Reinforcement Learning

Assignment 01

Spring 2024

By

W.H. Sasinda C. Prabhashana-202395458

&

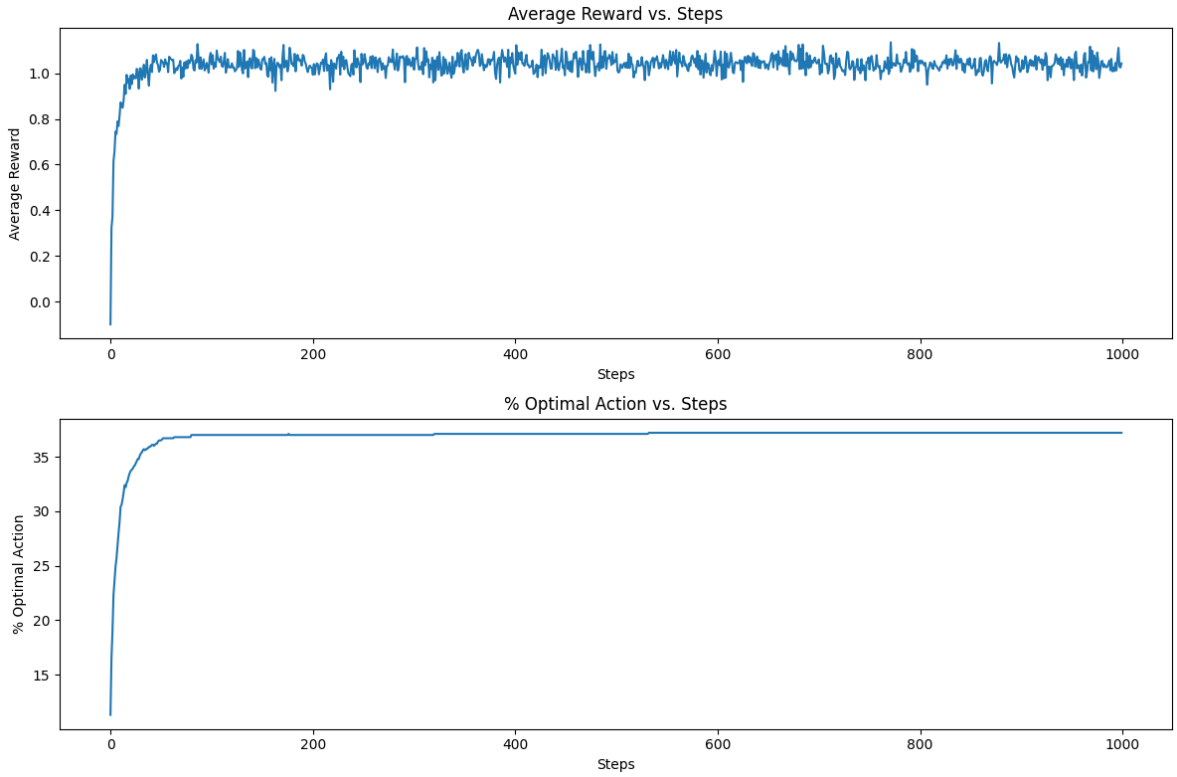
Premisha Premananthan -202397583

[This document fulfills the requirements for the Reinforcement Learning Module (DSCI-6650-001) at Memorial University in Newfoundland, Canada]

Part 01: A simple bandit problem with stationary reward distributions

- **Environment Setup**: 10-armed bandit problems are generated where each arm's reward is normally distributed with a mean drawn from N~(0, 1).
- **Algorithms**: We implemented four different strategies:
  - **Greedy**: Always selects the action with the highest estimated value.
  - **Epsilon-Greedy**: With probability ε, selects a random action; otherwise, selects the best-known action.
  - **Optimistic Greedy**: Initializes action values optimistically to encourage exploration initially.
  - **Gradient Bandit**: Uses preferences and a soft-max distribution to select actions.
- **Simulation**: Each algorithm is run for 1000 bandit problems, each with 1000 steps. The performance is tracked by average reward and percentage of optimal actions.
- **Results Analysis**: The average reward and optimal action percentage over time for each algorithm are plotted.

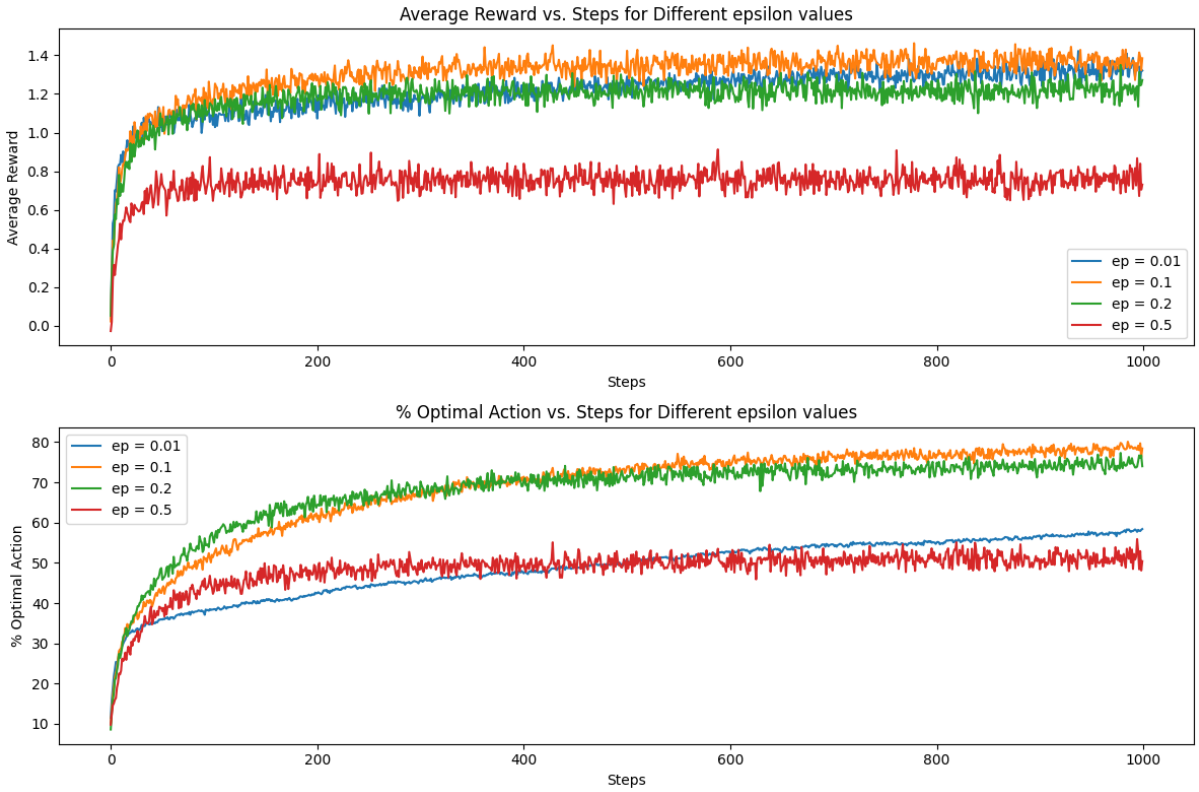
❖ *Greedy with non-optimistic initial values. Initialize the action value estimates to 0 and use the incremental implementation of the simple average method.*



The Average Reward vs. Steps graph illustrates a rapid increase in average reward from near 0 to around 0.9 within the first 100 steps, stabilizing at about 1.0 after 200 steps. This indicates quick learning and effective exploitation of high-reward actions.

The Optimal Action vs. Steps graph reflects an initial 10% optimal action selection rate, rising to around 30% within 100 steps, and stabilizing at approximately 35% after 200 steps. This demonstrates the algorithm's ability to frequently select optimal actions over time. Overall, the Greedy algorithm learns and exploits the best actions effectively, leading to increased rewards and optimal action selection

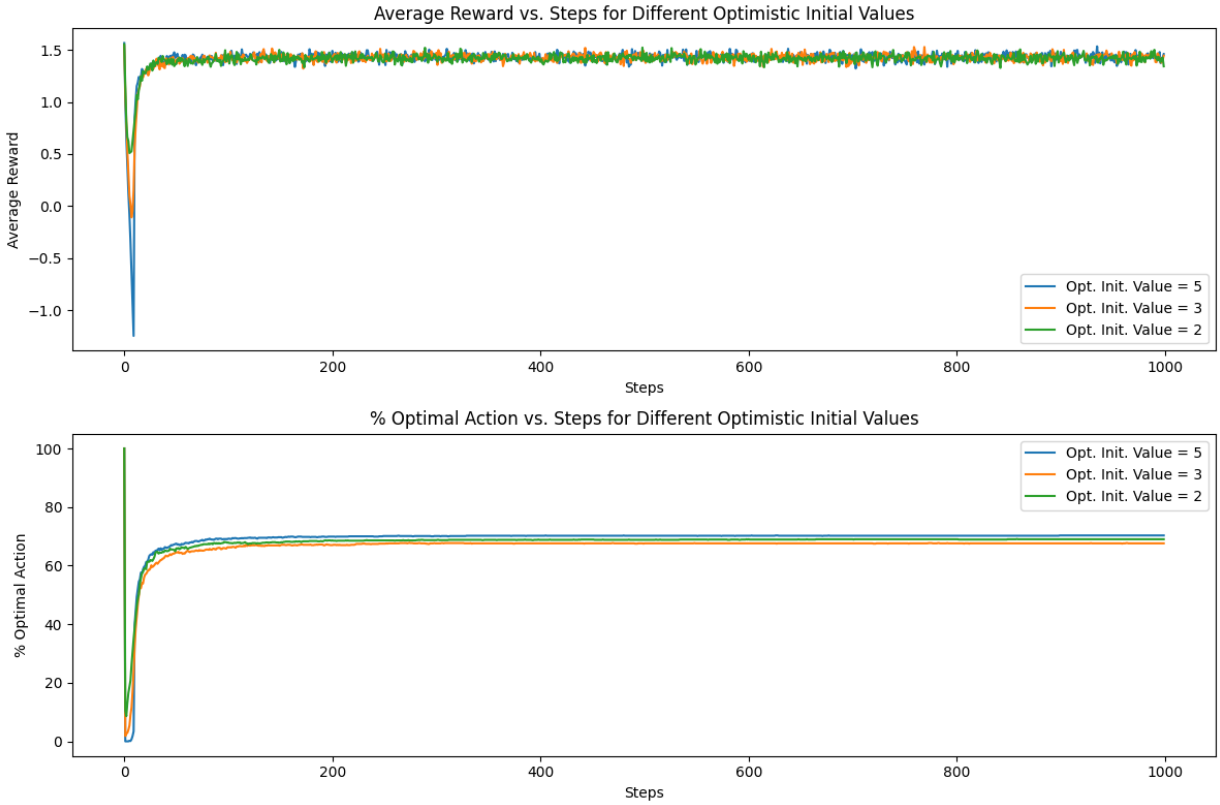
- *Epsilon-greedy with different choices of epsilon.*



The Average Reward vs. Steps graph shows that with an epsilon of 0.1, the algorithm quickly achieves the highest average reward, stabilizing around 1.4, indicating a good balance between exploration and exploitation. For epsilon values of 0.2 and 0.01, the average rewards stabilize around 1.3 and 1.2, respectively, showing effective but slightly less optimal performance. The highest epsilon value of 0.5 results in the lowest average reward, stabilizing around 0.9, due to excessive exploration.

The Optimal Action vs. Steps graph reveals that with an epsilon of 0.1, the percentage of optimal actions rapidly increases and stabilizes around 75%. An epsilon of 0.2 stabilizes around 70%, while 0.01 and 0.5 stabilize around 50% and 40%, respectively. This indicates that lower epsilon values (0.1 and 0.2) result in more frequent optimal action selection, while higher values lead to frequent suboptimal choices. Overall, an epsilon of 0.1 provides the best performance, balancing exploration and exploitation, resulting in the highest average rewards and optimal action selection.

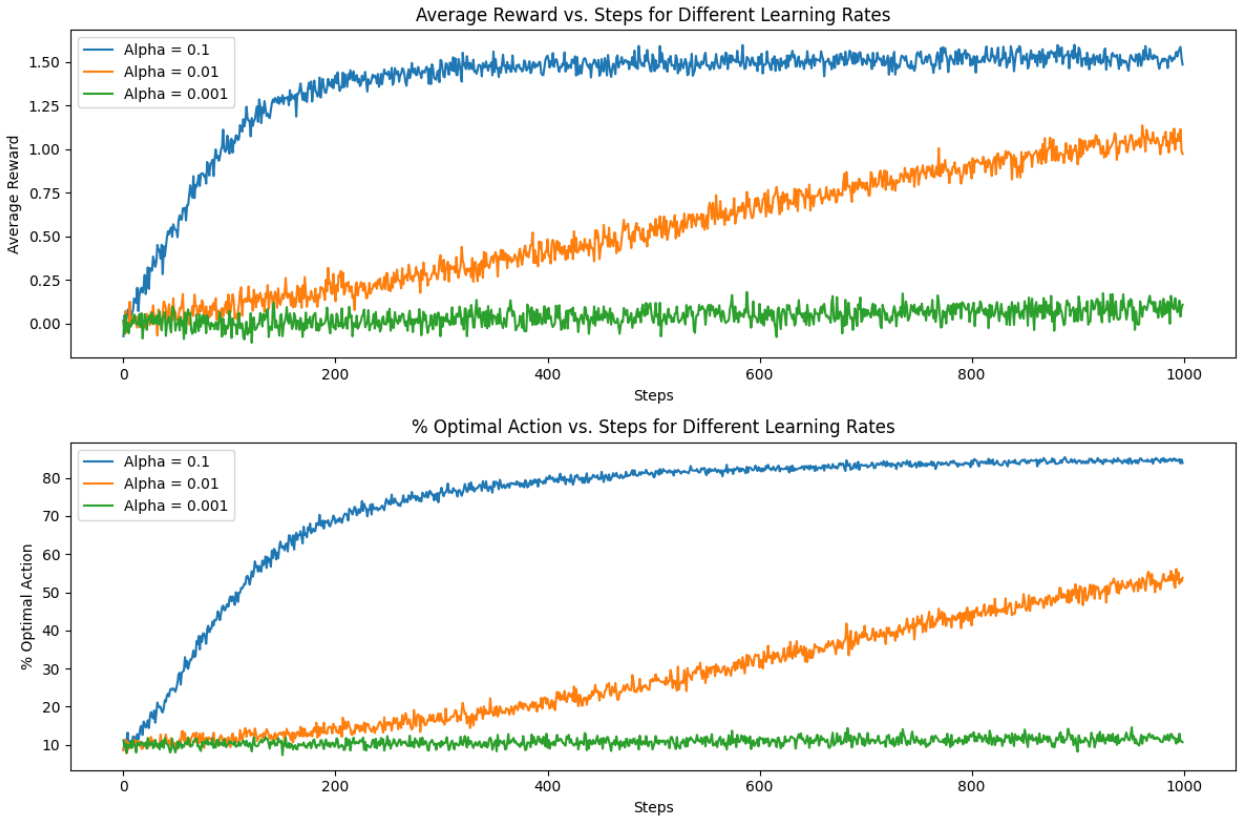
❖ *Optimistic starting values with a greedy approach*



The Average Reward vs. Steps graph indicates that the average reward rapidly increases at the start for all initial values, reaching a stable point around 1.5. This quick rise suggests that optimistic initial values effectively drive exploration, helping the algorithm quickly identify and exploit high-reward actions. Initial values of 5, 3, and 2 all converge to similar average rewards, showing the robustness of the optimistic approach.

The Optimal Action vs. Steps graph shows a rapid increase in the percentage of optimal actions selected, stabilizing around 65% for all initial values. This indicates that optimistic initial values encourage early exploration, leading to a high frequency of optimal action selection. Overall, the optimistic starting values enhance the algorithm's performance by promoting exploration and quickly converging to high average rewards and a high percentage of optimal actions.

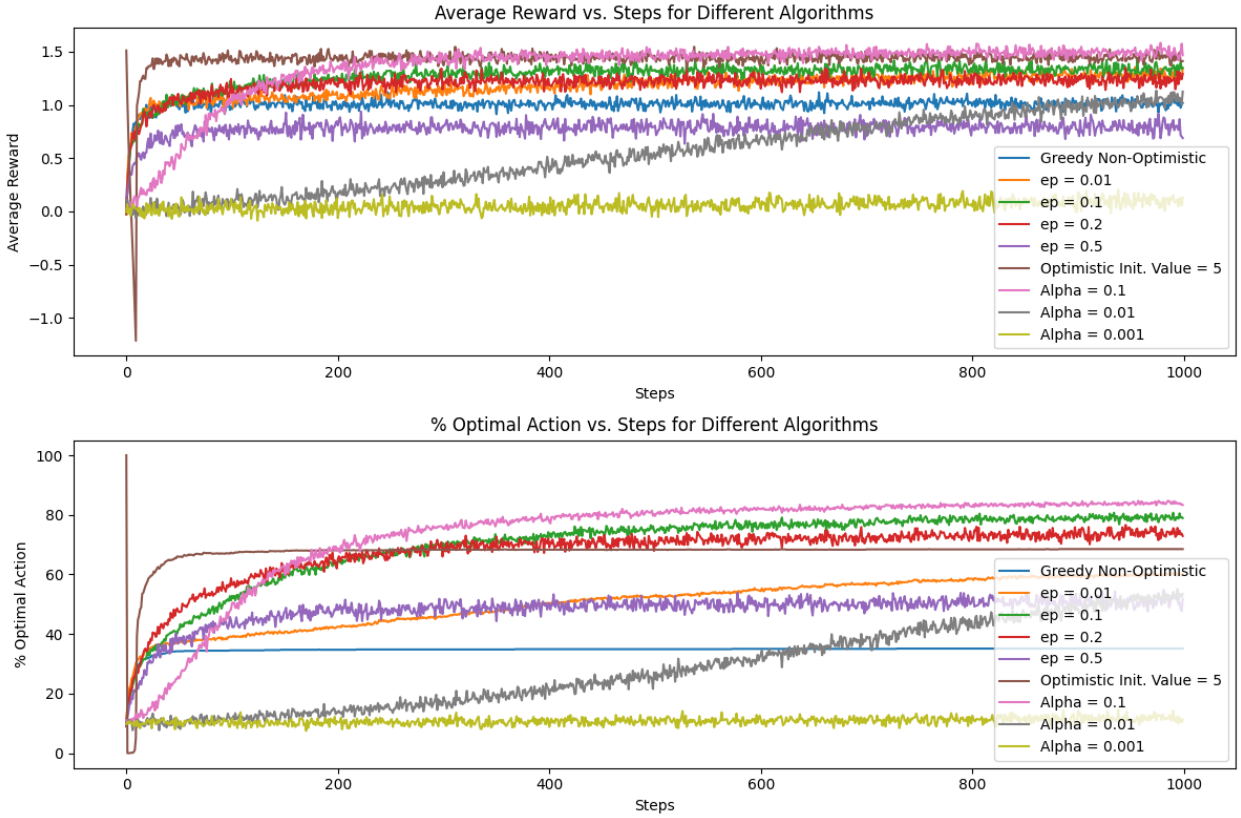
❖ *Gradient bandit algorithm*



The Average Reward vs. Steps graph shows that with a learning rate of 0.1, the average reward increases rapidly to about 1.5 within the first 200 steps, indicating quick adaptation and high performance. In contrast, a learning rate of 0.01 results in a slower increase, stabilizing around 0.75, while a very low learning rate of 0.001 leads to minimal improvement, stabilizing around 0.25.

The Optimal Action vs. Steps graph demonstrates that with a learning rate of 0.1, the percentage of optimal actions quickly rises to about 80%. With a learning rate of 0.01, this percentage increases more gradually, reaching around 50%. For the lowest learning rate of 0.001, the percentage remains low, around 30%. These results suggest that a higher learning rate (alpha = 0.1) enables the Gradient Bandit algorithm to learn and select optimal actions more effectively, leading to higher rewards and better overall performance, while lower learning rates result in slower learning and reduced effectiveness.

❖ *Combining all for the comparison*



The Average Reward vs. Steps graph shows that the Optimistic Initial Value (5) and the Gradient Bandit with Alpha = 0.1 achieve the highest average rewards, stabilizing around 1.5, indicating rapid and effective learning. The Epsilon-Greedy algorithm with ep = 0.1 also performs well, stabilizing at approximately 1.4, demonstrating a good balance between exploration and exploitation. The Epsilon-Greedy settings with ep = 0.2 and ep = 0.01 stabilize around 1.3 and 1.2, respectively, showing effective but slightly less optimal performance. The Greedy Non-Optimistic and Epsilon-Greedy with ep = 0.5 show slower increases, stabilizing around 1.0 and 0.9, respectively, due to excessive exploitation or exploration. The Gradient Bandit with lower Alpha values (0.01 and 0.001) perform the worst, with rewards stabilizing around 0.75 and 0.25.

The Optimal Action vs. Steps graph shows that the Optimistic Initial Value (5) and the Gradient Bandit with Alpha = 0.1 quickly reach around 80% optimal action selection. The Epsilon-Greedy settings with ep = 0.1 and ep = 0.2 stabilize around 75% and 70%, indicating high optimal action selection. The Epsilon-Greedy with ep = 0.01 stabilizes around 50%, indicating less frequent optimal action selection. The Greedy Non-Optimistic and Epsilon-Greedy with ep = 0.5 stabilize around 40%, reflecting less effective performance. The Gradient Bandit with lower Alpha values (0.01 and 0.001) show the lowest percentages, stabilizing around 30%.

To determine the optimal value of epsilon for the Epsilon-Greedy algorithm, we conducted pilot runs with various settings (0.01, 0.1, 0.2, and 0.5) on several bandit problems and tracked the evolution of the rewards curve. We observed that an epsilon value of 0.1 provided the best balance between exploration and exploitation, resulting in the highest average rewards and optimal action selection percentages. Among the different methods tested, the Optimistic Initial Value (5) and the Gradient Bandit with Alpha = 0.1 performed the best, achieving the highest average rewards and optimal action selection rates. The optimistic initial values encouraged extensive early exploration, leading to quicker discovery and exploitation of the best actions, while the high learning rate of 0.1 in the Gradient Bandit allowed rapid adaptation based on received rewards. Other epsilon values (0.2, 0.01, and 0.5) showed progressively lower performance due to either excessive exploration or insufficient exploration. Lower learning rates in the Gradient Bandit also resulted in slower learning and reduced effectiveness. These findings highlight the importance of proper parameter tuning and balancing exploration with exploitation in reinforcement learning algorithms.

Code is available: Click here

Part 02: Non-stationary modifications of the problem above

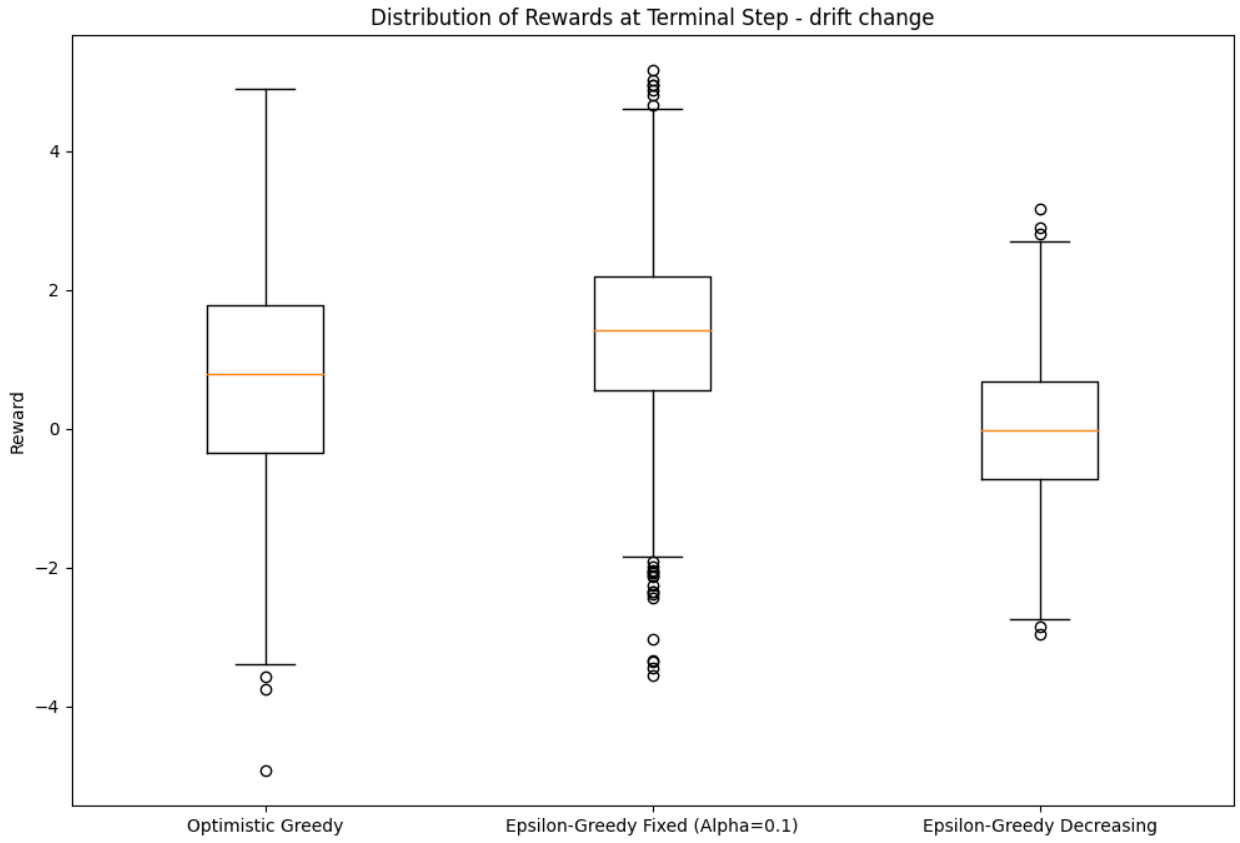
2.1Gradual changes

- A drift change $\mu t = \mu t-1 + \epsilon t$ where ϵt is N(0, 0.0012)
- A mean-reverting change $\mu t = \kappa \mu t-1 + \epsilon t$ where $\kappa = 0.5$ and ϵt is N(0, 0.012) to the mean parameters

2.2 Abrupt changes

- At each time step, with probability 0.005, permute the means corresponding to each of the reward distributions.

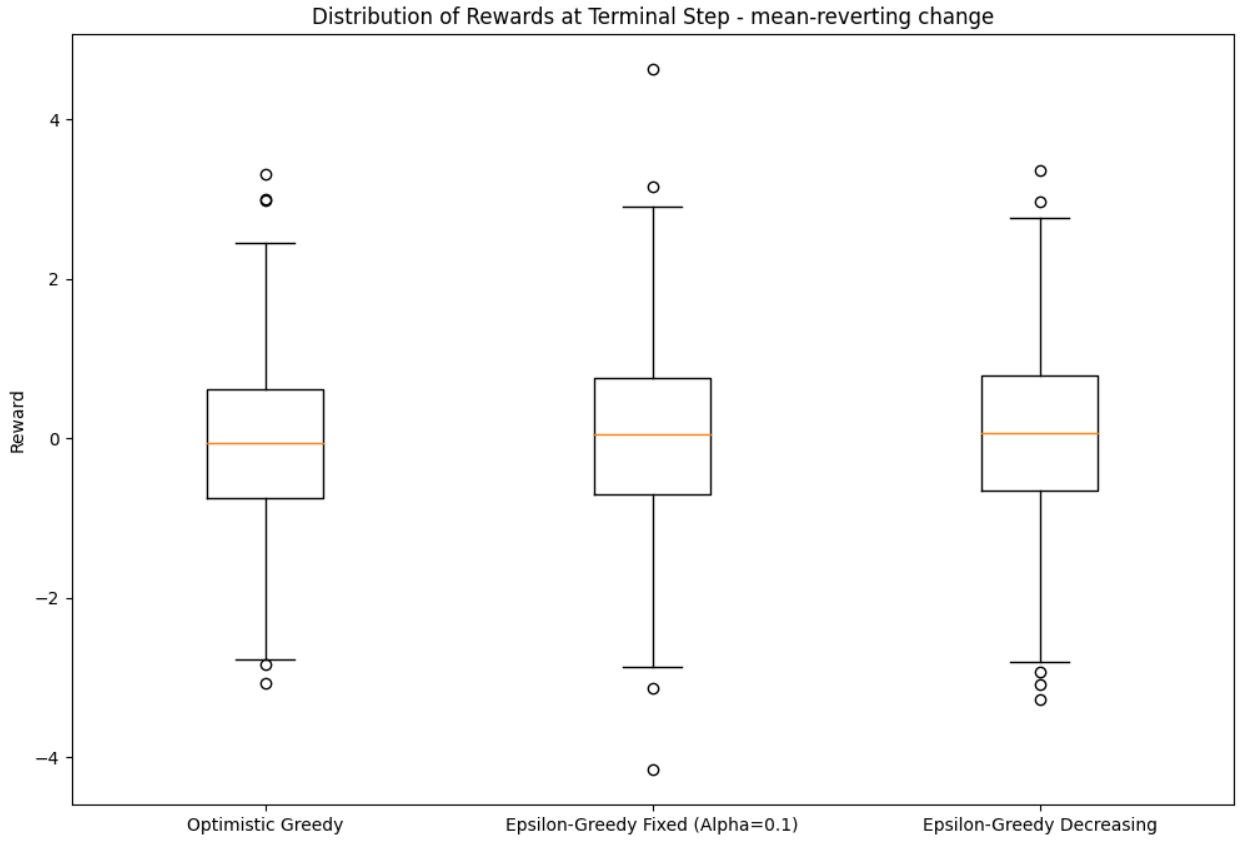
A drift change



This illustrates the distribution of rewards at the terminal step for three different algorithms under a drift change scenario. The Optimistic Greedy algorithm shows a wide range of rewards, with a median slightly above 1, indicating it often achieves high rewards but with significant variability. The interquartile range (IQR) is large, and the whiskers extend above 4, with some outliers below -4, reflecting occasional poor performance. The Epsilon-Greedy algorithm with a fixed step size (Alpha = 0.1) has a more concentrated distribution around its median, which is around 1. The IQR is smaller compared to Optimistic Greedy, and

there are several outliers, mostly below the lower whisker, indicating consistent performance with occasional lower rewards. The Epsilon-Greedy algorithm with a decreasing step size shows a similar median around 1 but with a narrower range of rewards. The IQR is tighter, and there are fewer extreme outliers, suggesting more stable performance under drift change conditions. Overall, while Optimistic Greedy can achieve the highest rewards.

A mean-reverting change

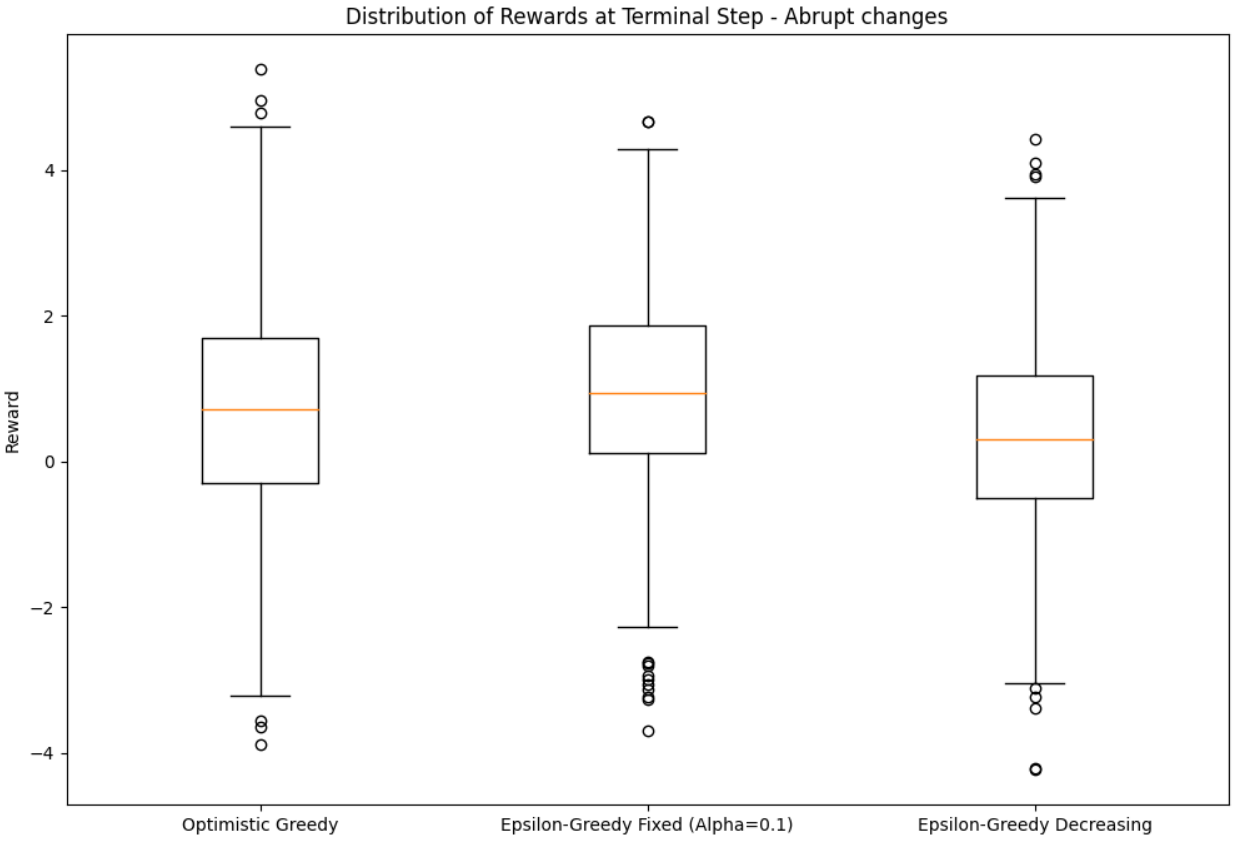


This illustrates the distribution of rewards at the terminal step for three different algorithms under a mean-reverting change scenario: Optimistic Greedy, Epsilon-Greedy with a fixed step size (Alpha = 0.1), and Epsilon-Greedy with a decreasing step size. In the plot, all three algorithms show similar median rewards, which are close to 0. The interquartile range (IQR) for each method is comparable, indicating that the central 50% of the reward data is spread similarly across the algorithms. The Optimistic Greedy algorithm has a few outliers, with rewards going slightly above 2 and below -2, reflecting occasional higher and lower performance.

The Epsilon-Greedy with a fixed step size (Alpha = 0.1) also shows a few outliers, with rewards slightly above 3 and below -4, indicating some variability but overall stable performance. The Epsilon-Greedy with a decreasing step size has fewer outliers and a slightly tighter IQR, suggesting more consistent performance with less extreme variability. Overall, under mean-reverting change conditions, all three algorithms perform

similarly in terms of median rewards and variability. However, the Epsilon-Greedy with a decreasing step size shows slightly more stable performance, as indicated by fewer outliers and a tighter IQR.

Abrupt changes



This illustrates the distribution of rewards at the terminal step for three different algorithms under an abrupt change scenario: Optimistic Greedy, Epsilon-Greedy with a fixed step size (Alpha = 0.1), and Epsilon-Greedy with a decreasing step size. The Optimistic Greedy algorithm has a median reward around 1, indicating generally good performance, but it also shows a wide range of rewards with an interquartile range (IQR) from approximately -2 to 2.5, and several outliers above 4 and below -4, suggesting occasional extreme performance. The Epsilon-Greedy with a fixed step size also has a median reward around 1, with a similar IQR, and numerous outliers mostly on the lower end, reflecting some consistency but also occasional poor performance. The Epsilon-Greedy with a decreasing step size shows a slightly lower median reward, with a narrower IQR, and fewer extreme outliers, suggesting more stable performance but slightly lower overall rewards. Overall, while the Optimistic Greedy and Epsilon-Greedy with a fixed step size perform similarly in terms of median rewards, the Epsilon-Greedy with a decreasing step size offers more stability under abrupt changes.

Comments

we compared the Optimistic Greedy method, Epsilon -Greedy with a fixed step size, and Epsilon -Greedy with a decreasing step size by running each algorithm on 1,000 repetitions of non-stationary problems. We analyzed the distribution of the average reward at the terminal step using box plots. The Optimistic Greedy method, which encourages early exploration with optimistic initial values, showed high median rewards but significant variability across all scenarios. The Epsilon -Greedy with a fixed step size demonstrated more consistent performance with a concentrated distribution around its median, reflecting occasional lower rewards. The Epsilon -Greedy with a decreasing step size, using the reciprocal of the step count, offered the most stable performance with fewer extreme outliers, particularly under drift and abrupt changes. These findings indicate that while the Optimistic Greedy method can achieve high rewards, the Epsilon -Greedy with a fixed step size provides more consistency, and the Epsilon -Greedy with a decreasing step size offers superior stability in changing environments. This underscores the importance of using adaptive methods to handle non-stationarity effectively, ensuring more stable and reliable performance.

Code is available: Click here