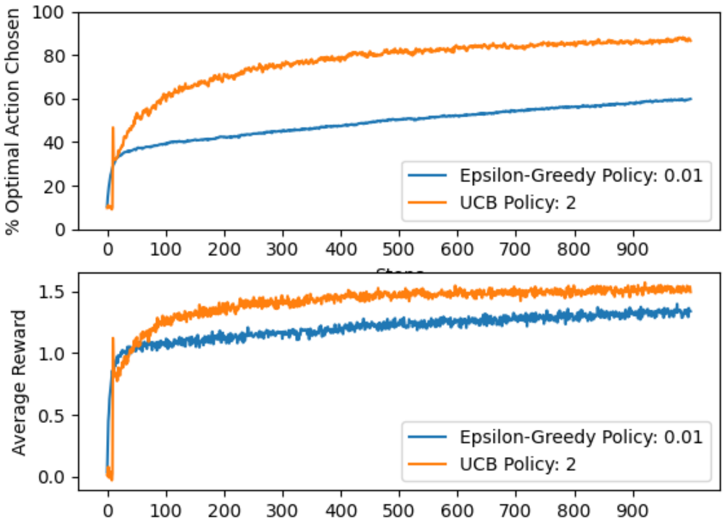
# Assignment 1: The K-Armed Bandit

## UCB vs Epsilon-Greedy

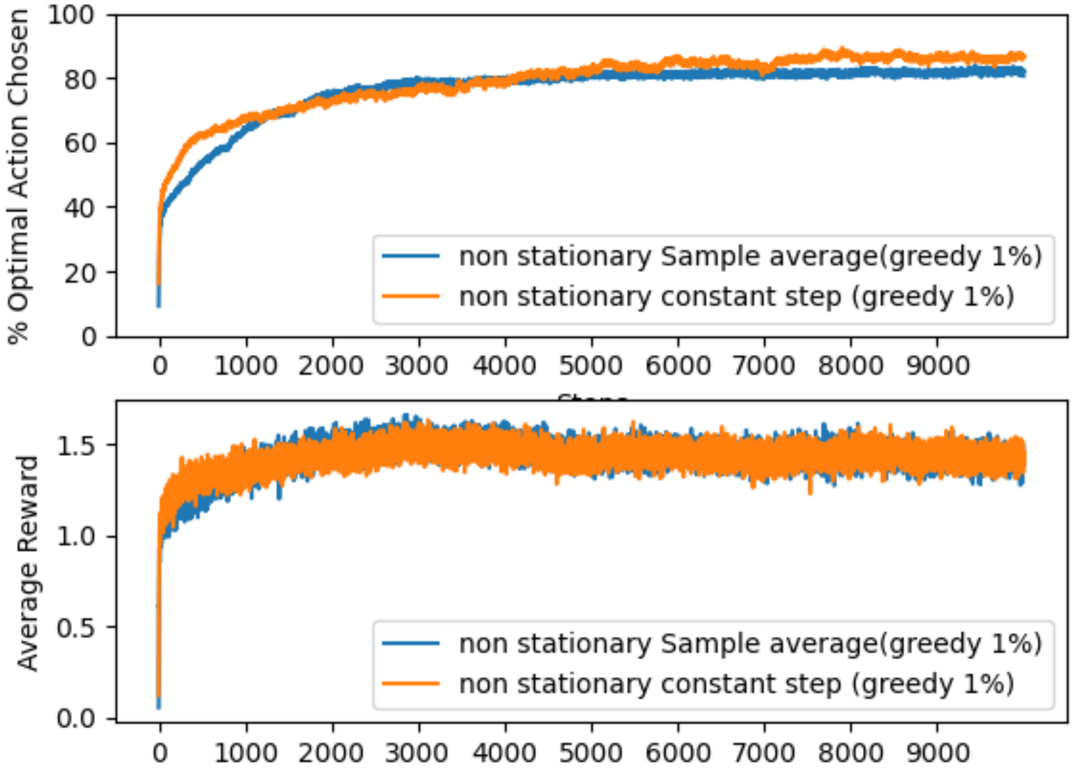
We set up the bandit environment as k = 10, u = 10 sigma =1.the environment is stationary, This time we compare two different agents, one is greedy epsilon 1% agent the other is UCB agent. Then we have 1000 timesteps, run for 2000 times and eventually get two plots.



From the graph, we can easily see that the UCB(C=2) will always outperform epsilon greedy (0.01) as timestep increase. The reason behind it might be instead of randomly assigning probability. The UCB more focus on encourage the agents choose the actions which has never been chosen. So, it is the better way for exploration.

Random Walks(sigma = 0.01)

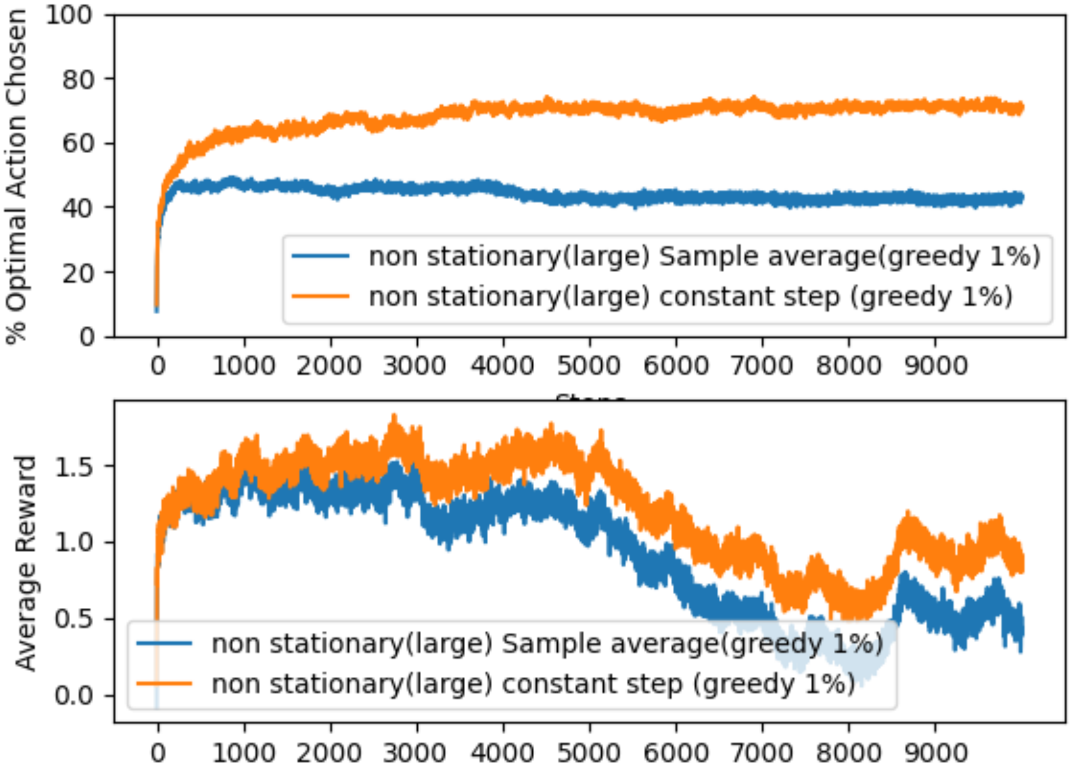
When environment is nonstationary. We try to add an incremental random variable N (0,0.01) at first. And then we try to construct two agent they both use the greedy epsilon 1% but one of them use sample average but another take fixed learning rate (1%). we set the timestep =10000 and runs = 500 and compare those two agents behave.



When We can see that those two models behave basically the same.

## Random Walks (sigma = 0.1)

Other conditions keep the same, just increases the sigma of random variable from 0.01 to 0.1.



We can see that the constant step method behaves way much better than Sample average method. The reason why it might happen is because there is no constant q* but the Sample average method try to find it under large number. What it learned from past always get updated by this disturbance.