Q1: (a).

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

~~$Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$~~

Loop forever (for each episode): $N(s) \leftarrow 0$

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

~~Append G to $Returns(S_t)$~~

~~$V(S_t) \leftarrow$ average$(Returns(S_t))$~~

$N(S_t) \leftarrow N(S_t)+1$

$V(S_t) \leftarrow V(S_t) + (G - V(S_t)) / N(S_t)$

(b).

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:

$\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

~~$Returns(s,a) \leftarrow$ empty list,~~ for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode): $N(s,a) \leftarrow 0$.

Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0

Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

~~Append G to $Returns(S_t, A_t)$~~

~~$Q(S_t, A_t) \leftarrow$ average$(Returns(S_t, A_t))$~~

$\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$

$$\begin{cases} N(S_t, A_t) \leftarrow N(S_t, A_t) + 1 \\ Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \dfrac{G - Q(S_t, A_t)}{N(S_t, A_t)} \end{cases}$$

Q2: (a) For Blackjack task, the state in each episode is constantly changing (it appears only once). Even if every-visit MC is used, since the state appears only once, it gets the same result as using first-visit.

(b). $T=10.\quad \gamma=1$

$G_t = R_{t+1} + \gamma G_{t+1}$

$$\left\{\begin{aligned} G_0 &= R_1 + G_1 = 10 \\ G_1 &= R_2 + G_2 = 9 \\ G_2 &= R_3 + G_3 = 8 \\ &\ \ \cdots \\ G_9 &= R_{10} + G_{10} = 1 \\ G_{10} &= 0. \end{aligned}\right\} \Rightarrow$$

For first-visit MC:

$V(s) = 10.$

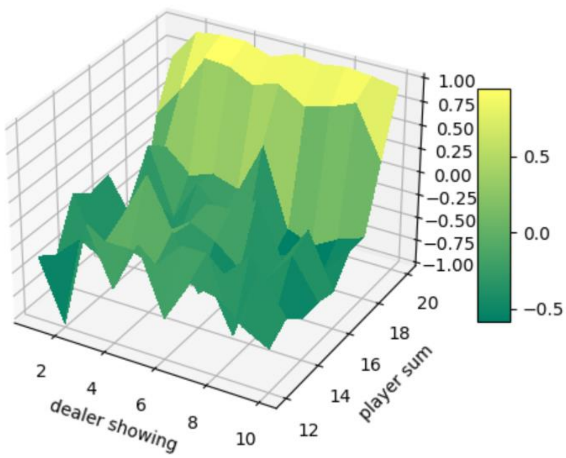
For every-visit MC.

$$V(s) = \frac{G_{10} + G_9 + \cdots + G_0}{10}$$

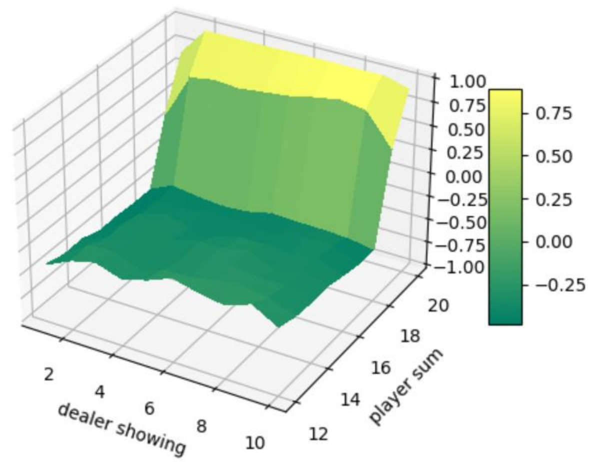
$$= \frac{0 + 1 + \cdots + 10}{10} = 5.5$$

Q3: (a)

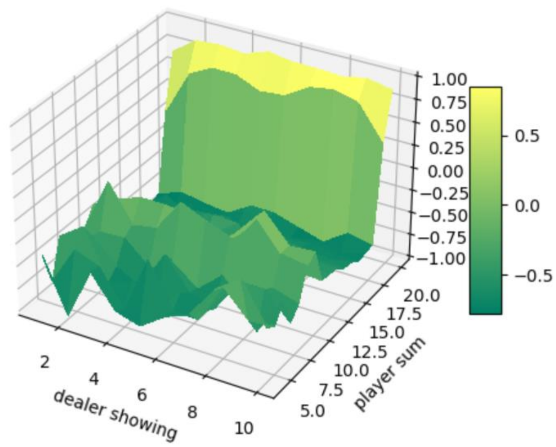
Usable Ace after 10000 episodes



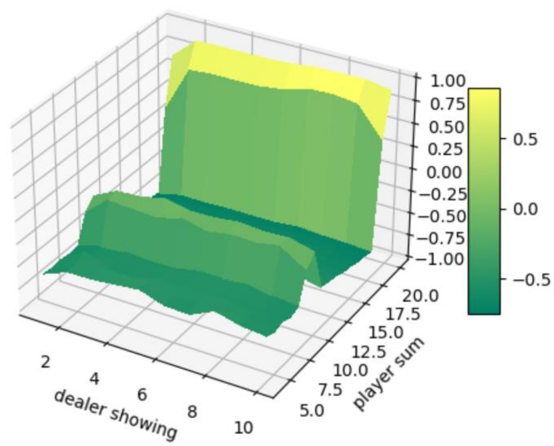
Usable Ace after 500000 episodes



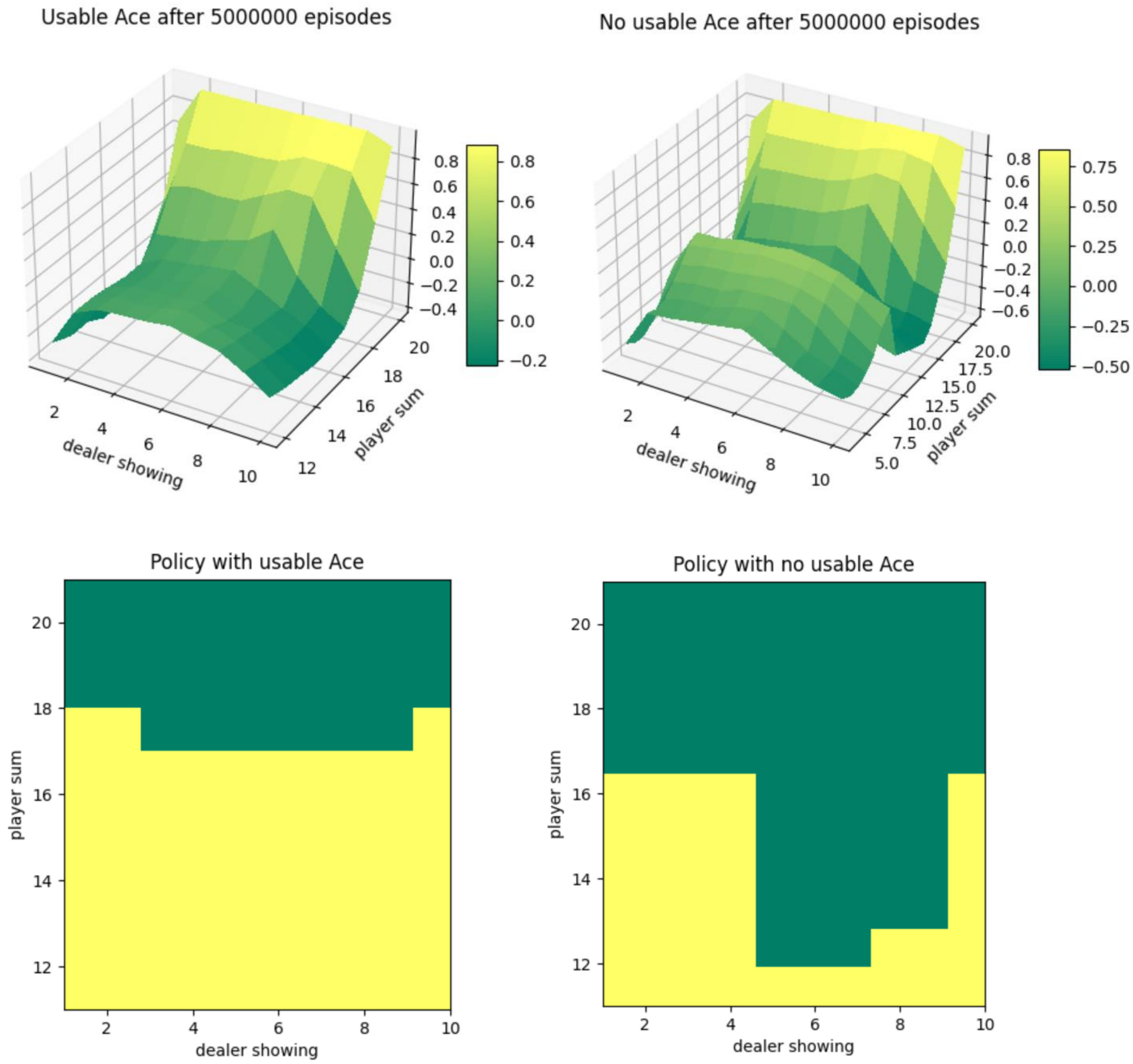
No usable Ace after 10000 episodes



No usable Ace after 500000 episodes

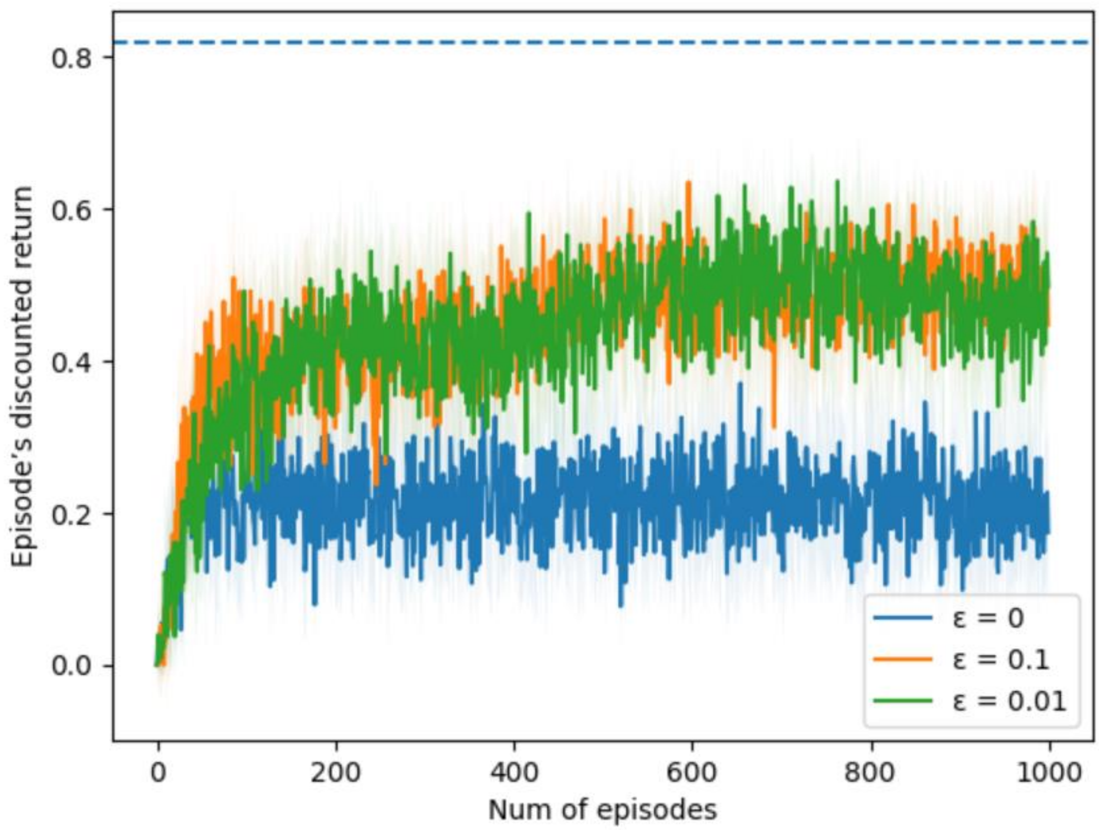


(b)



Usable Ace after 5000000 episodes
No usable Ace after 5000000 episodes
dealer showing
player sum
Policy with usable Ace
Policy with no usable Ace
dealer showing
player sum

Q4:(b)



(c) When $\varepsilon = 0$, without exploring start, the policy will always choose the state with the first positive result and refuse to make a new attempt. However, with exploring start, this is prevented. It will have chances to try other states. Thus, it will still be useful for improving the policy.

Q5: (a) $$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \quad n \geq 2$$

$$V_{n+1} = \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k}, \quad n \geq 1$$

$$= \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n} W_k} + \frac{W_n G_n}{\sum_{k=1}^{n} W_k}$$

$$= \frac{\sum_{k=1}^{n-1} W_k}{\sum_{k=1}^{n-1} W_k} \times \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n} W_k} + \frac{W_n G_n}{\sum_{k=1}^{n} W_k}$$

$$= V_n - \frac{V_n \times W_n}{\sum_{k=1}^{n} W_k} + \frac{W_n G_n}{\sum_{k=1}^{n} W_k}$$

$$= V_n + (G_n - V_n) \times \frac{W_n}{\underbrace{\sum_{k=1}^{n} W_k}_{= C_n}}$$

$$= V_n + \frac{W_n}{C_n} \times (G_n - V_n).$$

(b). When the policy π is greedy, the probability of action (A_t) under state (S_t) is 1. So we should expect W involves $\frac{1}{b(A_t|S_t)}$.