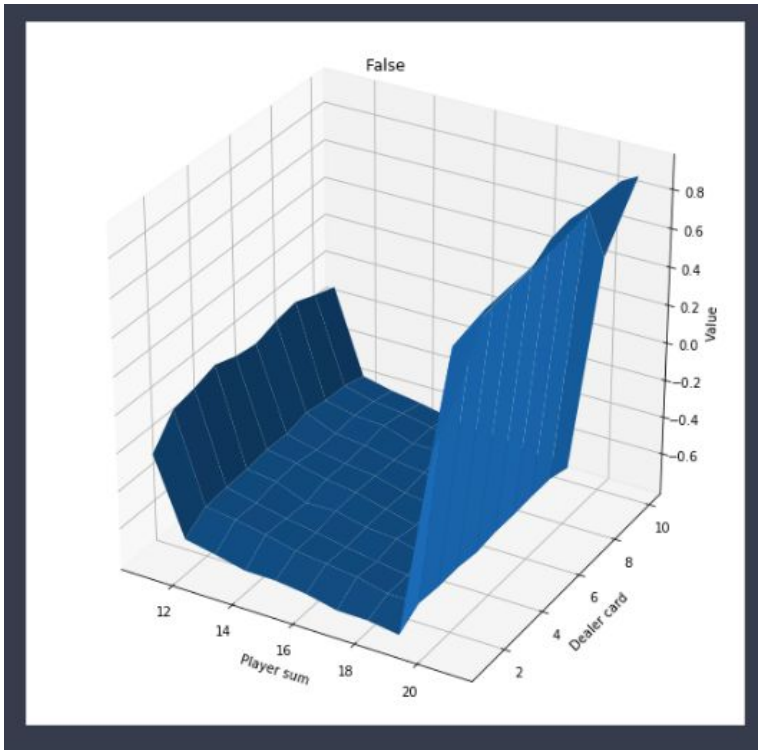
REPORT

PHD17008
Aradhya Neeraj Mathur

Q1.
NOTE: the pseudocode has been used to implement the question 4 5.1 , 5.2.
The pseudocode for question 4 is as follows
G += R(t+1)
Q(s, a) = (Q(s, a) * count(S,a) + G) / (count(S,a) + 1)
While the pseudocode in book maintains a list of the returns and then calculates the average my method keeps track of the count of each state action pair and the multiplies the older Q value with the old count and adds the current updated return and then divides by the new count which is essentially equivalent to the mean.

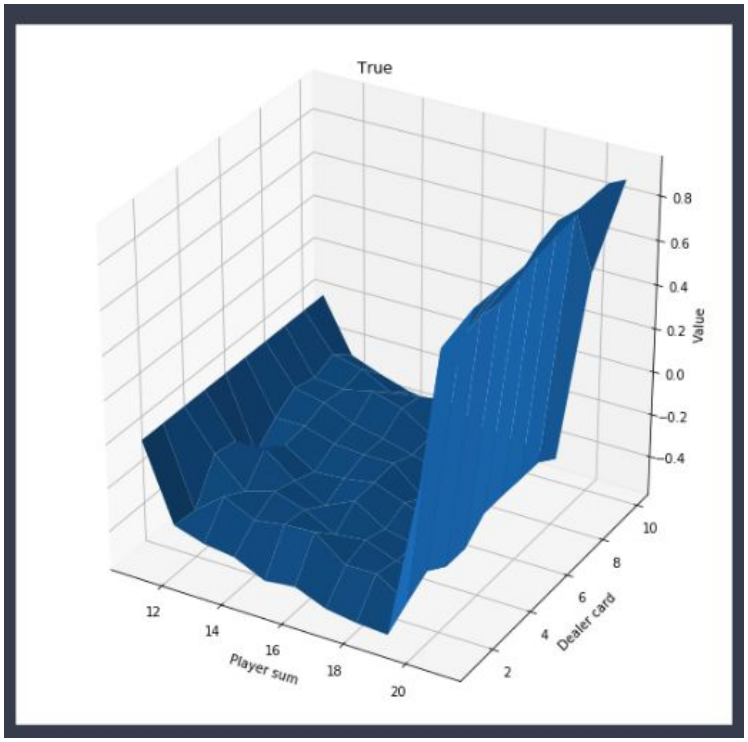
Q4 Value diagrams
Fig 5.1
For 500000 episodes

Non usable Ace



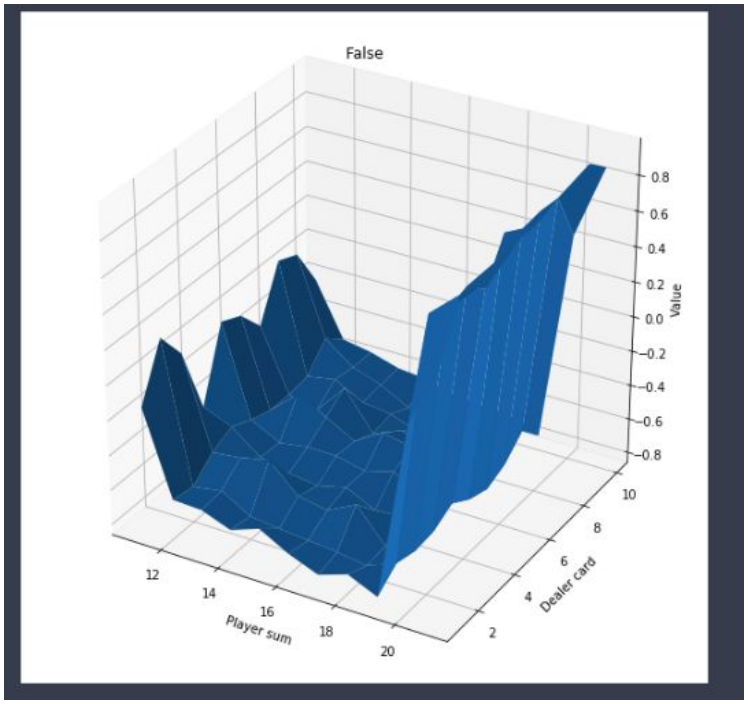
The code has been explained in the notebook

Usable Ace



For 10000 episodes

Non usable Ace



Usable Ace

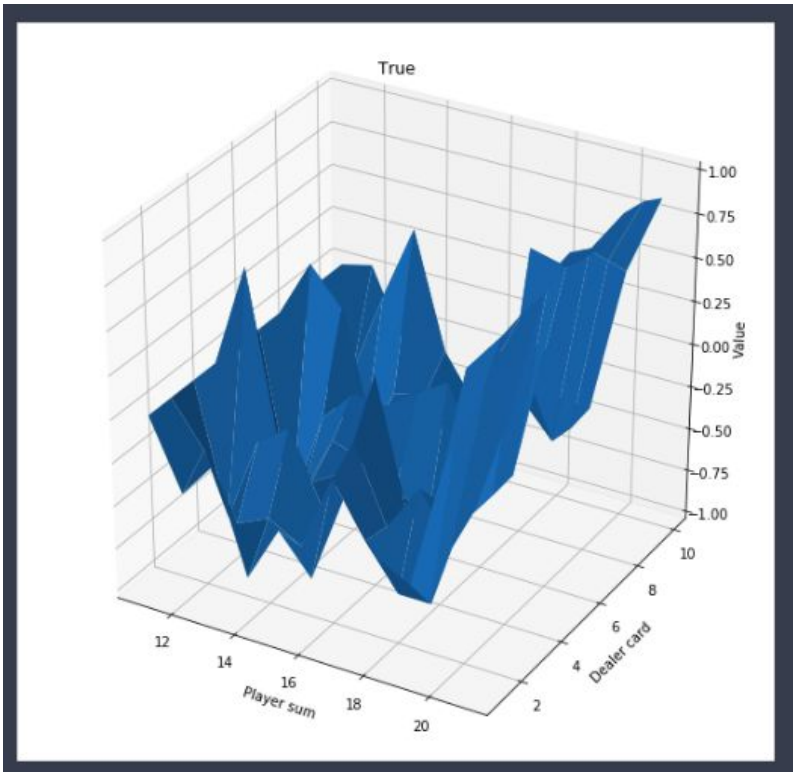
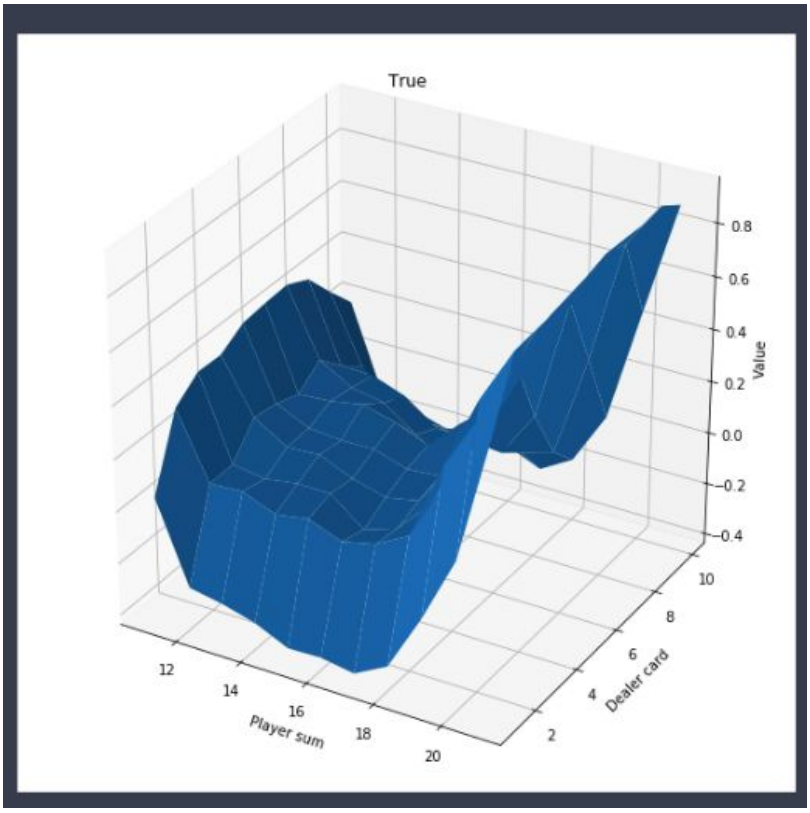
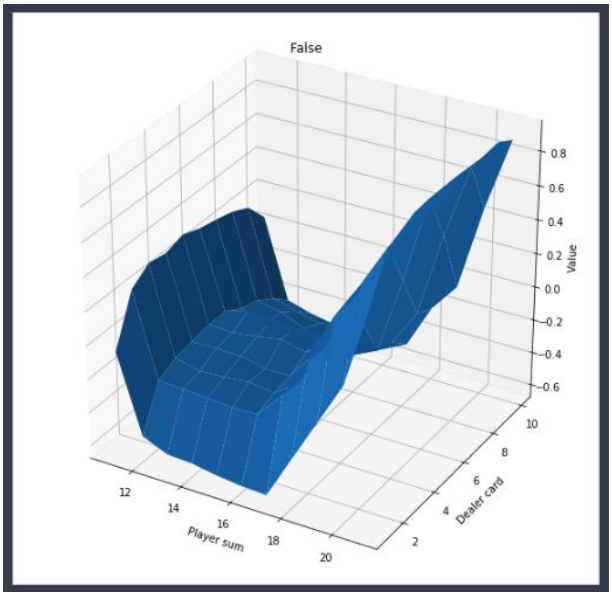


Fig 5.2

Usable Ace

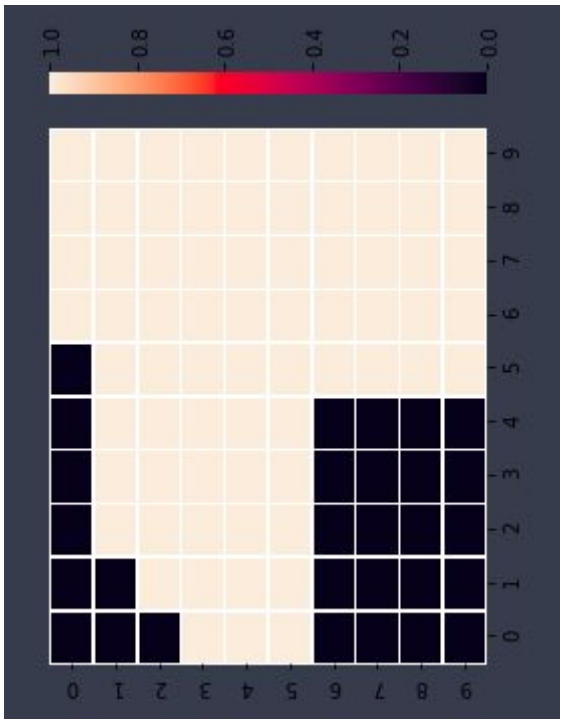


Non Usable Ace

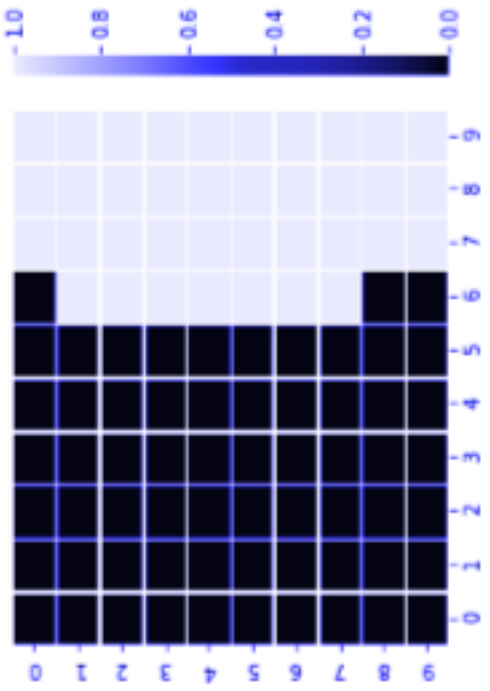


Policy diagram

Non Usable Ace



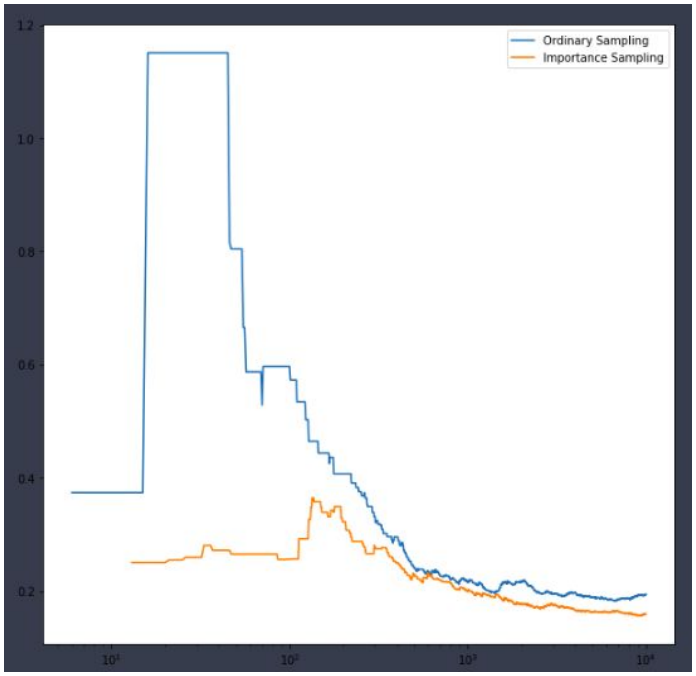
Non Usable Ace Policy Hit:0, Stick:1



Usable Ace Policy Hit:0, Stick:1

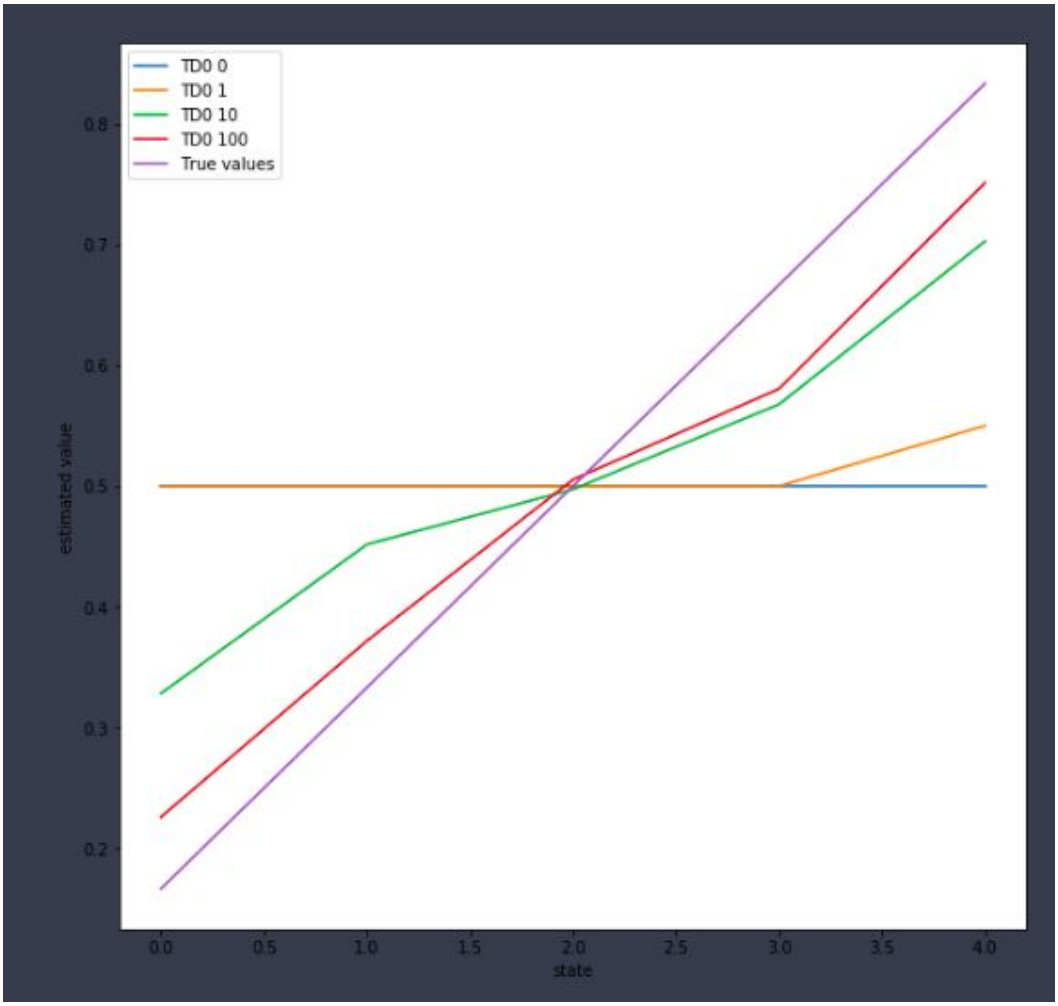
Fig 5.3

X axis : number of episodes y axis: MSE error in value

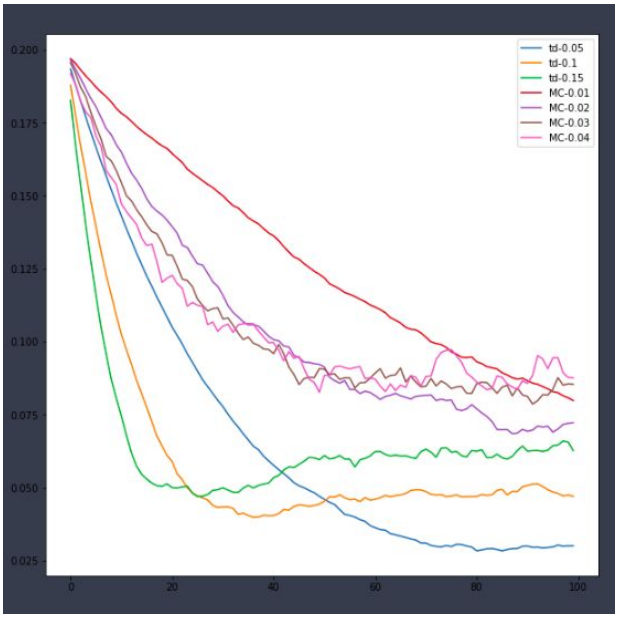


It can be observed from the graphs that the error of ordinary sampling is more than importance sampling.

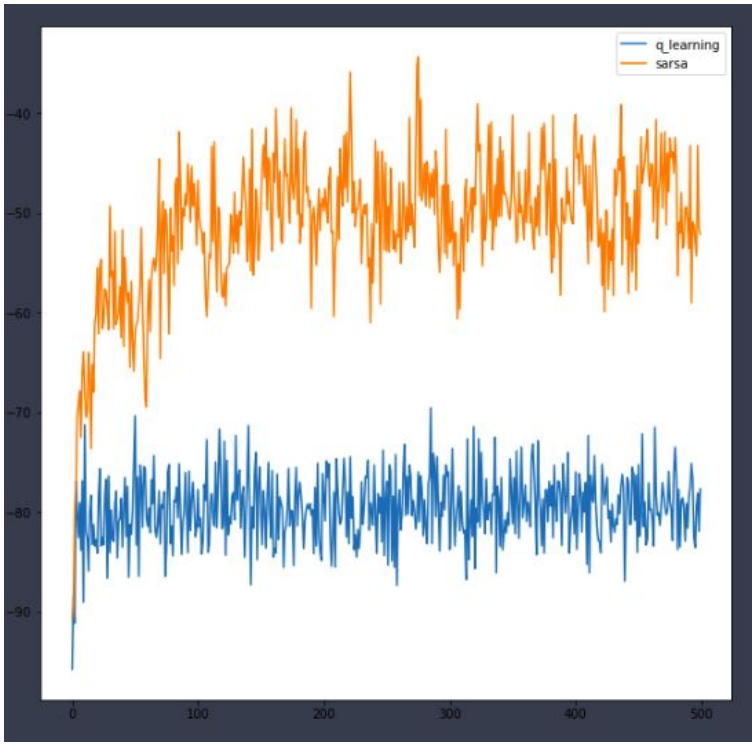
Q6.



Empirical RMS error,
averaged over states for TD0 and MC for different alpha values



Q7.



Q Learning Vs sarsa for sum of rewards y axis:sum of rewards x axis:episodes

Q2:

Ex. 5.6

equation 5.6 calculates weighted importance sampling using the following eqn.

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

Equation analogous for $Q(s,a)$.

If we collect and cover the state action pairs in $\mathcal{T}(s,a)$ we can express $Q(s,a)$ as

$$Q(s,a) = \frac{\sum_{t \in \mathcal{T}(s,a)} \rho_{t+1:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s,a)} \rho_{t+1:T(t)-1}}$$

$$\rho_{t+1:T(t)-1} = \prod_{i=t+1}^{T-1} \frac{\pi(A_i|S_i)}{b(A_i|S_i)}$$

$T(t)$ is the first time of termination after time t and G_t represents returns from time $t+1$ till $t(t)$