Reinforcement Learning: Assignment 1

Sameer Ahamed (202381922)

June 13, 2024

1 Stationary reward distributions

1.0.1 Bandit setup

- **Arms (k)**: 10
- **True Means**: Each arm's reward distribution mean is drawn from a normal distribution $\mathcal{N}(0, 1)$.
- **Rewards**: Rewards for each selected action are drawn from a normal distribution centered around the true mean of the selected arm.

1.0.2 Action selection

A greedy strategy is utilized, wherein the action with the greatest estimated value is chosen at each phase. The estimations of the action values are initialized to zero. This environment is the same for all the experiments in the first question.

1.0.3 Simulation Parameters

- **Steps**: 1000
- **Runs**: 1000

The results are obtained by calculating the average of 1000 runs in order to establish statistical significance.

1.1 Greedy with non-optimistic initial values

1.1.1 Results and Inferences

- **Initial Phase**: The average reward increases rapidly during the initial phase (first 100 steps), indicating that the greedy method quickly identifies high-reward actions. Additionally, the percentage of optimal actions selected starts very low (around 10%) and increases sharply in the initial phase.
- **Stabilization and Convergence**: After approximately 100 steps, the average reward stabilizes around 1.0. This suggests that the algorithm consistently selects actions with high rewards, though there is some fluctuation due to the variability in the reward distributions. By approximately 200 steps, the percentage of optimal actions selected converges to around 35%. This indicates that the greedy method frequently selects the optimal action but does not always do so, possibly due to the initial exploration and inherent variability in the reward estimates.

1.2 Epsilon greedy with different choices of epsilon

1.2.1 Results and Inferences

- **Average Reward over Time**:

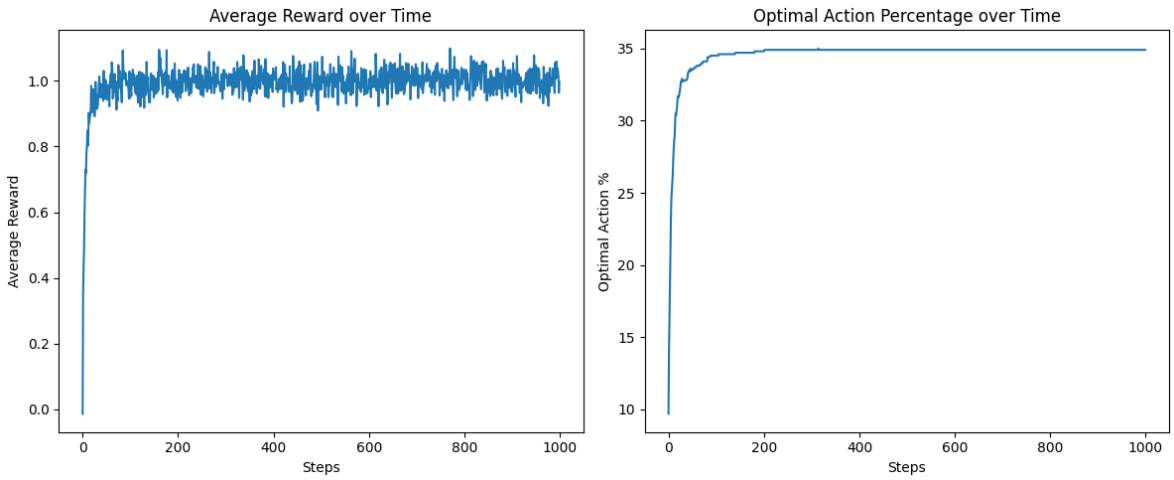

Figure 1: Greedy with Non Optimistic intial values

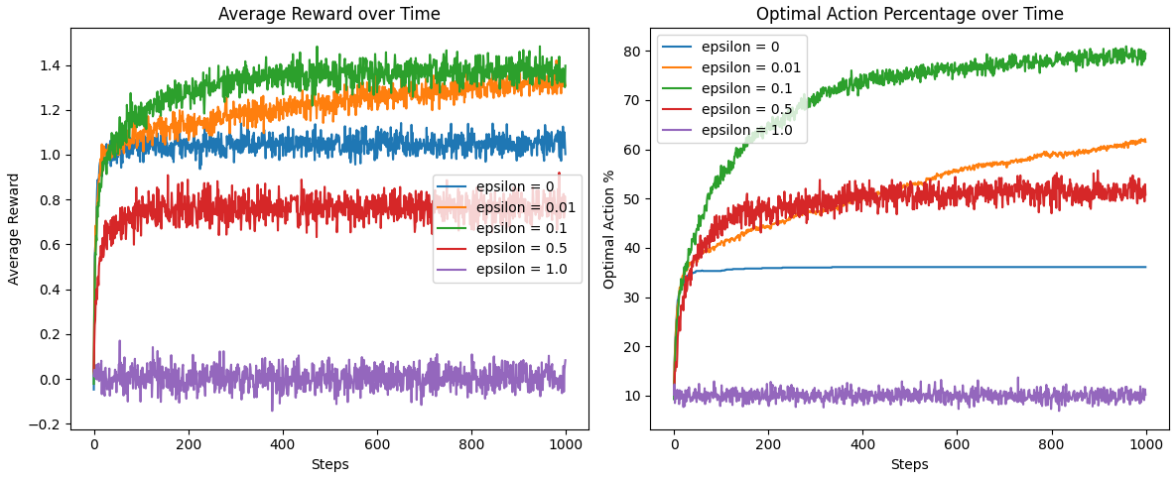

Figure 2: Epsilon greedy with different choices of epsilon

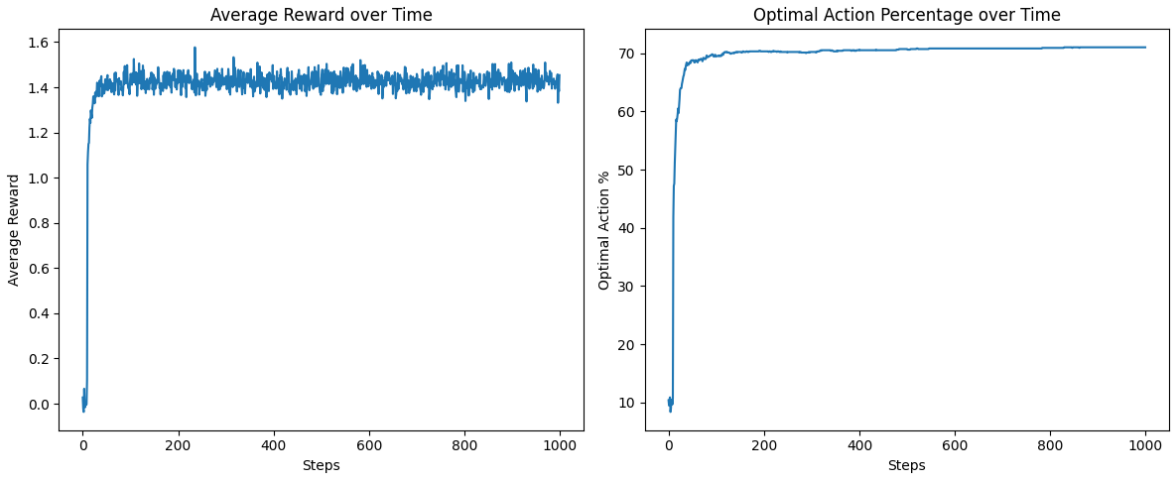

Figure 3: optimistic starting values with a greedy approach

- $\epsilon = 0$: The average reward quickly stabilizes around 0.7, indicating that with no exploration, the greedy method may not find the optimal action if it doesn't initially select high-reward actions.
- $\epsilon = 0.01$: The average reward stabilizes around 1.0, showing that minimal exploration helps find better actions compared to no exploration.
- $\epsilon = 0.1$: The average reward stabilizes around 1.2, indicating a good balance between exploration and exploitation, leading to higher average rewards.
- $\epsilon = 0.5$ and $\epsilon = 1.0$: The average rewards are lower compared to $\epsilon = 0.1$, suggesting that excessive exploration results in sub-optimal rewards as the algorithm frequently tries random actions.

- **Optimal Action Percentage over Time**:
  - $\epsilon = 0$: The optimal action percentage remains low (10%) throughout, indicating that without exploration, the greedy method rarely identifies the optimal action.
  - $\epsilon = 0.01$: The optimal action percentage increases to around 50%, indicating some improvement with minimal exploration.
  - $\epsilon = 0.1$: The optimal action percentage increases to around 80%, demonstrating a good balance of exploration and exploitation.
  - $\epsilon = 0.5$ and $\epsilon = 1.0$: The optimal action percentage remains lower (30% for $\epsilon = 0.5$ and 10% for $\epsilon = 1.0$), as excessive exploration prevents the algorithm from consistently selecting the optimal action.

Overall, the results indicate that an epsilon value of 0.1 provides a good balance between exploration and exploitation, leading to higher average rewards and a higher percentage of selecting the optimal action.

1.3 optimistic starting values with a greedy approach

1.3.1 Results and Inferences

- **Average Reward over Time**: - The average reward rapidly increases and stabilizes around 1.4, indicating that the optimistic initial values help the greedy method to quickly find and exploit high-reward actions. - The reward remains consistently high with minor fluctuations, showing that the method is effective in maintaining high rewards over time.
- **Optimal Action Percentage over Time**: - The optimal action percentage starts very low (around 10%) but rapidly increases to around 70% within the first 100 steps. - After this initial

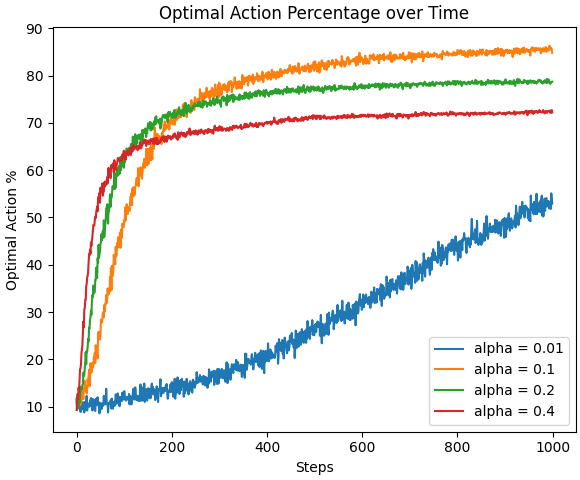

Figure 4: Gradient bandit with different learning rates

phase, the percentage of optimal actions selected remains stable around 70%, indicating that the optimistic initialization leads to frequent selection of the optimal action over time.

Overall, the results indicate that using optimistic initial values for the greedy algorithm significantly improves both the average rewards and the percentage of optimal actions selected. The high initial values encourage exploration in the early steps, allowing the algorithm to discover high-reward actions, which it then exploits effectively.

1.4 Gradient Bandit with Different Learning Rates (Alpha)

1.4.1 Results and Inferences

- **Optimal Action Percentage over Time**:
  - $\alpha = 0.01$: The optimal action percentage increases steadily and reaches around 60% after 1000 steps. This indicates slow but consistent learning.
  - $\alpha = 0.1$: The optimal action percentage increases more rapidly, reaching around 80% within 200 steps and stabilizing around this value. This suggests a good balance between learning speed and stability.
  - $\alpha = 0.2$: The optimal action percentage increases quickly and reaches around 75% within 200 steps, stabilizing slightly lower than for $\alpha = 0.1$. This indicates faster learning but slightly less stability.
  - $\alpha = 0.4$: The optimal action percentage increases rapidly but stabilizes around 70%, indicating that a higher learning rate can lead to faster initial learning but may cause more fluctuations and less stable long-term performance.

Overall, the results suggest that a moderate learning rate ($\alpha = 0.1$) provides the best balance between learning speed and stability, leading to a higher percentage of optimal actions selected over time.

1.5 Which method performs the best and why?

Using the optimal action percentage over time as the metric, the best performing model is:

- **Gradient Bandit with** $\alpha = 0.1$: Stabilizes slightly above 80%, showing the best performance in selecting the optimal action consistently.

The **gradient bandit with** $\alpha = 0.1$ demonstrates slightly better performance in terms of the percentage of optimal actions selected over time compared to epsilon-greedy with $\epsilon = 0.1$. It achieves and stabilizes just above 80% optimal actions, indicating a more effective balance between exploration and exploitation. Thus, the gradient bandit model with $\alpha = 0.1$ is the best performing model among the tested options.

The gradient bandit algorithm with an appropriate learning rate ($\alpha = 0.1$ in this case) outperforms other methods in terms of optimal action percentage for several reasons:

- **Action Preference Mechanism**: The gradient bandit method updates the preferences for each action in a set, according to the rewards obtained. This method enables the algorithm to directly manipulate the likelihood of choosing each action. The gradient bandit algorithm, in contrast to the epsilon-greedy technique, does not rely on estimated values and a set exploration probability. Instead, it dynamically adjusts the probabilities of each action based on their performance relative to the average reward.
- **Softmax Action Selection**: Utilizing a softmax function to transform preferences into probabilities guarantees a seamless and uninterrupted modification of action selection probabilities. This approach encourages the implementation of actions that have demonstrated success, while also permitting a certain degree of research. The softmax function provides a more efficient way to manage the trade-off between exploration and exploitation, as compared to using a fixed epsilon parameter in epsilon-greedy approaches.
- **Use of Average Reward Baseline**: By incorporating an average reward baseline into the preference update calculation, the gradient bandit algorithm is able to standardize the rewards it receives. By normalizing the data, the algorithm becomes more capable of differentiating between actions based on their relative performance rather than their absolute rewards. This can result in more reliable and consistent learning.

2 Non-stationary modifications to the problem

2.1 Drift Change with Optimistic Greedy Method

- The optimistic greedy agent performs reasonably well in an environment with drifting means, achieving a median average reward of around 1.5 at the terminal step.
- The consistency in the agent's performance is indicated by the narrow IQR, but the presence of outliers shows that the agent's performance can vary significantly in some cases.
- The optimistic initial values help the agent to explore initially, but the agent might still struggle in some runs to adapt to the ongoing drift in the reward distributions.
- Overall, the optimistic greedy agent provides a good balance between exploration and exploitation, allowing it to handle the non-stationary environment with drifting means reasonably well.

2.2 Drift Change with epsilon-greedy with a fixed step size

- The epsilon-greedy agent with a fixed step size shows moderate performance in an environment with drifting means, achieving a median average reward of around 1.0 at the terminal step.
- The wider IQR indicates more variability in the agent's performance, suggesting that the fixed step size may not be as effective in adapting to the ongoing drift in the reward distributions.
- The presence of outliers shows that while the agent can occasionally perform very well, it also has a tendency to perform poorly in some runs.
- Overall, the epsilon-greedy agent with a fixed step size provides a moderate balance between exploration and exploitation but struggles to consistently handle the non-stationary environment with drifting means.

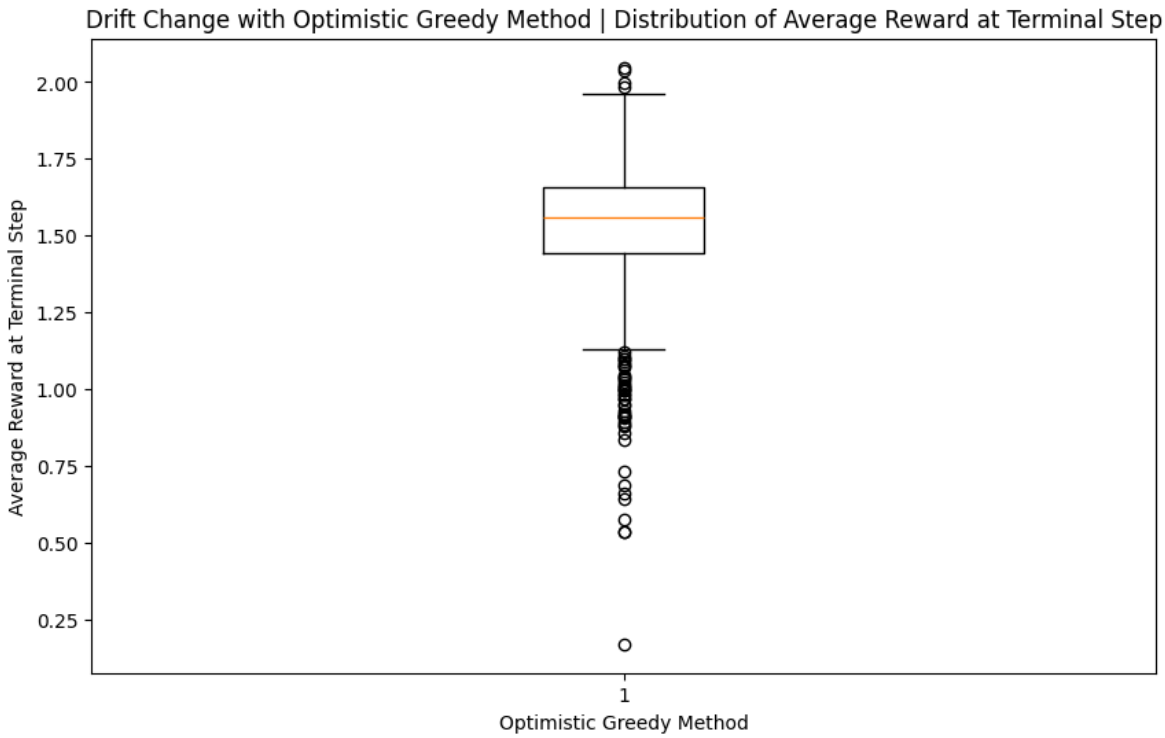

Figure 5: Drift Change with Optimistic Greedy Method

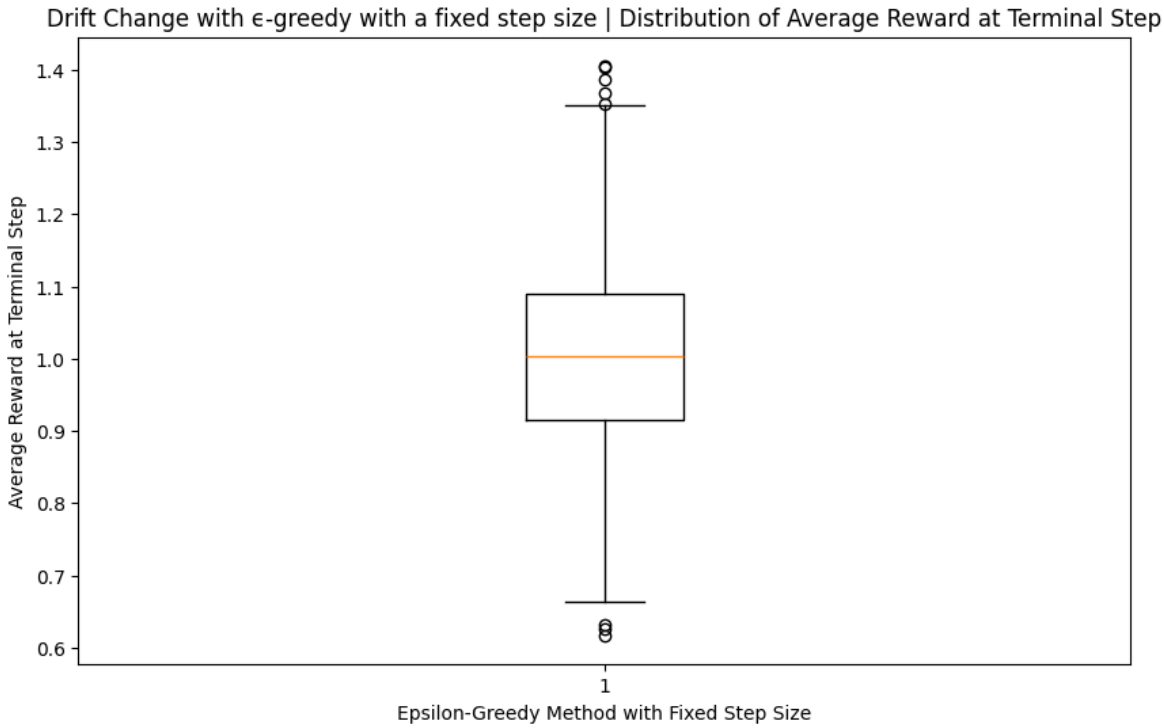

Figure 6: Drift Change with epsilon greedy with fixed step size

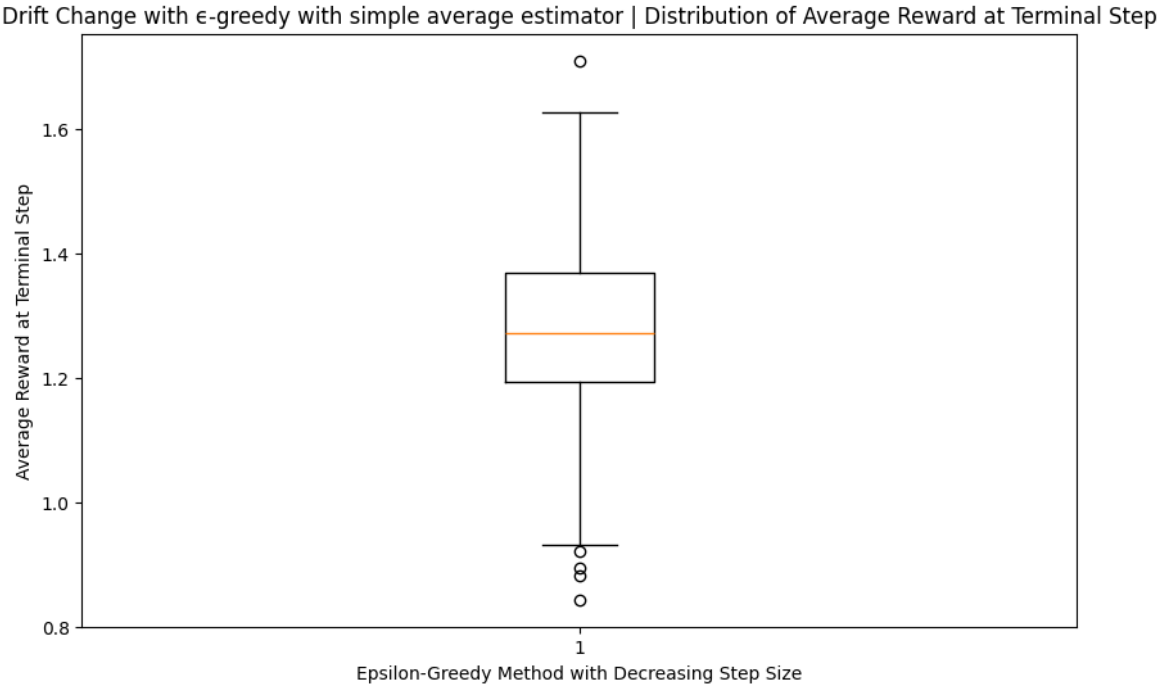

Figure 7: Drift Change with epsilon-greedy with simple average estimator

2.3 Drift Change with epsilon-greedy with a decreasing step-size (e.g., simple average estimator)

- The epsilon-greedy agent with a decreasing step size shows good performance in an environment with drifting means, achieving a median average reward of around 1.25 at the terminal step.
- The moderate IQR indicates reasonable consistency in the agent's performance, although there is still some variability.
- The presence of outliers shows that while the agent can occasionally perform very well, it also has a tendency to perform poorly in some runs.
- Overall, the epsilon-greedy agent with a decreasing step size provides a good balance between exploration and exploitation, allowing it to handle the non-stationary environment with drifting means relatively well.

2.4 Optimistic Greedy Method with mean-reverting change

- The optimistic greedy agent performs poorly in an environment with mean-reverting changes, achieving a median average reward of around 0.0 at the terminal step.
- The narrow IQR indicates that the agent's performance is consistently low, with most runs producing average rewards close to the median.
- The presence of outliers shows that while the agent can occasionally perform slightly better or worse, these instances are not common.
- Overall, the optimistic greedy agent struggles to adapt to the mean-reverting changes in the reward distributions, resulting in low overall performance.

2.5 epsilon-greedy with fixed step size model with mean-reverting change

- The epsilon-greedy agent with a fixed step size performs poorly in an environment with mean-reverting changes, achieving a median average reward of around 0.0 at the terminal step.

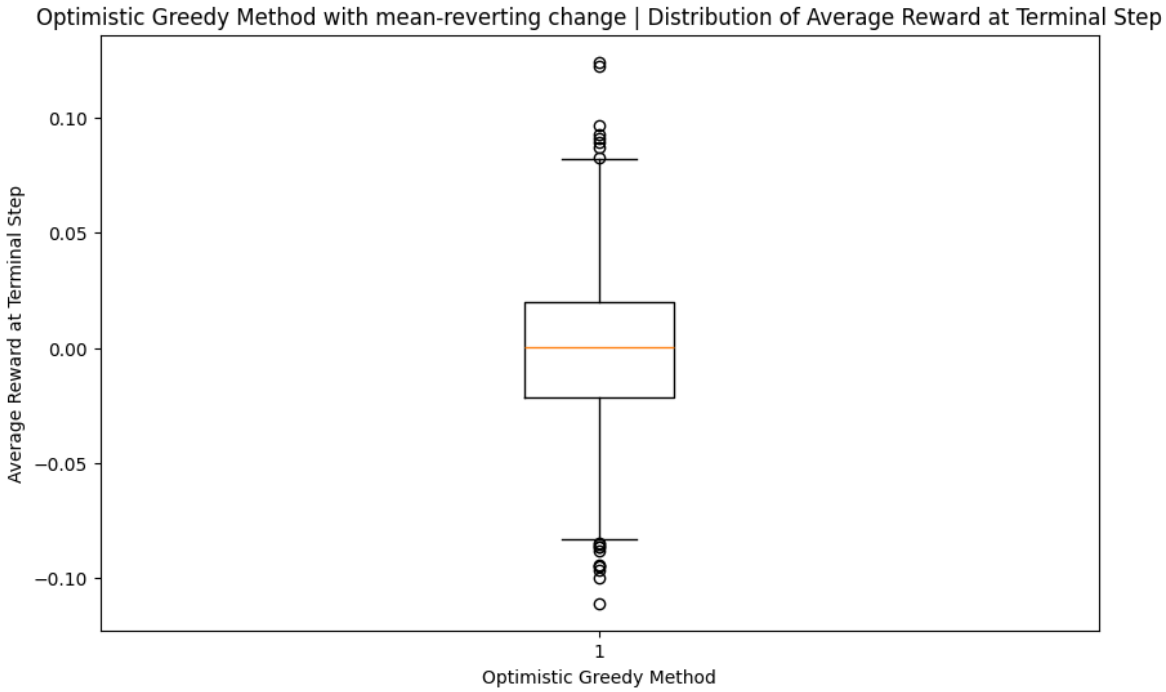

Figure 8: Optimistic Greedy Method with mean-reverting change

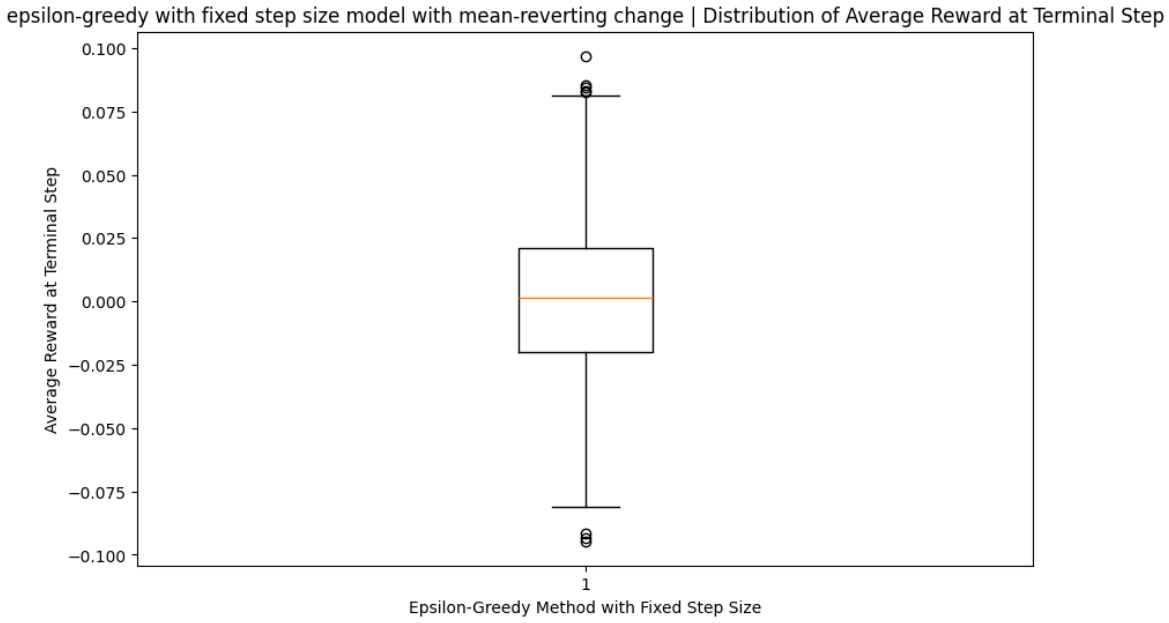

Figure 9: epsilon-greedy with fixed step size model with mean-reverting change

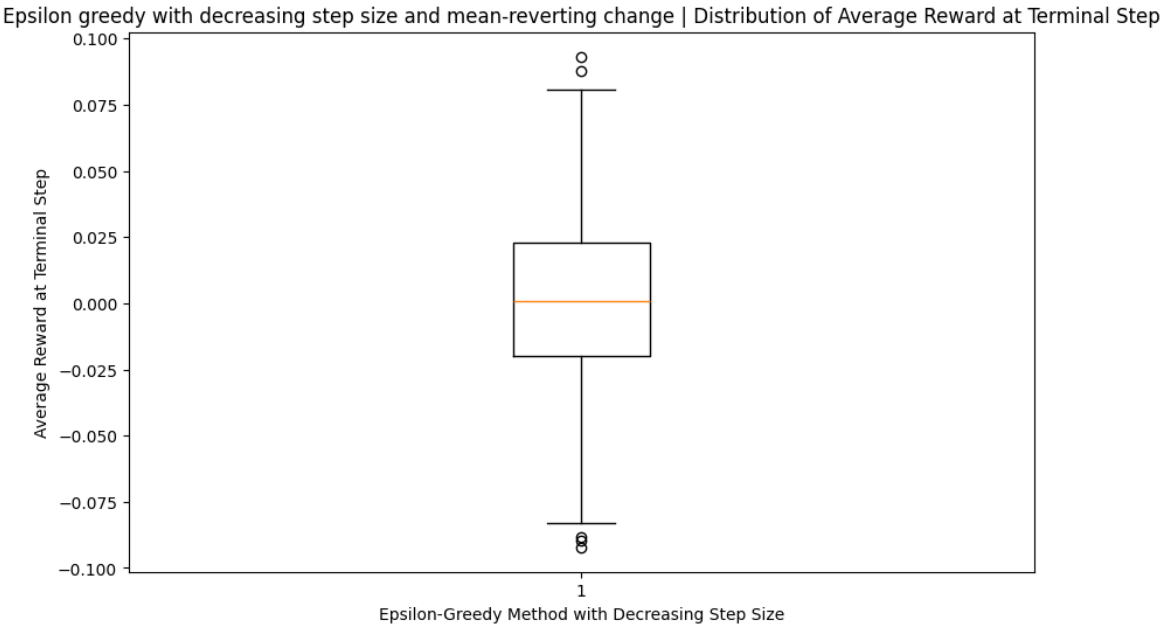

Figure 10: Epsilon greedy with decreasing step size and mean-reverting change

- The narrow IQR indicates that the agent's performance is consistently low, with most runs producing average rewards close to the median.
- The presence of outliers shows that while the agent can occasionally perform slightly better or worse, these instances are not common.
- Overall, the epsilon-greedy agent with a fixed step size struggles to adapt to the mean-reverting changes in the reward distributions, resulting in low overall performance.

2.6 Epsilon greedy with decreasing step size and mean-reverting change

- The epsilon-greedy agent with a decreasing step size performs poorly in an environment with mean-reverting changes, achieving a median average reward of around 0.0 at the terminal step.
- The narrow IQR indicates that the agent's performance is consistently low, with most runs producing average rewards close to the median.
- The presence of outliers shows that while the agent can occasionally perform slightly better or worse, these instances are not common.
- Overall, the epsilon-greedy agent with a decreasing step size struggles to adapt to the mean-reverting changes in the reward distributions, resulting in low overall performance.

2.7 Abrupt Changes and Optimistic Greedy Method

- The optimistic greedy agent performs poorly in an environment with abrupt changes, achieving a median average reward of around 0.0 at the terminal step.
- The wide IQR indicates that the agent's performance is highly inconsistent, with significant variability across different runs.
- The presence of many outliers shows that while the agent can occasionally perform better or worse, these instances are relatively common.
- Overall, the optimistic greedy agent struggles to adapt to the abrupt changes in the reward distributions, resulting in low overall performance and limited adaptability.

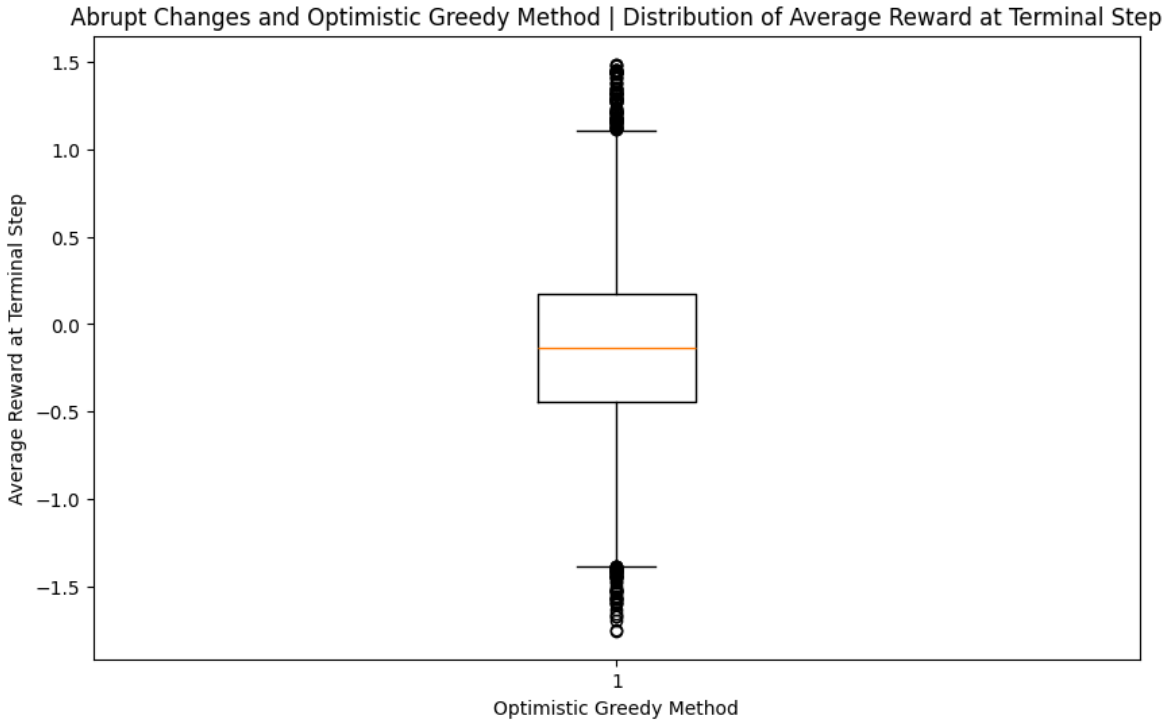

Figure 11: Abrupt Changes and Optimistic Greedy Method

2.8 Abrupt Changes and epsilon-Greedy with Fixed Step Size

- The epsilon-greedy agent with a fixed step size performs relatively well in an environment with abrupt changes, achieving a median average reward of around 1.5 at the terminal step.
- The wide IQR indicates that the agent's performance can vary significantly across different runs, showing both high and low extremes.
- The presence of outliers demonstrates that while the agent can perform very well in some runs, it can also perform poorly in others.
- Overall, the epsilon-greedy agent with a fixed step size shows adaptability to the abrupt changes in the reward distributions, resulting in relatively high overall performance despite the variability.

2.9 Abrupt Changes and epsilon-Greedy with Decreasing Step Size

- The epsilon-greedy agent with a decreasing step size performs poorly in an environment with abrupt changes, achieving a median average reward of around 0.0 at the terminal step.
- The moderate IQR indicates variability in the agent's performance across different runs, with performance spanning a range of rewards.
- The presence of many upper outliers shows that while the agent can occasionally perform well, these instances are not common.
- Overall, the epsilon-greedy agent with a decreasing step size struggles to consistently adapt to the abrupt changes in the reward distributions, resulting in low overall performance despite occasional high rewards.

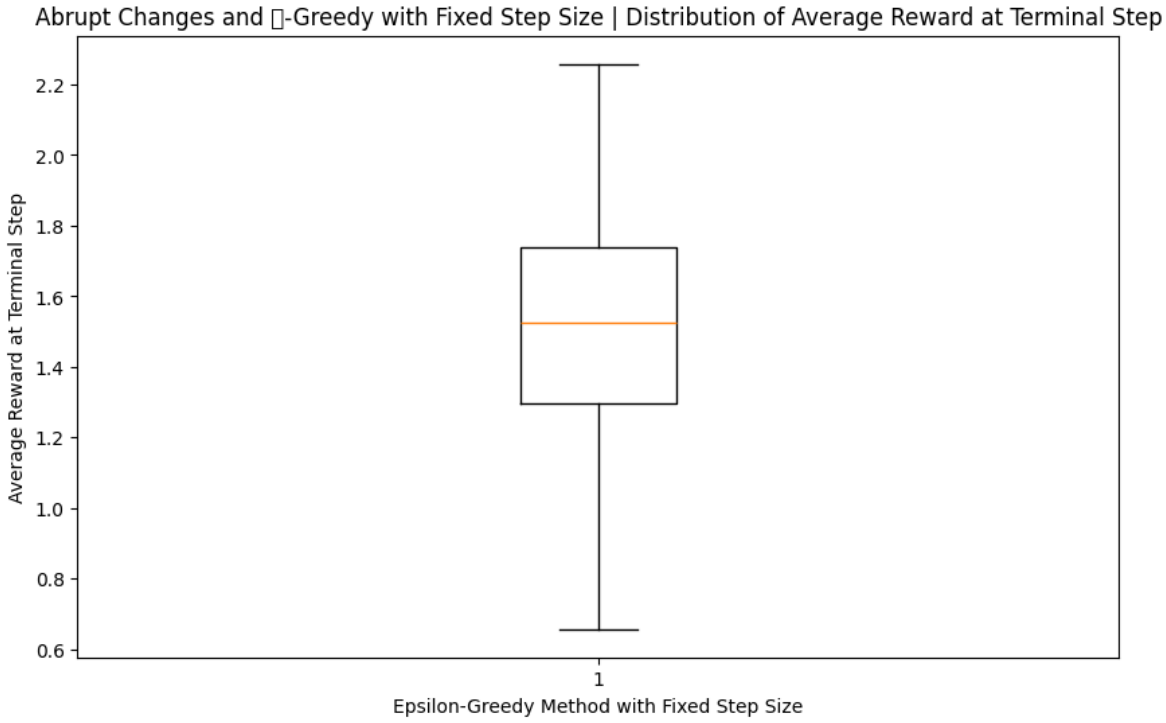

Figure 12: Abrupt Changes and epsilon Greedy with Fixed Step Size

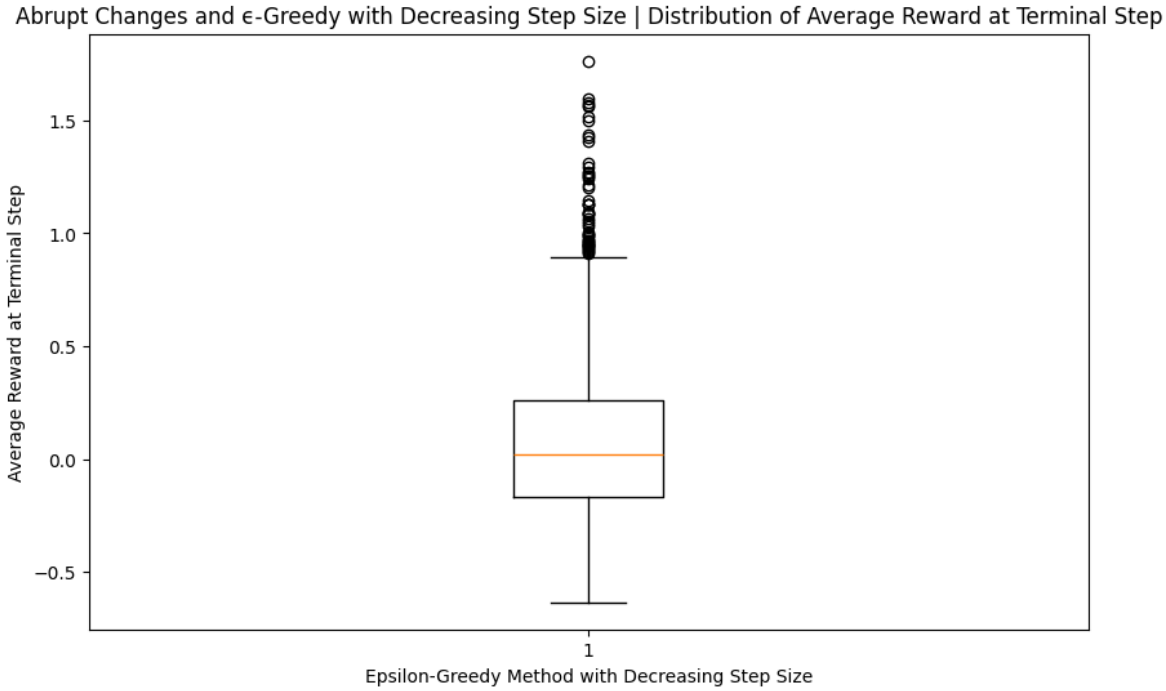

Figure 13: Abrupt Changes and epsilon-Greedy with Decreasing Step Size