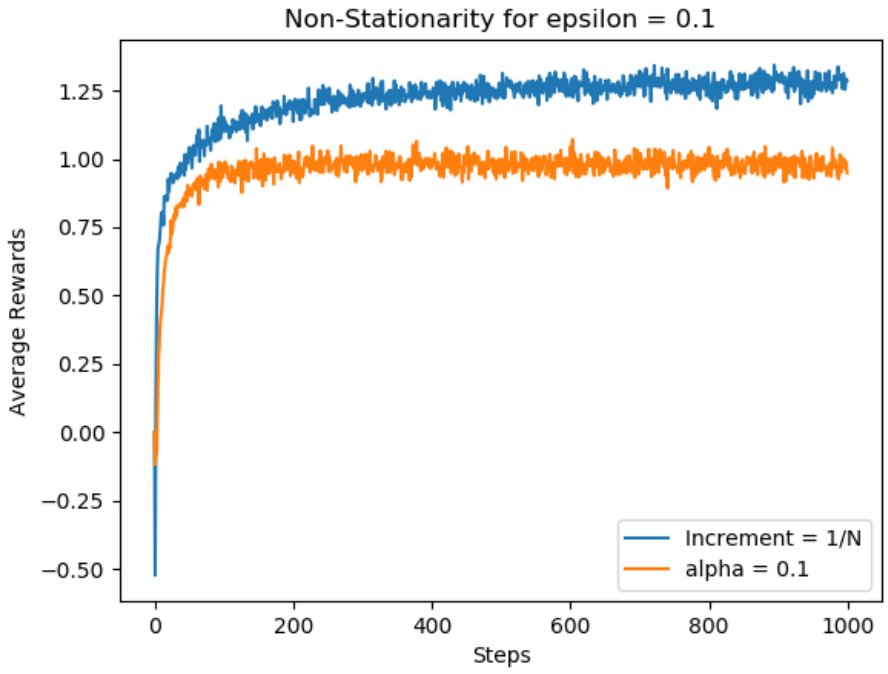
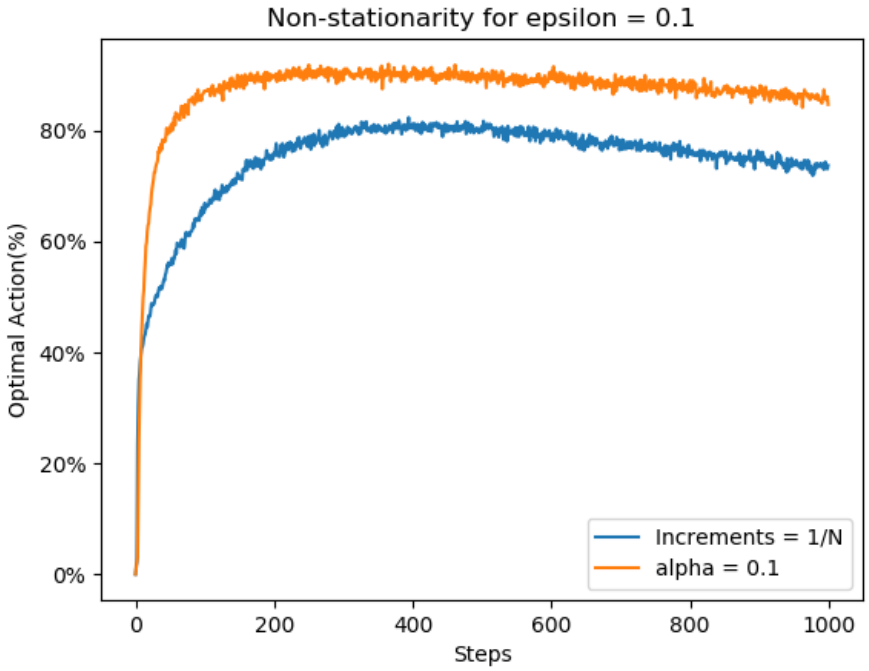
Assignment 1

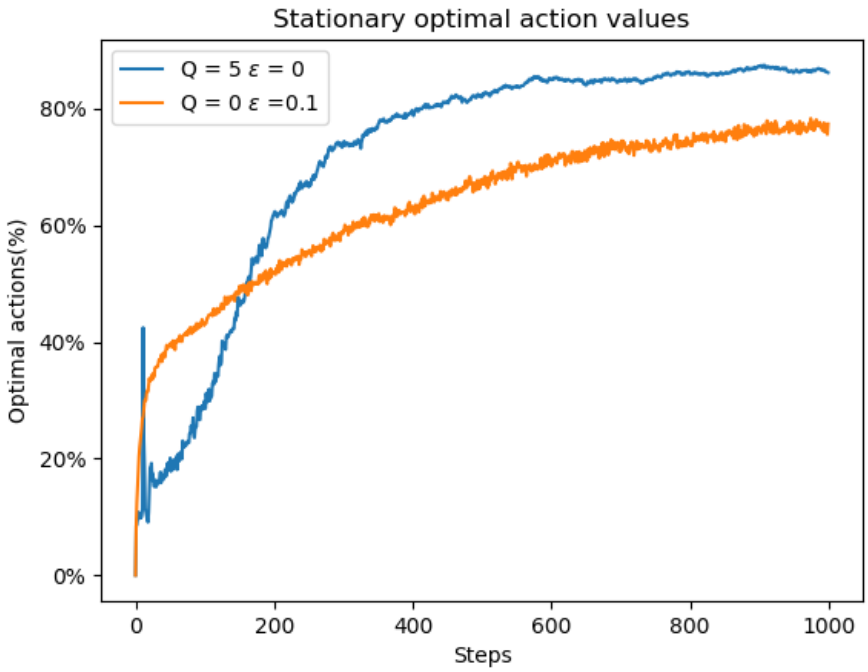
Q1 - Exercise 2.5, Non stationarity for average rewards and optimal actions.

- Number of bandits = 2000
- Time steps = 1000

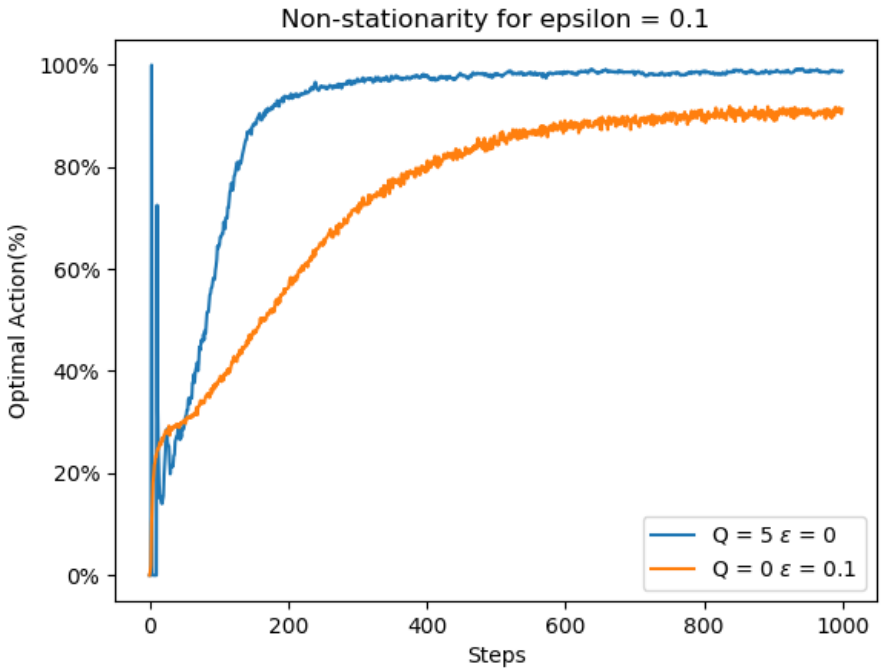




Q2- Optimistic Initial Values

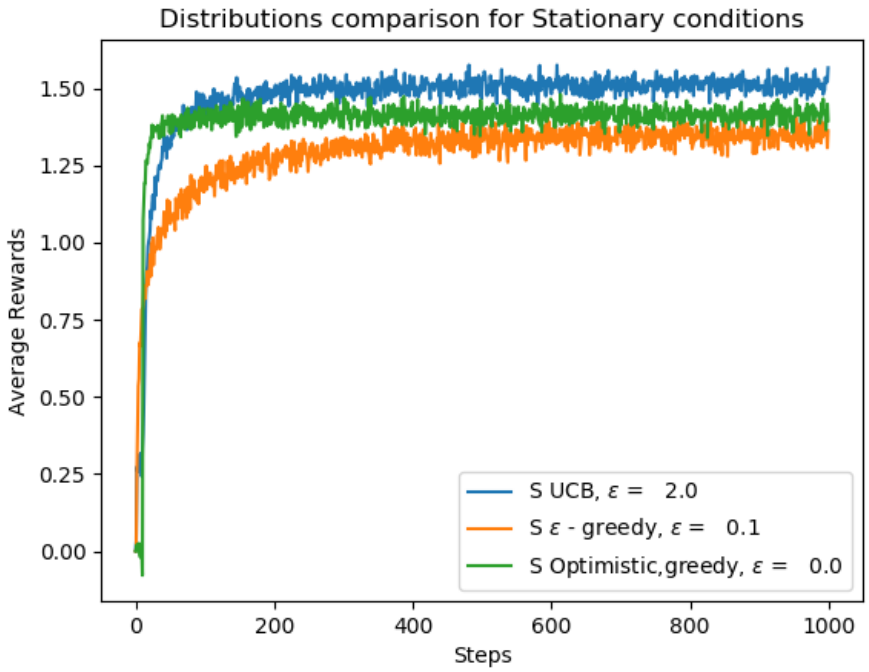


We could see initial spikes due to exploration by all the agents in initial steps but resulting into failure in rewards they see, trying all rewards for at-least 40% times, they try other actions too. In particular, due to high optimistic initial value agents explore more but as they approach closer to average rewards, their exploration decreases and lesser spikes in later parts of time steps. Since in optimistic initial values we start with high values of Q but as they become closer to Q*, exploration reduces and we get convergence.



In non-stationary conditions, due to same values of Q*(a) for all machines, and high values of Q, exploration goes on to maximum in initial time steps and then decreases due to randomness in each arm. But for Q=0, the actions are increasing as time steps increase uniformly unlike optimistic initial values where the exploration first increases and then converges as they near the average Q values.

Q4. Comparison of optimistic initial values, upper confidence bound and epsilon greedy for stationary and non-stationary cases.



The above graph is for stationary conditions for comparing all the 3 distributions for average rewards. As described in book upper confidence bound performs the best.