Subject: Reinforcement Learning
Experiment 2
The Upper Confidence Bound (UCB) Bandit Algorithm

Name: Kresha Shah
SAP ID: 60009220080

AIM:
To solve the Bandit problem using the Upper Confidence Bound (UCB1) algorithm

THEORY:

A highly effective multi-armed bandit strategy is the Upper Confidence Bounds (UCB1) strategy. Rather than performing exploration by simply selecting an arbitrary action, chosen with a probability that remains constant, the UCB algorithm changes its exploration-exploitation balance as it gathers more knowledge of the environment.

Using the UCB1 strategy, we select the next action using the following:

$$A_t = \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$$

Where;

Q (a) is the estimated value of action ‘a’ at time step ‘t’.

N (a) is the number of times that action ‘a’ has been selected, prior to time ‘t’. ‘c’ is

a confidence value that controls the level of exploration. **Exploitation:**

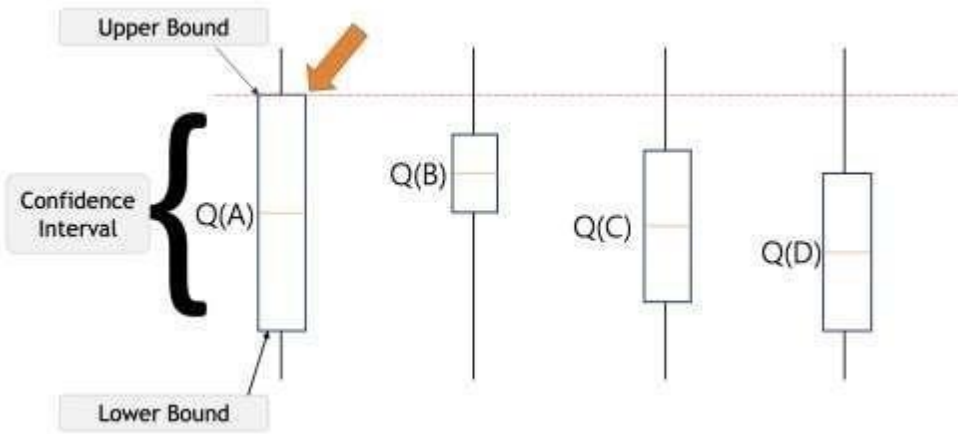
- Q (a) represents the exploitation part of the equation. UCB is based on the principle of “optimism in the fact of uncertainty”, which basically means if you don’t know which action is best then choose the one that currently looks to be the best. Taking this half of the equation by itself will do exactly that: the action that currently has the highest estimated reward will be the chosen action.

Exploration:

- The second half of the equation adds exploration, with the degree of exploration being controlled by the hyper-parameter ‘c’. Effectively this part of the equation provides a measure of the uncertainty for the action’s reward estimate.

- If an action has not been tried very often, or not at all, then N (a) will be small. Consequently, the uncertainty term will be large, making this action more likely to be selected. Every time an action is taken, we become more confident about its estimate. In this case N (a) increments, and so the uncertainty term decreases, making it less likely that this action will be selected as a result of exploration (although it may still be selected as the action with the highest value, due to the exploitation term).

  For example, let us say we have these four actions with associated uncertainties in the picture below, our agent has no idea which is the best action. So according to the UCB algorithm, it will optimistically pick the action that has the highest upper bound i.e., A. By doing this either it will have the highest value and get the highest reward, or by taking that we will get to learn about an action we know least about.



ALGORITHM:

Input: N arms, number of rounds T ≥ N

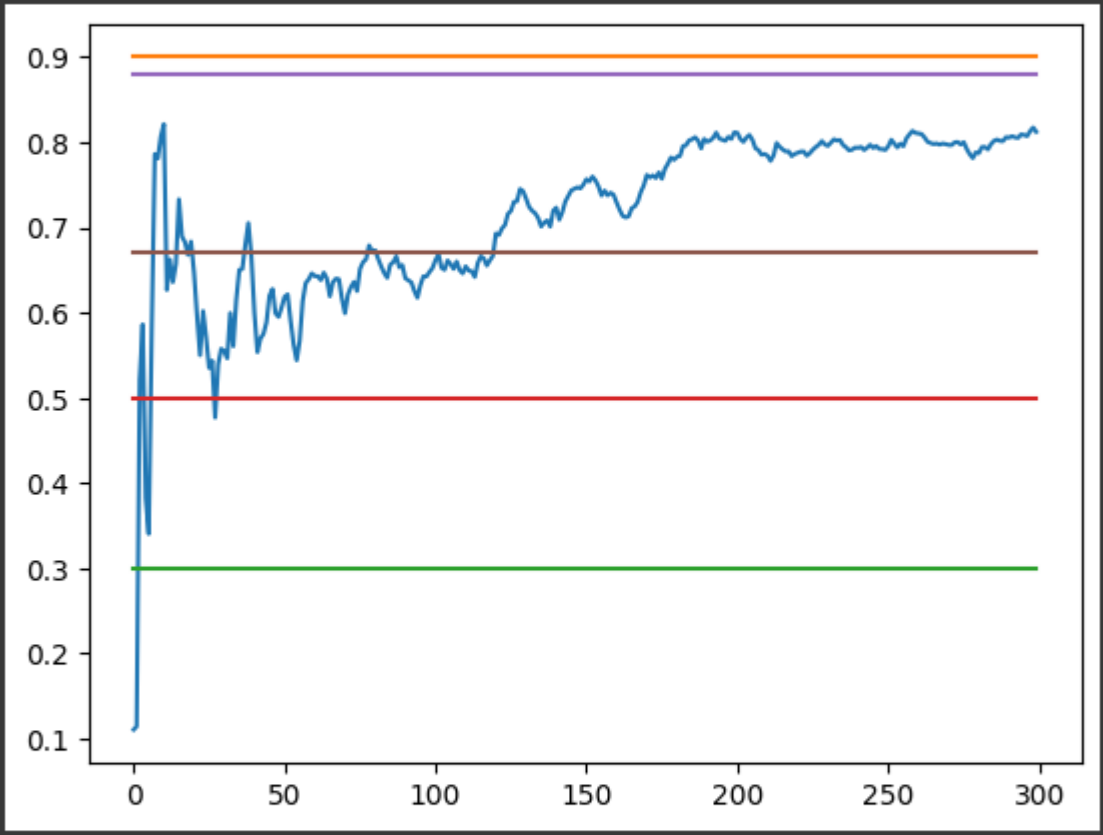
1. For t = 1 . . . N, play arm t.
2. For t = N + 1 . . . T, play arm

$$I_t = \underset{i \in \{1 \ldots N\}}{\arg\max} \, \mathrm{UCB}_{i,t-1}.$$

LAB ASSIGNMENT TO DO:

1. For ‘n’ arms, implement the UCB1 algorithm and calculate the overall reward.

```
Enter the number of arms: 5
Enter mean for arm 1: 0.9
Enter mean for arm 2: 0.3
Enter mean for arm 3: 0.5
Enter mean for arm 4: 0.88
Enter mean for arm 5: 0.67
Enter the value of c: 3
Enter the number of iterations: 300
```

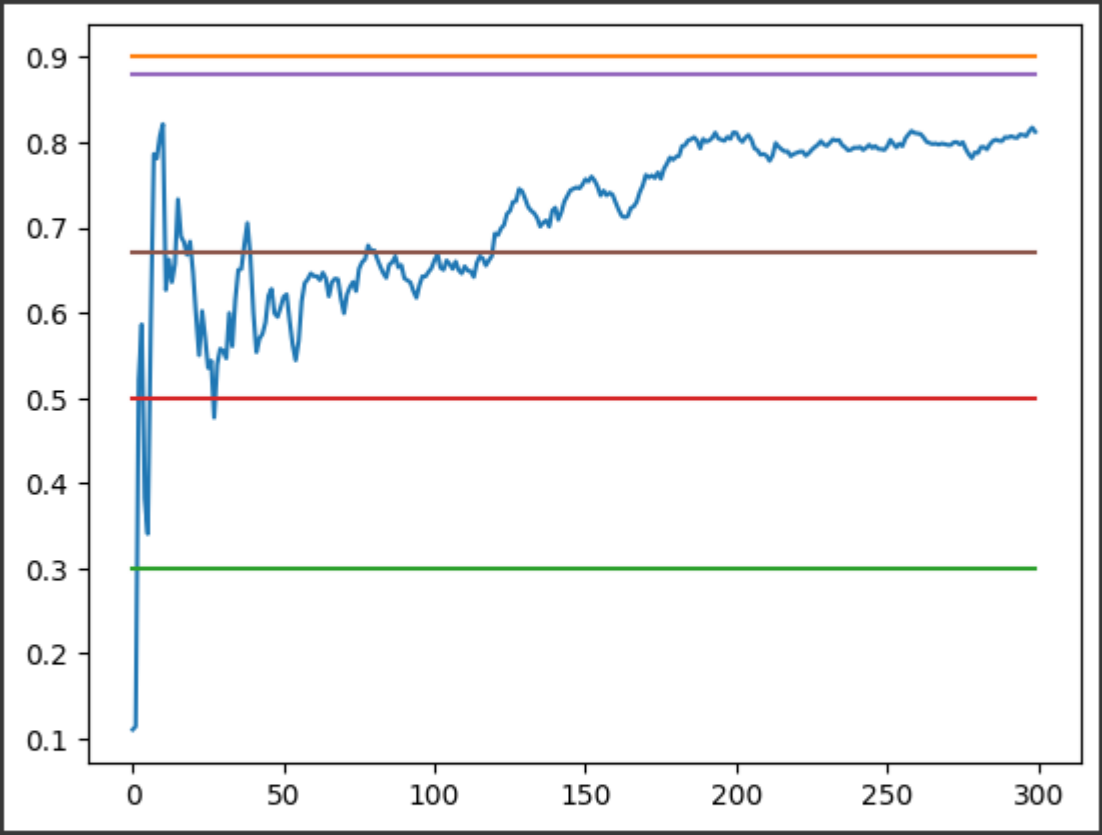



```
Number of times arm 1 selected: 64
Mean of rewards from arm 1: 0.8209249271173806
Number of times arm 2 selected: 29
Mean of rewards from arm 2: 0.34463732264878566
Number of times arm 3 selected: 35
Mean of rewards from arm 3: 0.4990639170192255
Number of times arm 4 selected: 128
Mean of rewards from arm 4: 1.0734476171484866
Number of times arm 5 selected: 44
Mean of rewards from arm 5: 0.5982799773852967
```

2. Plot a graph of the UCB values for ‘n’ arms across k time steps and record the observations.

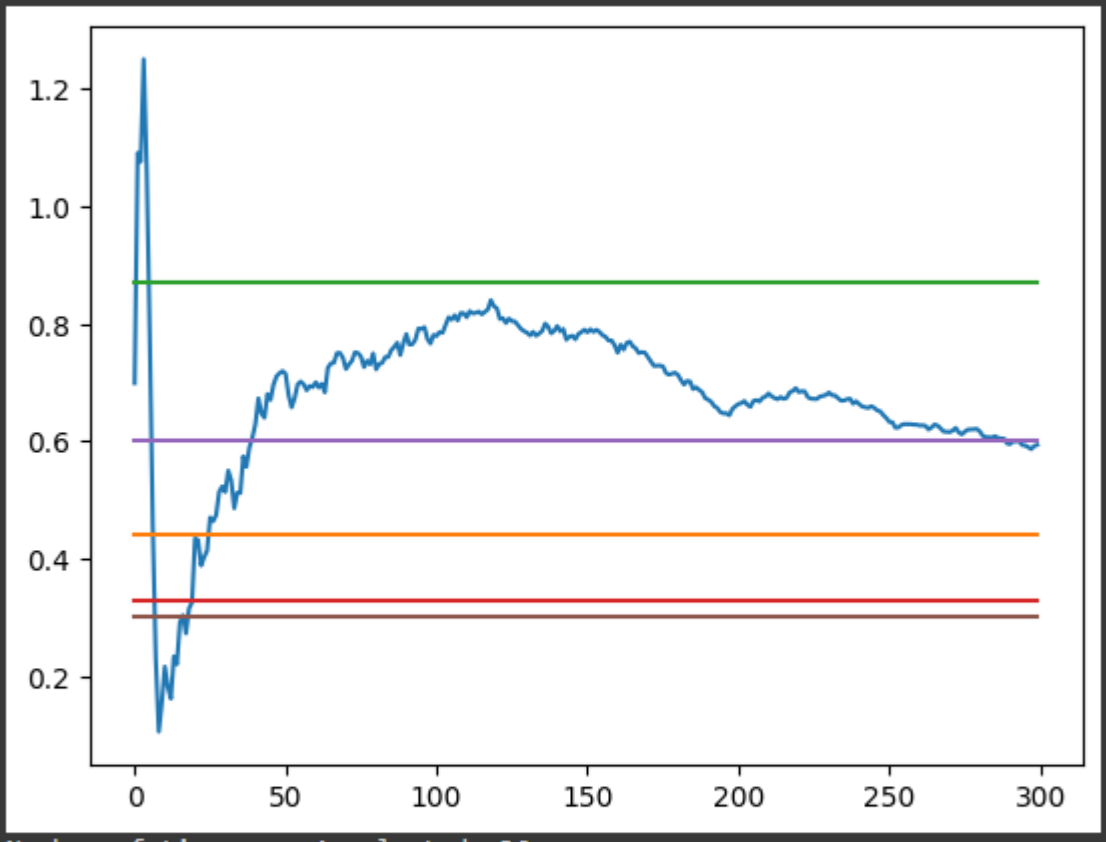
Same c value:

```
Enter the number of arms: 5
Enter mean for arm 1: 0.9
Enter mean for arm 2: 0.3
Enter mean for arm 3: 0.5
Enter mean for arm 4: 0.88
Enter mean for arm 5: 0.67
Enter the value of c: 3
Enter the number of iterations: 300
```




```
Number of times arm 1 selected: 64
Mean of rewards from arm 1: 0.8209249271173806
Number of times arm 2 selected: 29
Mean of rewards from arm 2: 0.34463732264878566
Number of times arm 3 selected: 35
Mean of rewards from arm 3: 0.4990639170192255
Number of times arm 4 selected: 128
Mean of rewards from arm 4: 1.0734476171484866
Number of times arm 5 selected: 44
Mean of rewards from arm 5: 0.5982799773852967
```

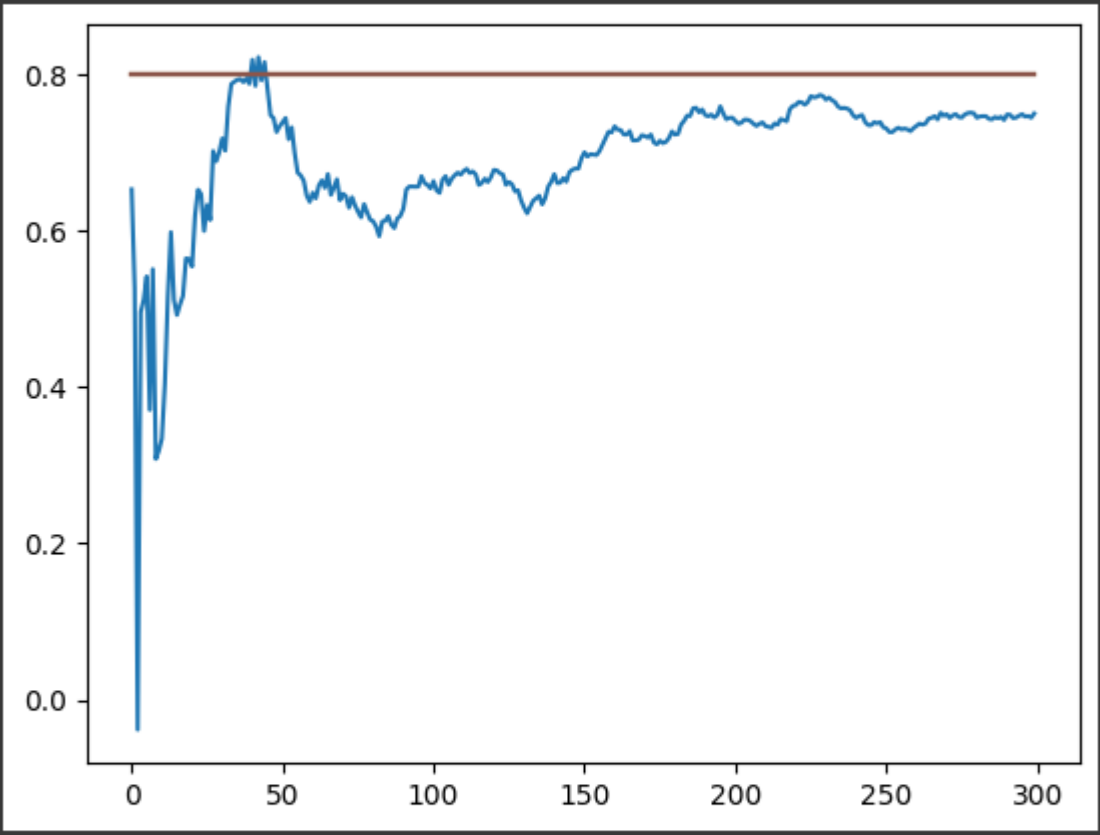
```
Enter the number of arms: 5
Enter mean for arm 1: 0.44
Enter mean for arm 2: 0.87
Enter mean for arm 3: 0.33
Enter mean for arm 4: 0.6
Enter mean for arm 5: 0.3
Enter the value of c: 3
Enter the number of iterations: 300
```




```
Number of times arm 1 selected: 36
Mean of rewards from arm 1: 0.29566272229272195
Number of times arm 2 selected: 109
Mean of rewards from arm 2: 0.8235145228781544
Number of times arm 3 selected: 39
Mean of rewards from arm 3: 0.3239241949923681
Number of times arm 4 selected: 78
Mean of rewards from arm 4: 0.6803610551065326
Number of times arm 5 selected: 38
Mean of rewards from arm 5: 0.31643317125716186
```

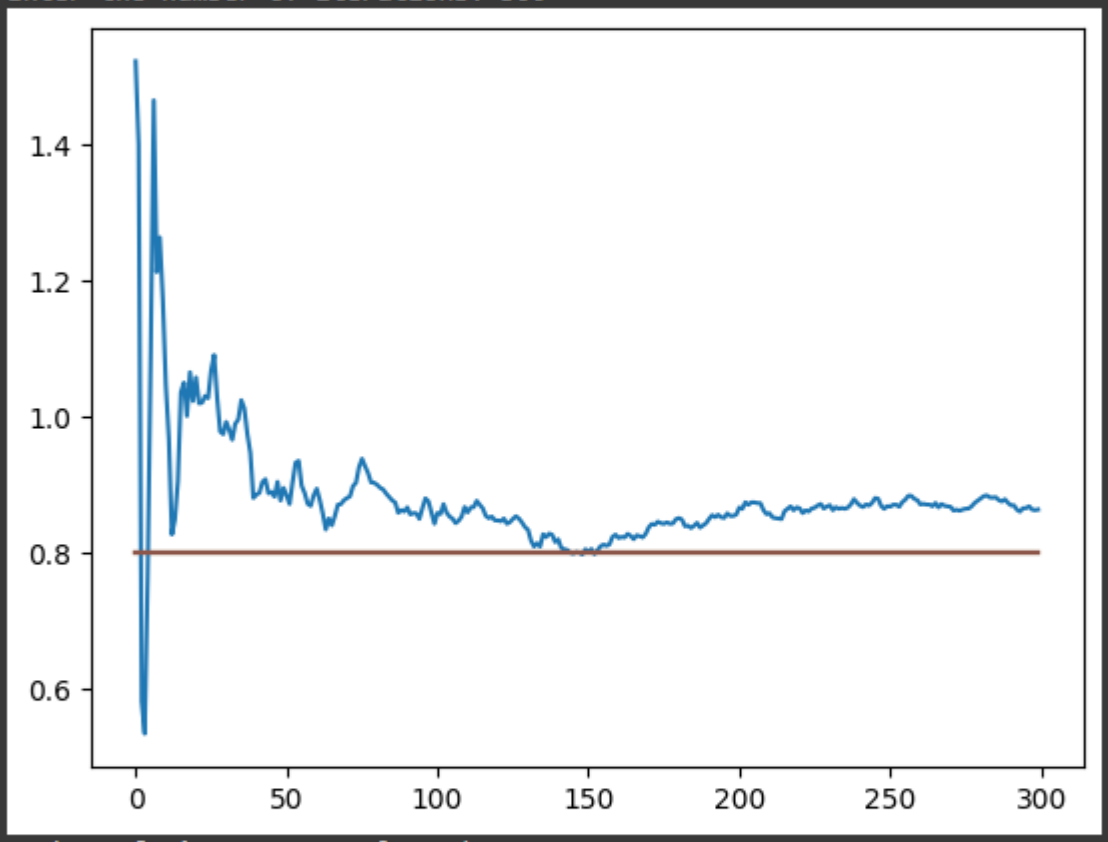
Same mean value:

```
Enter the number of arms: 5
Enter mean for arm 1: 0.8
Enter mean for arm 2: 0.8
Enter mean for arm 3: 0.8
Enter mean for arm 4: 0.8
Enter mean for arm 5: 0.8
Enter the value of c: 5
Enter the number of iterations: 300
```




```
Number of times arm 1 selected: 42
Mean of rewards from arm 1: 0.4243568447873006
Number of times arm 2 selected: 63
Mean of rewards from arm 2: 0.7813528959836441
Number of times arm 3 selected: 73
Mean of rewards from arm 3: 0.8920954659684889
Number of times arm 4 selected: 63
Mean of rewards from arm 4: 0.8000403163228853
Number of times arm 5 selected: 59
Mean of rewards from arm 5: 0.7180840790127199
```

```
Enter the number of arms: 5
Enter mean for arm 1: 0.8
Enter mean for arm 2: 0.8
Enter mean for arm 3: 0.8
Enter mean for arm 4: 0.8
Enter mean for arm 5: 0.8
Enter the value of c: 3
Enter the number of iterations: 300
```




```
Number of times arm 1 selected: 70
Mean of rewards from arm 1: 0.9331454041178588
Number of times arm 2 selected: 48
Mean of rewards from arm 2: 0.7491247202372139
Number of times arm 3 selected: 52
Mean of rewards from arm 3: 0.772307592871107
Number of times arm 4 selected: 57
Mean of rewards from arm 4: 0.8429203246497415
Number of times arm 5 selected: 73
Mean of rewards from arm 5: 0.9522235655944559
```

Colab link: https://colab.research.google.com/drive/1m3tH6XY02f6bLJEQmPtJyAG1ITNNUc_6?usp=sharing