1 Question 4

1.1 Figure 5.1

Below are the plots for the blackjack game as replicated from Figure 5.1 from book.

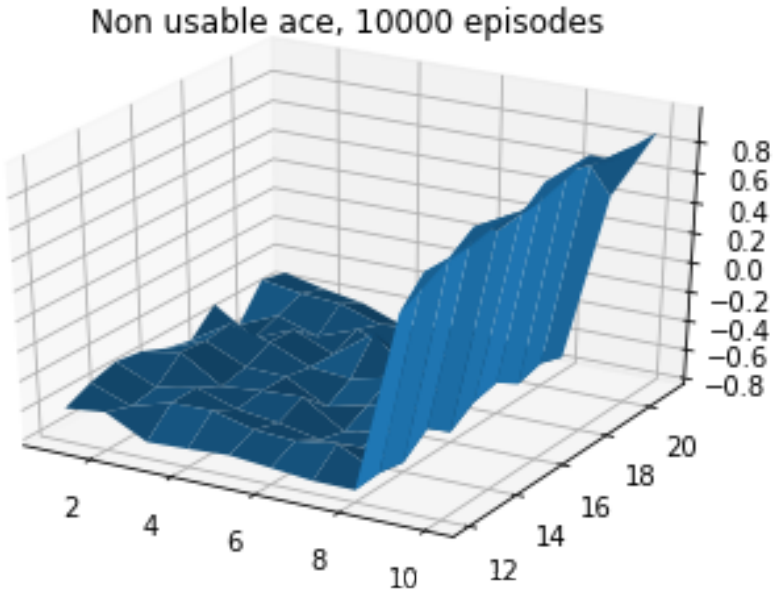

Figure 1: Value Function for non-usable ace, 10000 episodes

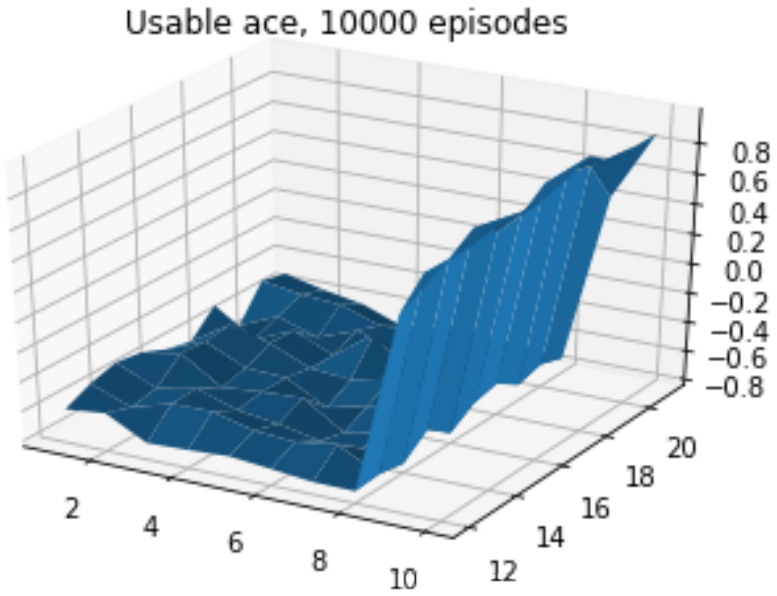

Figure 2: Value Function for usable ace, 10000 episodes

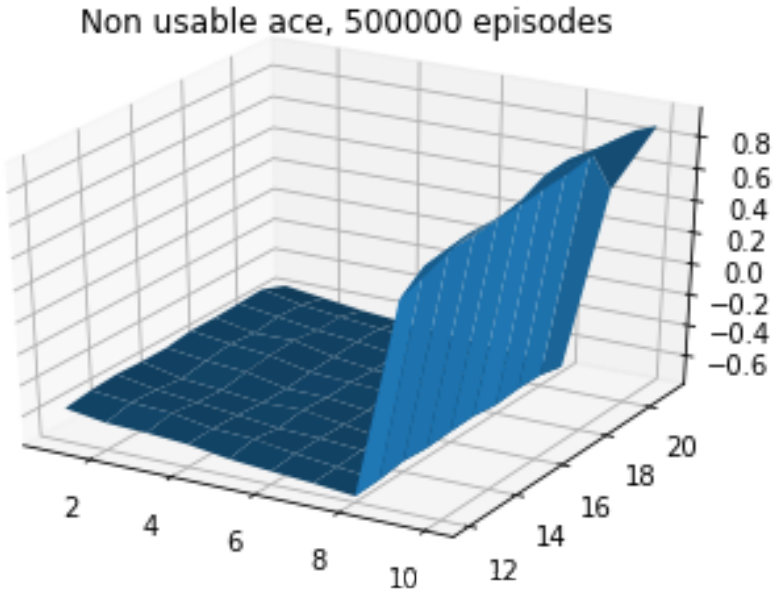

Figure 3: Value Function for non-usable ace, 500000 episodes

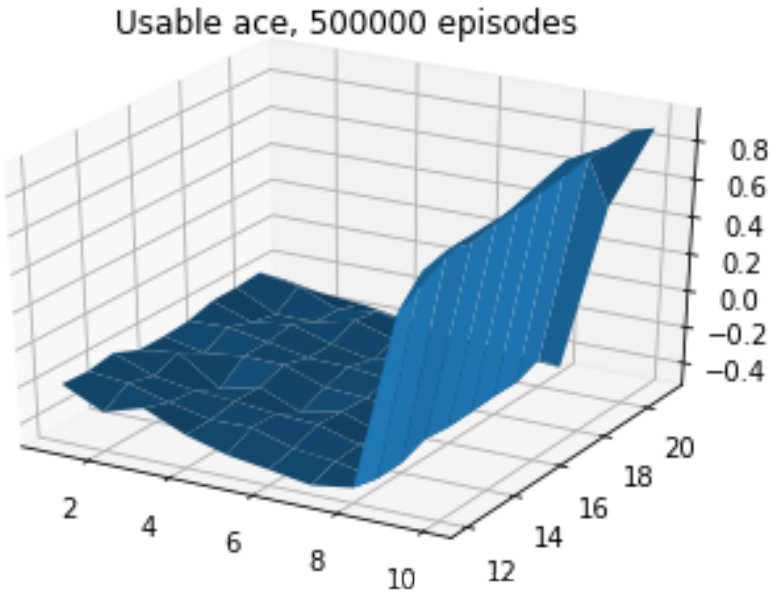

Figure 4: Value Function for usable ace, 500000 episodes

1.2 Figure 5.2

For exploring starts, for each of the episodes, the first state and action pair are chosen randomly and then an episode is generated following the policy pi accordingly.

Below are the plots as replicated from figure 5.2

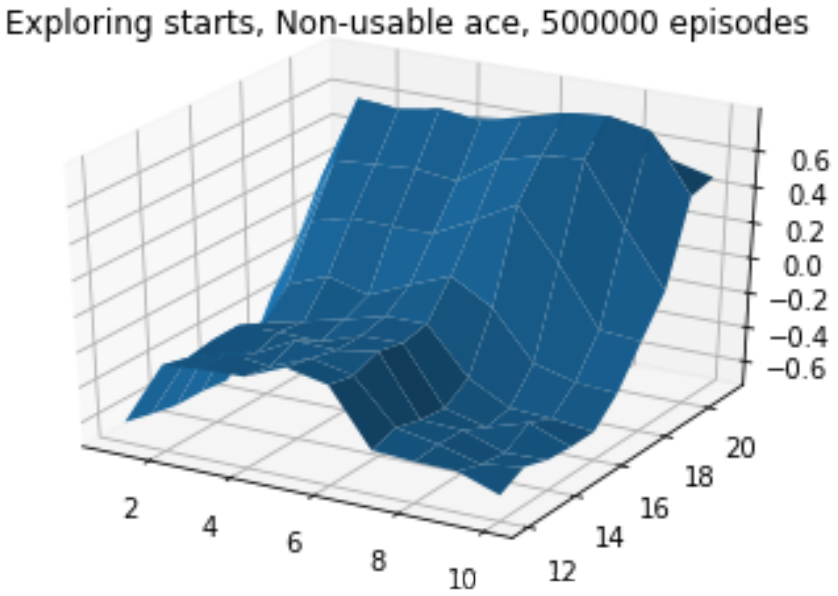

Figure 5: Value Function for non-usable ace, 500000 episodes

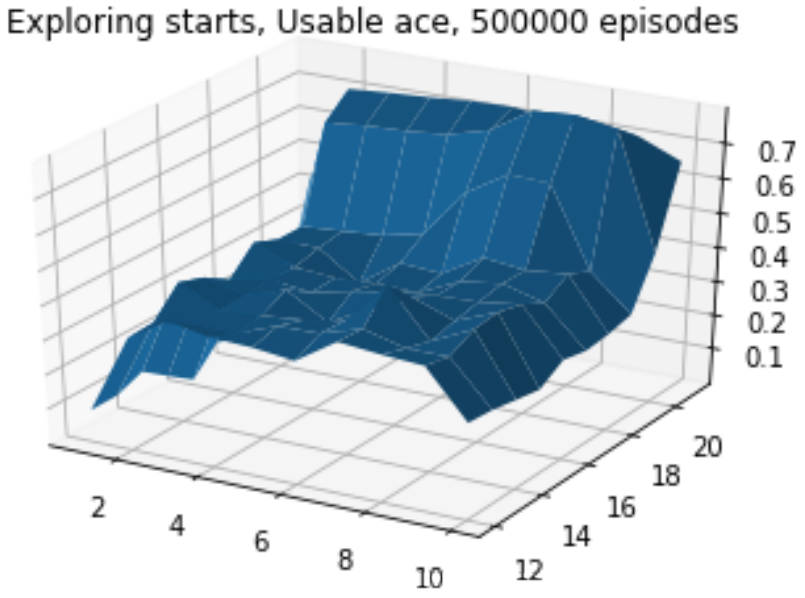

Figure 6: Value Function for usable ace, 500000 episodes

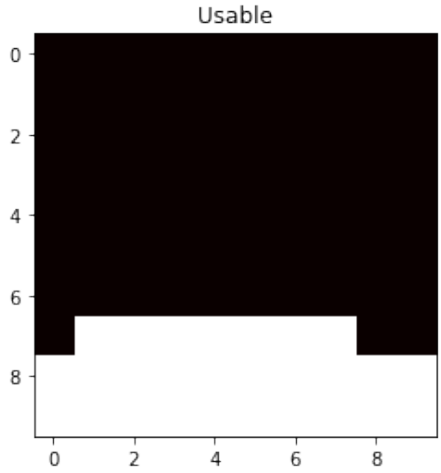

Figure 7: Policy Function for usable ace, 500000 episodes

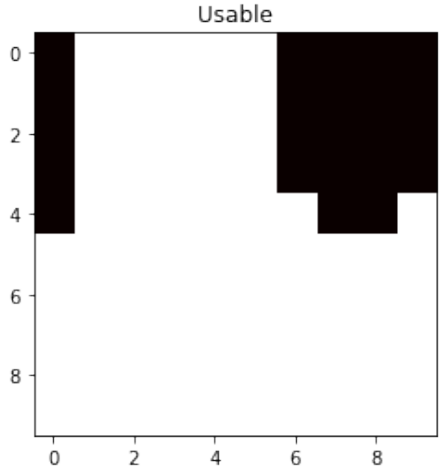

Figure 8: Policy Function for non-usable ace, 500000 episodes

1.3 Figure 5.3

For importance sampling, although both ordinary and weighted sampling lead to the same convergence point, yet the ordinary one starts with a much higher value as compared to the weighted importance sampling case and therefore takes longer time to converge.

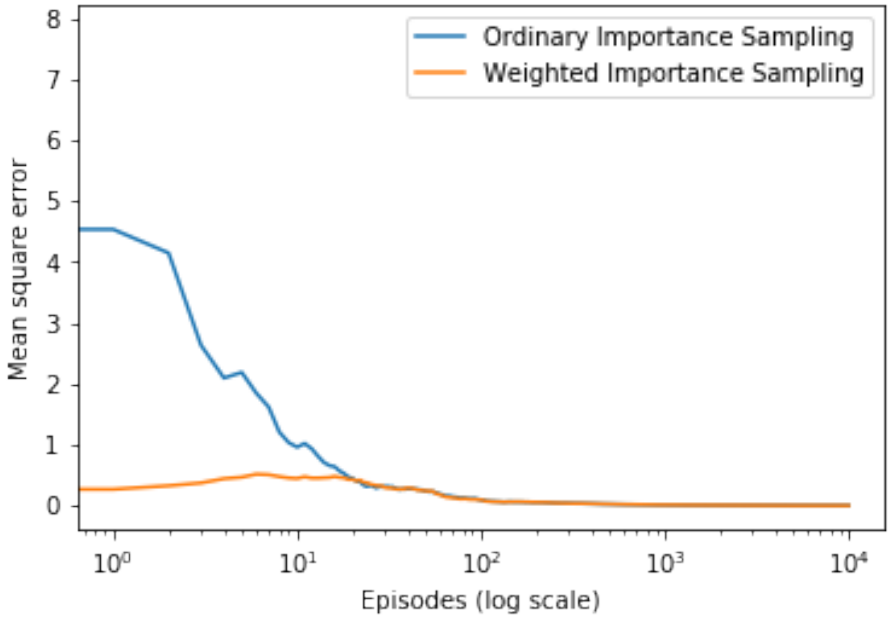

Figure 9: Comparison between Ordinary Importance Sampling and Weighted Importance Sampling

2 Question 6

State vs Estimated value plot for different number of episodes. As we can see that for a large number of episodes say 100, the estimated values comes very close to the true values.

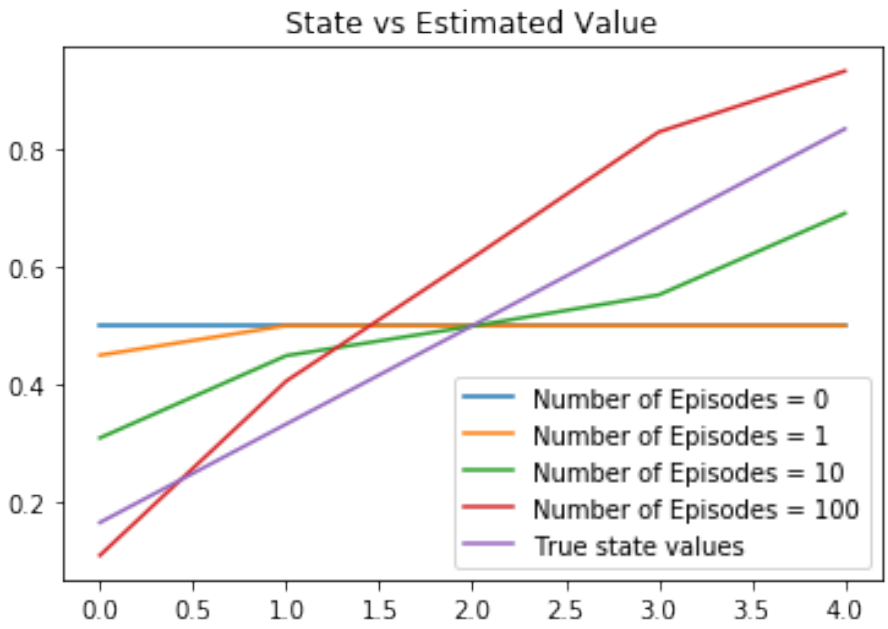

Figure 10: State vs Estimated Value

The RMSE for Monte Carlo are more as compared to the TD algorithm for the given set of values of the alpha. This shows that TD actually performs better than MC.

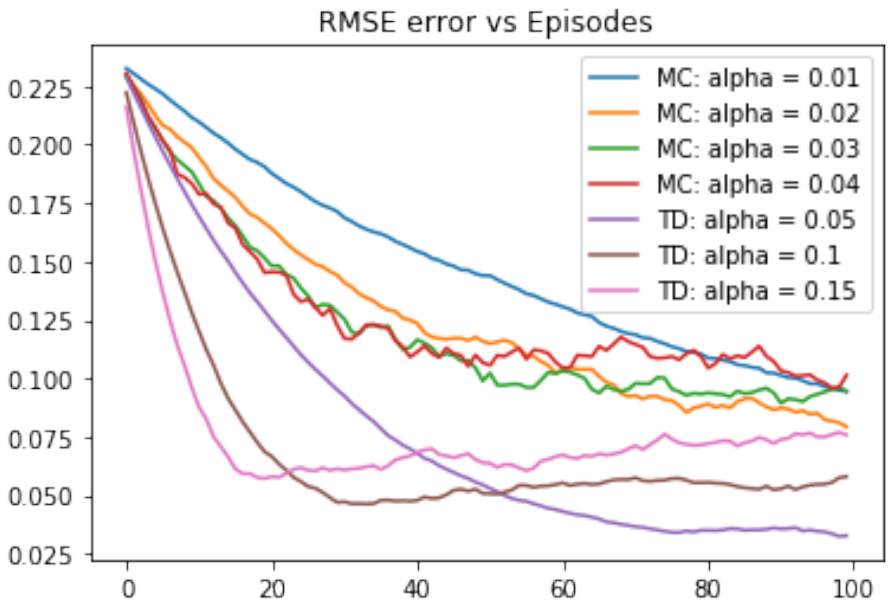

Figure 11: RMSE error vs Episodes

3 Question 7

As seen in the graph, SARSA performs better than Q-Learning as per the sum of rewards plotted.

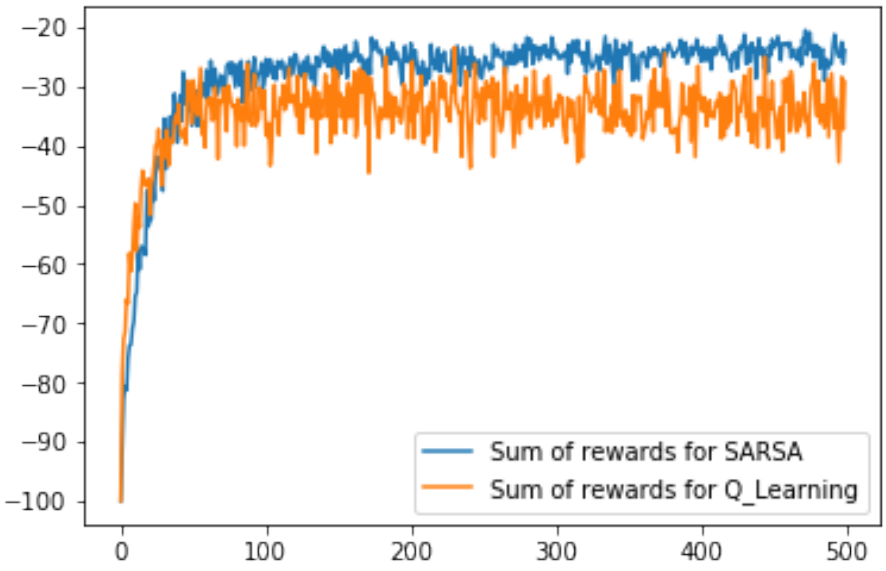

Figure 12: Sum of rewards for SARSA vs Sum of rewards for Q-Learning