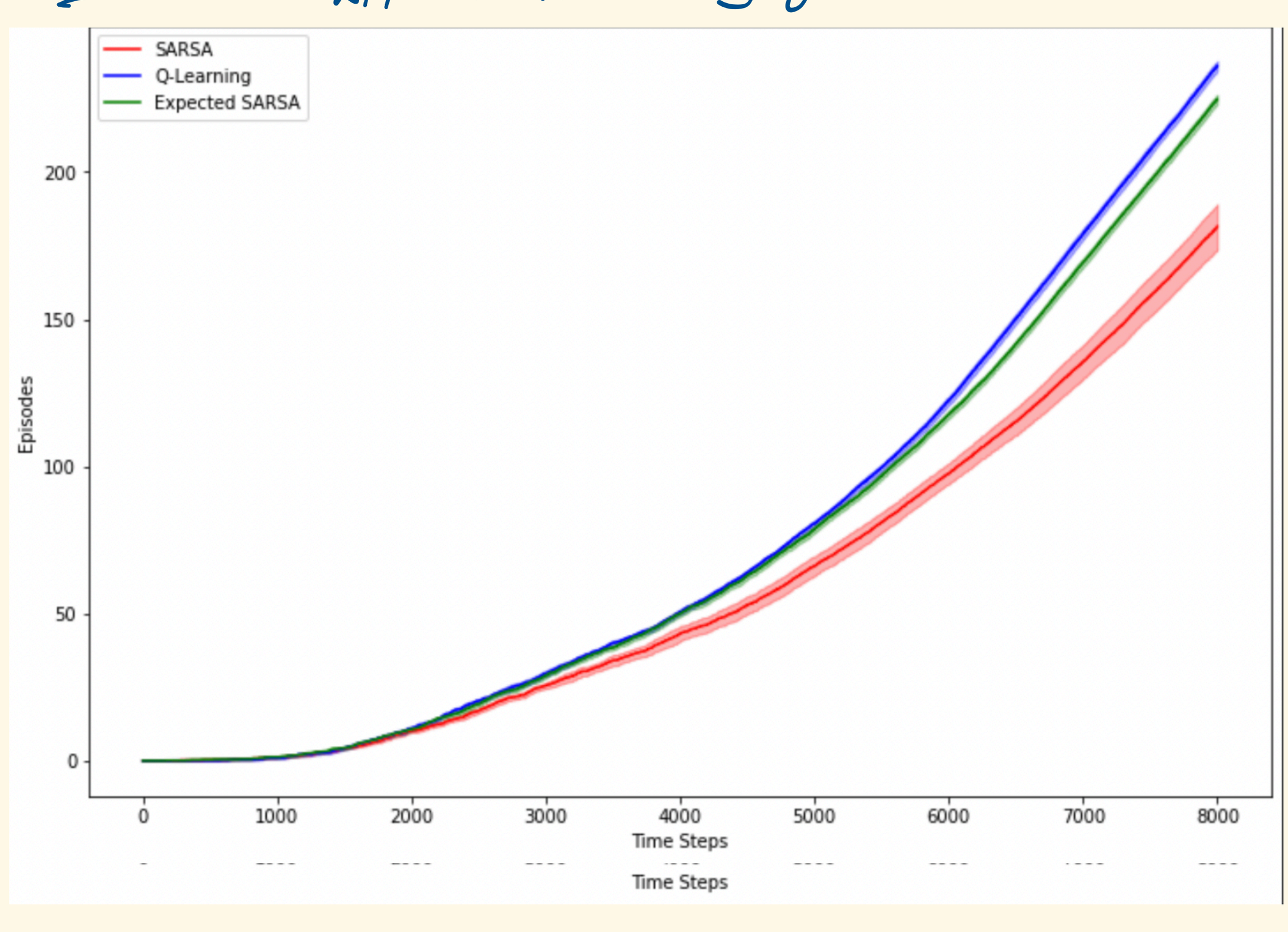
Even with wider range of α the MC would become any better, as we update MC only after the episode termination

3b

· Higher α would Cause the TD error to exaggerate, In turn would Cause the high oscillation for the estimated Values, & that estimate will shifting Causing high RMS error.

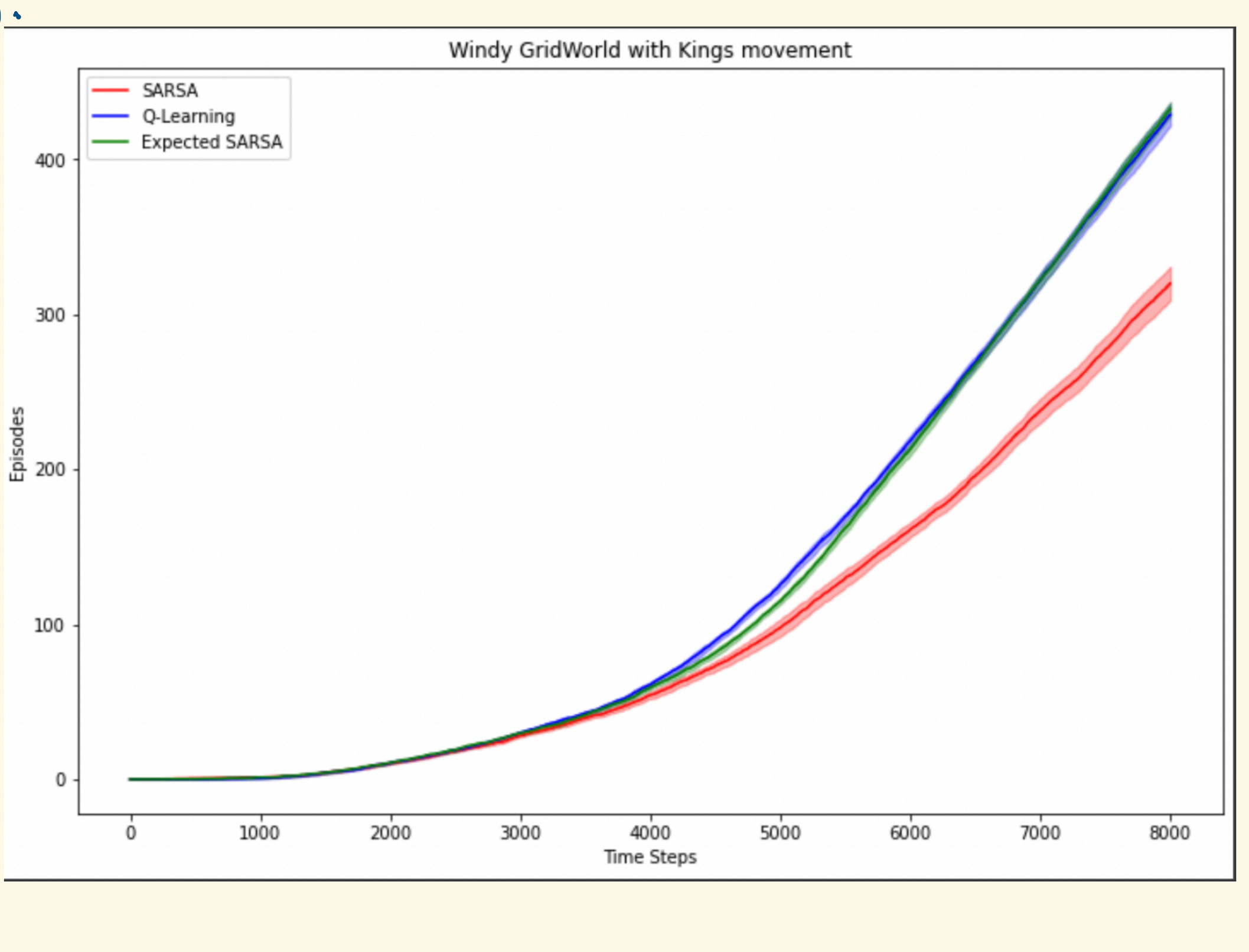
Q.4

4.a Applied on windy gridworld, No King Movement



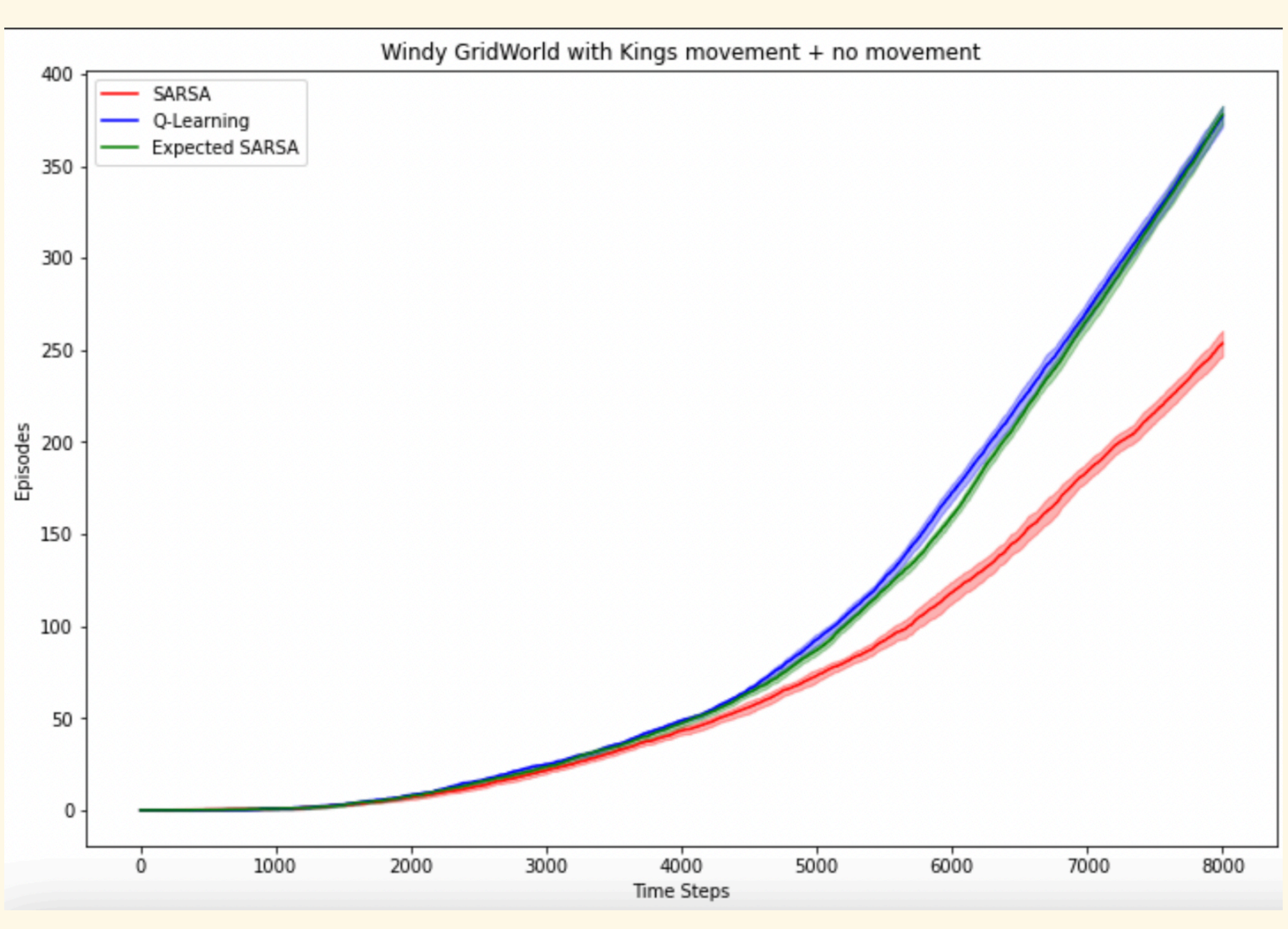
Around (200 – 250) episodes generated

4b.



Plot is better with the King's Movement

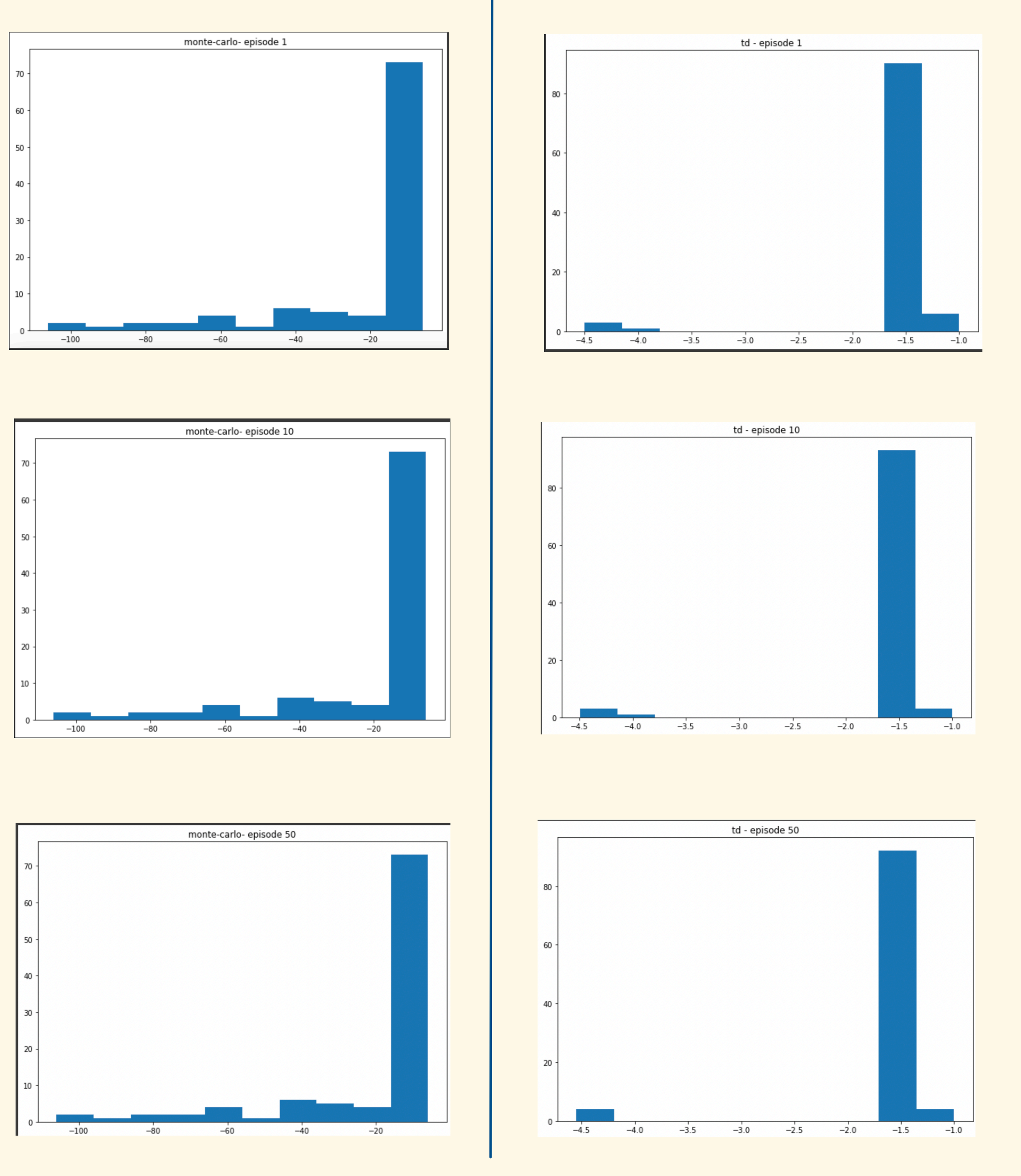
More Episodes generated More than 400



Plots is better than the 4a but worse the previous plot with only King's Movement

episodes More than 350 less than 400

Q.5 Monte Carlo | TD (0)



5.b

Monte Carlo give better value estimation with the lower bias but has higher Variance On the other hand, TD(0) has more stable estimation but shows high Variance, as the State value Converges TD(0) bias reduces.

When we use the the state value function for the Testing of the 100 episodes generated the Monte Carlo State Value estimation does not affect the result as the bootstrapping not needed for Monte Carlo.

Whereas for the TD(0) as bootstrapping is there it performs better with increasing $N = \{1, 10, 50\}$, as TD(0) has bias, more episodes will make it Converge to true Value & reduce bias.