Comp138 Reinforcement Learning: K-armed Bandits

Junsung Tak

September 2021

1 Goals

The goal of this assignment is to explore the inadequacies of the sample average method in non-stationary problems. Additionally, multiple feedback techniques will be used (ie, e-greedy, greedy, optimistic initial value, and Upper bound selection) on both stationary and non-stationary settings in order to illuminate the differences between the feedback techniques.

2 Introduction

We will define the stationary K-bandit problem as follows: There are k different actions to take. These actions will return some reward value which is chosen from a constant, normal Gaussian distribution with mean = 0 and standard deviation = 1. Action is chosen based on metrics derived from several approaches (sample mean, recency bias). Furthermore with multiple feedback techniques we will see a wide range of balance between exploration and exploitation which hopefully will provide insights into approaches in (non) stationary environments.

3 Different Action Selections

ϵ-greedy:
Agent balances exploration and exploitation in this method. There is ϵ chance that the agent will try a completely random action. With 1-ϵ chance the agent will pick the optimal action (defined by optimal action being the highest q* value) based on expected reward.

Greedy:
Agent only exploits. From the start agent will choose the action with the highest q* value and will continue to do so without exploring.

Optimistic Initial Value (OIV):

Similar to ϵgreedy in action selection. The important difference in this action selection is that expected rewards start at 5 instead of 0.

Upper Confidence Bound action selection:
Agent considers the expected reward but also considers the number of times an action has been taken. This methodology takes recency into account allowing for a clean balance between exploitation and exploration.

$$A = argmax(Q(a) + c\sqrt{\frac{lnt}{N(a)}}) \tag{1}$$

where A is the action chosen, t is the number of total steps, and N(a) being the number of times the specific action was taken. c is a constant size parameter and is set at the value of 0.1.

4 Specifics about experiment and parameters

General parameters of each method: (q*(initial) denotes the true values of each arm at initialization)

ϵ-greedy: $\epsilon = 0.1$, steps = 10000, q*(initial) = [0, 0, ..., 0]

greed: $\epsilon = 0$, steps = 10000, q*(initial) = [0, 0, ..., 0]

OIV: $\epsilon = 0.1$, steps = 10000, q*(initial) = [5, 5, ..., 5]

UCB: ϵ = NaN, steps = 10000, q*(initial) = [0, 0, ..., 0], $\alpha = 0.1$

4.1 Expected Reward Updates

Stationary:
The stationary bandit's true values are randomized from a normal distribution with mean 0 and 1 variance. Its true value stays stationary as the experiment is conducted. Since the environment is stationary, the expected reward update is calculated by this formula:

$$Q(a) = Q(a) + \frac{1}{K_n(a)}(reward - Q(a)) \tag{2}$$

where Q(a) is the expected reward given certain action a is chosen, $K_n(a)$ is the number of times the spe-
cific action was taken, and reward is the actual reward acquired from the bandit.

Non-stationary:
The non-stationary bandit's true values are all initialized to the same value (in

this case, 0) and on each step all the true values are incremented by a random walk. The random walk itself is a normal distribution with mean = 0 and standard deviation = 0.01. Through this, the true value of each arm is not constant. Since the problem is non-stationary, the calculation of expected reward is a bit different and is given by this formula:

$$Q(a) = Q(a) + \alpha(reward - Q(a)) \tag{3}$$

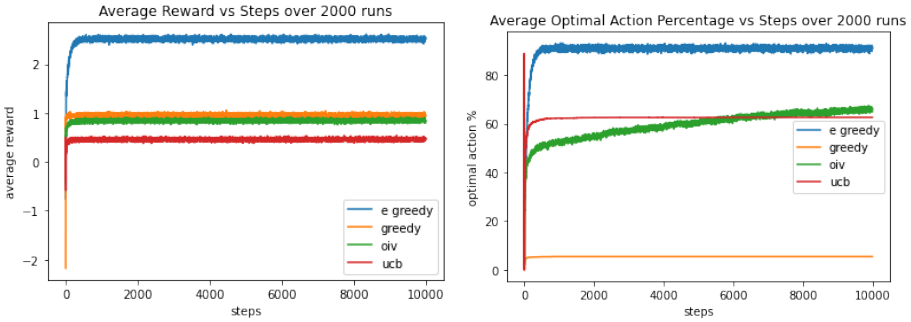
where α denotes a constant step size parameter between 0 and 1.

5 Experiment

Four different action selection methods will be tested in both stationary and non-stationary settings to compare and contrast results. A test bed of 2000 instances will be created for the experiment. As a result, 1 run or experiment will run for 10,000 steps and this will be run for 2000 different instances.

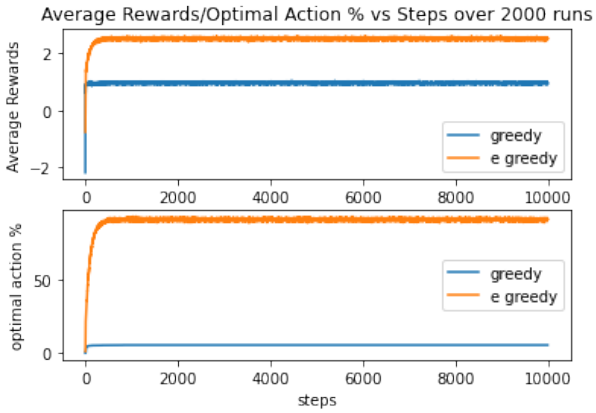
6 results

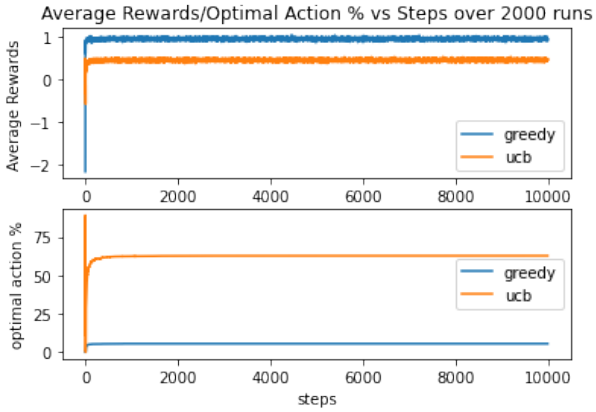
6.1 Stationary

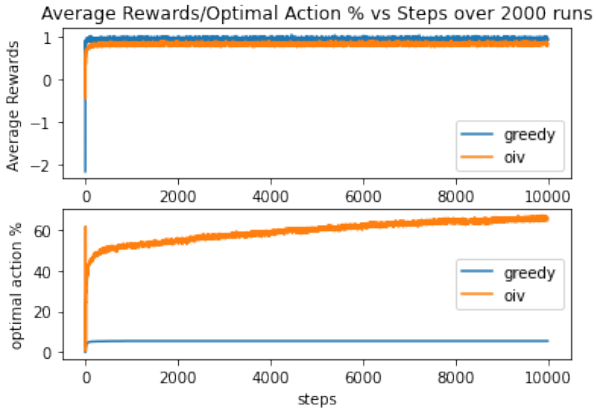


In the stationary enviorment we can see that ϵ-greedy, incremental method of action selection is the most optimal. The ϵ-greedy method reaches an average optimal action percentage of above 80% only after a few runs. This argument is further bolstered by the impressive difference between ϵ-greedy method and the other 3 methods in the average rewards graph.

Below, each action selection method is compared to the greedy method. The greedy method will be seen almost as a reference for comparison sake.



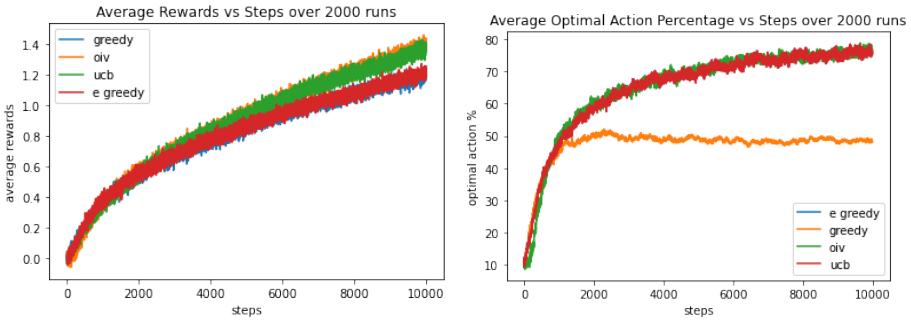




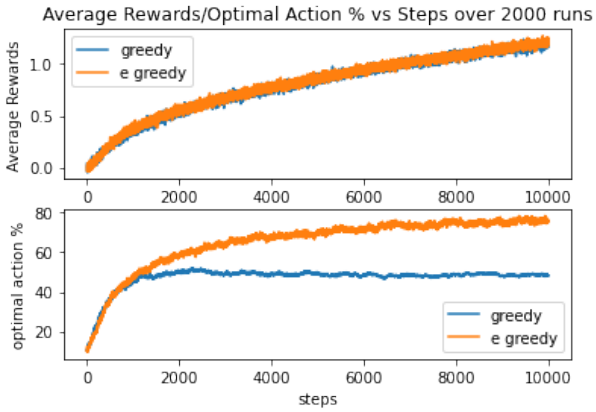
We see that the ucb method is performing subpar compared to the greedy method. Interestingly, we see that the optimal action % reaches very high during the first few steps. This is shown by the almost vertical line in the optimal action % graph for ucb and greedy methods. This is due to how if $K_n(a) = 0$, that action is considered to be the most optimal action. This

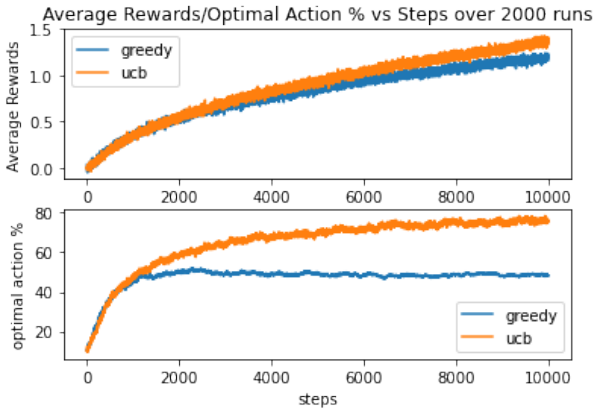
encourages early exploration, which enables the agent to learn at a more rapid rate. The agent attempted every action (since their $K_n(a) = 0$ at the few initial steps) and then found the most optimal one, achieving a percentage that is greater than 50%.
Furthermore, the average reward difference between the greedy and oiv method is surprisingly small, yet oiv reaches a significantly higher optimal action percentage. .

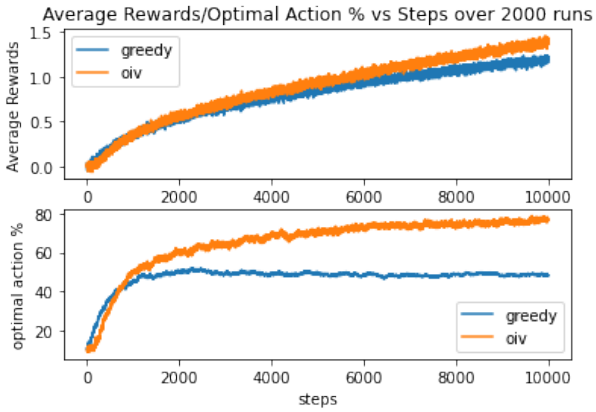
6.2 Non-stationary



In the unstationary enviorment we see that there are smaller differences between the action selection methods. We see that the OIV and UCB action selection methods are performing the best. Furthermore, the three action selection methods seem to be attaining similar optimal action percentages.







Between the greedy and ϵ-greedy methods their average rewards are similar but we see that ϵ-greedy method picks more optimal actions as evinced by the graph. There is some similarity between the oiv and ucb trends. Both their average rewards and optimal action percentages are quite similar and grow at similar rates as well.

7 Conclusion

There are bigger differences in performance between the action selection methods in the stationary enviorment compared to the non-stationary one. While I expected the ucb action selection method to be significantly better than other methods the data shows that oiv is also a valid method in a non-stationary setting. One thing we can be sure of is that sample average methods such as the greedy method is inadequate for a non-stationary setting. Its optimal action percentage falls far behind the 3 other methods. While ucb has shown itself to be adequate perhaps tweaking some constants and the enviorment may lead to more obvious differences. For example, if the standard

deviation of the random walks were higher ucb would be able to make better use of its recency component in the action selection process.