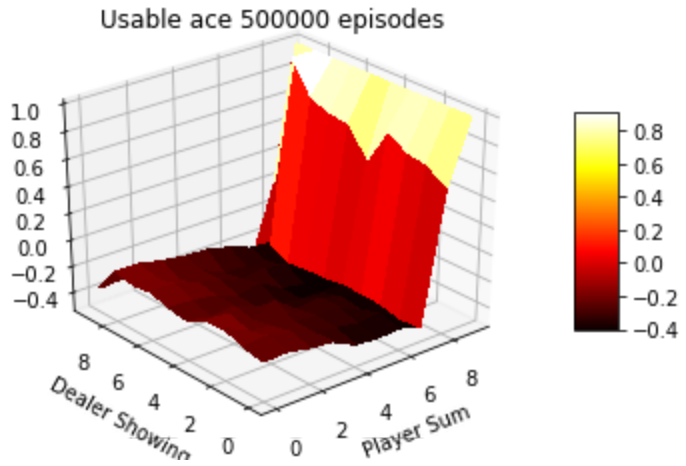
REINFORCEMENT LEARNING

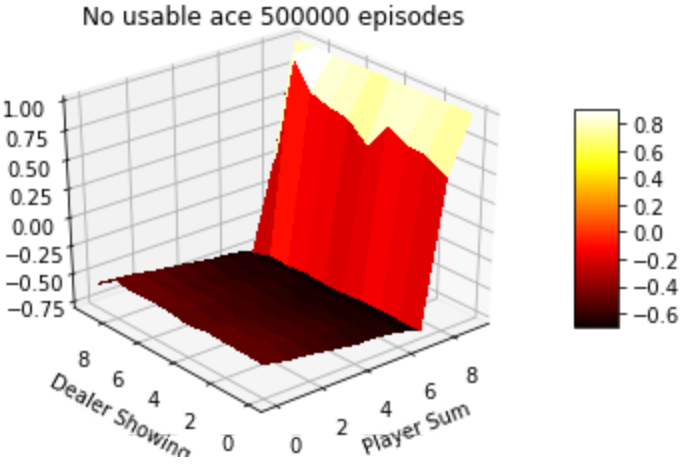
Homework 3

Surabhi S Nath
2016271

Q4. BLACKJACK

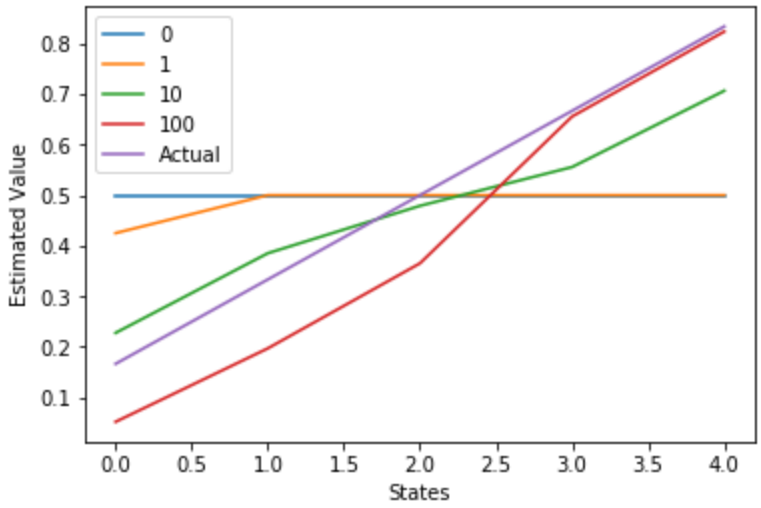
In this problem, we need to implement the Blackjack game and generate plots for Monte Carlo policy evaluation. The following 3D plots are obtained, which match the ones in the book

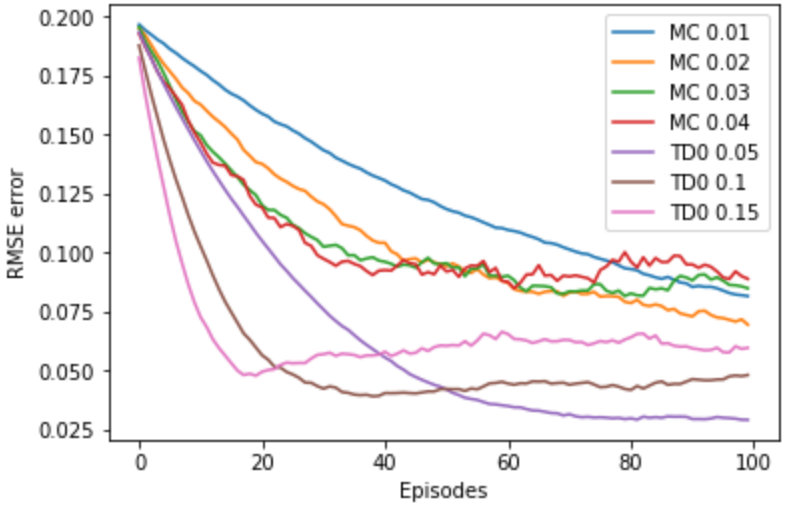




Q6. RANDOM WALK

In this problem, we had to simulate the above random walk. The V values and RMSE error exactly match the representation in the book.





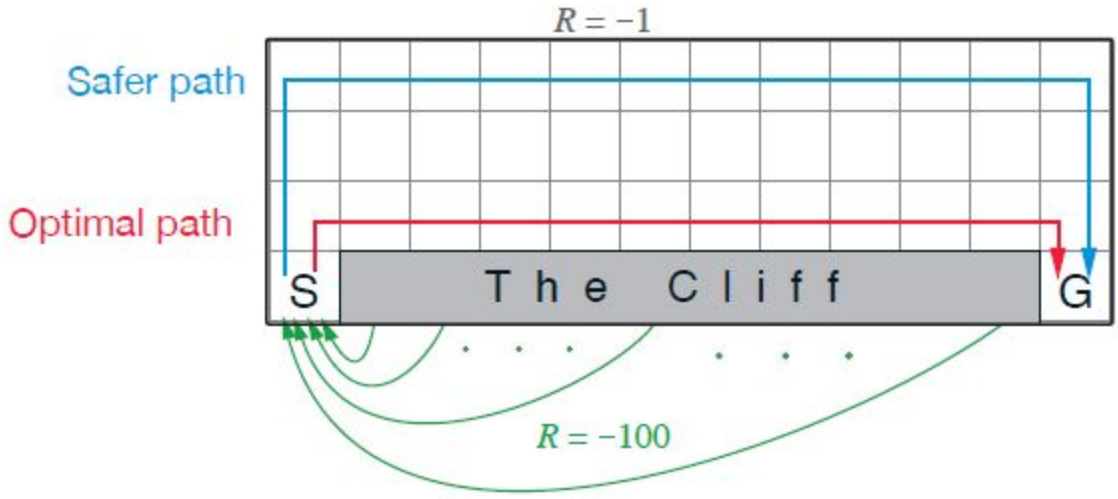
5.3. Since only V(A) has changes, it means that from C, the agent reached A then reach the terminal state to the left of A. The difference in V value for terminal state and A led to a

decreased V'(A). No other state was visited which has a difference in V values of s and s' hence no other state value is updated.

5.4. The value of alpha is in some sense a tradeoff how quickly you want to reach near the optimal values vs how accurately you want to reach near them. At a lower alpha, the V values obtained in the long run will be much nearer to the optimal values but the time taken will be very high. A smaller alpha will hence significantly perform better than the plots shown but time taken (number of episodes) will be much more.

5.5. For a large alpha, initially when V values and actual values are very far apart, the RMSE will be very high and will decrease with episodes. When is is almost near convergence, if the alpha is high, it may happen that is misses the optimal V values and keeps fluctuating about non optimal values as a result of which the RMSE increases.

Q7. CLIFF WALKING



For comparing the SARSA and Q learning performance, the best example is Cliff walking. The agent needs to travel from S to G. Falling into the cliff kills the agent and hence yields a high negative reward. The Q learning approach performs greedily hence chooses the red coloured path while SARSA chooses a non greedy and perhaps longer path from S to G. However, since the greedy choice is associated with a high probability of falling into the cliff, the performance of Q learning is worse as compared to SARSA in this example. This can we verified from graph below.

Graph using:
alpha = 0.5
epsilon = 0.1
discount = 1