Assignment1
Reinforcement Learning
Ferdowsi University of Mashhad Spring2023

Arya Ebrahimi 9822762175

Question1:

The exploration rate grows with increasing the ϵ, so the agent acts less greedy. If the ϵ is zero, the agent will always choose the greedy action and never explores. Suppose the ϵ is n which is a number between zero and one. In that case, the agent will explore n percent of the time and exploits the current best action otherwise.

If the ϵ is one, the agent acts randomly at each timestep. In this case, the average reward would be around zero because the agent selects uniform random actions. The true reward equals zero, so selecting random actions would not provide a positive reward for the agent.
On the other hand, small ϵ may exploit a wrong action at first, but after some timesteps, it will find better choices and select them most of the time, so the average reward is probably a positive value.

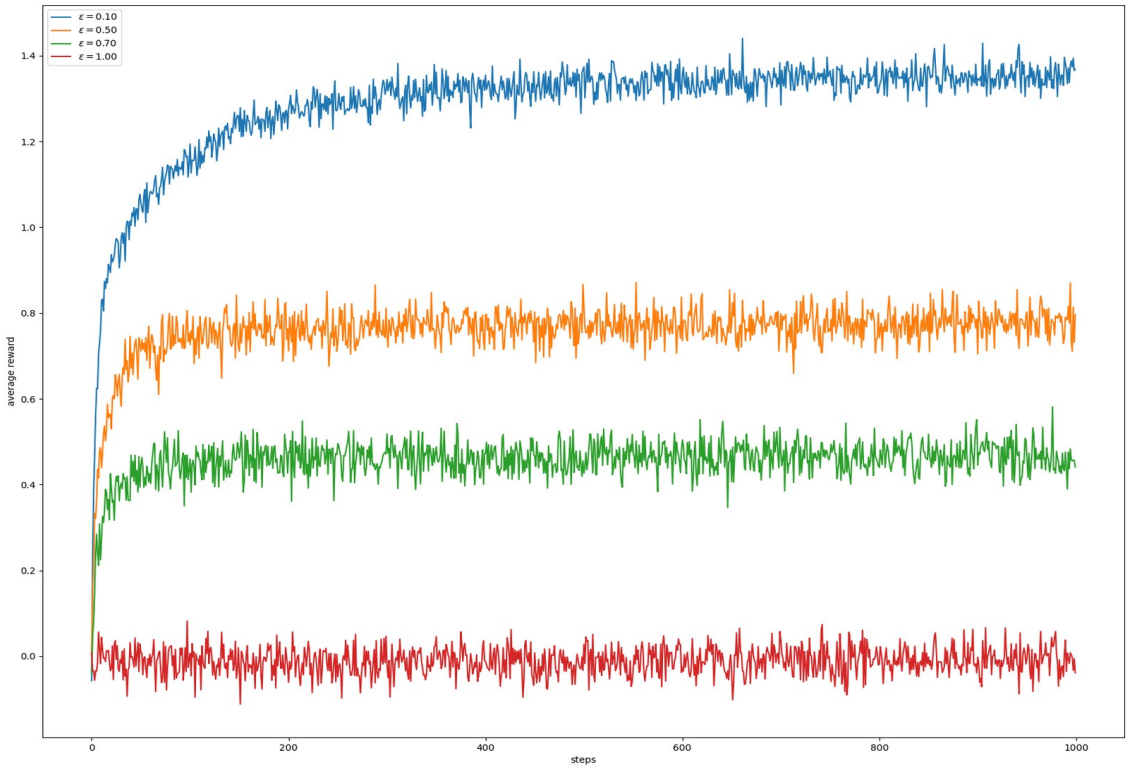

Figure 1: Average performance of ϵ-greedy method on the 10-armed testbed.

Question2:

The ϵ-greedy method provides a way to explore at some timesteps and exploit otherwise. If the number of timesteps increases and approaches infinity, then every action will be sampled an infinite number of times, ensuring that $Q_t(a)$ converges to $q_*(a)$. Thus, using the ϵ-greedy method, the probability of selecting the optimal action will be greater than $1 - \epsilon$.

Question3:

The UCB method has a term that is a measure of uncertainty and is responsible for exploring.
At each timestep, the agent will choose the action with the highest upper confidence bound. By selecting an action, its upper confidence bound gets one step closer to the actual reward distribution of that action.
At first, upper confidence bounds are the same, and the agent should choose randomly between actions. After selecting an action, its upper confidence bound will change. In the next step, the agent should choose the action

with the highest upper bound. Since the upper bound is probably an optimistic value, the previous action will not be selected, and the agent should choose from the other actions.
If an action has a lower reward, then its upper confidence bound decreases more, and for the optimal action, the upper bound will converge to the actual reward.

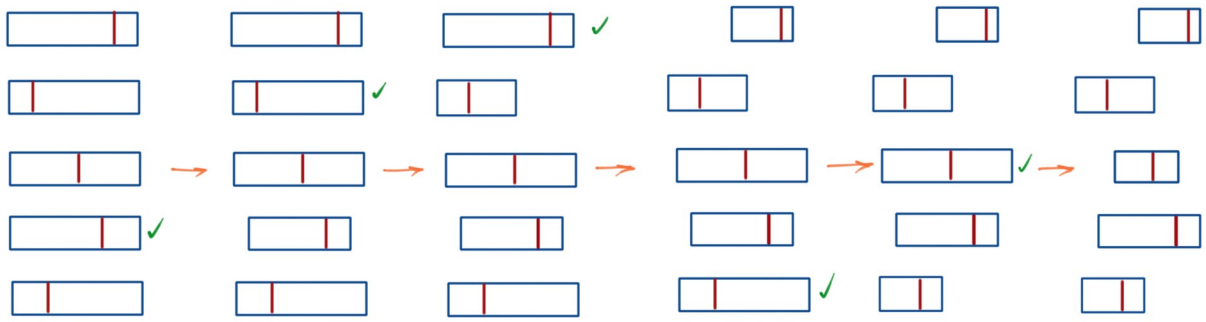

Figure 2: UCB method example for a 5-armed bandit.

As can be seen in Figure 2, after five timesteps, all the actions are explored, and the upper bound confidence gets closer to the actual value.

Question4:
As can be seen in Figure 2, for a 5-armed testbed, all the actions are selected after five timesteps, and for the 6th timestep, the action with the highest upper bound should be selected. Since all the upper bounds are updated and are closer to their actual action value, the probability of selecting optimal action is very high at this timestep. The same scenario can be seen for the 10-armed bandit. After ten timesteps, all the actions are selected once, and for the 11th timestep, many runs will select the optimal action, and this causes a spike in the plot.