RL HW EX4:

Question 1:

a. The results would be pretty much the same. This is because in this blackjack example their is only one way that the same state can be reached twice in the same episode is if an ace is flipped from 11 to 1 and you end up with a total value that is the same as you have already encountered. This is a rare case and thus wouldn't have much of an effect on the state-value functions.

b. First visit:

G = 10/1 = 10

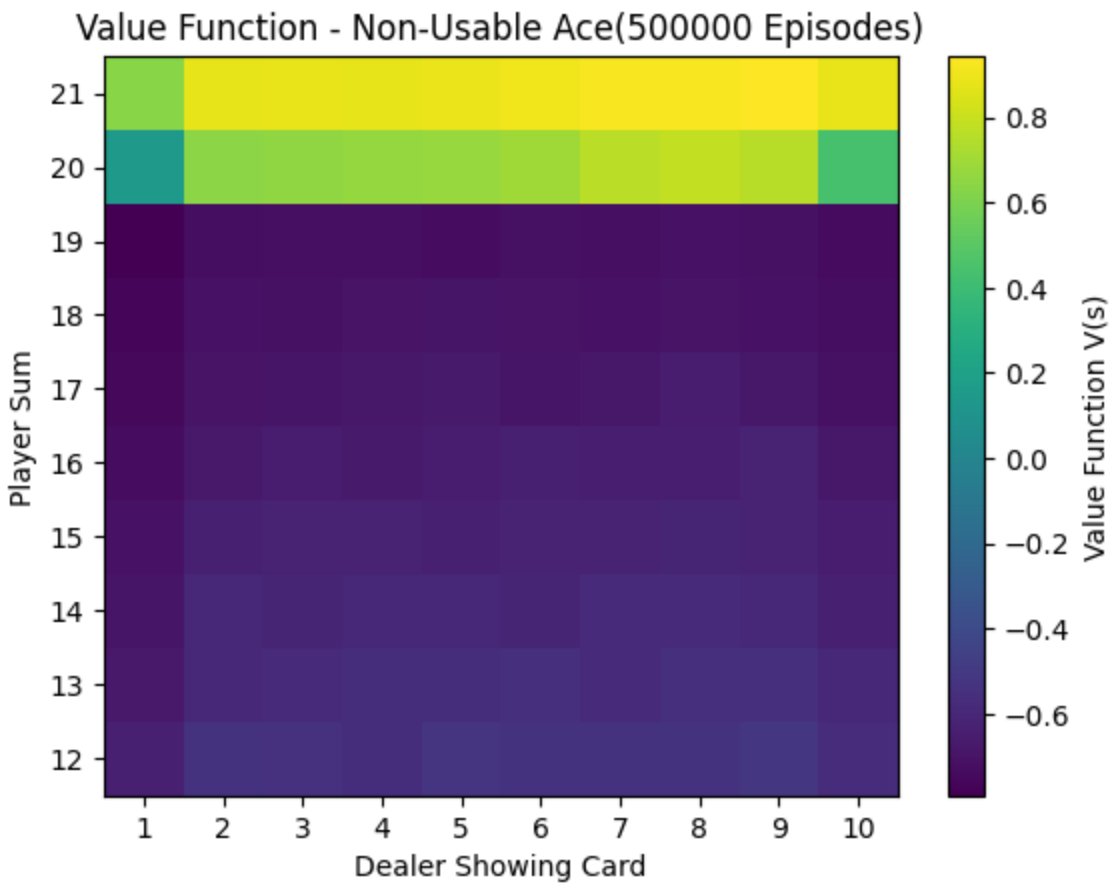
Every visit:

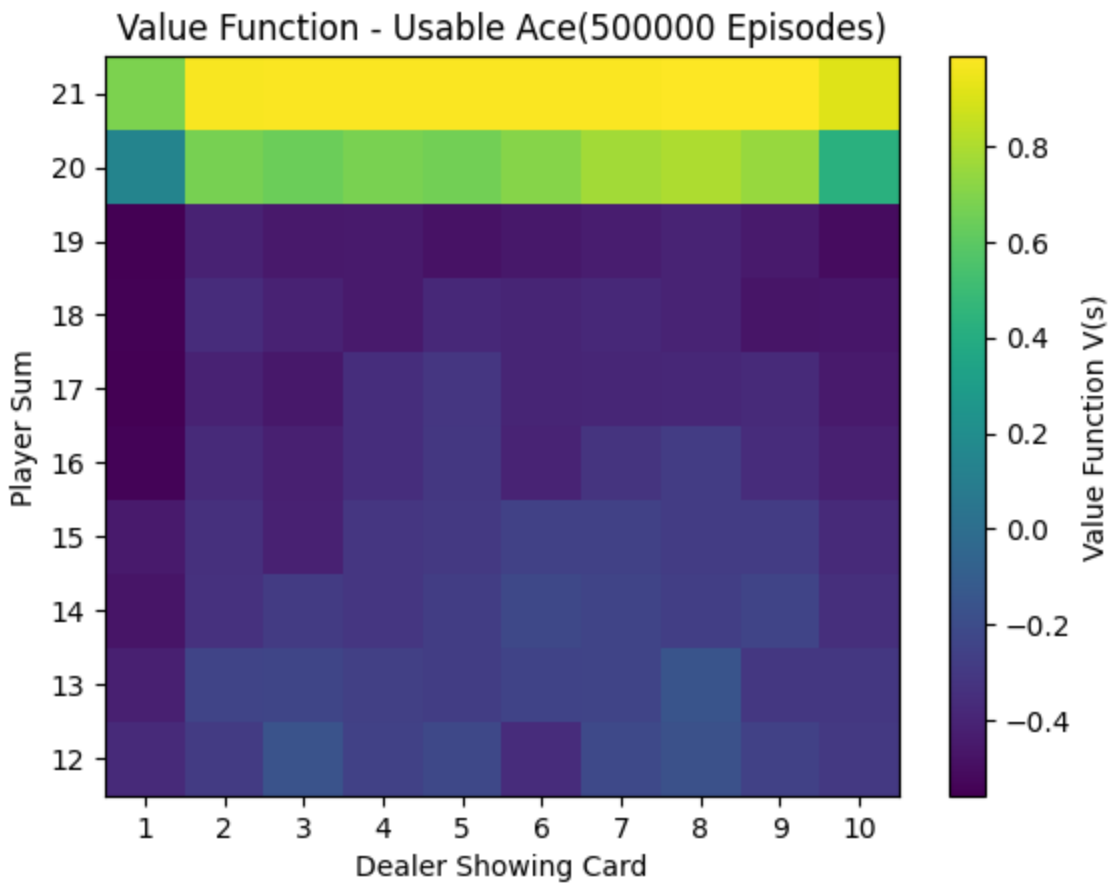
G = avg(1,2,3,4,5,6,7,8,9,10)

G = 5.5

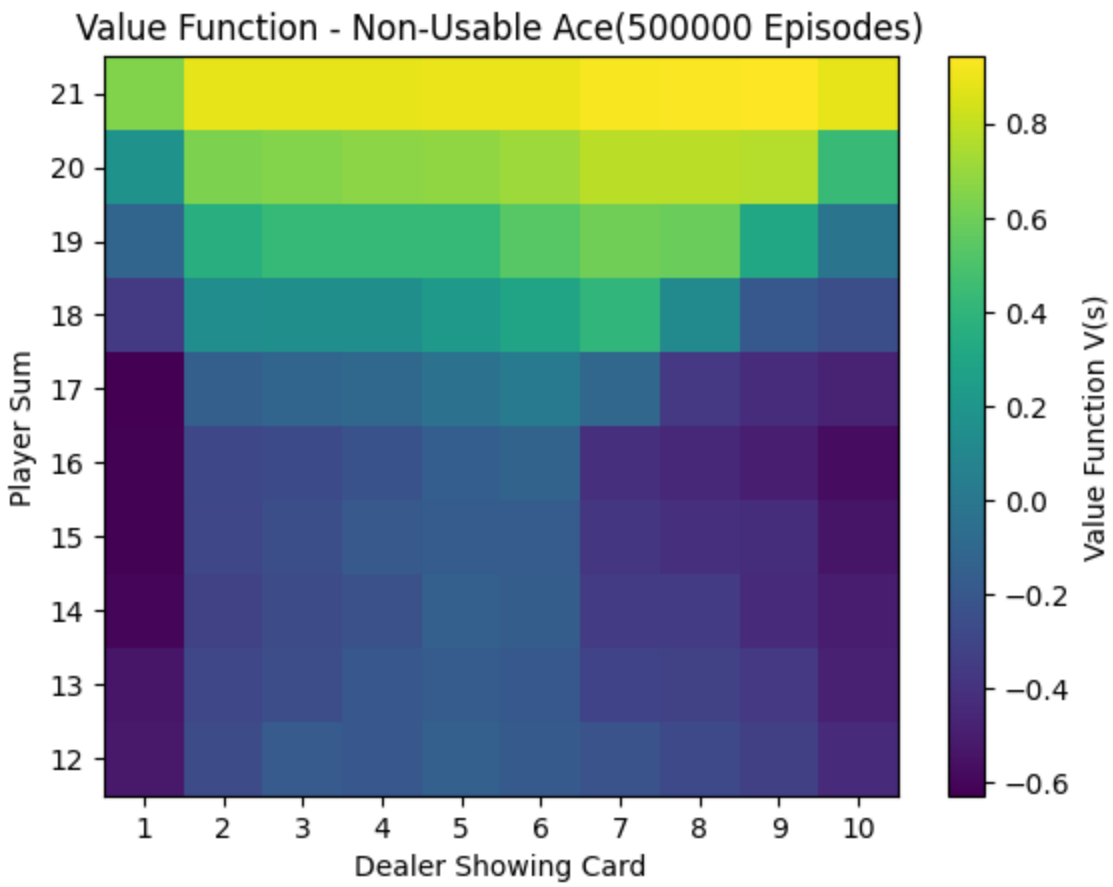
Question 2:

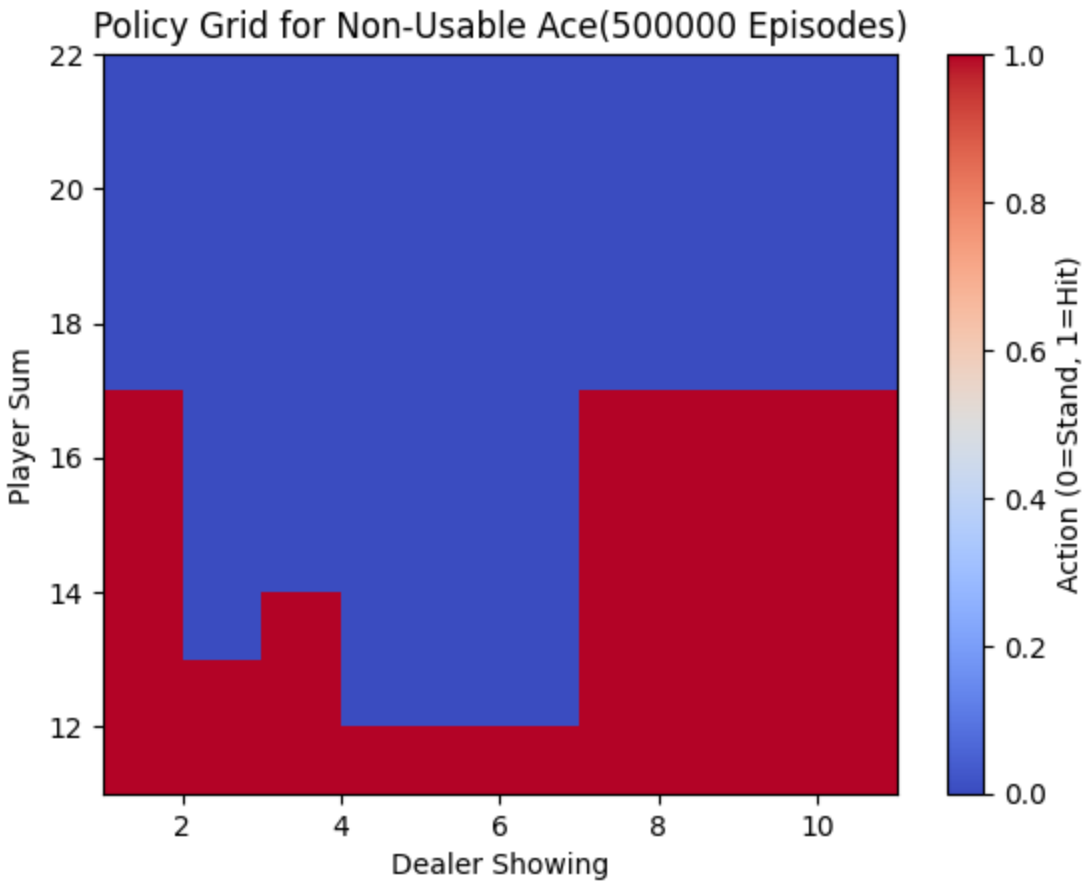
a.

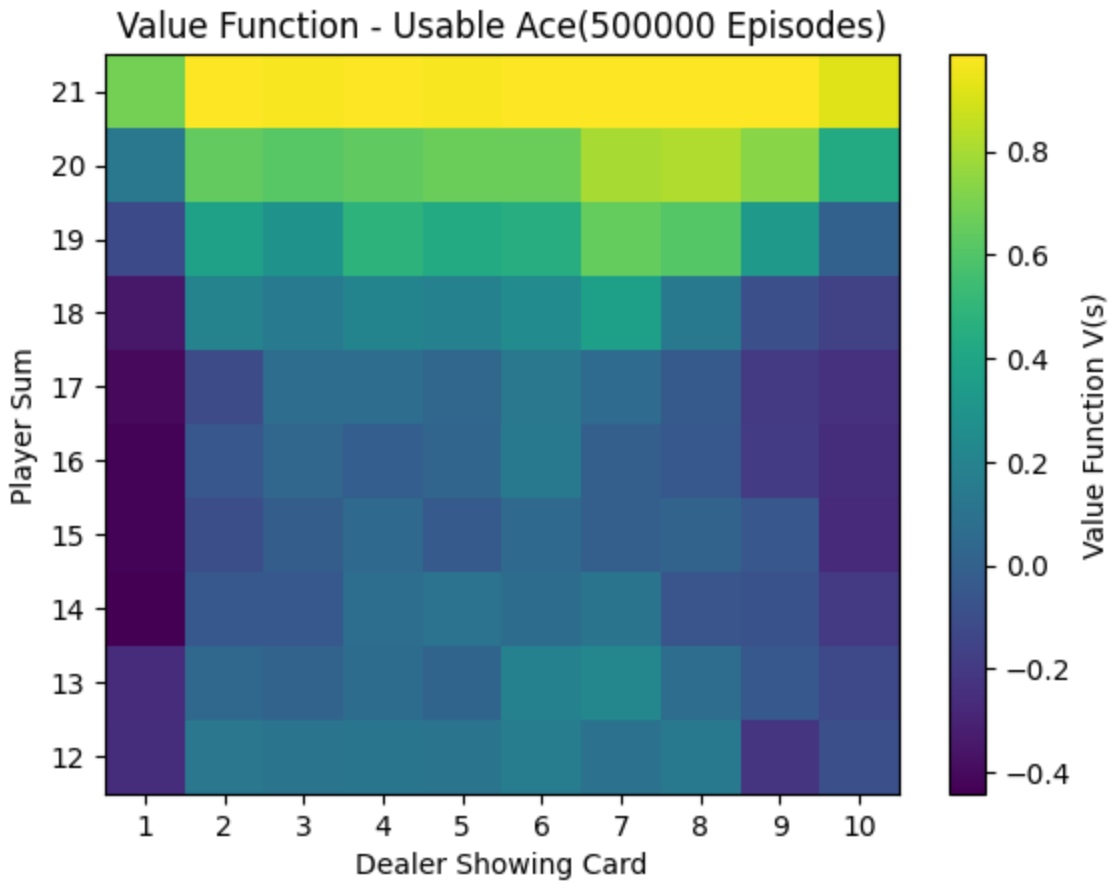

Value Function - Non-Usable Ace(500000 Episodes)
Player Sum
21
20
19
18
17
16
15
14
13
12
1
2
3
4
5
6
7
8
9
10
Dealer Showing Card
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Value Function V(s)

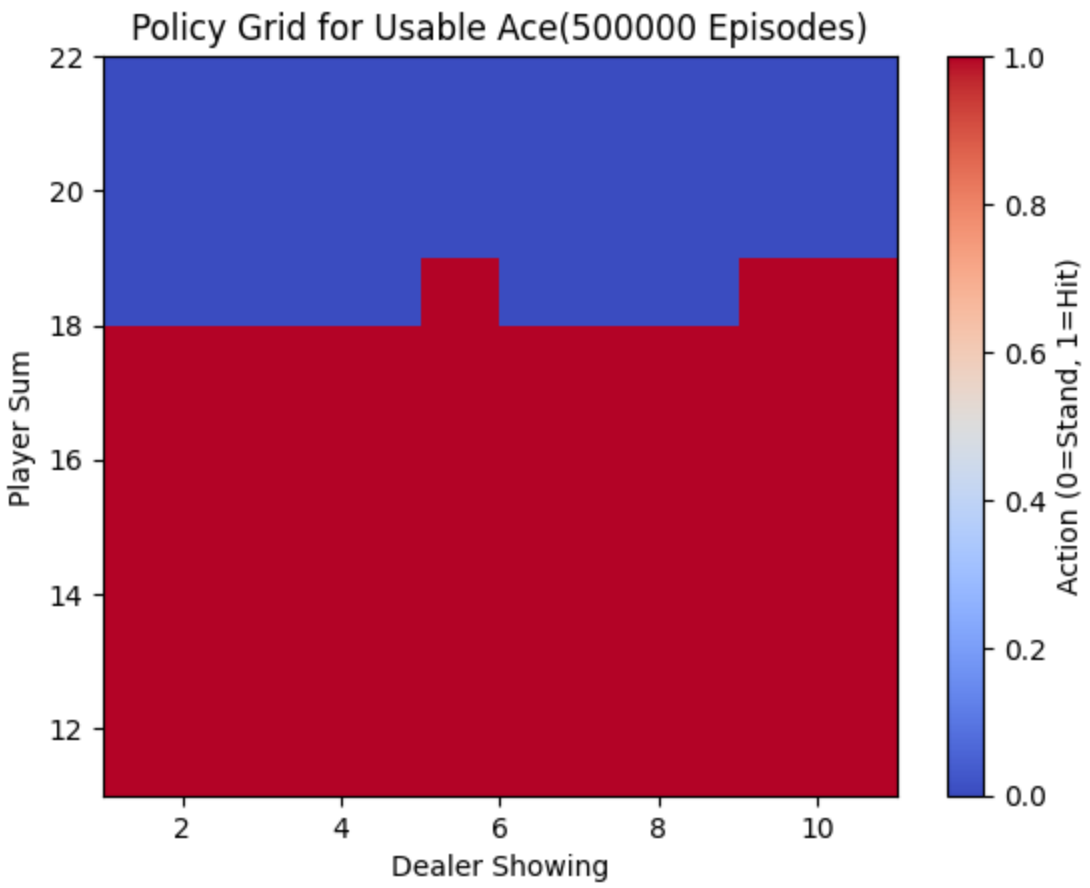

Value Function - Usable Ace(500000 Episodes)
Player Sum
21
20
19
18
17
16
15
14
13
12
1
2
3
4
5
6
7
8
9
10
Dealer Showing Card
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Value Function V(s)

b.


Value Function - Non-Usable Ace(500000 Episodes)
Player Sum
21
20
19
18
17
16
15
14
13
12
1
2
3
4
5
6
7
8
9
10
Dealer Showing Card
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Value Function V(s)

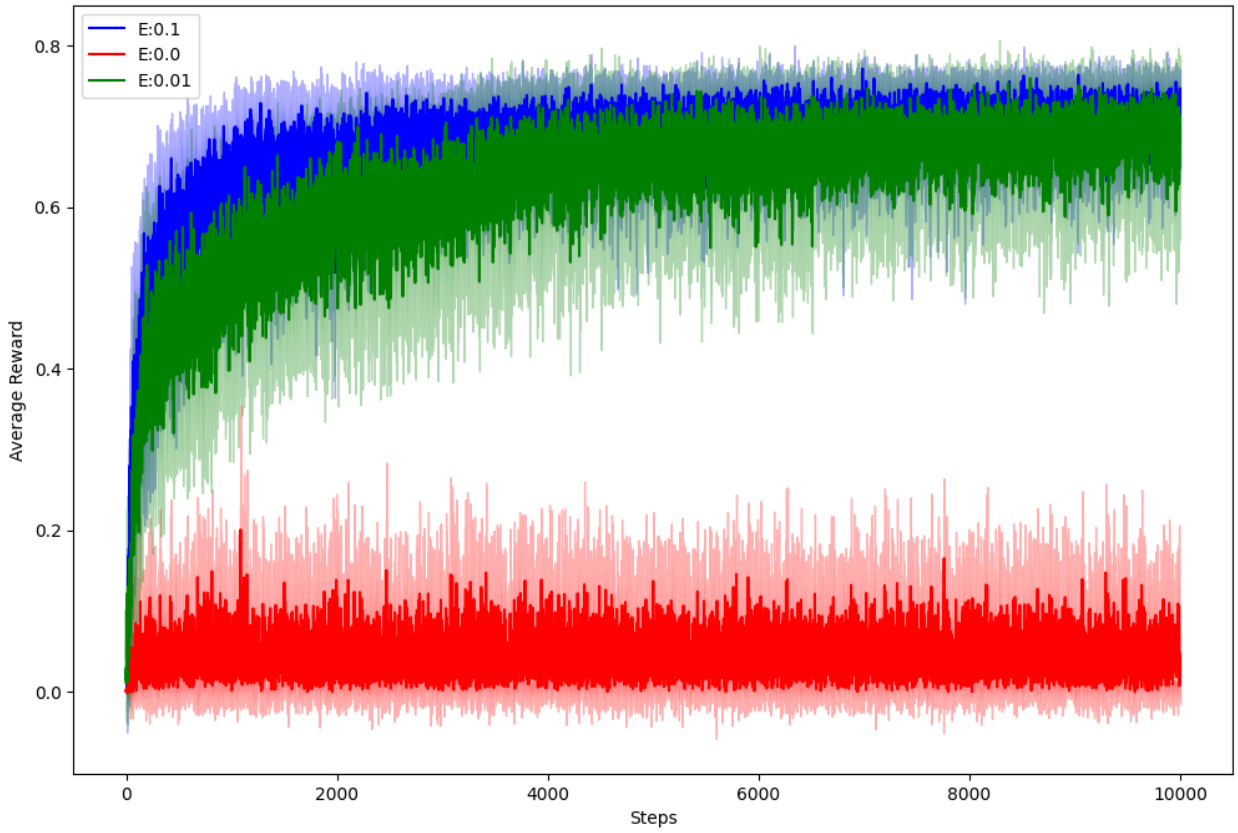

Policy Grid for Non-Usable Ace(500000 Episodes)
Player Sum
Dealer Showing
Action (0=Stand, 1=Hit)
22
20
18
16
14
12
2
4
6
8
10
1.0
0.8
0.6
0.4
0.2
0.0


Value Function - Usable Ace(500000 Episodes)
Player Sum
21
20
19
18
17
16
15
14
13
12
1
2
3
4
5
6
7
8
9
10
Dealer Showing Card
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Value Function V(s)


Policy Grid for Usable Ace(500000 Episodes)
Player Sum
Dealer Showing
Action (0=Stand, 1=Hit)
22
20
18
16
14
12
2
4
6
8
10
1.0
0.8
0.6
0.4
0.2
0.0

Question 3:

a.



a. The results for $\epsilon = 0$, highlight the importance of doing exploring starts. It highlights this as exploring starts ensure optimality by guaranteeing that all state-action pairs are explored, whcih lets the agent to accurately assess the value of each action in every state. This exploration is necessary because, without trying all possible actions, the agent cannot know if its current policy is truly optimal. Similarly, an ϵ-greedy policy ensures exploration by occasionally selecting random actions, helping avoid getting stuck in suboptimal behavior. If neither exploring starts nor ϵ-greedy exploration are used, the agent will only exploit its current knowledge and may never discover better actions, meaning it can never guarantee reaching the optimal policy.

Question 4:

b. The policy π is a greedy and deterministic policy, we only observe the trajectories where the probability of action A_t is selected in state S_t is 1(

$\pi(A_t|S_t) = 1)$ so $1/b(A_t, S_t)$ is correct for estimating the importance-sampling ratio in this case.