Experiment No. 2
Implementing a multi-armed bandit problem using UCB.
Date of Performance:
Date of Submission:

Aim: Implementing a multi-armed bandit problem using UCB.

Objective: objective of UCB is to learn how the algorithm balances exploration by choosing arms with uncertain outcomes (based on confidence intervals) and exploitation by favoring arms with high estimated rewards.

Theory:

The Multi-Armed Bandit Problem (MAB) is a fundamental problem in the field of reinforcement learning and decision-making under uncertainty. The problem involves a gambler who has to choose among several slot machines (also called "one-armed bandits") with unknown payout probabilities. The gambler's objective is to maximize his or her total payout by choosing the best slot machine to play.
The MAB problem is commonly used in various fields, such as clinical trials, online advertising, and recommendation systems. The MAB problem can be seen as a trade-off between exploration and exploitation. Exploration refers to the gambler trying out different slot machines to learn about their payout probabilities, while exploitation refers to the gambler sticking to the slot machine that has shown the highest payout probability so far.

A bandit is defined as someone who steals your money. A one-armed bandit is a simple slot machine wherein you insert a coin into the machine, pull a lever, and get an immediate reward. But why is it called a bandit? It turns out all casinos configure these slot machines in such a way that all gamblers end up losing money!

A multi-armed bandit is a complicated slot machine wherein instead of 1, there are several levers that a gambler can pull, with each lever giving a different return. The probability distribution for the reward corresponding to each lever is different and is unknown to the gambler.

Upper Confidence Bounds(UCB)

Upper Confidence Bounds (UCB) is a popular algorithm for solving the multi-armed bandit problem. It is similar to the ε-greedy algorithm but uses a different strategy for balancing exploration and exploitation.

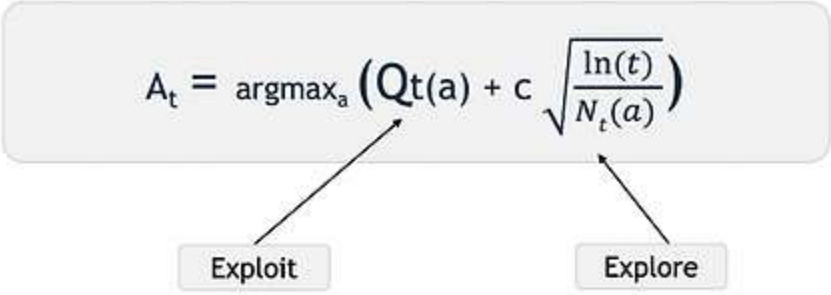
In UCB, the algorithm selects the arm with the highest upper confidence bound at each time step. The upper confidence bound is a measure of how uncertain the algorithm is about the reward distribution for that arm. It is calculated as follows:

UCB = estimated mean reward + exploration bonus

The exploration bonus is a function of the number of times the arm has been selected and the total number of times all arms have been selected. It is designed to encourage exploration of arms that have been selected less frequently, while still selecting arms with high estimated rewards.

The UCB algorithm starts by selecting each arm once to establish initial estimates of the reward distribution. It then selects the arm with the highest upper confidence bound at each subsequent time step. As more data is collected, the estimates of the reward distribution become more accurate and the algorithm becomes more confident in its arm selection.

With UCB, '*A*$_t$*'*, the action chosen at time step '*t*', is given by:



$$A_t = \text{argmax}_a \left(Q_t(a) + c \sqrt{\frac{\ln(t)}{N_t(a)}} \right)$$

where;

- *$Q_t(a)$ is the estimated value of action 'a' at time step 't'.*
- *$N_t(a)$ 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.*

One advantage of UCB over the ε-greedy algorithm is that it does not require a parameter to specify the level of exploration. The algorithm automatically adjusts the level of exploration based on the uncertainty of the reward distribution for each arm.

In summary, Upper Confidence Bounds (UCB) is an algorithm for solving the multi-armed bandit problem

that balances exploration and exploitation by selecting the arm with the highest upper confidence bound at each time step. It is an effective algorithm for solving the problem and does not require a parameter to specify the level of exploration.

Conclusion:

1. Explain Example of Multi-arm Bandit Problem.
2. Compare ε-greedy and UCB.