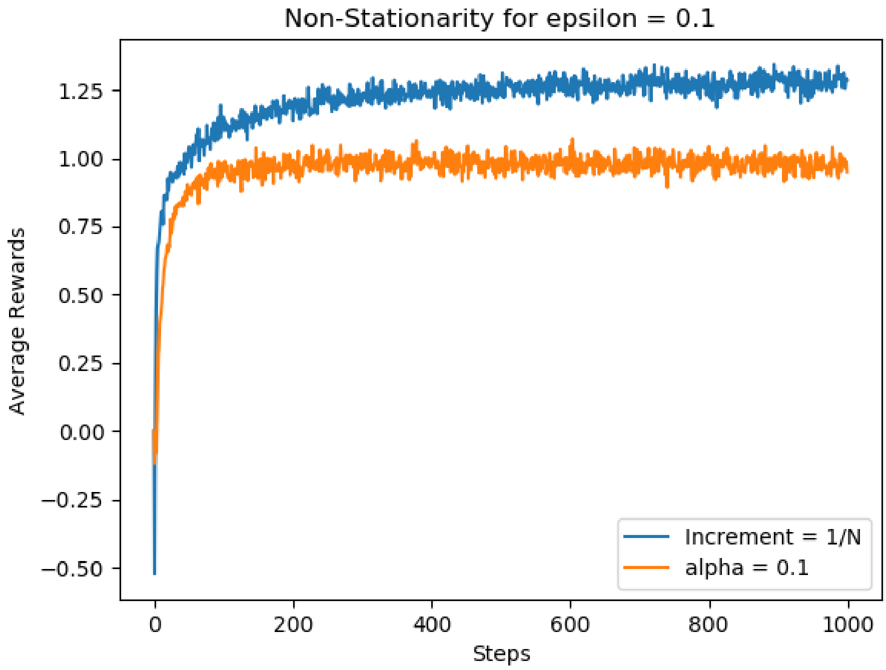
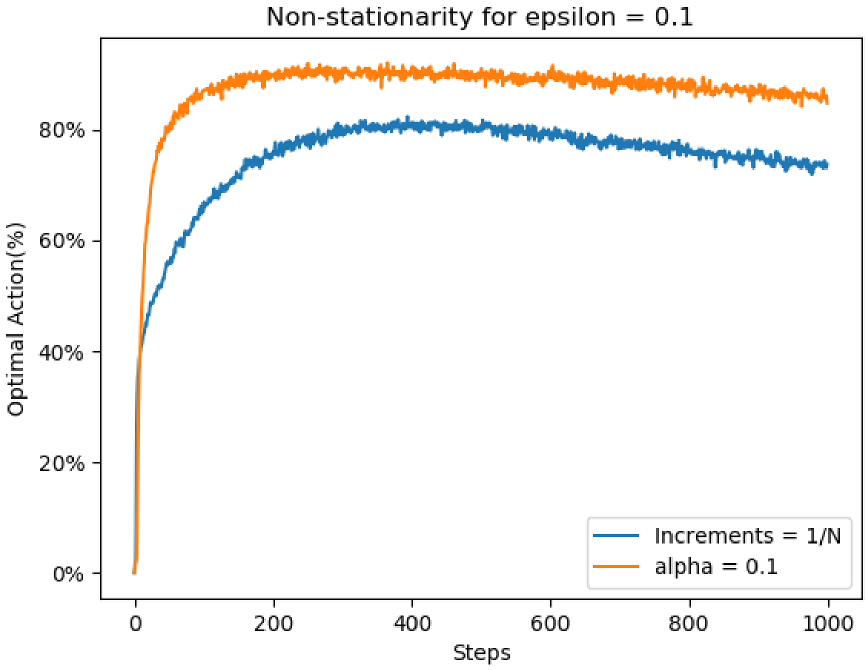
Assignment 1

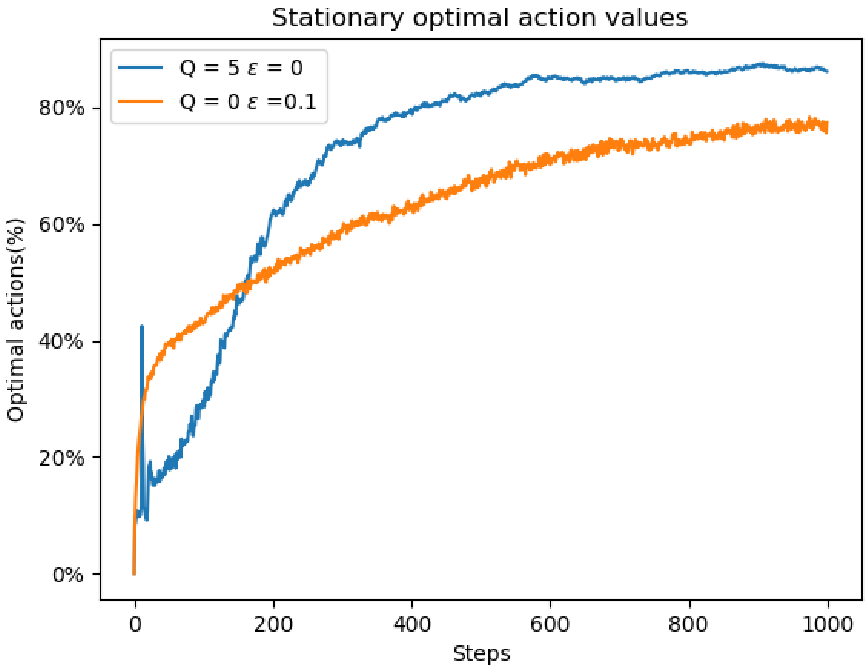
Q1 - Exercise 2.5, Non stationarity for average rewards and optimal actions.

- Number of bandits = 2000
- Time steps = 1000

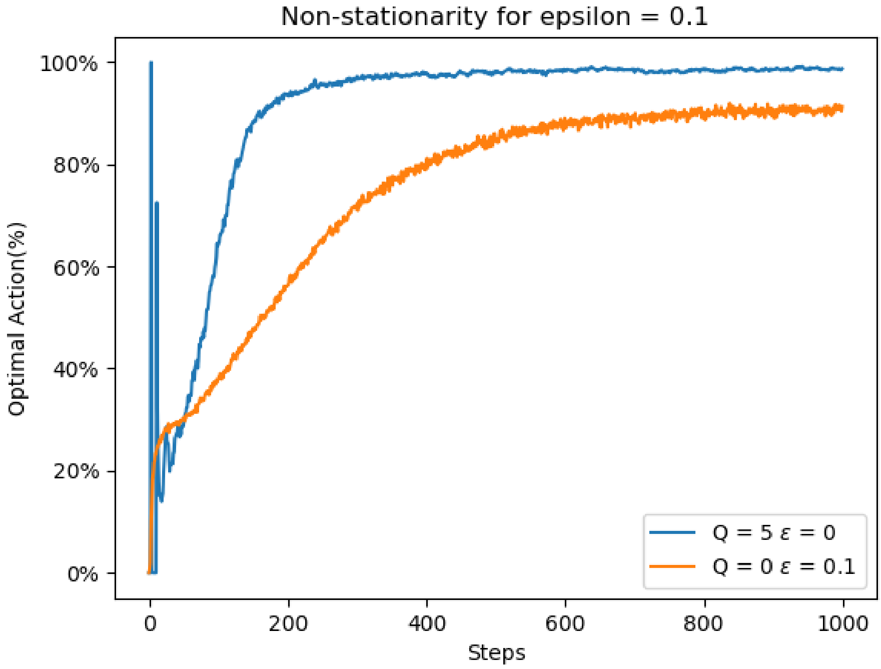




Q2- Optimistic Initial Values



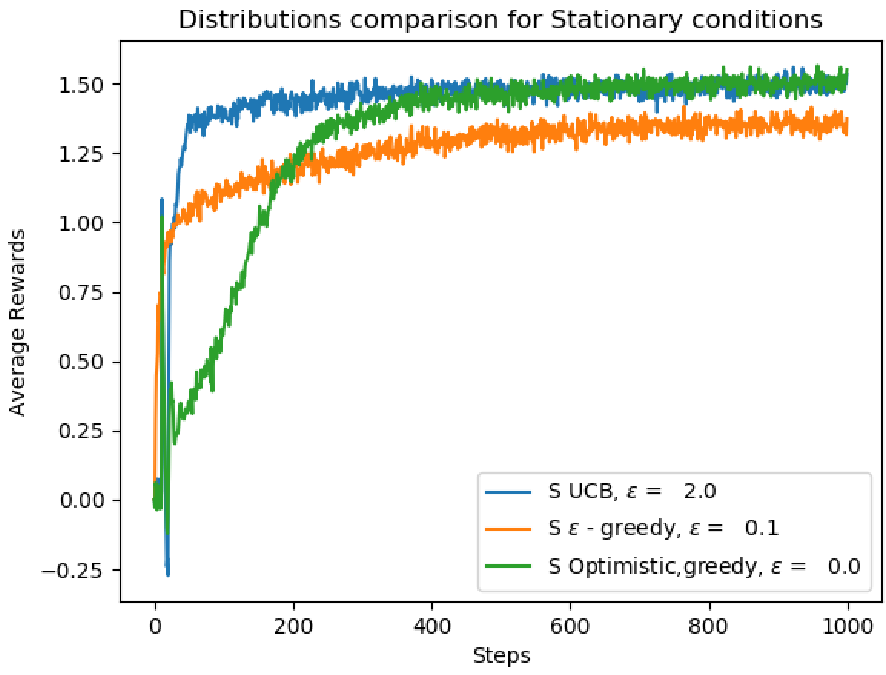
We could see initial spikes due to exploration by all the agents in initial steps but resulting into failure in rewards they see, trying all rewards for at-least 40% times, they try other actions too. In particular, due to high optimistic initial value agents explore more but as they approach closer 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.

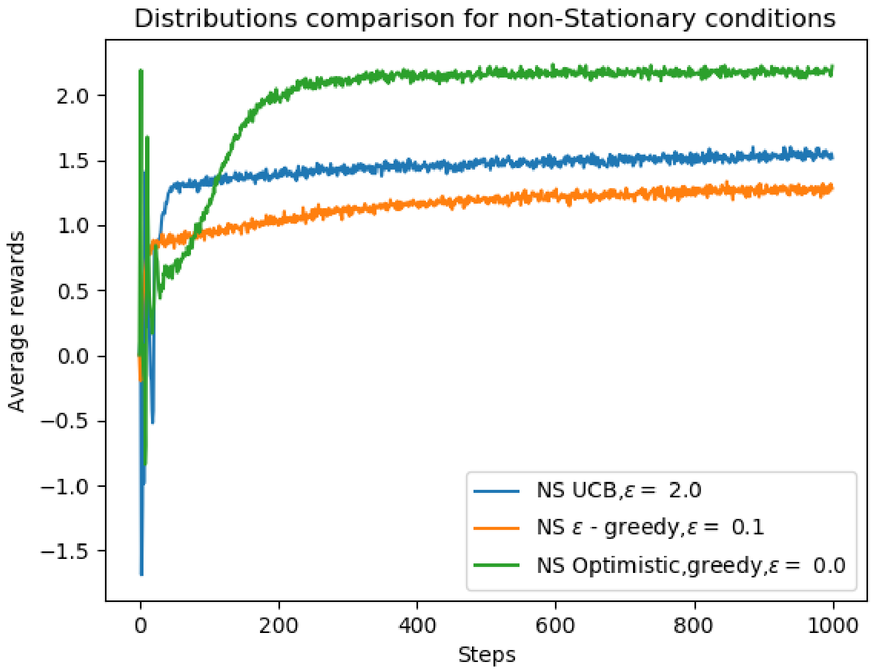


In non-stationary conditions, due to same values of Q*(a) for all machines, and high values of Q, exploration goes on to maximum in initial time steps and then decreases due to randomness in each arm. But for Q=0, the actions are increasing as time steps increase uniformly unlike optimistic initial values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

A not too high optimistic initial value might make optimal actions perform better for stationary condition, due to a very high true q values and choosing actions greedily gives high spikes, an optima value of q values will make exploration of optimal actions better.

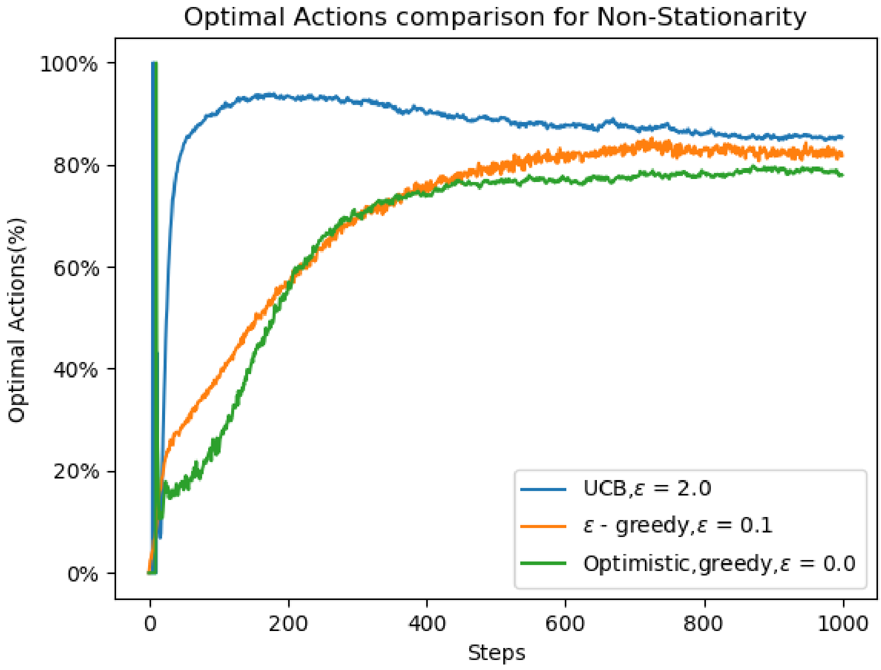
Q4. Comparison of optimistic initial values, upper confidence bound and epsilon greedy for stationary and non-stationary cases.

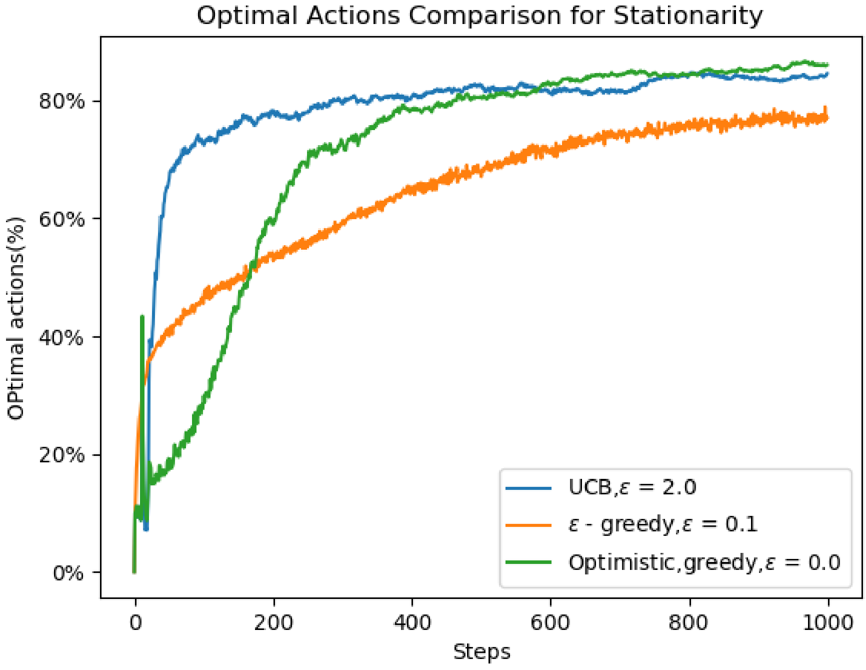




Above graphs are the comparisons among ‘Optimistic Initial Values’, ‘Upper confidence Bound’ and ‘Epsilon Greedy’. The top graph is for stationary conditions where we can observe ‘UCB’ performing the best due to selection of non greedy actions on the account of assuming they might perform better in further conditions (as explained in book). OIV also performs almost as good as UCB and both of them better than epsilon greedy.

The next graph is for non-stationary conditions where UCB can't perform as good as Optimistic initial value and epsilon greedy due to large state spaces and complex methods of choosing an action. Thus, UCB explores more in the beginning like other conditions but optimistic greedy gives best rewards for non-stationary conditions.





The above 2 graphs are for optimal actions for all distributions for stationary and non-stationary conditions. The first graph is for optimal actions in non-stationary conditions. All the distributions as expected explore in the starting owing to same initial values for arms in all machines (q*(a)) but optimal actions for UCB decreases drastically as time increases after the rewards approach close to the average rewards for true values. All the distributions thus converge to same value.

The last graph is for optimal actions for stationary conditions. UCB performs the best for stationary conditions owing to choosing all actions not greedily but exploring the ones which might give best rewards in future. Similarly, Optimistic initial values perform better than epsilon greedy as shown in text too.