RL

Homework - 1

Ans 2. The oscillations and spikes in the early part of the curve for the optimistic method represents that the agent explores an action and then gets disappointed by the reward for that action and explore a new action for the next time step. So, after trying out all the actions, the percentage of selecting an optimal action rises. (around 35 - 40 %). The spikes in the graph are caused because that many bandit agents selected an optimal action at a particular time step.

Ans 4. As can be seen from the graph of Average Rewards in stationary case, the slope of graph of UCB is more than the graphs of ϵ-greedy & optimal initial values. for time steps greater than 1000. So, the optimal action taken by UCB is more than the rest.

For non-stationary case, the percentage of optimal action is reduced to 8 - 10 % but UCB is still the best method among the 3. Since UCB explores more, it has higher optimal action percentage.