**CSC 580 AI II| HW 1-2** | **Name: Om Prakash Gunja** | **Student ID: 2131025**

**Analysis of Epsilon-Greedy Algorithm Results**

## Cell (3): Run the Algorithm with Various Epsilons

1. **Differences in Q-values**:

* **Epsilon = 0.0**: Q values are heavily influenced by greedy reward approach essentially no exploration but the random selection of max values when there were more than 1 same q values made initial values to pick random action cause all of them are [0,0,0,0] but from the N values we can see how the action 1 being selected a lot compared to others

Q = [0.0, 0.0, 12.4039, 0.0]

N = [429.734, 58.294, 251.0225, 260.9495]

* **Epsilon = 0.01**: A small degree of exploration allowed slightly, Action 3 started to get picked up many times compared to previous iteration because of reward and epsilon

Q = [8.0, 10.7143, 9.3398, 10.9265]

N = [195.6405, 82.367, 396.3965, 325.596]

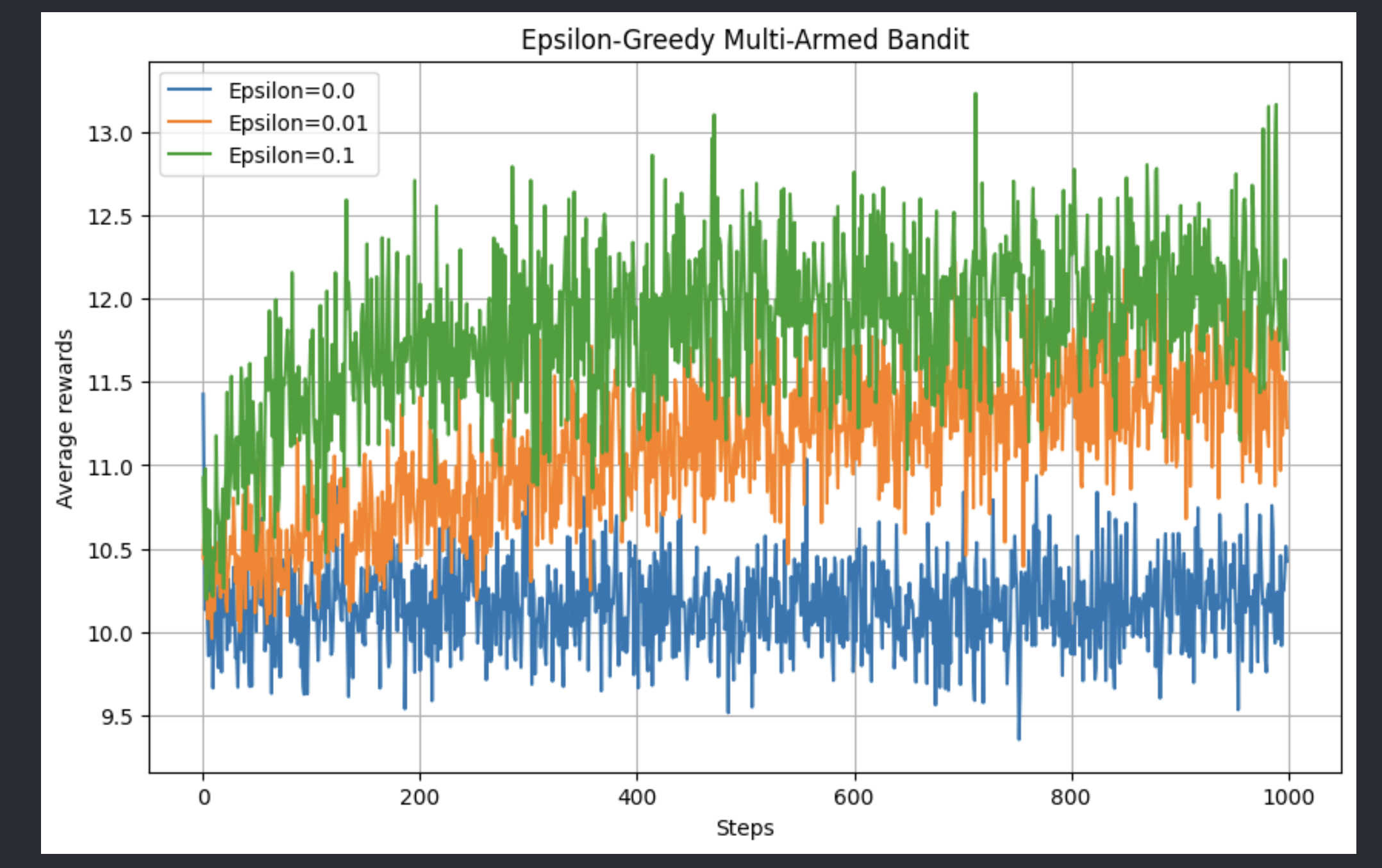
* **Epsilon = 0.1**: Higher exploration significantly diversified the actions taken, allowing a more robust evaluation of each action's actual reward distribution.

Q = [8.0, 10.0, 12.9023, 10.0777]

N = [52.7735, 168.704, 534.7975, 243.725]

1. **Differences in Action Selection Frequencies (N):**

* Lower epsilon values favored actions that seemed best since the start, causing a focus on less exploration
* Higher epsilon values spread out action choice more, which helped with learning over time but made it take longer to settle on the best action.



## Cell (4): Plot of Average Reward

**Observations:**

The plot shows average reward fluctuating across time, At low epsilon values, the blue line is wide spread due to both Epson and picking random max and the fluctuations are dues to exploration picking various rewarded actions. Continues picking of variety of actions made the lines more fluctuations

Green line is more fluctuated compared to orange which is more fluctuated than blue primarily because of epsilon making the algo to explore other options

But the high exploration rate made the algorithm to explore action 3 more which is optimal action for higher reward lower exploration ( e = 0.0 ) made the algo explore less options sticking to the few options based on initial steps

## Cell (5): Plot of Optimal Actionpasted-movie.png

**Observations:**

The more often the best choice is made, the more this improves over time, but it depends on epsilon:

Low epsilon: Chooses what seems best at first, but it might not actually be the best.

High epsilon: Starts slowly in choosing the best, but eventually gets better as it learns more.High epsilon helps explore a lot, so it finds the best choice in the long run, but it doesn't perform as well in the short term.

## Cell (6): Initial Q-values

The Q-values were initialized with [10, 0, 0, 0] for this step. The following results were obtained:

1. **Differences in Q-values**:

* **Epsilon = 0.0**: Exploitation makes the agent to always select the first action, cause the initiated q has 10 at the first place making it the only option to explore as epsilon is low ( no exploration )

Q = [8.0, 0.0, 0.0, 0.0]

N = [1000.0, 0.0, 0.0, 0.0]

* **Epsilon = 0.01**: Slight exploration made the agent to get better rewards for Action 3 while still picking Action 1 initially:

Q = [8.0, 0.0, 12.3683, 10.2]

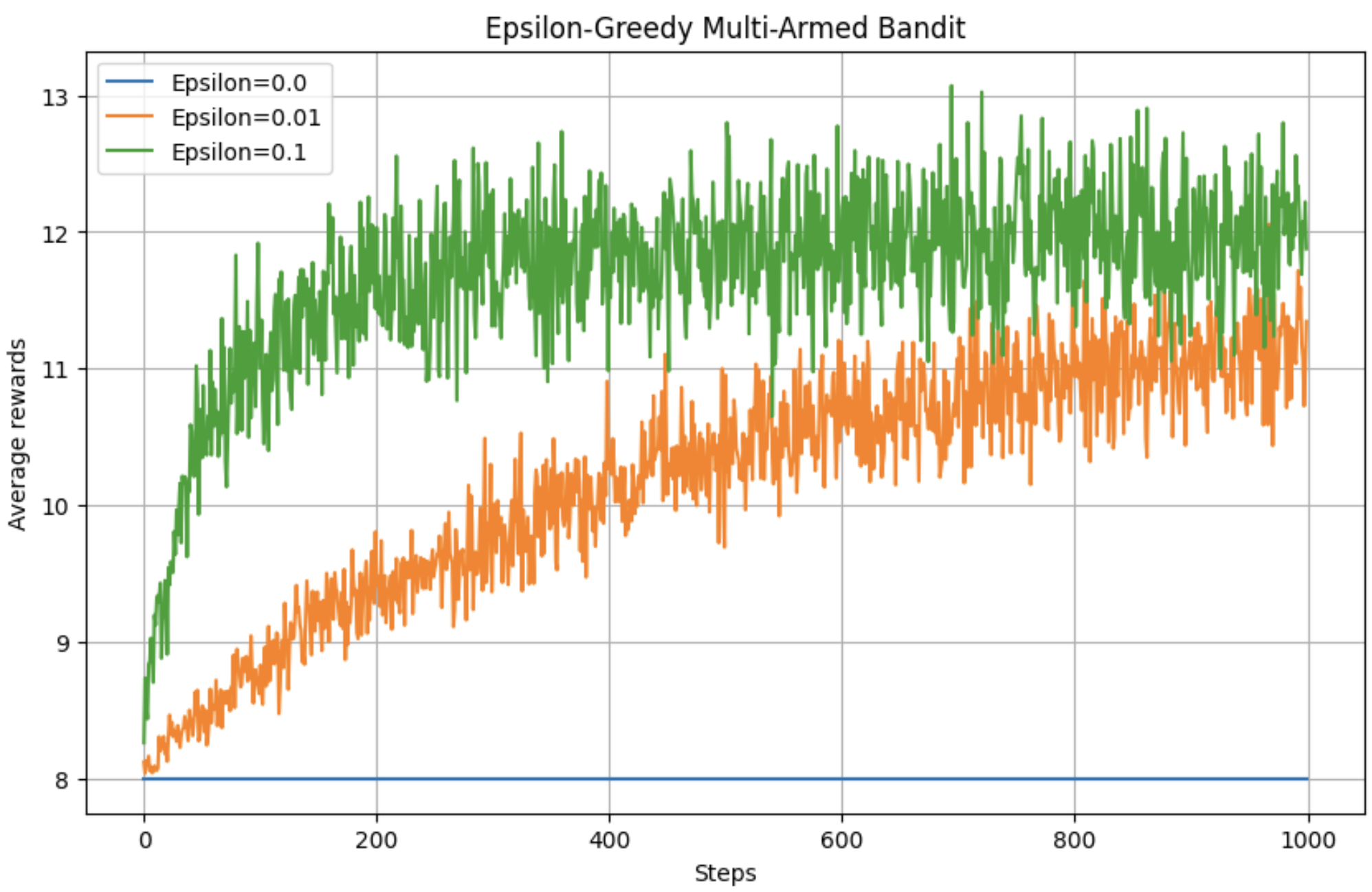
N = [436.818, 64.4155, 274.7955, 223.971]

* **Epsilon = 0.1**: Higher exploration significantly favored Action 3, with additional attention given to other actions:

Q = [8.0, 4.7619, 12.6384, 10.3953]

N = [81.3015, 181.322, 489.885, 247.4915]

## Additional Average Rewards Plot (Initialized Q-values)



**Observations:**

The plot of average rewards when Q-values are initialized with [10, 0, 0, 0] shows the following trends:

* **Epsilon = 0.0**: The average reward stabilizes at the constant value of Action 1 (8), as there is no exploration.
* **Epsilon = 0.01**: The rewards gradually increase over time as limited exploration allows the agent to discover higher reward actions.
* **Epsilon = 0.1**: Higher exploration results in faster and higher convergence of average rewards, aligning with the agent's ability to effectively evaluate all actions.

**Comparison with Expected Results:**

The expected plot assumes that the agent would initially explore and learn faster across actions, leading to a smoother convergence. However, in my case:

* The initialized Q-values had a big impact on early choices, especially when epsilon was low. This made it harder to find the best actions.
* But higher epsilon values helped fix this, green and orange seems to be converging at the end as observed which indicated the moderate exploration took some steps to reach the optimal action picking but the higher exploration made the agent to discover the optimal action in less steps

Reflections:

1. I learned how much the exploration and exploitation matters in agents overall outcomes and how important it is to balance them while execution
2. The algorithm building took a lot of time due to variety of outputs and graphs
3. Next time I would write down the basic terms of the algorithm and implement efficiently