Reinforcement Learning
Spring 2023
HW1

Sara Ghavampour
9812762781

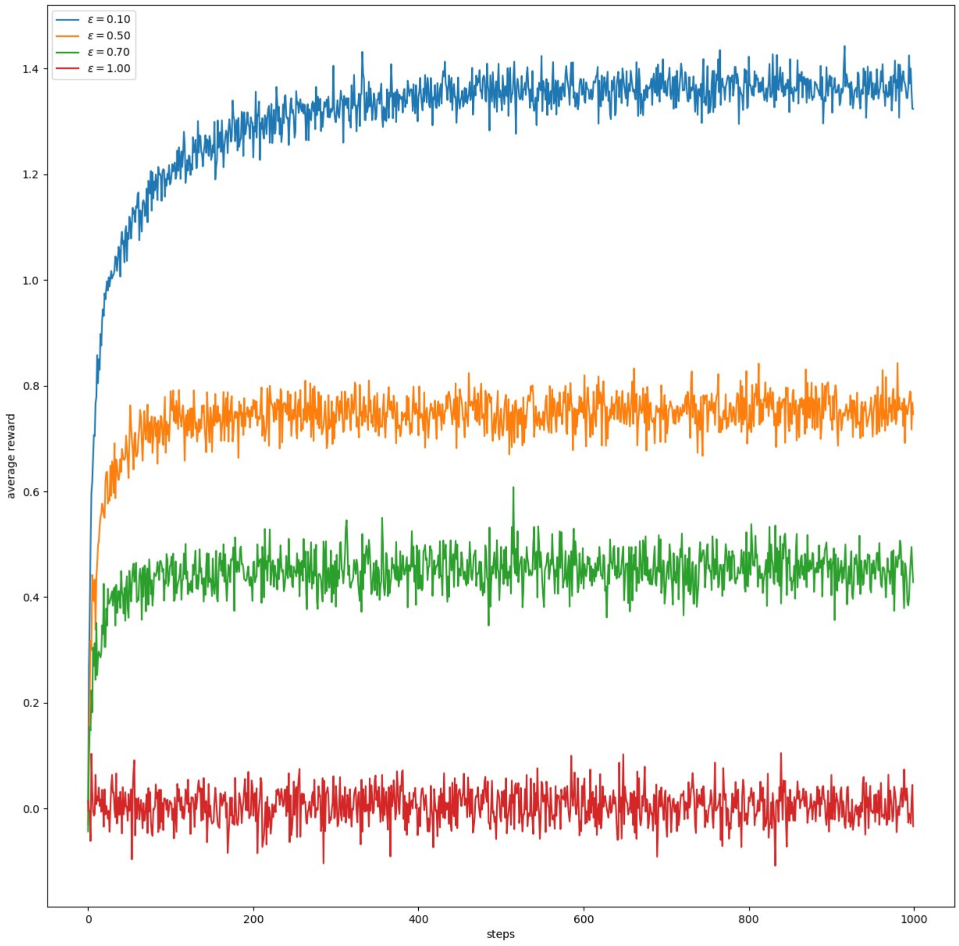
Problem 1

Uncertainty caused by lack of sufficient information about env model, makes exploration a key element of RL systems to solve both stationary and non-stationary problems. As being said to balance exploration and exploitation, epsilon greedy algorithm can be used.

Parameter ϵ determines amount of exploration in epsilon-greedy algorithm. ϵ can be in range [0,1]. If ϵ is too small it means that most of the time (1- ϵ) algorithm decides to exploit current knowledge of model and actions and does not sample unseen actions to gain more knowledge which causes maximum short-term reward and converging to suboptimal.

By increasing ϵ algorithm explore and sample more actions, and by exploring more uncertainty is reduced and better actions can be taken in long run and eventually performs better .

If ϵ is too big like 1, it simply means that algorithm explore most of the time and never uses the gained knowledge by all this exploration so it basically acts in a random manner and average reward can be around 0.(q* has a normal distribution(0,1))



Problem 2

This algorithms balance eplore-exploit dilemma. By probability ϵ it explores in some time steps and exploits otherwise. By increasing number of time steps, each action will be sampled infinite times and by rule of large numbers in sample-average function, value estimations converge to q* . It explores in all time steps and not just early stages so it can be used for both stationary and non-statationary problems.

Problem 3

UCB uses uncertainty in action selection and mix exploration and exploitation together. It does not choose an action just by it’s value estimation and also uses another term named UCB exploration term. The latter term provides a confidence interval (uncertainty) that if it is small shows that we are certain that true value of this action is in this interval(value estimates are close to true values) and if it is big shows that we are not certain about true values (value estimates are distant from true values).

UCB follows principal of optimisim in case of uncertainty meaning that, it always chooses action with maximum upper bound and assumes that this is a good thing. At each step,UCB chooses action with maximum upper bound and and update maximum upper bound based on reward received by the action and maximum upper bound gets closer to true values.

UCB exploration term decreases as number of times that a particular action is selected increases(uncertainty decreases by many selections of an action) and because of this feature it is not suitable for non stationary tasks.

Problem 4

In 10 armed bandit problem(10 avtions), UCB sample all actions in first 10 time steps(each action once) and updates all maximum upper bound so for 11’th timestep agent tends to choose optimal action and causes 11’th step’s spike.

z\xsc\sx