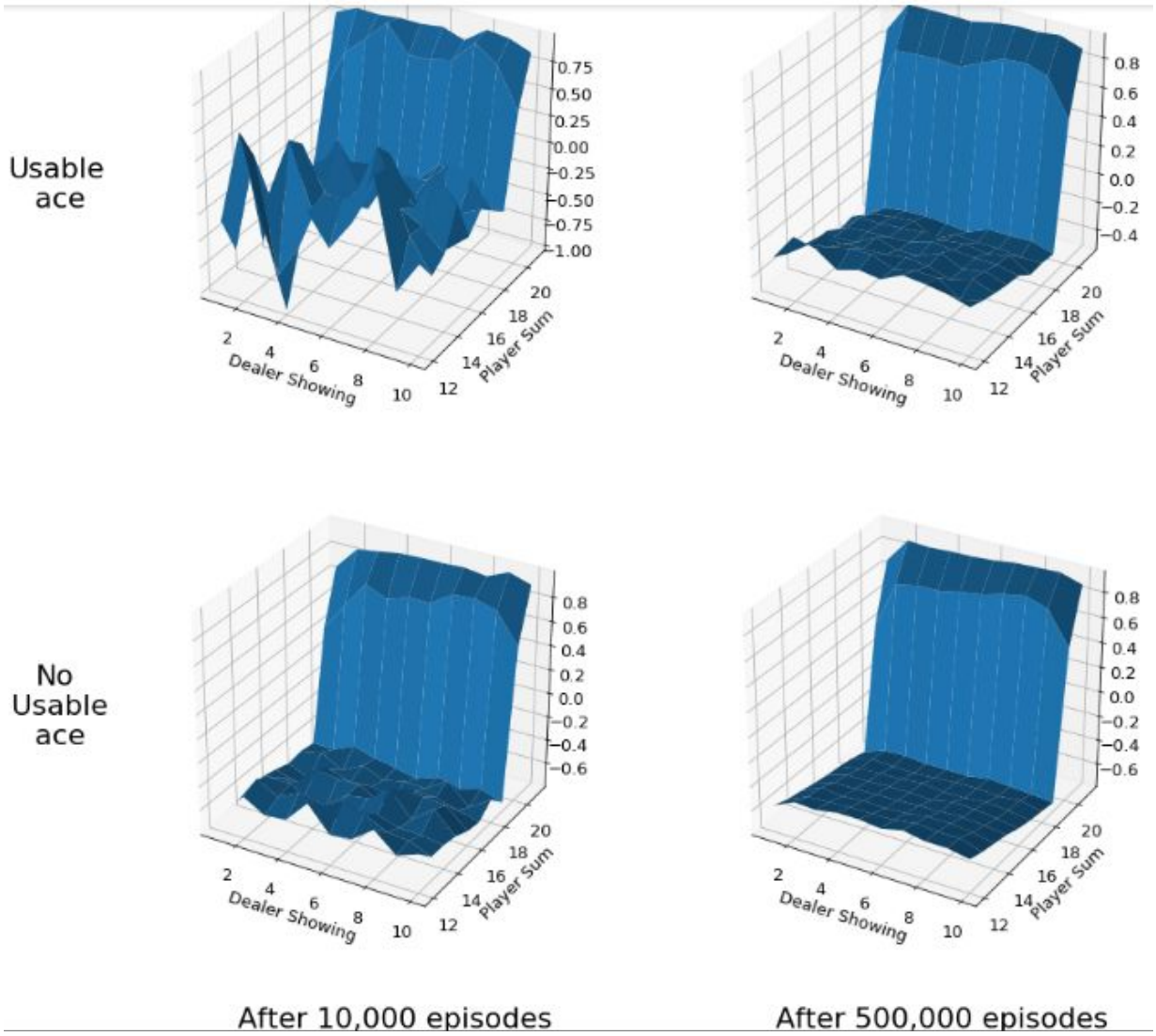
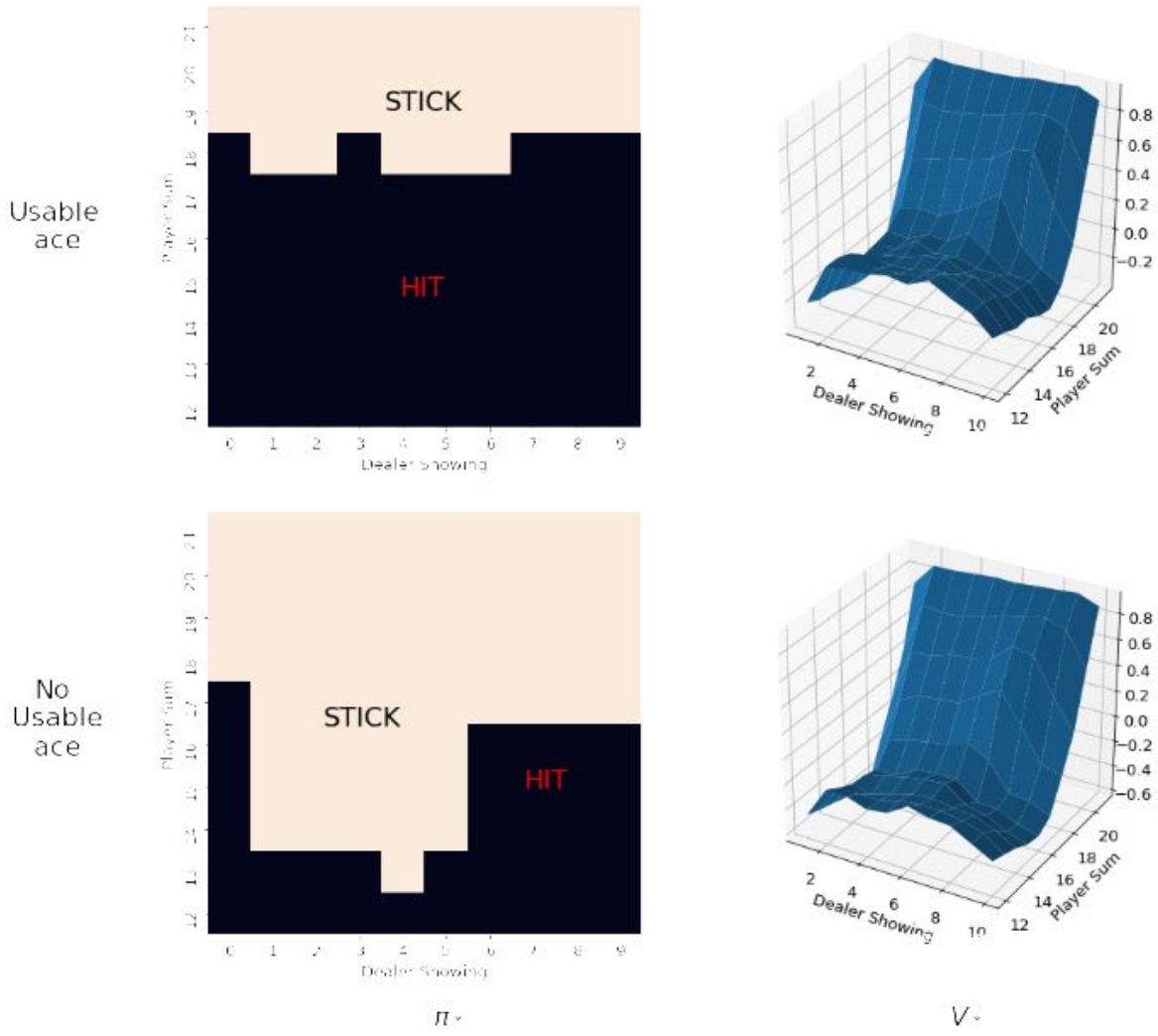
RL Assignment 3

Shaurya Goel
2016194

ANS-4

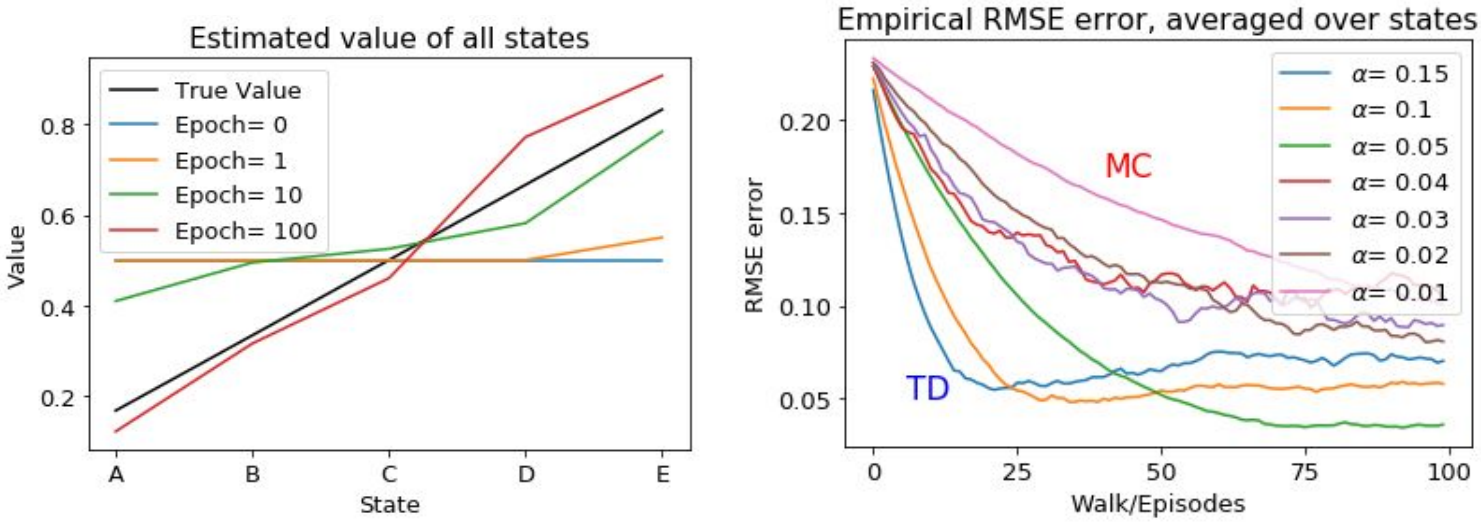


The usable ace case is noisier as getting an ace is less likely. Hence, we need more estimates to remove this noise. At, 5,00,000 iterations we get a smooth surface.



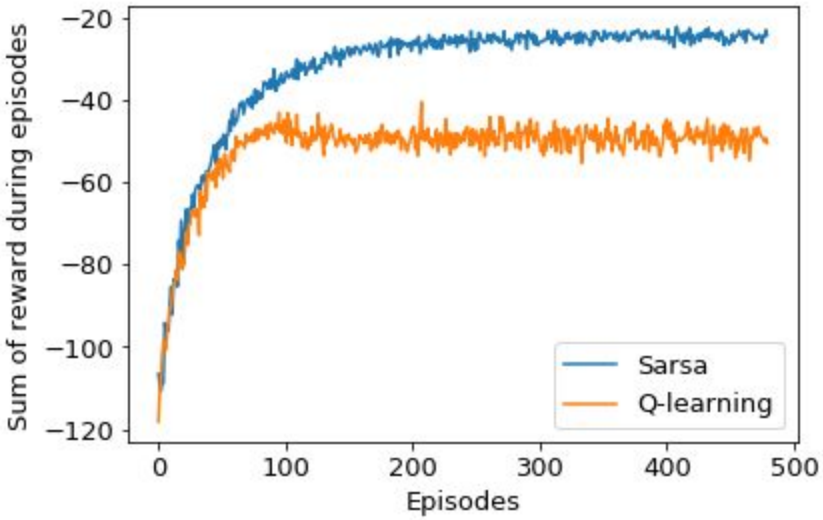
Here we see the optimal policy and optimal value function.

ANS-6



The left figure shows that TD0 algorithm converges to optimal value function very quickly. The right figure shows that TD algorithm is in general better than MC for the given values of alpha.

ANS- 7



We see that Q learning gives lower reward than SARSA for the given example. But, Q-learning gives a better policy than SARSA. This is because Q-learning walking alongside of the cliff, which increases the agent's chance to fall off the cliff due to epsilon-greedy policy.