1-a)

Initialize:

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

$C(s) = 0 \quad s \in \mathcal{S}$

Loop forever (each episode)

Generate an episode π: $S_0 \ldots R_T$

$G \leftarrow 0$

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

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

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

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

$V(S_t) \leftarrow V(S_t) + \frac{1}{C(S_t)}(G - V(S_t))$

b) Replace Initialization of return map with

$C(s,a) \leftarrow 0 \quad \forall s \in \mathcal{S}, a \in A(s)$

Replace "Append G in Return" with

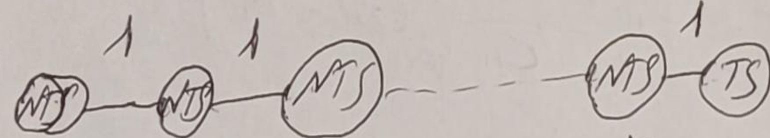
$C(S_t, a_t) \leftarrow C(S_t, a_t) + 1$

Replace "$Q \leftarrow$ average( )" with

$Q(S_t, a_t) \leftarrow Q(S_t, a_t) + \frac{1}{C(S_t, a_t)}(G - Q(S_t, a_t))$

2-a) No, by the nature of blackjack we will essentially see a state once in each episode so first-visit and every-visit make the same result.

b)



$\Rightarrow$ first-visit: $V(s=NTS) = \sum_{i=1}^{10} 1 = 10$

every-visit: $V(s=NTS) = \frac{1}{10}(1+2+3+ \cdots 10) = \frac{1}{10}\sum_{i=1}^{10} i = 5.5$

c) Yes, according to the book which says in

$\bar{V}[X] = E[X^2] - \bar{X}^2$ if $E[X^2] \to \infty \Rightarrow \bar{V}[X] \to \infty$

$\bar{X}^2$: finit

for first-visit: $E_b\left[\left(\prod_{t=0}^{T-1} \frac{\pi}{b} G_0\right)^2\right] \to \infty$ we know $\bar{V}(X) \to \infty$ for every-visit

for every-visit we will have:

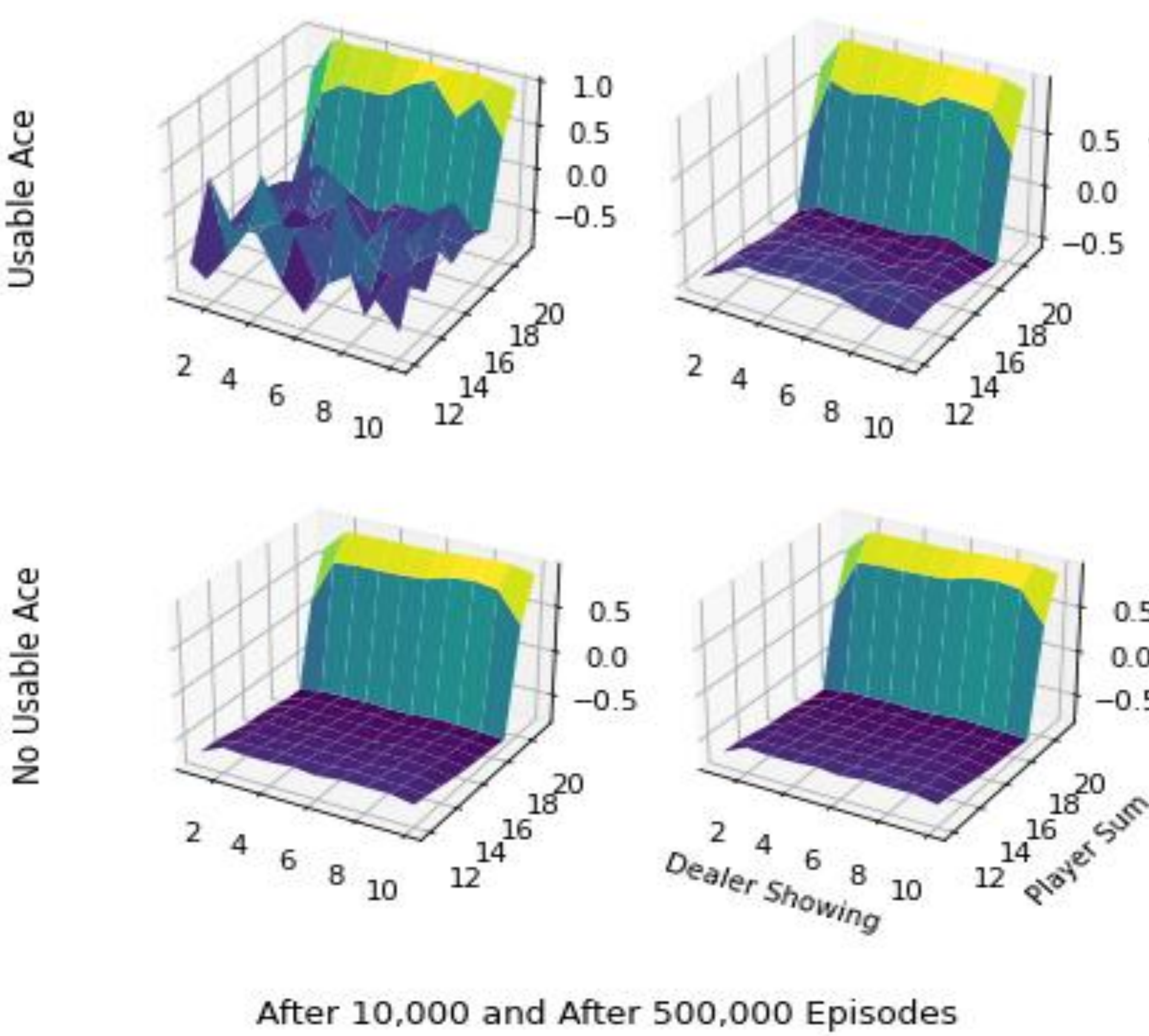
$E_b\left[\left(\frac{1}{T-1}\sum_{K=1}^{T-1} \prod^{K} \frac{\pi}{b} G_0\right)^2\right] = \frac{1}{T-1} E_b\left[\left(\sum_{K=1}^{T-1} \prod_{t=0}^{K} \frac{\pi}{b} G_0\right)^2\right] \to \infty$

$1/T-1$ constant, because $E_b\left[\left(\sum_{K=1}^{T-1} \prod_{t=0}^{K} \frac{\pi}{b} G_0\right)^2\right] \to \infty$

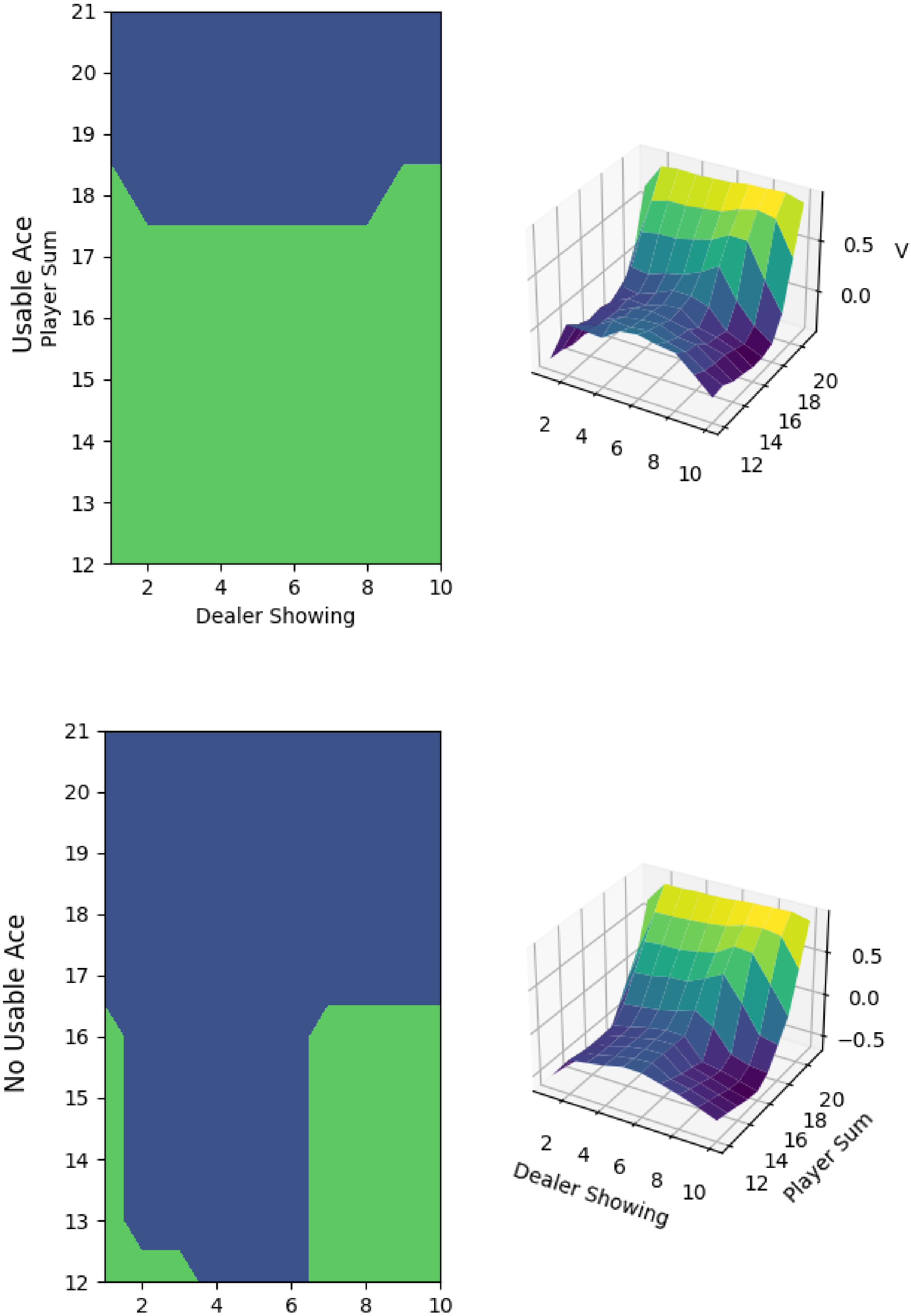
and $\bar{X}^2$ for every-visit is finit too

For running Q3 and Q4, in algorithm.py, there is a main function in which you need to uncomment each question and part you like to run

Q3- a



Q3- b

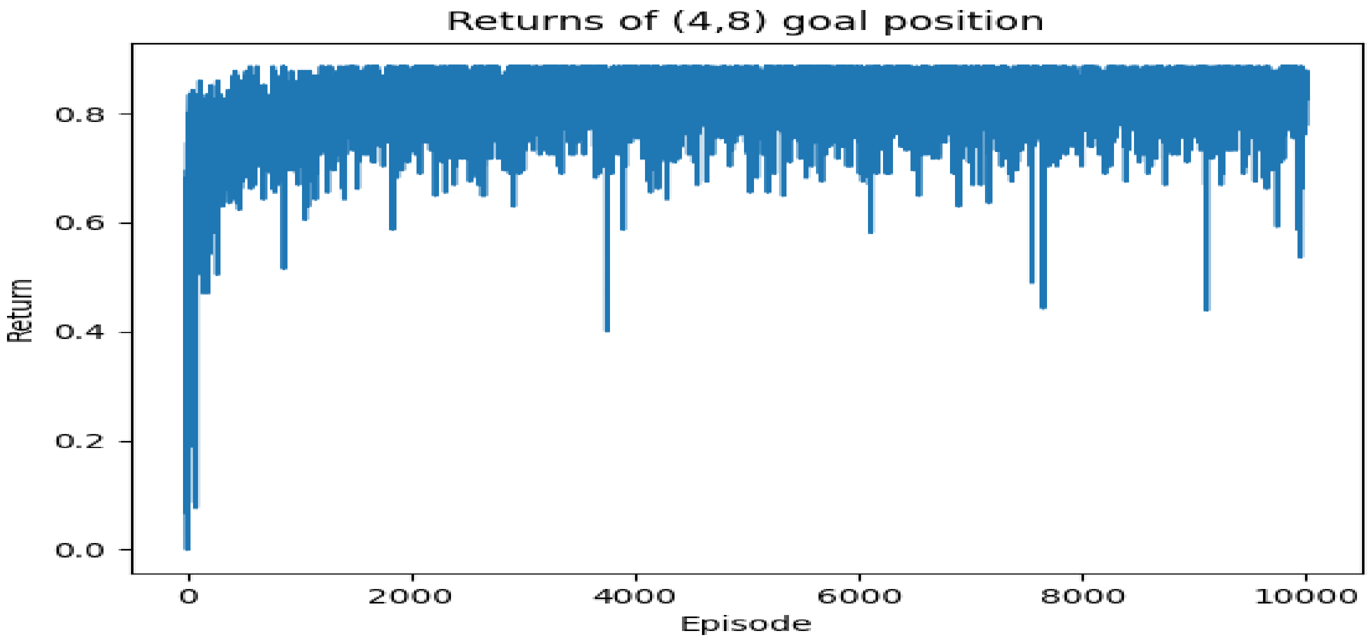


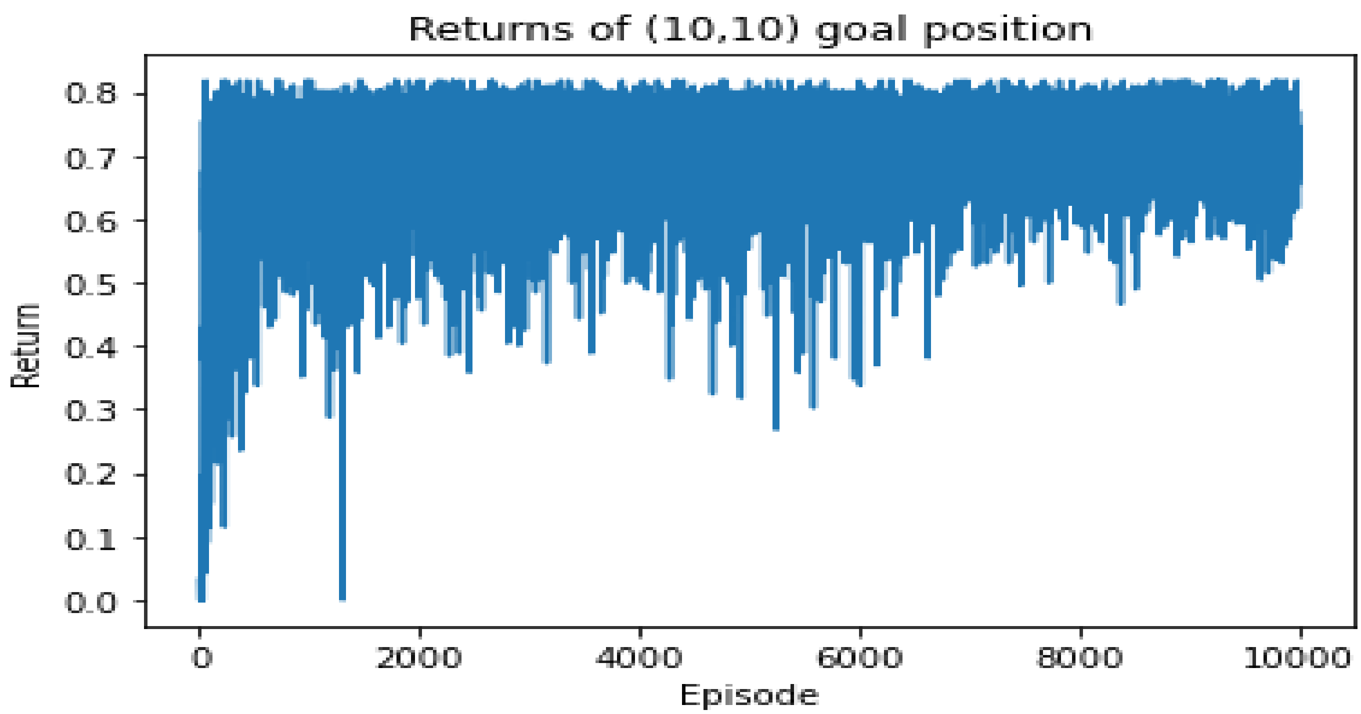
Optimal Policy and Value Function

Green represents hit and blue represents stay.

Q4- a -

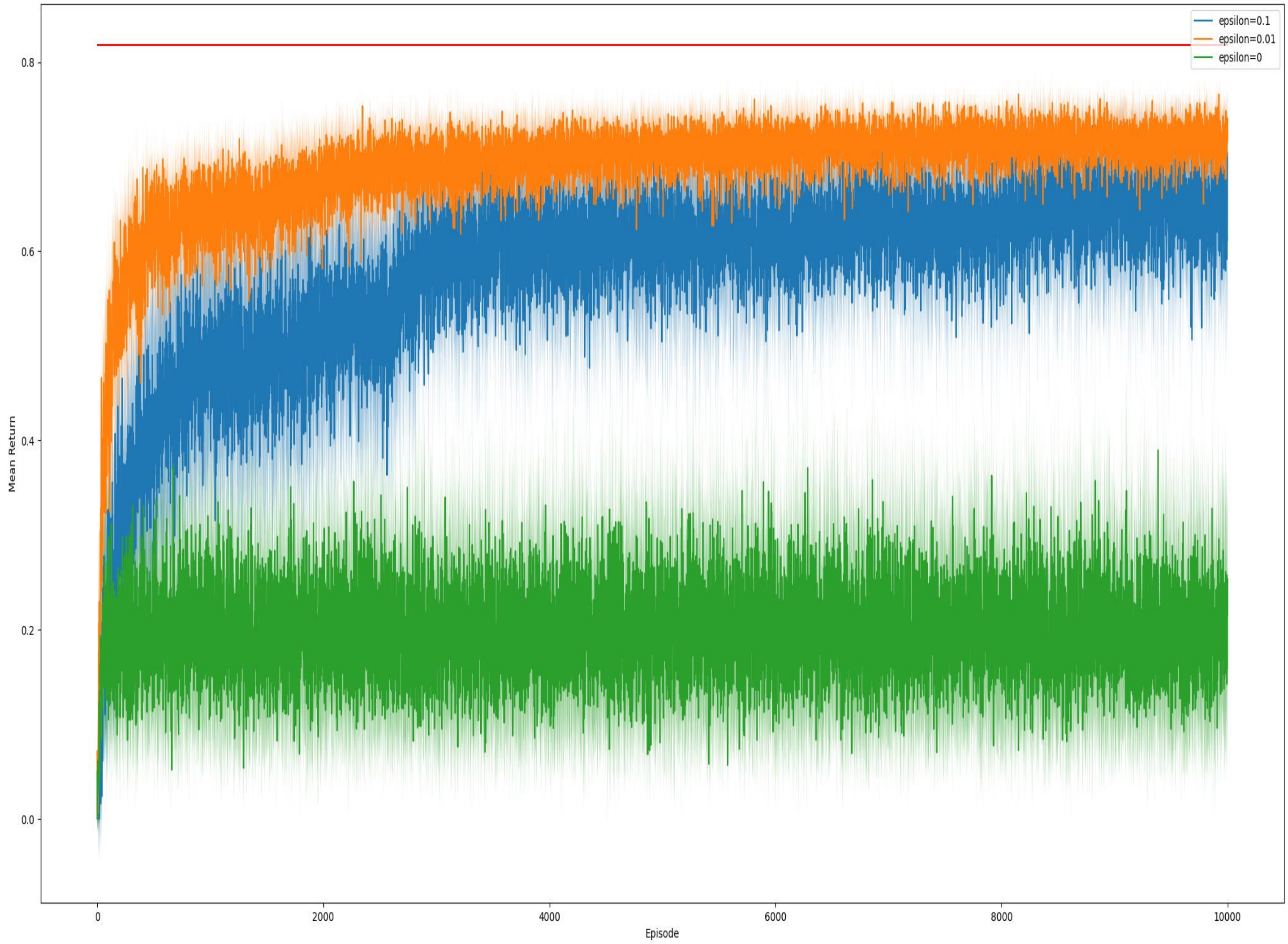
The following plot shows returns when the goal is at (4,8), showing that it works for non-original goal states



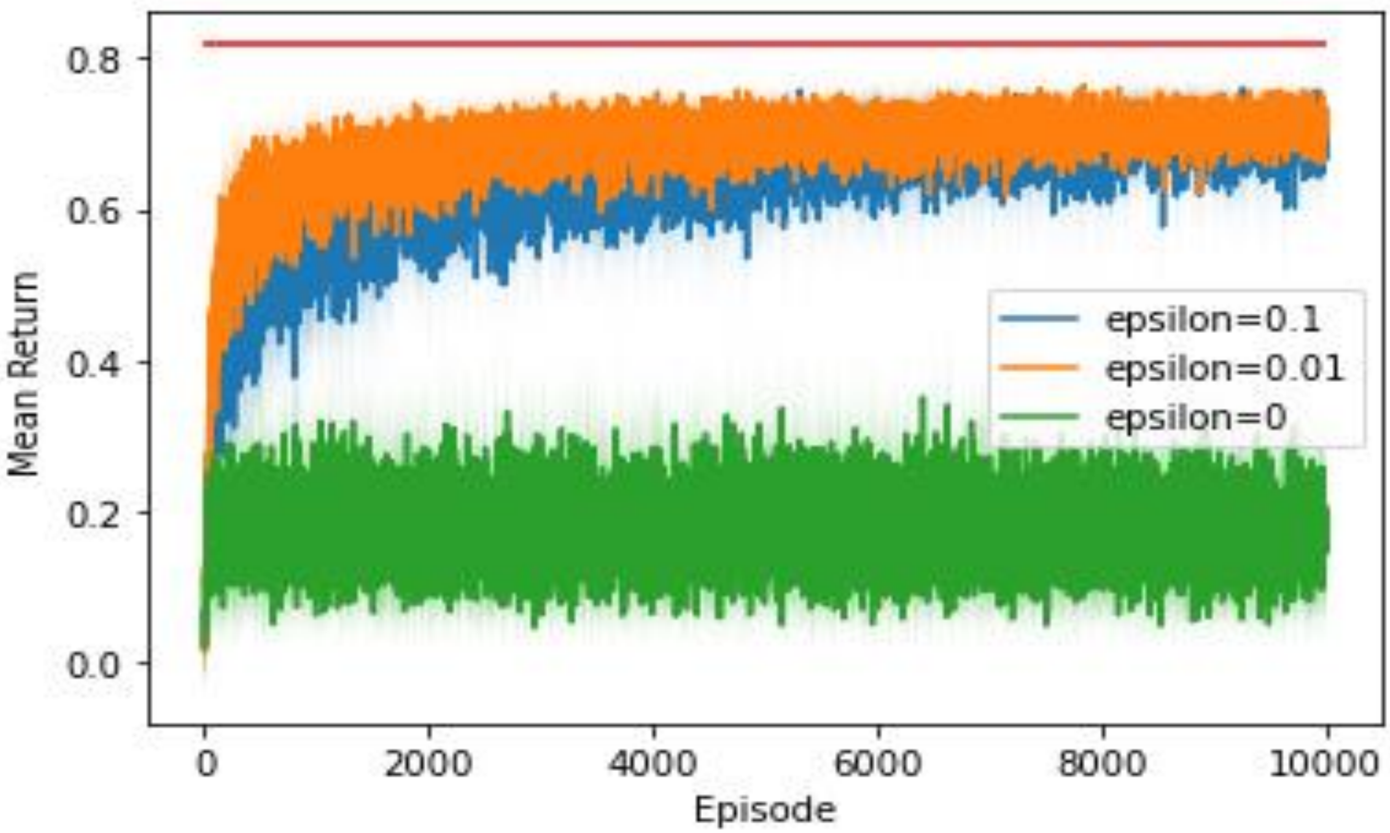


b-

Learning curve of (4,8) goal position



Learning curve of (10,10) goal position



C) Without exploring starts, the agent will follow the path it first finds to the goal as the only exploration. Because the initial starter is random, there is a large chance that this path will be very inefficient, far from optimal answer, which leads to the mean return of only 0.2

5-a) $V_{n+1} = \frac{\sum^{n} W_k G_k}{\sum^{n} W_k} = \frac{W_n G_n + \sum^{n-1} W_k G_k}{\sum^{n} W_k} \frac{\sum^{n-1} W_k}{\sum^{n-1} W_k}$

$= \left[\frac{W_n G_n}{C_{n-1}} + V_n\right] \frac{C_{n-1}}{C_n} = \frac{W_n G_n}{C_n} + \frac{V_n C_{n-1}}{C_n}$ $\quad C_i = \sum_{k=1}^{i} W_k$

$= V_n + \frac{W_n G_n}{C_n} + \frac{V_n C_{n-1}}{C_n} - V_n = V_n + \frac{W_n G_n}{C_n} + \frac{V C_{n-1} - V_n C_n}{C_n}$

$= V_n + \frac{W_n G_n}{C_n} + \frac{-V_n W_n}{C_n} = V_n + \frac{W_n}{C_n}\left[G_n - V_n\right]$

b) because we are assuming that the target policy is greedy and determinstic so we could consider $\pi(A_t|S_t) = 1$

$\Rightarrow \frac{\pi}{b} = \frac{1}{b}$

Q6) I am still working on question 6, when ever it finish I will send it.