HW1

Q2. Ex2.6 from Reinforcement Learning - An Introduction (S+B) 2018

For stationary problems, after setting the optimistic initial values the agent would try to explore all the possible actions. After they have tried all the possible bandits the agent would choose the action with the highest reward it got and there is a high probability that the action with the highest probability at this time step is also the optimal action.
For non-stationary problems, optimistic initial values perform worse because the drive to explore here is temporary and if the task keeps changing which needs more exploration we cannot facilitate that using optimistic initial values. Also, the initial spike in non-stationary problems is due to the fact that the true rewards are initialized to the same value and take random walks (using the same setting as in ex2.5).

Q3. Ex2.7 from Reinforcement Learning - An Introduction (S+B) 2018

The analysis is present in a separate PDF named ex2.7_analysis.

Q4. For stationary case, UCB performs better than e-greedy with optimistic initial values, this is because of the exploration promoted by the factor Nt in the equation for UCB. Due to the inverse relation of exploration and the number of times an action is selected promotes exploration at first and then promotes greedy selection to converge faster and also give higher rewards overall.
For non-stationary case, UCB can't explore enough as it is limited by the number of times an action is selected but as the goal is changing every time step in the non-stationary problem we need more exploration than UCB can provide, thus e-greedy performs better as it keeps exploring with a constant probability.

Note:- Please ignore 10_armed_testbed.py it was used to create the testbed initially for testing

Name:- Vishal Samal
Roll No:- 2017209