**Vivekanand Education Society’s Institute of Technology**

**Department of AIDS Engineering**



**Subject: Reinforcement Learning**

**Class: D16AD**

| **ROLL NO: 30** | **NAME: SUHANEE KANDALKAR** | | |
| --- | --- | --- | --- |
| **EXP NO: 05** | T**ITLE: UCB and Optimal Initial Value** | | |
| **DOP:** |  | **DOS: 08/03/25** |  |
| **GRADES:** | **LOs MAPPED:** | | **SIGNATURE:** |

## 

## Aim**:**

Experimenting with Upper Confidence Bound and Optimistic Initialization strategy and analyzing its impact on the learning performance of an agent.

## Theory**:**

Reinforcement Learning (RL) involves an agent that interacts with an environment to maximize cumulative rewards over time. Two popular exploration strategies used in RL are:

* Upper Confidence Bound (UCB)
* Optimistic Initialization

These methods help balance the exploration-exploitation trade-off—a fundamental challenge in RL.

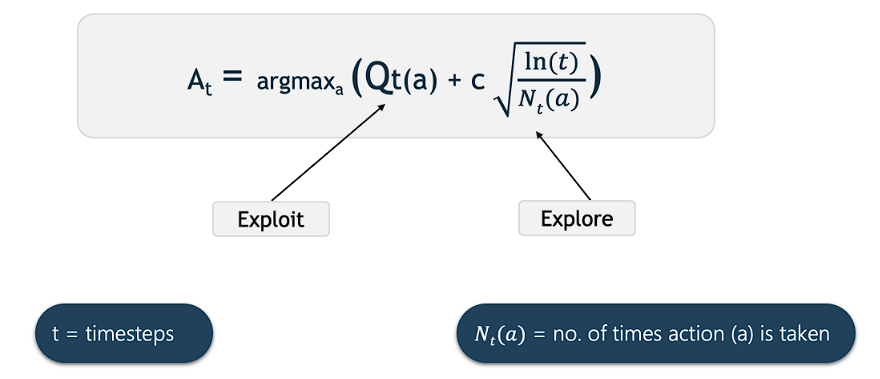
**Exploration:** It allows the agent to improve its knowledge about each action. Hopefully, leading to a long-term benefit.

**Exploitation:** It allows the agent to choose the greedy action to try to get the most reward for short-term benefit. A pure greedy action selection can lead to suboptimal behaviour.

### 1. Upper Confidence Bound (UCB)

Upper-Confidence Bound action selection uses uncertainty in the action-value estimates for balancing exploration and exploitation. Since there is inherent uncertainty in the accuracy of the action-value estimates when we use a sampled set of rewards, UCB uses uncertainty in the estimates to drive exploration.

The Upper Confidence Bound follows the principle of optimism in the face of uncertainty which implies that if we are uncertain about an action, we should optimistically assume that it is the correct action.



Where:

* Qt(a) is the estimated reward for action 𝑎.
* Nt(a) is the number of times action
* 𝑎 has been chosen.
* 𝑡 is the total number of steps taken.
* 𝑐 is a confidence parameter controlling exploration.

The UCB strategy ensures that actions with higher uncertainty (less explored) get sampled more frequently.

### 2. Optimistic Initialization

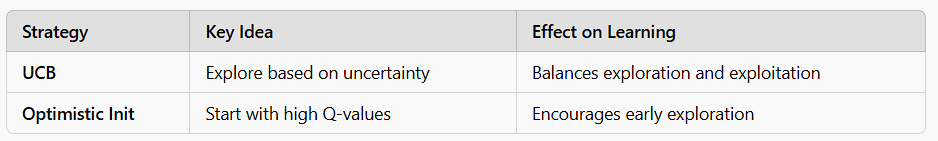
Optimistic Initialization is another exploration strategy where the agent starts with high initial values for action-value estimates. This encourages the agent to explore initially rather than exploiting suboptimal actions.



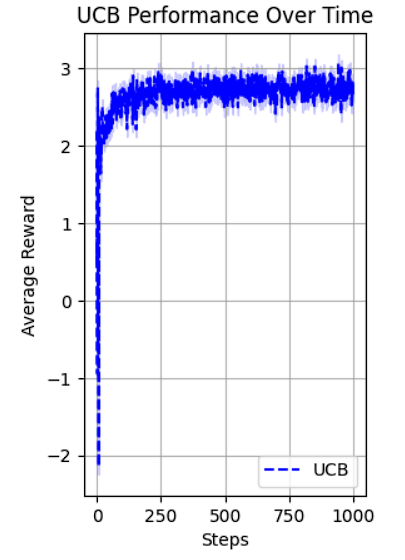
Where:

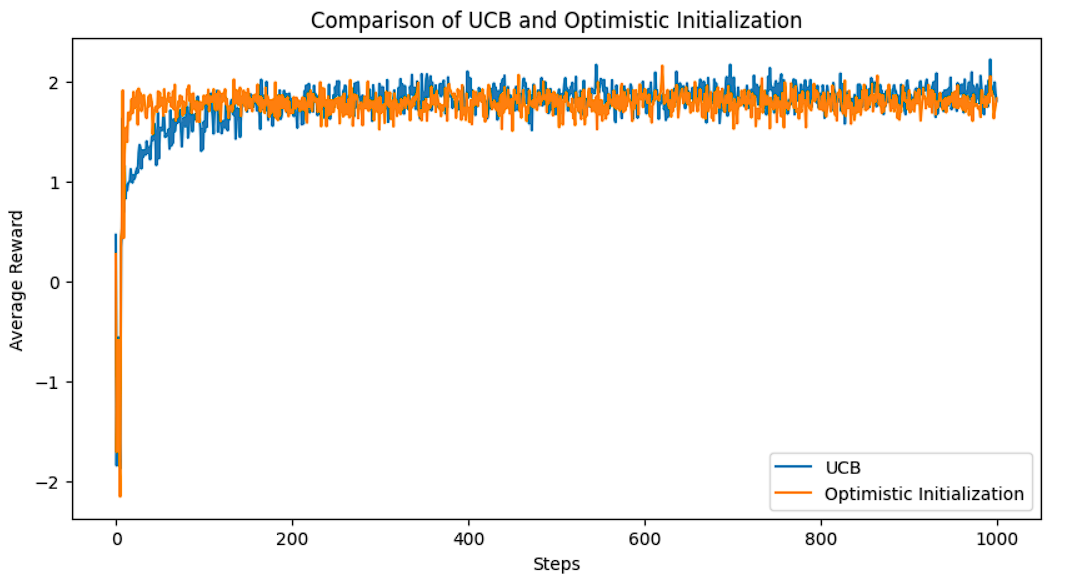
* 𝑄0(a) is the initial Q-value of action 𝑎
* 𝑄init is set to a high value (e.g., 1 or 10) to encourage exploration.

Since early rewards are usually lower, the agent will explore different actions until it finds a better-performing one.



# Code:





[**https://colab.research.google.com/drive/1cZ8gUgtnf9HH5lopbqdfSh2UXhwx0ww\_?usp=sharing**](https://colab.research.google.com/drive/1cZ8gUgtnf9HH5lopbqdfSh2UXhwx0ww_?usp=sharing)

# Conclusion :

The experiment showed that UCB balances exploration and exploitation, leading to steady performance improvement. Optimistic Initialization encourages early exploration but can be biased toward initial assumptions. UCB adapts better over time, making it more effective for long-term learning. Optimistic Initialization works well initially but may struggle with dynamic environments.