Analysis of Epsilon-Greedy and UCB Algorithms

Experiment Setup

The multi-armed bandit problem was set up with 5 arms. The true mean rewards for these arms were defined as `[1.0, 1.5, 1.2, 0.8, 1.1]`. The algorithm was run for 1000 steps in each experiment. For simplicity, the actual rewards were sampled from a normal distribution with the specified mean and a standard deviation of 1.

Epsilon-Greedy Analysis

The Epsilon-Greedy algorithm was tested with different values of epsilon (ε), the probability of exploration. The following epsilon values were used: 0.0, 0.01, 0.1, and 0.5. The total reward obtained for each epsilon value is shown below:

- Epsilon = 0 : Total Reward = 1038.14
- Epsilon = 0.01: Total Reward = 1452.19
- Epsilon = 0.1: Total Reward = 1430.86
- Epsilon = 0.5: Total Reward = 1308.89

When epsilon is 0 the algorithm is purely greedy, always choosing the arm with the highest estimated value. This approach quickly converges on an arm, but if the initial estimates are inaccurate, it may get stuck exploiting a suboptimal arm, as seen in the relatively lower total reward. With a small amount of exploration (epsilon = 0.01), the algorithm has a chance to discover better arms, leading to a significantly higher total reward. Increasing epsilon further (0.1 and 0.5) increases exploration. While exploration is necessary to find the optimal arm, excessive exploration can lead to pulling suboptimal arms more often, thus reducing the total reward. The results suggest that for this specific problem and number of steps, a small amount of exploration (epsilon = 0.01 or 0.1) yields better results than no exploration or excessive exploration.

UCB Algorithm Analysis

The UCB algorithm was tested with different values of the exploration parameter 'c'. The 'c' parameter controls the balance between exploitation (based on the estimated mean reward) and exploration (based on the uncertainty of the estimate). The following 'c' values were used: 0.1, 0.5, 1.0, and 2.0. The total reward obtained for each 'c' value is shown below:

- UCB with c = 0.1: Total Reward = 1427.95
- UCB with c = 0.5: Total Reward = 1503.87
- UCB with c = 1.0: Total Reward = 1446.13
- UCB with c = 2.0: Total Reward = 1427.40

The UCB algorithm inherently balances exploration and exploitation by selecting the arm that maximizes the sum of its estimated value and a confidence bound. A higher 'c' value increases the influence of the confidence bound, leading to more exploration. Similar to epsilon-greedy, there appears to be an optimal range for the exploration parameter. In this experiment, a 'c' value of 0.5 resulted in the highest total reward, indicating a good balance between exploring arms with uncertain estimates and exploiting arms with high estimated values. Lower or higher 'c' values led to slightly lower total rewards, suggesting either insufficient or excessive exploration, respectively.

Thompson Sampling Analysis

The Thompson Sampling algorithm was implemented for a Bernoulli bandit problem. The true success probabilities for the 5 arms were set to `[0.3, 0.7, 0.5, 0.2, 0.6]`. The algorithm was run for 1000 steps. The total reward obtained was 695.00.

Thompson Sampling is a Bayesian approach that maintains a probability distribution over the likely reward for each arm. For Bernoulli bandits, the Beta distribution is a common choice as it is the conjugate prior to the Bernoulli likelihood. The algorithm samples from these distributions in each step and chooses the arm with the highest sampled value. This naturally balances exploration and exploitation: arms with higher uncertainty (fewer pulls) or higher estimated success probabilities are more likely to be sampled as the best, leading to exploration, while arms with consistently high rewards and low uncertainty are more likely to be chosen for exploitation.

The total reward of 695.00 over 1000 steps, with a maximum possible reward of 1 per step, indicates that the algorithm effectively identified and exploited the arm with the highest true success probability (arm 1 with probability 0.7). The accumulated reward is close to the maximum possible reward (700 if only the best arm was pulled), suggesting efficient learning and convergence towards the optimal policy.

Preliminary Comparison (Epsilon-Greedy, UCB, and Thompson Sampling)

Based on the initial experiments with 1000 steps:

- **Epsilon-Greedy:** Best total reward around 1452.19 (with epsilon 0.01 or 0.1). This was for a bandit with rewards sampled from a normal distribution.
- **UCB:** Best total reward around 1503.87 (with c=0.5). Also for a bandit with rewards sampled from a normal distribution.
- **Thompson Sampling:** Total reward of 695.00. This was for a Bernoulli bandit (rewards 0 or 1).

Directly comparing the total rewards across these experiments is not entirely fair due to the different reward distributions (Normal vs. Bernoulli). However, we can observe the relative

performance within their respective problem settings. Both UCB and Epsilon-Greedy showed sensitivity to their exploration parameters. Thompson Sampling, on the other hand, adaptively balances exploration and exploitation based on the uncertainty in the reward estimates, without requiring a manually tuned exploration parameter like epsilon or c. This adaptiveness is a key advantage of Thompson Sampling, particularly in scenarios where the optimal exploration rate is unknown or changes over time.

Comparative Analysis of Epsilon-Greedy, UCB, and Thompson Sampling on Bernoulli Bandits

To provide a fair comparison, all three algorithms (Epsilon-Greedy, UCB, and Thompson Sampling) were evaluated on the same Bernoulli bandit problem with 5 arms and true success probabilities `[0.3, 0.7, 0.5, 0.2, 0.6]`. Each algorithm was run for 1000 steps, and the experiment was repeated for 100 trials to average out the randomness and obtain more reliable performance estimates.

The average total rewards obtained over 100 trials are summarized below:

- **Epsilon-Greedy:**
    - Epsilon = 0.01: Average Total Reward = 538.64
    - Epsilon = 0.1: Average Total Reward = 645.20
    - Epsilon = 0.2: Average Total Reward = 633.30
    - *Best Epsilon-Greedy performance:* Approximately 645.20 (with epsilon = 0.1)

```
True success probabilities: [0.3, 0.7, 0.5, 0.2, 0.6]
Epsilon = 0.01: Average Total Reward over 100 trials = 538.64
Epsilon = 0.1: Average Total Reward over 100 trials = 645.20
Epsilon = 0.2: Average Total Reward over 100 trials = 633.30

[Done] exited with code=0 in 1.126 seconds
```

- **UCB:**
    - UCB with c = 0.1: Average Total Reward = 644.27
    - UCB with c = 0.5: Average Total Reward = 675.03
    - UCB with c = 1.0: Average Total Reward = 644.23
    - UCB with c = 2.0: Average Total Reward = 592.38
    - *Best UCB performance:* Approximately 675.03 (with c = 0.5)

```
True success probabilities: [0.3, 0.7, 0.5, 0.2, 0.6]
UCB with c = 0.1: Average Total Reward over 100 trials = 644.27
UCB with c = 0.5: Average Total Reward over 100 trials = 675.03
UCB with c = 1.0: Average Total Reward over 100 trials = 644.35
UCB with c = 2.0: Average Total Reward over 100 trials = 592.38

[Done] exited with code=0 in 3.115 seconds
```

- **Thompson Sampling:**
    - Average Total Reward over 100 trials = 676.40

```
True success probabilities: [0.3, 0.7, 0.5, 0.2, 0.6]
Thompson Sampling: Average Total Reward over 100 trials = 676.40

[Done] exited with code=0 in 1.357 seconds
```

Based on these results, the UCB algorithm with c=0.5 achieved the highest average total reward, closely followed by Thompson Sampling. The Epsilon-Greedy algorithm, even with tuned epsilon values, performed slightly worse than the best configurations of UCB and Thompson Sampling on this specific problem instance and number of steps.

Total Reward: UCB (with optimal c) and Thompson Sampling generally accumulated higher total rewards compared to Epsilon-Greedy. This indicates that UCB and Thompson Sampling were more effective at identifying and exploiting the optimal arm (arm with probability 0.7) within the given number of steps.

Exploration Efficiency:

- **Epsilon-Greedy:** Explores with a fixed probability epsilon. This means it continues to explore even when it has a good estimate of the optimal arm, which can lead to pulling suboptimal arms unnecessarily in later steps.
- **UCB:** Balances exploration and exploitation by considering the uncertainty in the estimates. Arms that have been pulled fewer times have a higher confidence bound, making them more likely to be explored. As an arm is pulled more often, its confidence bound shrinks, and the algorithm shifts towards exploiting arms with higher estimated values. This approach ensures that all arms are explored sufficiently initially, and exploration naturally decreases over time.
- **Thompson Sampling:** Explores in proportion to the probability that an arm is optimal, based on the current belief distribution. Arms with higher uncertainty or higher estimated probabilities are more likely to be sampled and chosen. This Bayesian approach provides a more intuitive and often more efficient way to balance exploration and

exploitation, as exploration is driven by the need to reduce uncertainty about potentially optimal arms.

Convergence Speed: While not directly measured in these experiments, the higher total rewards achieved by UCB and Thompson Sampling suggest faster convergence to near-optimal performance compared to Epsilon-Greedy, especially in the later steps where the fixed exploration rate of Epsilon-Greedy can be detrimental. UCB and Thompson Sampling are better equipped to focus on the most promising arms as they gather more data.

In summary, for this Bernoulli bandit problem, UCB and Thompson Sampling demonstrated superior performance in terms of total reward compared to Epsilon-Greedy. UCB's confidence bound mechanism and Thompson Sampling's probability matching approach provide more sophisticated and often more effective strategies for balancing exploration and exploitation than the simple fixed-probability exploration of Epsilon-Greedy. The choice between UCB and Thompson Sampling can depend on the specific problem characteristics and prior knowledge, but both are generally strong performers in multi-armed bandit settings.