Assignement: ex4

Ian Steenstra

October 2024

1 Problem 1

- **(a):** The results would be identical because the same state is never revisited within an episode.
- **(b):** For the one-state MDP, the first-visit MC estimate is 10, while the every-visit MC estimate is 5.5 (the average of the returns 10, 9, 8, ..., 1).

2 Problem 2

- **(a):** See code and Figures 1 & 2.

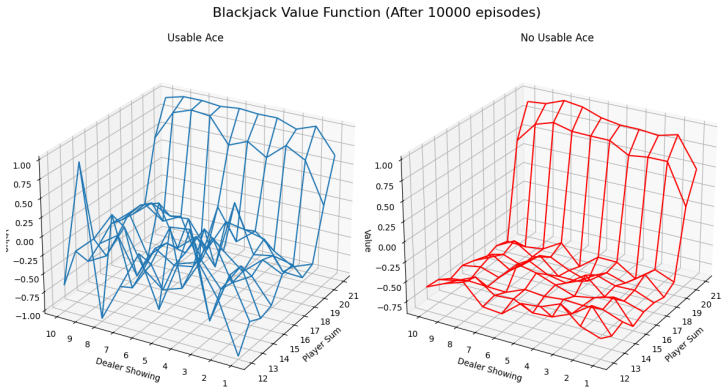

Figure 1: Problem 2(a): Blackjack Value Function (After 10k episodes)

- **(b):** See code and Figures 3 & 4.

3 Problem 3

- **(a):** See code and Figure 5.

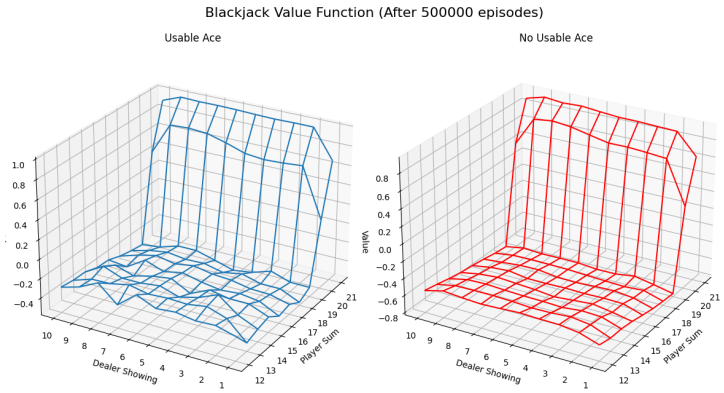

Figure 2: Problem 2(a): Blackjack Value Function (After 500k episodes)

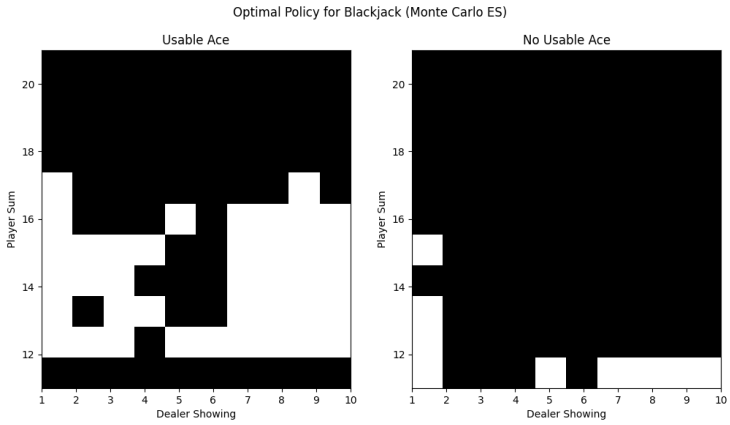

Figure 3: Problem 2(b): Optimal Policy for Blackjack (Monte Carlo ES)

- **(b):** The results show the importance of exploring starts in Monte-Carlo ES because it allows for exploration of other possible optimal valleys. Without exploration, the model may settle into a local optimal.

4 Problem 4

- **(a):** Start with Equation 5.7:

$$V_n = \frac{\sum_{i=1}^{n-1} W_i G_i}{\sum_{i=1}^{n-1} W_i} \quad \text{for } n \geq 2$$

Convert to V_{n+1}:

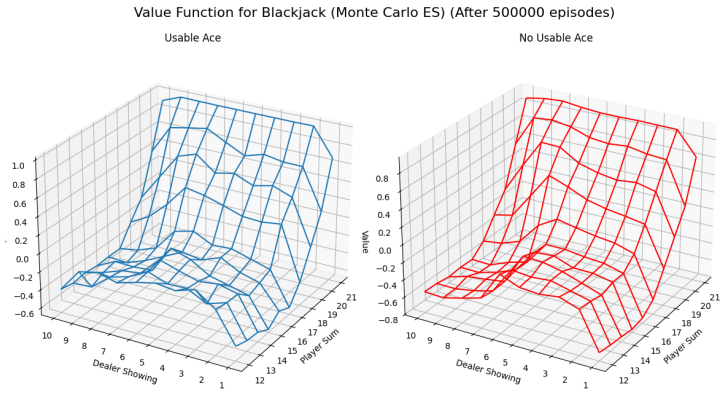

Figure 4: Problem 2(b): Value Function for Blackjack (Monte Carlo ES)

$$V_{n+1} = \frac{\sum_{i=1}^{n} W_i G_i}{\sum_{i=1}^{n} W_i}$$

Separate out the nth term:

$$V_{n+1} = \frac{W_n G_n + \sum_{i=1}^{n-1} W_i G_i}{W_n + \sum_{i=1}^{n-1} W_i}$$

Substitute using equations from the book:

$$\begin{aligned} V_{n+1} &= \frac{W_n G_n + V_n C_{n-1}}{C_n} \\ &= \frac{W_n G_n + V_n (C_n - W_n)}{C_n} \\ &= \frac{W_n G_n + V_n C_n - W_n V_n}{C_n} \\ &= \frac{V_n C_n + W_n (G_n - V_n)}{C_n} \\ &= V_n + \frac{W_n}{C_n}(G_n - V_n) \end{aligned}$$

- **(b):** This is correct because, at time t, only the probability of the current action A_t matters; future actions haven't happened yet and are irrelevant for weighting the observed return at this time step. The probabilities of subsequent actions are incorporated incrementally in later updates as the episode unfolds, thus achieving the effect of the full importance sampling ratio without the need to wait until the end of the episode.

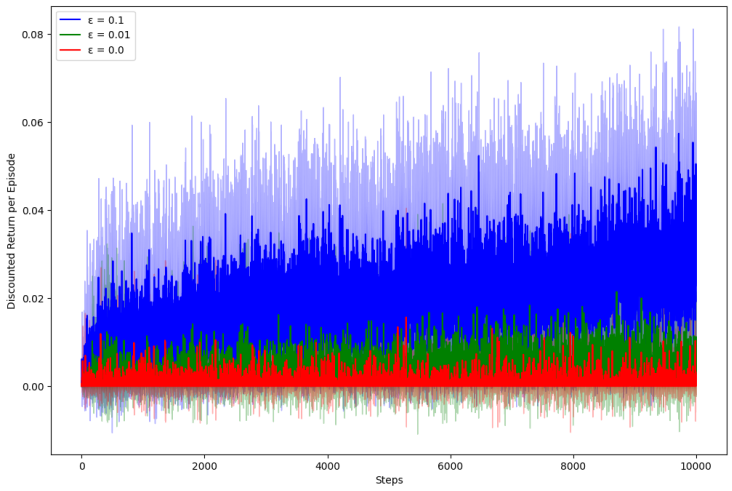

Figure 5: Problem 3(a): Four Rooms plot

5 Problem 5

- **(a):** See code and Figures 6 & 7.
- **(b):** See code and Figures 8 & 9.
- **(c):** The off-policy methods rapidly increased in discounted reward initially and fluctuated as the number of episodes increased, while the on-policy more slowly increased is discounted reward but fluctuated less. Lastly, the off-policy performed better than the on-policy in general.

  As for the two versions of the race track, V2 seems harder than V1 based on the results of the on-policy method. My best guess is because of the variance in the starting positions. For V1, the method of getting to the goal states for any starting position is relatively the same. Unlike V2, where each starting position leads to a different boundary right away given the edge starting in the bottom left and up towards the goal states.

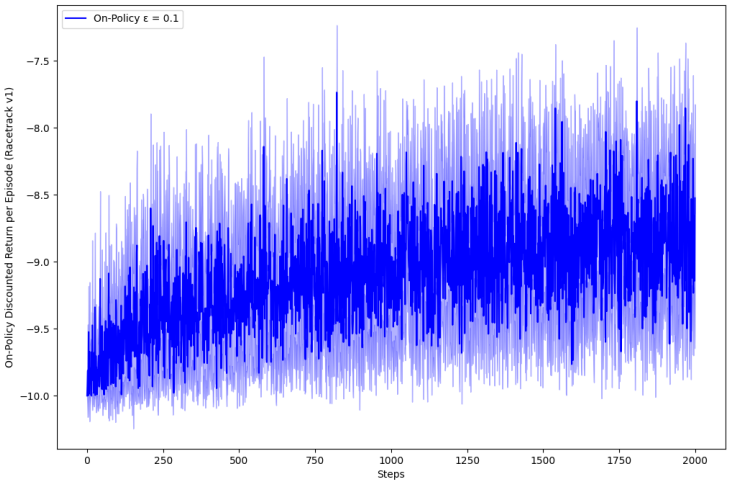

Figure 6: Problem 5(a): On-Policy Racetrack V1

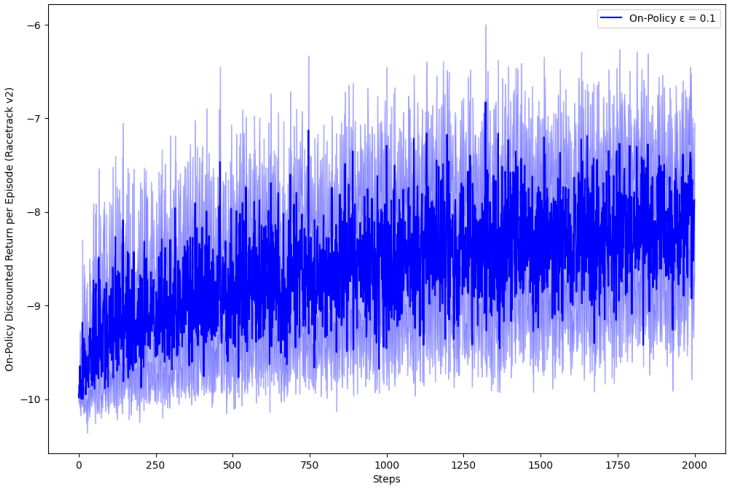

Figure 7: Problem 5(a): On-Policy Racetrack V2

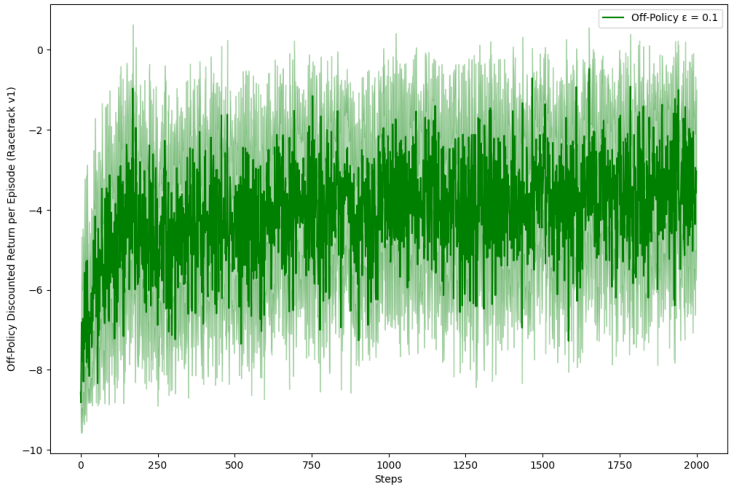

Figure 8: Problem 5(b): Off-Policy Racetrack V1

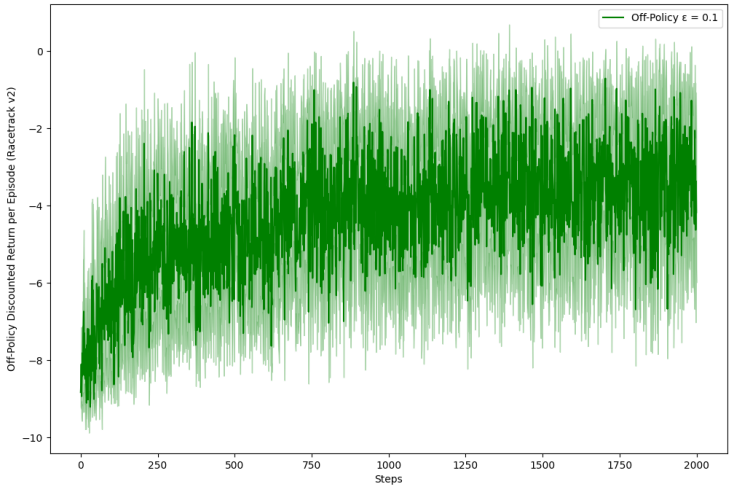

Figure 9: Problem 5(b): Off-Policy Racetrack V2