'''

CS 5180 Fall 2022

Exercise 4: Monte-Carlo Methods

Hongyan Yang

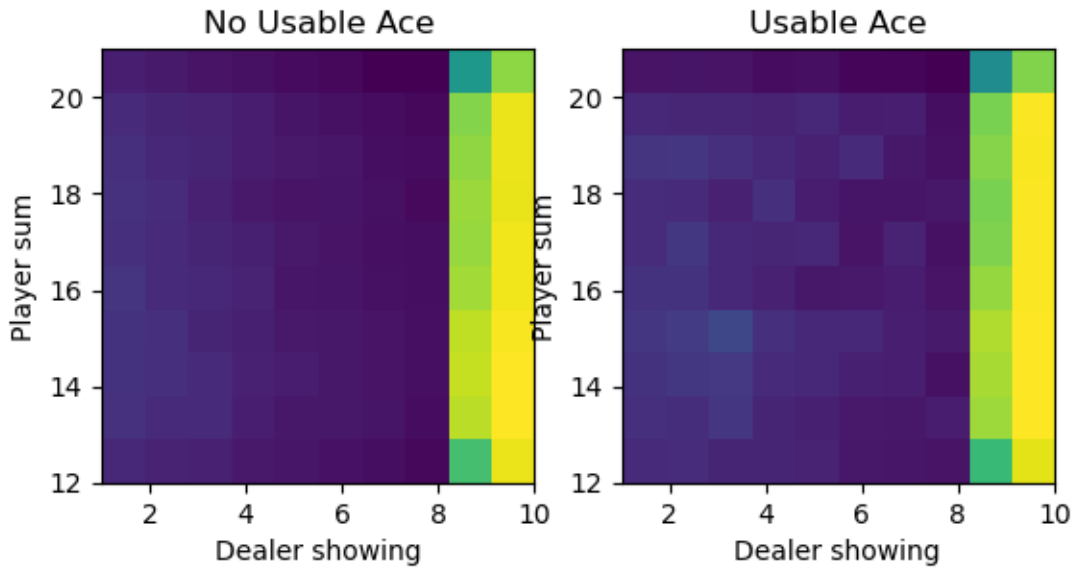
'''

1. 2 point. (RL2e 5.2, 5.5, 5.8) First-visit vs. every-visit.

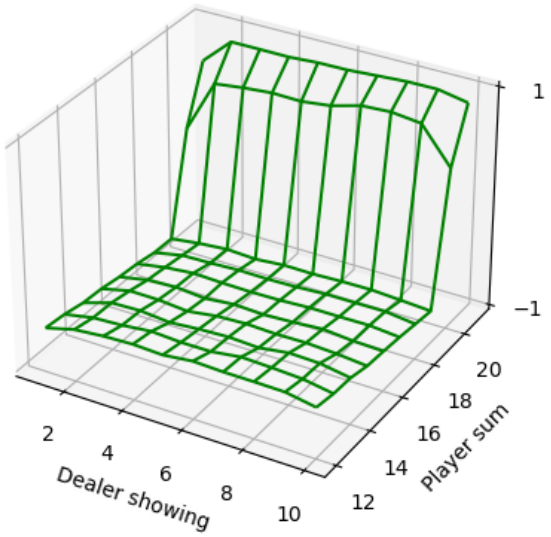
(a) If every-visit MC was used, the results will NOT be different. Because of the rule of Blackjack and the strategy to count usable ace as 11. In a certain trajectory, the play's states will not be duplicated, making it no different between every-visit and first-visit MC in this specific situation.

(b) Based on the info given, this 10 step trajectory consisting 9 visits to the nonterminal state and the last step of one visit to the terminal state.
First-visit estimate: $V(s) = Returns(St) = 10$
Every-visit estimate: $V(s) = average(Returns(St)) = \frac{10+9+\cdots+1}{10} = 5.5$

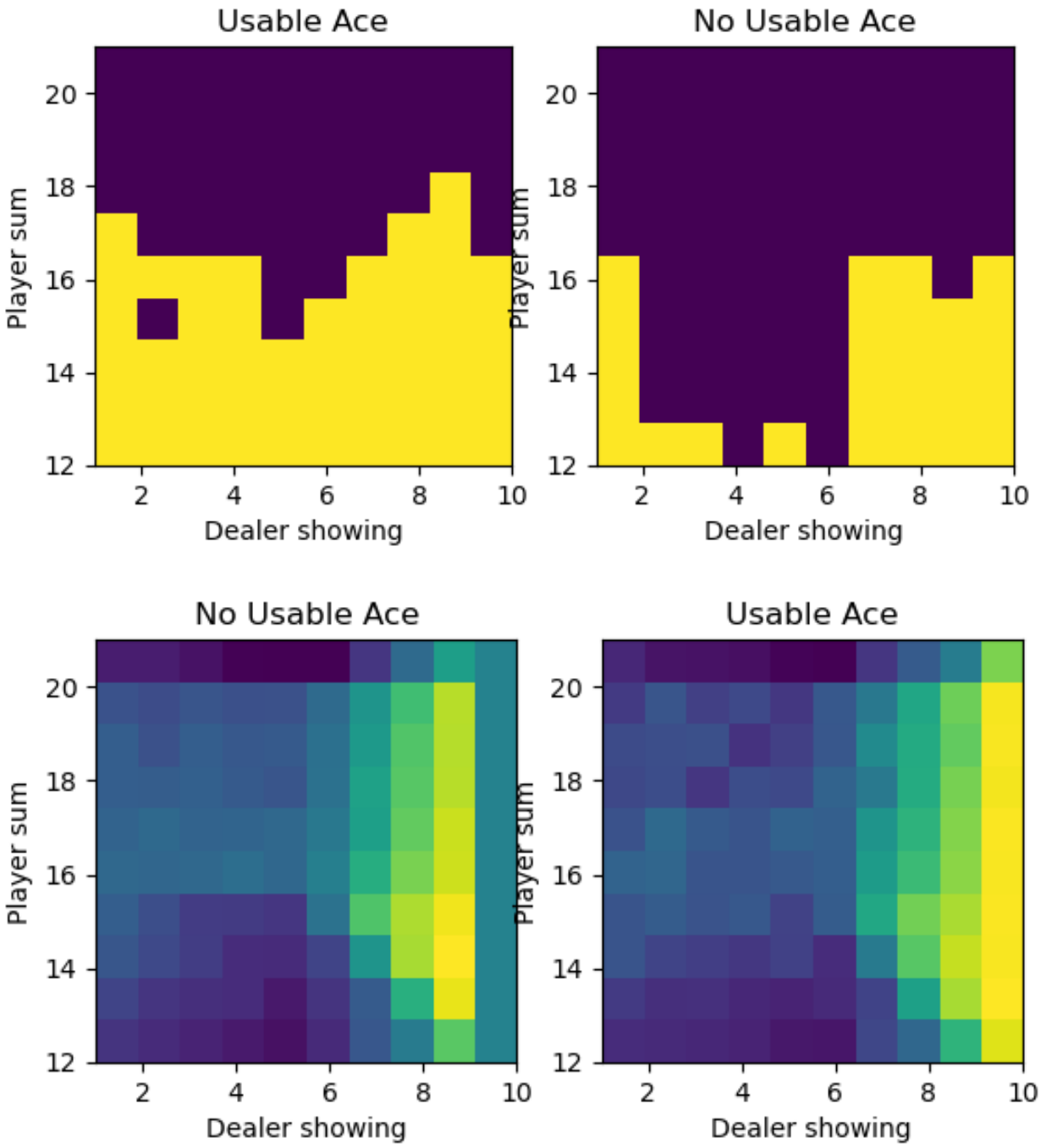
2. 2 points. Blackjack.
    a. Plot

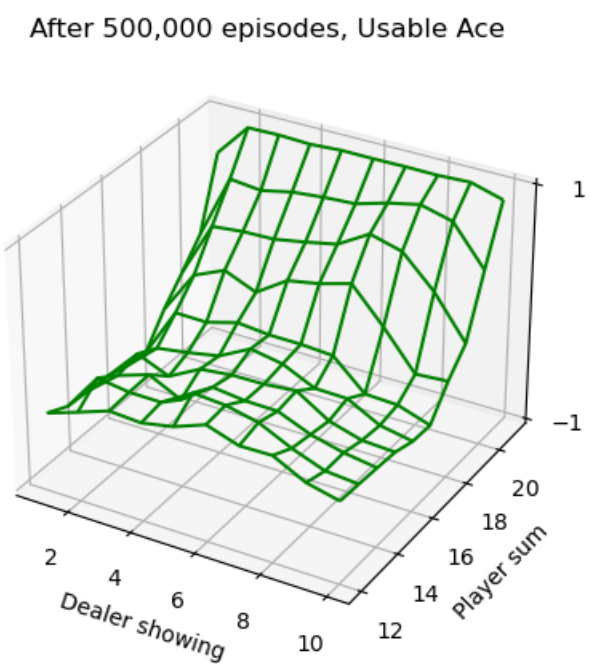


After 500,000 episodes, No Usable Ace

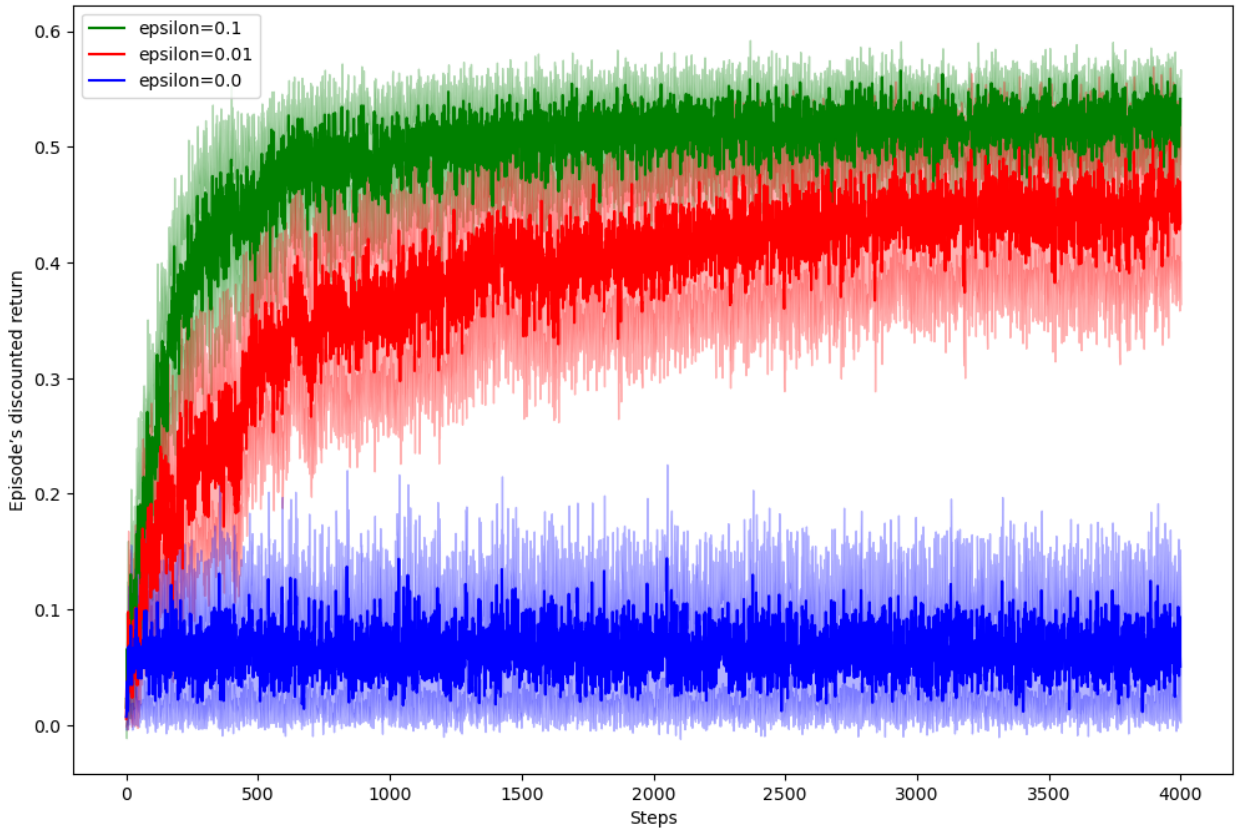


b. Plot


Usable Ace
No Usable Ace
No Usable Ace
Usable Ace
Player sum
Dealer showing
20
18
16
14
12
2
4
6
8
10


After 500,000 episodes, Usable Ace
1
−1
2
4
6
8
10
Dealer showing
12
14
16
18
20
Player sum

3. 2 points. Four Rooms, re-visited.
    a. Plot



    b. Written

With an epsilon = 0, in other words, without an exploration starts, the agent will stick to the initial "optimized" policy which is determined by the original Q_values setting. Then the agent will continue with the same policy and have no chance to try a different trajectory and update the Q values. Thus, it shows the importance of doing exploring starts in Monte-Carlo ES.

4. 1 point. (RL2e 5.10, 5.11) Off-policy methods

4. off-policy methods

(a) Given $V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$, $n \geq 2$,

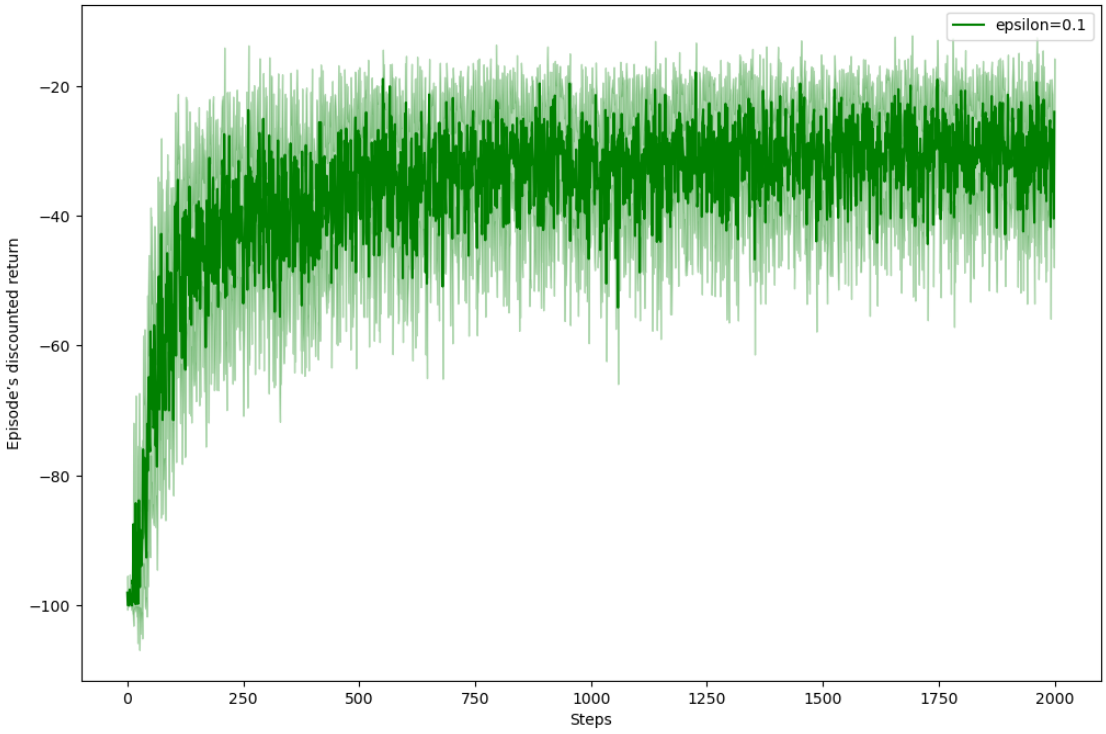
$$V_{n+1} = \frac{\left(V_n \times \sum_{k=1}^{n-1} W_k + W_n V_n\right) - W_n V_n + W_n G_n}{\sum_{k=1}^{n-1} W_k + W_n} = V_n + \frac{W_n (G_n - V_n)}{\sum_{k=1}^{n-1} W_k + W_n}, \quad n \geq 1$$

$$= V_n + \frac{W_n}{C_n}(G_n - V_n), \; n \geq 1 \text{ and } C_{n+1} = C_n + W_{n+1}$$

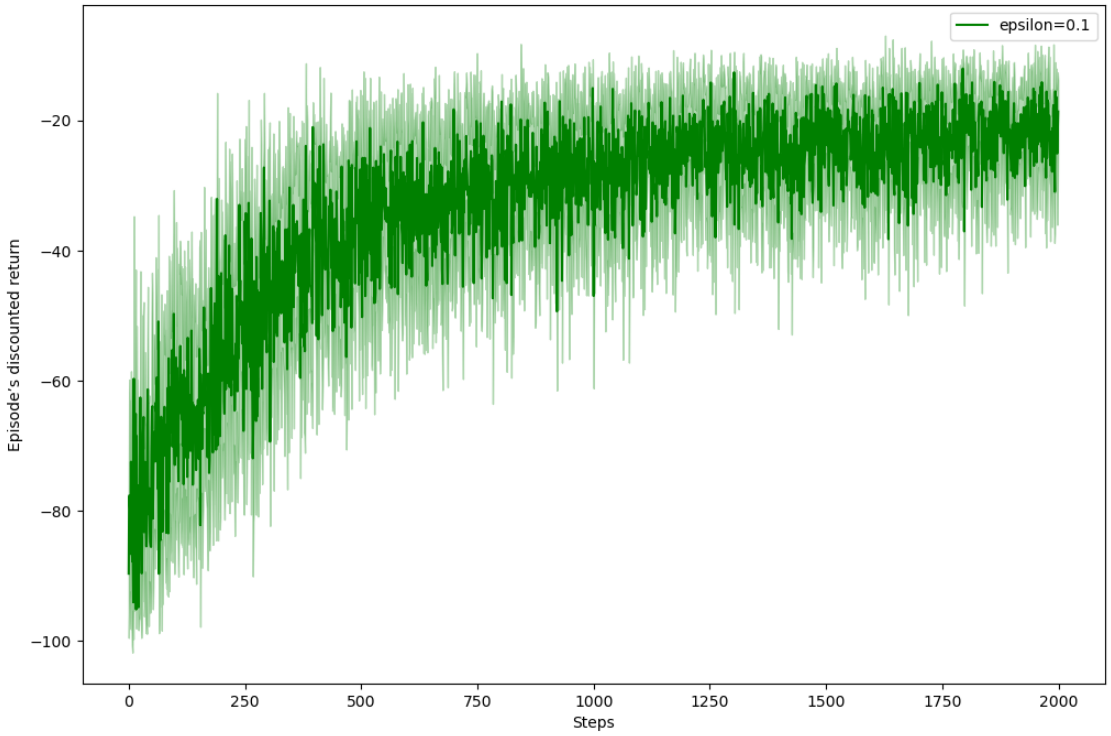
(b) According to the algorithm of off-policy MC control, the target policy is the greedy policy with respect to Q. In other words, $\pi(A_t|S_t) = 1$ for $t \in [0, T-1]$, ∴ the W update $= \frac{\pi(A_t|S_t)}{b(A_t|S_t)} = \frac{1}{b(A_t|S_t)}$

5. 3 points.[5180] (RL2e 5.12) Racetrack.
   a. Plot

Plot of V1

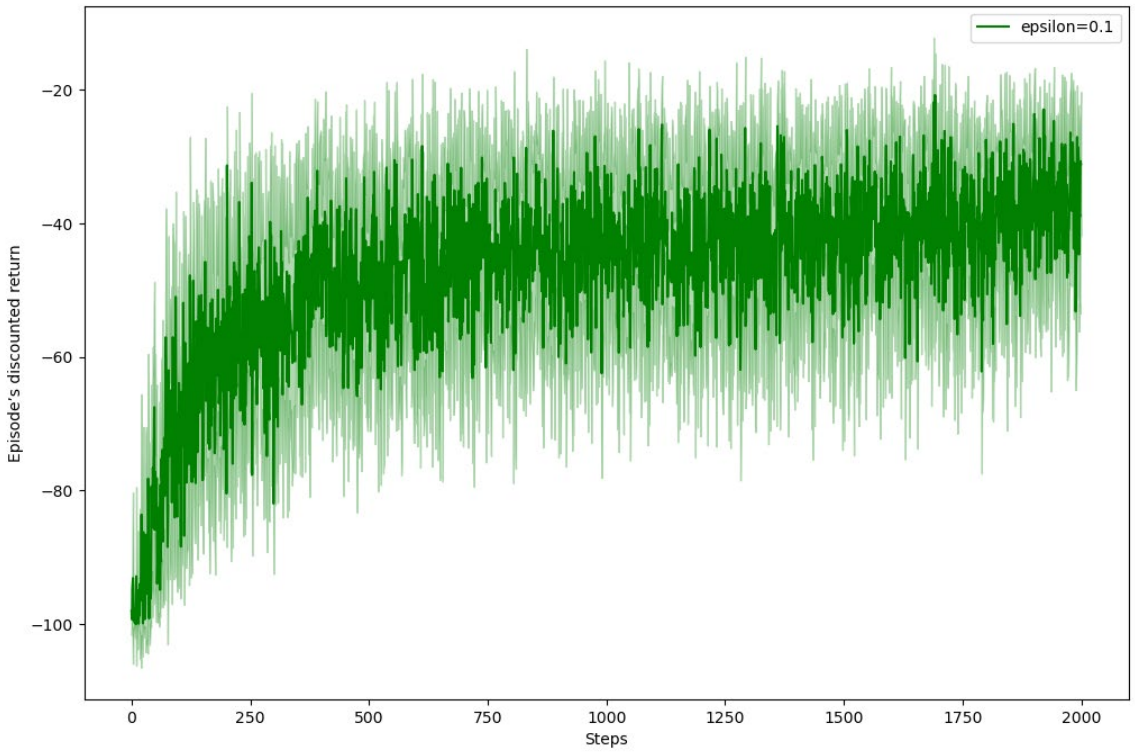


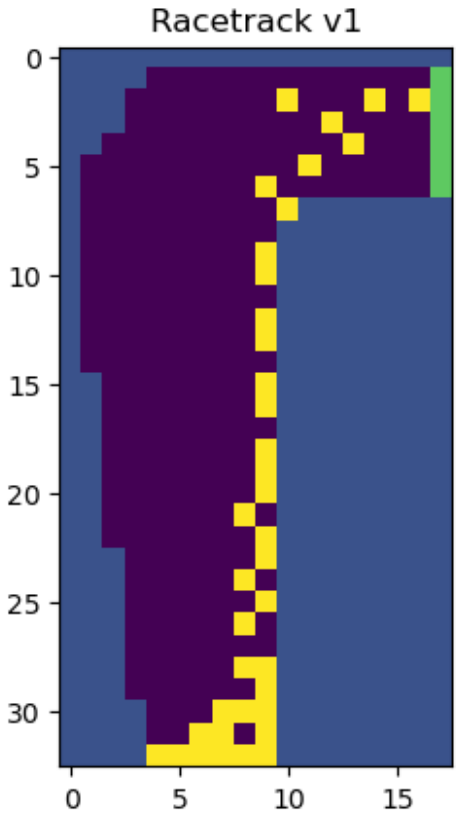
Plot of V2



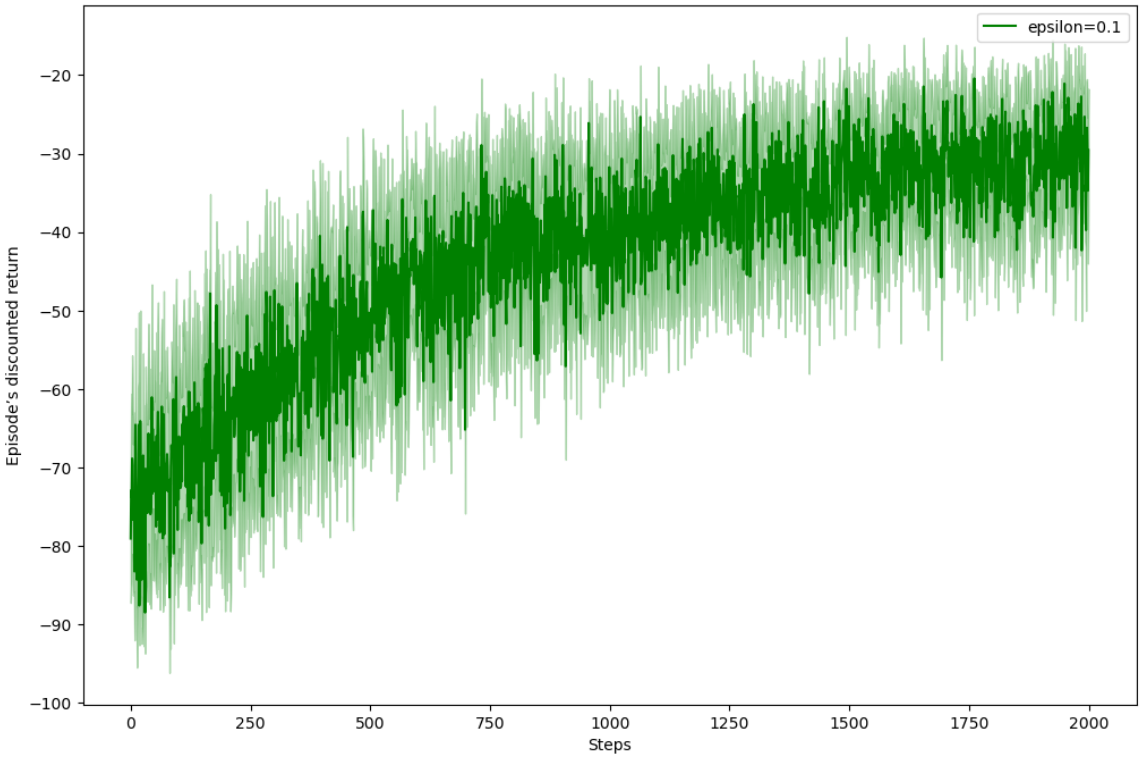
b. Plot

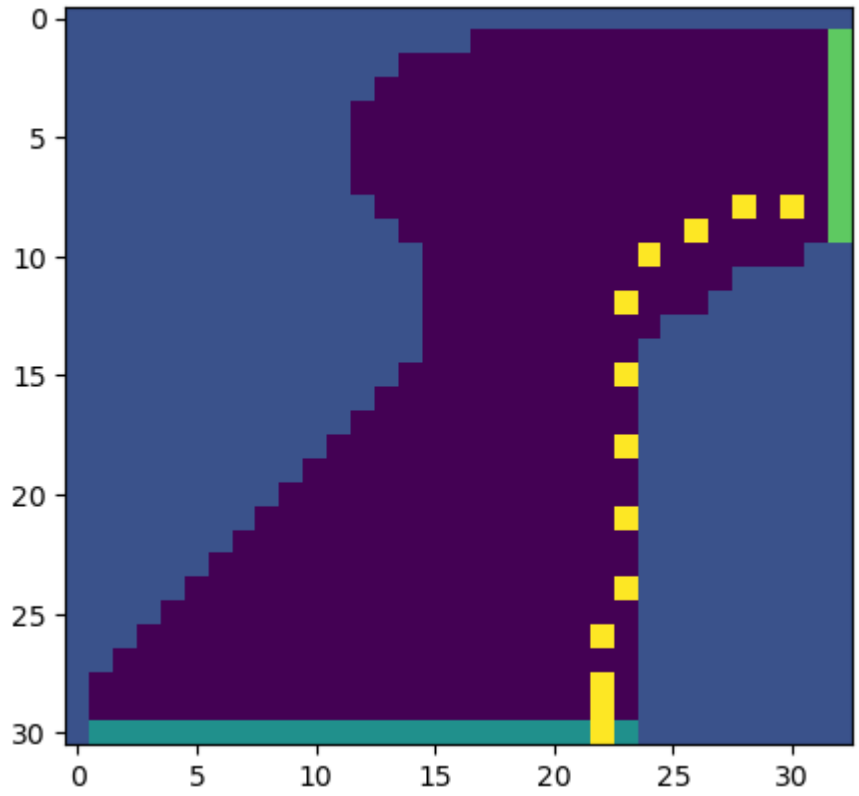
Plot of V1





Plot of V2


epsilon=0.1
Episode's discounted return
Steps
−20
−30
−40
−50
−60
−70
−80
−90
−100
0
250
500
750
1000
1250
1500
1750
2000


0
5
10
15
20
25
30

c. Written

There's no significant difference between the on-policy and off-policy methods, except that the off-policy method runs much more slowly than the on-policy one in this specific racetrack domain. The reason is that the greedy agent did not learn much from the behavior policy in the early stage and it learn from each trajectory inefficiently.

The interesting difference between the two tracks is that the v2 one is relatively quicker to learn from because it has a "wider" turn than v1, making the agent less likely to hit the track. So, each trajectory from v2 is more efficient.