Garrett Partenza
Advanced Reinforcement Learning
March 8, 2022

Exercise 5

Question 1. Read and understand Example 6.1. Is there any situation (not necessarily related to this example) where the Monte-Carlo approach might be better than TD? Explain with an example, or explain why not.

To answer this question, I tried to consider the main differences between Monte Carlo and TD learning. The main differences that come to mind is efficiency and bias. TD learning is more efficient considering it bootstraps initial values of states to make updates as opposed to averaging over the entire episode. As a result, if episodes are very long, you may still prefer TD. However, Monte Carlo is an unbiased model, and thus in situations where a biased model could prevent convergence you may prefer to use the Monte Carlo method.

Question 2. (a) Why is Q-learning considered an off-policy control method? (b) Suppose action selection is greedy. Is Q-learning then exactly the same algorithm as SARSA? Will they make exactly the same action selections and weight updates?

(a) Q-learning is considered an off-policy learning method because values of state-action pairs are estimated from assuming we follow a greedy policy, not by following the actual current policy of the agent. This contrasts with the SARSA method which estimates values of state-action pairs by following the current policy, and thus is considered to be on-policy.

(b) No, as I stated in my previous answer SARSA chooses actions by assuming we follow the current policy. This differs from following a greedy policy such as in Q-learning.

Question 3. (a) The specific results shown in the right graph of the random walk example are dependent on the value of the step-size parameter, α. Do you think the conclusions about which algorithm is better would be affected if a wider range of α values were used? (b) In the right graph of the random walk example, the RMS error of the TD method seems to go down and then up again, particularly at high α's. What could have caused this? Do you think this always occurs, or might it be a function of how the approximate value function was initialized?

(a) No, I do not believe that changing the values of α would effect the comparison of the two algorithms. In the MC method, increasing the step-size actually results in higher error. This is best shown by looking at the lines for TD step-size of 0.15 and MC step-size of 0.04. We can see that even with a lower step size, MC preforms worse, and since the chart indicates that increasing step size also increases MC RMS error, I do not see case in which a specific alpha would interchange the algorithms performance comparison.

(b) My guess would be that this effect is due to how the state-values were initialized. Prior to convergence, inaccurate state-value estimations caused the model to make decisions in a high-reward manner that was not a result of learning but from the fact that state-values were in an arbitrary state not yet converged.

(c) Larger walk problems allow the authors to show more granular analysis of how changing n can effect performance. A smaller walk problem would have shifted the value of n towards a smaller value. The change in the left terminal state reward to -1 allowed for a smaller n. I presume that if the left terminal state reward was 0 then a larger n would have resulted as the winner.

Question 4. Windy grid world.

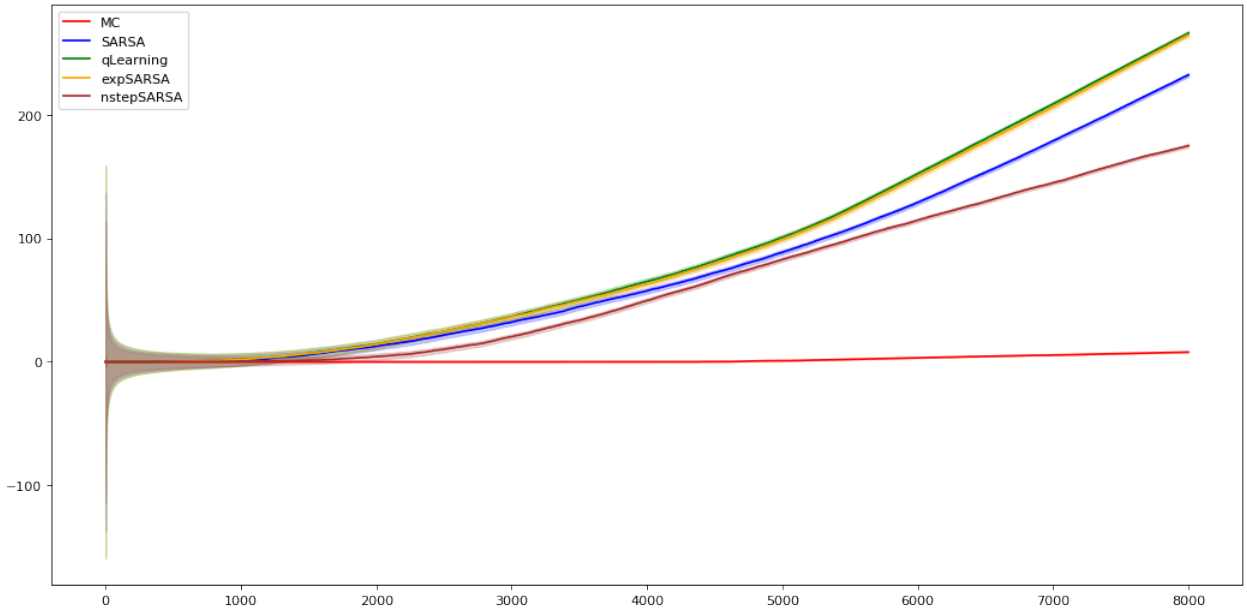


FIGURE 1. Windy

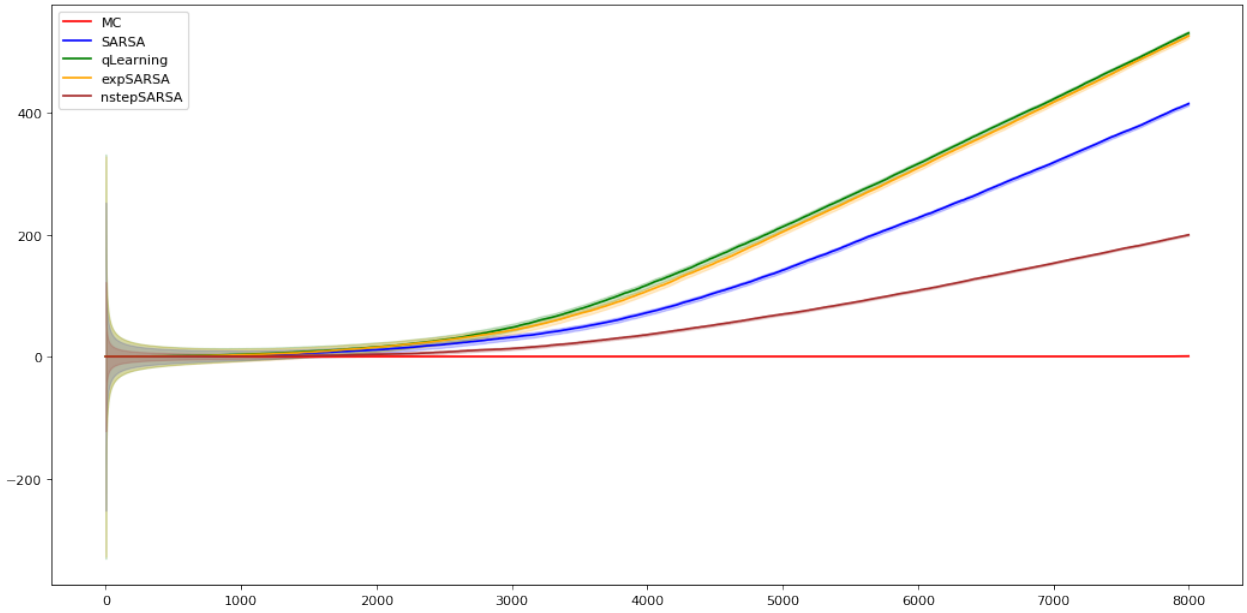


FIGURE 2. Windy + King

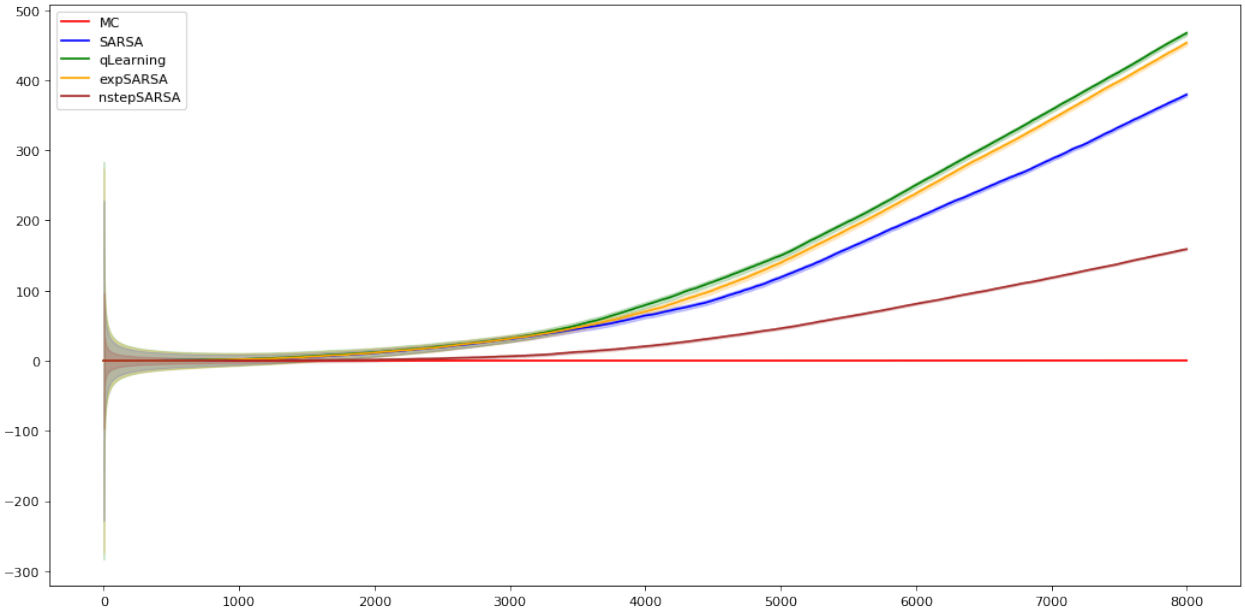


FIGURE 3. Windy + Nine

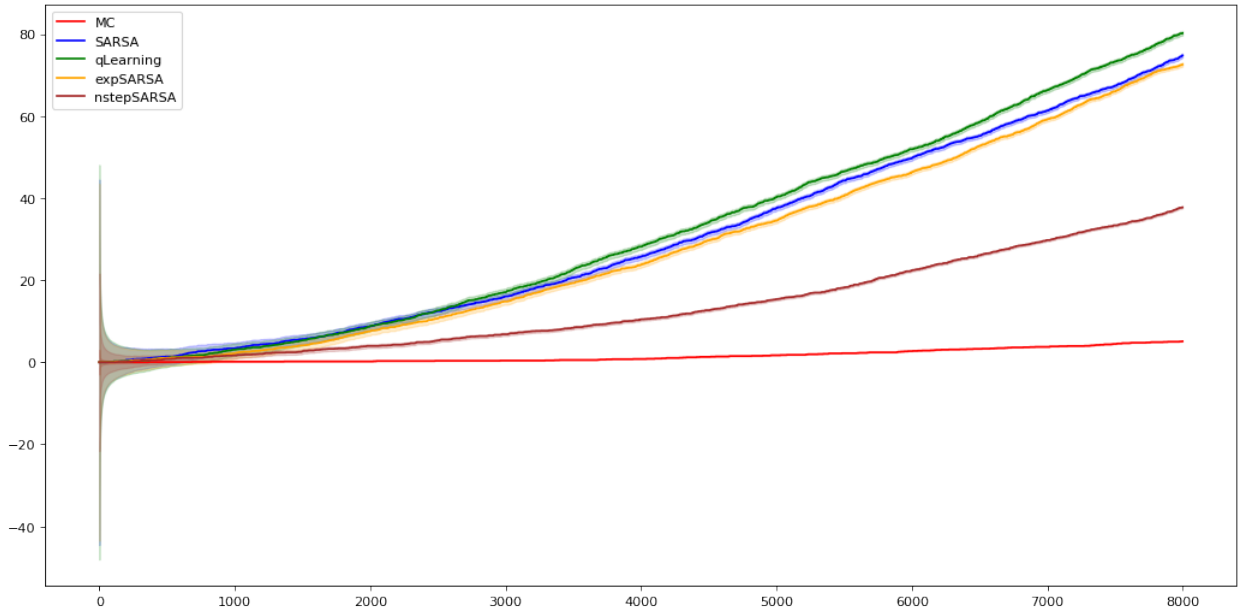


FIGURE 4. Windy + King + Stochastic

Question 5. Variance distributions.

Monte Carlo is a low bias method since it bases its value estimates off of truly observed episodes. This does result in high variance however. In TD and n-step SARSA, we initialize an arbitrary q-function and bootstrap our value estimates. As a result, we get a bias but lower variance and faster training time. We can see this demonstrated in my histogram plots, in that the distribution of G values for MC methods are much wider than that of the R estimates for TD(0) and n-step SARSA methods. Note that the x-axis range for all the histogram plots is fixed for each of the three trails. We can also see that increasing training time actually widens the distribution of G/R. Thus more training prior to evaluation results in lower bias higher variance for MC methods. We do not see that exaggerated of differences within the TD and N-step methods. It is certainly evident for MC however.

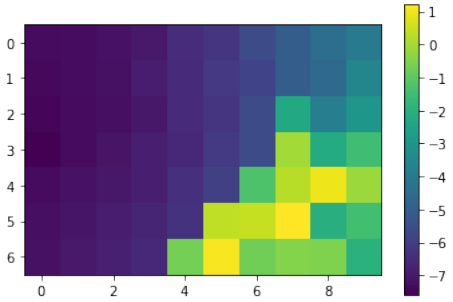
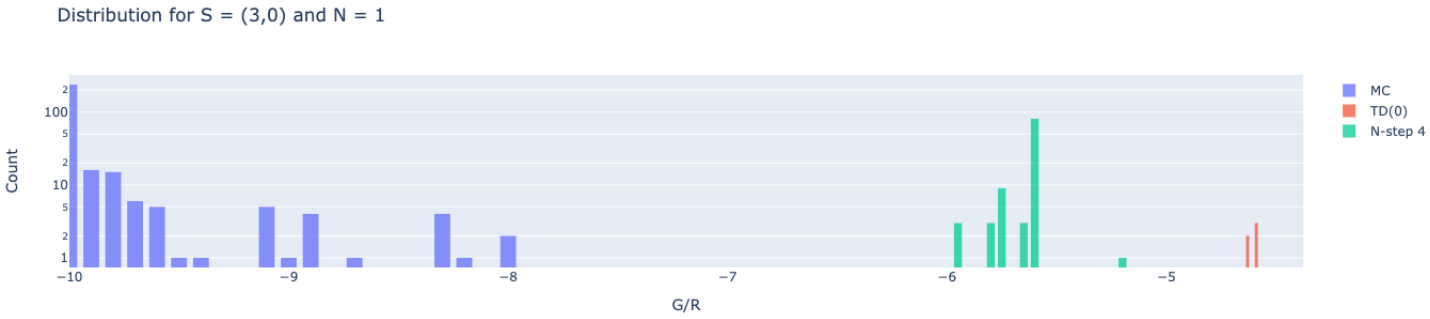
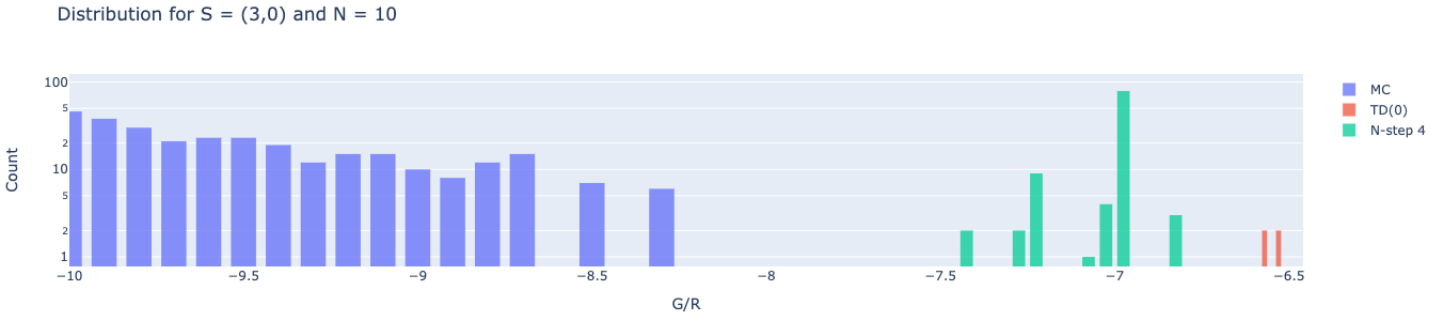
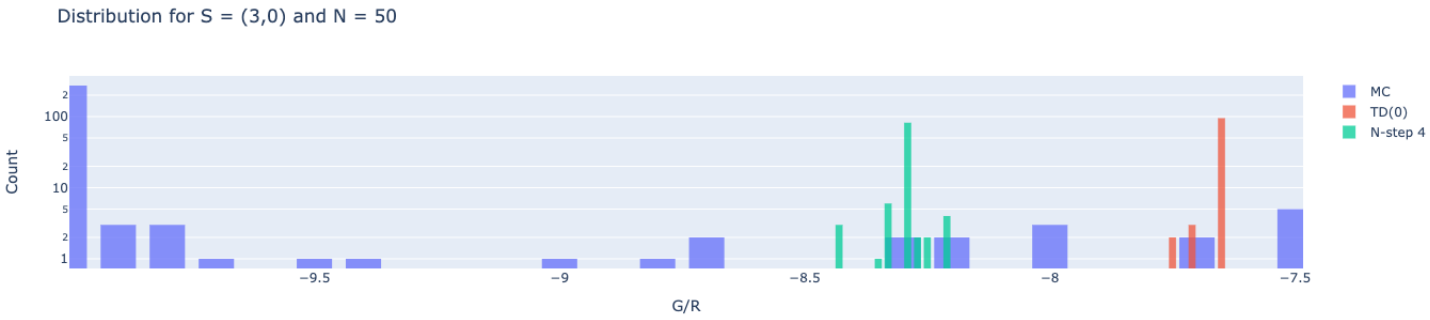


FIGURE 5. Learned Prior Policy by SARSA






Distribution for S = (3,0) and N = 50
Count
2
100
5
2
10
5
2
1
−9.5
−9
−8.5
−8
−7.5
G/R
MC
TD(0)
N-step 4