A Comparative Analysis of Reinforcement Learning Strategies: Epsilon-Greedy with Sample Average versus Constant Step-Size in a Non-Stationary, k-Armed Bandit Framework

Darcy Corson

September 26, 2023

1 Introduction and Motivation

Reinforcement Learning (RL) is a branch of machine learning that models the learning process of agents making sequential decisions over time in an environment. In RL, the agent observes the current state of the environment, selects an action, receives a reward, and transitions to a new state as a consequence. The agent's learning objective is to discover a policy – a strategic mapping from states to actions – that maximizes the expected cumulative reward over time.

Exercise 2.5, presented by Richard S. Sutton and Andrew G. Barto in their text, *Reinforcement Learning: an Introduction*, revolves around understanding and analyzing how well sample-average methods perform in environments where the underlying rules, probabilities, and dynamics are prone to change (Sutton & Barto, 2018). In these "non-stationary" environments, the ability to accurately estimate true q-values is central to the agent's ability to adapt and make optimal decisions. Thus, the methods used to estimate q-values must be both robust and accurate. If sample-average methods are unable to produce accurate estimates in non-stationary settings, it is imperative that other methods are identified that better cope with the changing environment. It is critical, therefore, to scrutinize the adaptability and accuracy of the sample-average method as compared to the constant step-size method in order to provide insightful conclusions about the methods' applicability and efficiency in non-stationary environments.

This study investigates the challenges associated with employing sample-average methods in non-stationary environments, emphasizing the dynamic nature of such problems where the true action value evolves over time. Through a modified version of the 10-armed testbed, the experiment serves to elucidate the difficulties inherent to sample-average methods in achieving accurate estimations of action values within fluctuating environments.

2 Background

2.1 The k-Armed Bandit Problem and Sample-Average Methods

The k-armed bandit problem introduces an agent confronted with k different options, or "arms", each associated with a reward drawn from a probability distribution that is unique to each arm. The agent's objective is to optimize the accumulated reward over a series of arm pulls, choosing actions that either exploit known options or explore the potentially more rewarding unknown options. This exploration vs. exploitation trade-off is central to RL, where an agent must decide whether to exploit the best-known option for immediate reward or explore less-known options to maximize the long-term benefits. In this context, the true value of each action, $q_*(a)$, represents the expected reward of an action, and is unknown to the agent. The estimated value, $Q_t(a)$, represents the agent's current estimated expectation of the action value at time t, based on the information that the agent has.

In the basic k-armed bandit problem, the agent uses the sample-average method to calculate the estimated action value, $Q_t(a)$. This method computes $Q_t(a)$ as the average of all the rewards that the agent has received so far from choosing action. Mathematically, if an action a has been chosen n times prior to time t, yielding rewards $r_1, r_2, \ldots, r_n$, then:

$$Q_{\mathrm{t}}(a) = \frac{1}{n} \sum_{i=1}^{n} r_{\mathrm{i}}$$

The basic k-armed bandit problem and the associated sample-average method traditionally assume a stationary environment. In a stationary environment, the true action values, $q_*(a)$, do not change over time, and the reward distributions for each arm remain constant. This assumption allows the sample-average method to converge towards the true action values as more samples are collected because it's based on the law of large numbers, which implies that, given a stationary distribution, the average of the sampled rewards will converge to the expected value of the distribution as the number of samples goes to infinity.

2.2 The Challenge of Non-Stationarity

In real-world scenarios, settings characterized by non-stationarity are pervasive, representing instances where change is inherent, such as in fluctuating stock markets and evolving user preferences. Given this, addressing non-stationarity becomes imperative for the development of learning systems. Within the framework of RL, environments deemed non-stationary are those where the foundational probability distributions, which direct the dynamics of the environment, most notably in rewards and state transitions, undergo changes over time.

As discussed, the sample-average method for determining Qt(a) assumes a stationary environment and posits that the average reward for an action serves as a reliable estimate for future rewards. This method presupposes the constancy of the environment's rewards and transition probabilities over time. However, to enable the implementation of meaningful modifications and enhancements to models, thereby allowing them to accurately determine Qt(a) in non-stationary environments, a nuanced understanding of the challenges intrinsic to applying sample-average methods in these dynamic contexts is crucial.

2.3 Constant Step-Size Methods

The constant step-size method of estimating Qt(a) in k-armed bandit problems is frequently preferred over the sample-average method when environments are non-stationary. This method employs a predetermined constant step-size parameter, denoted as:

$$Q_{\mathrm{new}}(a) = Q_{\mathrm{old}}(a) + \alpha[R_t - Q_{\mathrm{old}}(a)]$$

Here, $Q_{\mathrm{new}}(a)$ and $Q_{\mathrm{old}}(a)$ represent the updated and previous estimated values of the chosen action a, respectively, while R_{t} represents the actual reward received after performing the action. The constant step-size α, a number between 0 and 1, imparts a degree of recency bias to the model, making it particularly helpful in non-stationary environments where the true underlying distributions of rewards are subject to change. The constant α ensures that all received rewards, irrespective of the time of their receipt, are accorded some weight, albeit diminishing, in influencing the estimated action-values. As a result, the model maintains a perpetual learning state that enables it to adapt and respond to changes in the reward distributions as the environment changes, and makes it a valuable tool capable of navigating the landscapes of non-stationary problems.

3 Experiment

3.1 Purpose

The primary objective of this experiment is to evaluate and compare the performance of two methods, sample average and constant step-size, in a k-armed bandit problem under a non-stationary environment, using an ϵ-greedy strategy. The k-armed bandit problem is modeled as a class Bandit with k (10) arms, each representing an action with a real-valued reward drawn from a stationary probability distribution. The comparison evaluates the strategies on their accuracy, responsiveness, and convergence in estimating the optimal action-values over time, under varying conditions.

3.2 Method

The experiment utilizes a Bandit class to represent a 10-armed bandit, with each arm having an initial true action-value q_a^* and an estimated action-value Q(a). These true and estimated values are initialized to zero. The class also contains methods to decide actions, update true and estimated values, and simulate a time-step in a non-stationary environment. The non-stationary environment is modeled by adding a normally distributed incremental value with mean zero and standard deviation equal to 0.1, which is indicative of the degree of non-stationarity in the environment, to the true value of each arm at every time step.

$$q_t^*(a) = q_{t-1}^*(a) + \xi_t(a)$$

Where:
- $q_t^*(a)$ is the true action-value of action a at time t
- $q_{t-1}^*(a)$ is the true action-value of action a at time t-1
- $\xi_t(a) \sim N(0, \sigma_{ns})$ is the normally distributed incremental value with mean 0 and standard deviation σ_{ns} representing the degree of non-stationarity at time t for action a

At each time step, an action a_t is selected using an ϵ-greedy strategy, where with probability ϵ, a random action is selected, and with probability 1 - ϵ, the action with the highest estimated value is selected.

Once the action is selected, a reward R_t is received. The value of that reward is randomly selected from a normal distribution with mean equal to the true value of the action and a standard deviation of 1. Then, the estimated value of the selected action is updated using either the sample average method or the constant step-size method, depending on the experiment configuration.

For the sample average method, the new estimate is calculated as the old estimate plus a fraction of the difference between the received reward and the old estimate. This fraction is 1/n where N is the number of times the action has been selected.

$$Q_{new}(a) = Q_{old}(a) + \frac{1}{N}\left[R_t - Q_{old}(a)\right]$$

For the constant step-size method, the new estimate is calculated similarly, but the fraction is replaced by a constant step-size α.

$$Q_{new}(a) = Q_{old}(a) + \alpha\left[R_t - Q_{old}(a)\right]$$

The function, `run_experiment`, acts as the central mechanism to conduct operations for the experiment's specified instance of the 10-armed bandit problem across a predetermined number of steps. During this process, it logs pertinent data, including the rewards obtained, the incidence of optimal actions taken, and the temporal progression of both true and estimated values of the actions.

The parameter ϵ holds significance in this context as it determines the probability of selecting a random action, indicating exploration, versus the action boasting the highest estimated value, denoting exploitation. The choice between exploration and exploitation is made conditionally within the function. That is, if `np.random.rand() < epsilon`, a random action is selected, exemplifying exploration. If not, the action with the highest estimated value is selected, depicting exploitation, as realized by the line `action = np.argmax(q_estimate)` within the code.

To facilitate the recording of data at each step, `run_experiment` initializes multiple arrays, including `optimal_action_count` and rewards, to respectively store instances of optimal actions and the corresponding rewards received. Additionally, it manages arrays representing the evolving true and estimated values of each action, enabling comprehensive tracking and subsequent analysis of the progression of the simulation.

In every iterative step of the experiment, the action, dictated by the epsilon-greedy strategy, is selected and the associated reward is computed based on the true value of the selected action. This is expressed in the line `reward = q_true[action] + np.random.randn()`, where `np.random.randn()` imparts a normally distributed noise to the true action value. Subsequently, the `action_count` array is updated to mark the selection of the action, and the estimated value `q_estimate` of the executed action is recalculated, taking into account the newly received reward. This re-calibration employs either the

sample average method or the constant step-size method, depending upon the method chosen for that specific simulation run.

The true action values, denoted as `q_true`, undergo incremental adjustments post each step to embody the non-stationary nature of the environment. This is evidenced by the line `q_true += non_stationarity * (np.random.randn(k) * 0.01)`, where `non_stationarity` is a Boolean variable that manages the non-stationary behavior and `(np.random.randn(k) * 0.01)` introduces a subtle normally distributed noise to all true action values.

The simulation runs twice for each experiment, once using sample averages to update estimated action values and once using a constant step-size. These runs are aggregated over 2000 independent runs for the sake of robust analysis. Every run has 10,000 steps.

The results of the experiment are depicted using various plots. The first plot compares average rewards over time, with a t-test to statistically validate any observed differences. The second plot examines the percentage of optimal actions taken, a measure of how well the strategy learns the optimal action over time. The third plot represents the cumulative rewards comparison, which evaluates the overall performance and stability of the strategies.

The fourth and fifth plots illustrate the true and estimated Q-values for the first arm of the bandit using both the sample average method and constant step-size method. These plots give insights into the learning accuracy, convergence, stability, and responsiveness of the algorithms in adapting to the true Q-values in a non-stationary environment.

Finally, a sixth plot and subsequent analysis compare the convergence of the two methods by examining the absolute differences between true and estimated Q-values over time. This helps in understanding which strategy adapts better and learns the true values more accurately in the presence of non-stationarity.

3.3 Hypotheses

- There will be significant differences in the performance of the sample-average method and the constant step-size ($\alpha = 0.1$) method in a non-stationary k-armed bandit problem, measured in terms of average rewards, percentage of optimal action selected, and the convergence of estimated q-values to true q-values over time.

- Given the non-stationary nature of the environment, where the true values of the actions change over time, the method with constant step-size will adapt more effectively to these changes and hence will accumulate higher average rewards compared to the sample-average method.

- The constant step-size method will exhibit a higher percentage of optimal action selection over time compared to the sample-average method, due to its enhanced adaptability in non-stationary environments, enabling more frequent selection of the optimal action.

- The constant step-size method will exhibit faster and more stable convergence of estimated Q-values to the true Q-values. The constant learning rate (α) should allow the method to be more responsive to the changes in the environment, maintaining more accurate estimates of the q-values over time

3.4 Results

The first graph in Figure 1 graphically juxtaposes the cumulative rewards over time between the sample average and constant step-size methods, illustrating the evolution of rewards at each sequential step throughout the experiment. Initial trends in rewards improvement appear congruent between the two methods; however, the rate of improvement for the sample average method starts to plateau more swiftly, yielding lesser rewards over time in comparison to the constant step-size method. A t-test confirms the statistical significance of this discrepancy, affirming that the variation in rewards is intrinsic to the method's performance rather than being a manifestation of random chance (Figure 2).

The second graph of Figure 1 delineates the proportion of optimal actions executed over time by both methods. This trajectory is pivotal as it quantifies the learning efficiency of the agent, illustrating how frequently optimal actions are selected. A swift plateau is observed in the percentage of optimal actions with the sample average method, in stark contrast to the constant step-size method which exhibits a persistent increase in optimal actions as experience accrues.

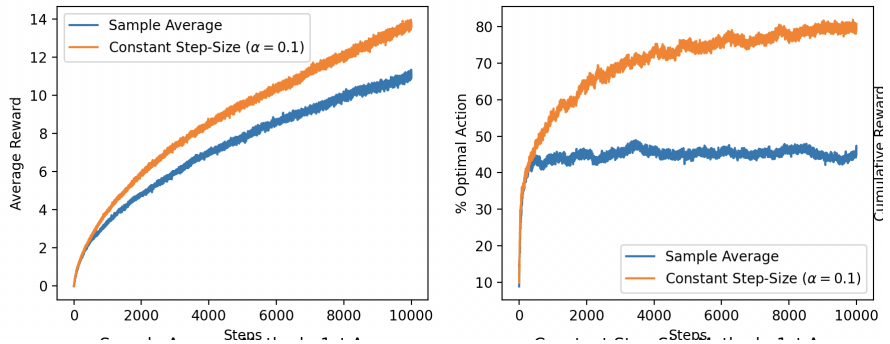

Figure 1: Comparative efficiency in obtaining rewards and learning efficacy in identifying optimal actions for the two methods.

```
T-stat: -37.356014799533504
P-Value: 2.6677665433809505e-295
Reject Null Hypothesis: There is a significant difference in rewards between the methods.
```

Figure 2: Results of t-test to assess the significance of the difference in average rewards between the two method

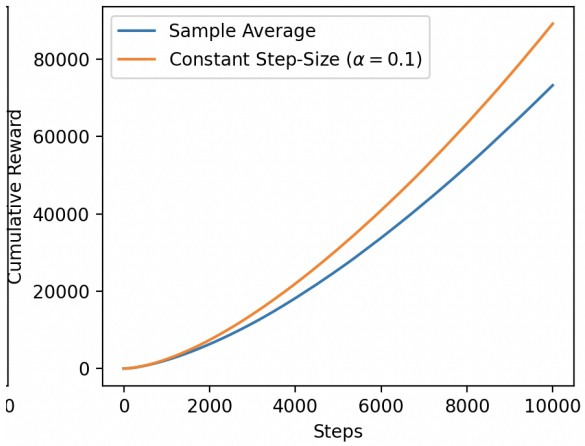

Figure 3: Illustration of the overall reward accumulation for comparative performance assessment for the two methods.

Figure 3 encapsulates the cumulative rewards of both methods, elucidating the diverging trajectories in reward accumulation rates. The discrepancy in accumulation rates between the sample average and the constant step-size methods becomes progressively pronounced as the experiment unfolds. That is, over the course of the experiment, the rate at which the step-size method accumulates rewards becomes increasingly smaller than the rate at which the constant step-size method accumulates rewards.

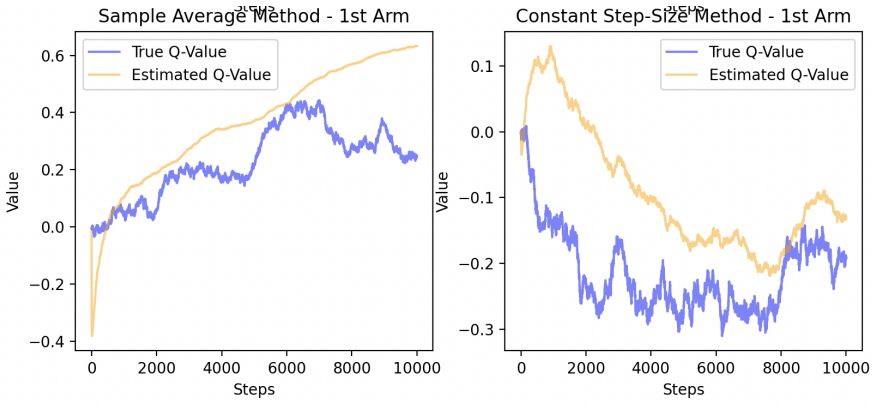

Figure 4: Illustration of the learning accuracy and adaptiveness of the methods.

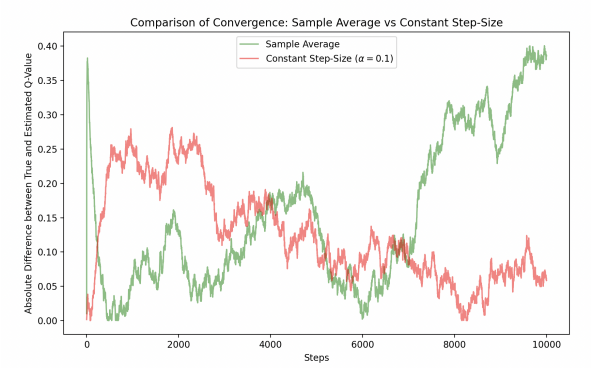

Figure 5: Representation of the absolute difference between true and estimated Q-values over steps for the two methods.

```
Mean Absolute Difference for Sample Average: 0.15316638968934726
Mean Absolute Difference for Constant Step-Size: 0.12617251758416417
```

Figure 6: Mean absolute difference between the true and estimated Q-values as an aggregate measure of the learning accuracy of the methods over the entire run.

In Figure 4, the congruence and discrepancies between true and estimated q-values are delineated, showcasing the learning precision, convergence rate, and adaptiveness of each method. It unveils the intermittent convergence of the sample-average method's estimates to the true values, interspersed with phases of substantial divergence, whereas the constant step-size method illustrates a consistent convergence trajectory. Figure 5, accentuating the absolute difference between the true and estimated q-values, reinforces the consistent convergence of the constant step-size method in contrast to the fluctuating trajectory of the sample average method. Comprehensively, the mean absolute difference calculated (Figure 6) corroborates the superior learning accuracy of the constant step-size method throughout the experiment. This analysis pertains exclusively to the instances involving the first lever of the bandit.

4 Conclusion

The results of this experiment accentuate the inherent deficiencies of the sample average method in non-stationary environments. Further, the comparative analysis with the constant step-size method, as elucidated through Figures 1 to 6, corroborates the hypothesis that the sample average method faces significant limitations in dynamic settings. The experimental findings demonstrate that, while

the sample average method may exhibit initial adaptability and learning efficiency comparable to the constant step-size method, it quickly plateaus in its learning curve, indicating a substantial reduction in its ability to optimize rewards and select optimal actions as the environment evolves. This stagnation is in stark contrast to the constant step-size method's performance. The constant step-size method was persistent in its improved performance and adaptability over the course of the experiment, thereby demonstrating its resilience and effectiveness in dynamic learning landscapes. The confluence of true and estimated q-values over time further reinforces the notion that the sample average method struggles to learn accuracy and adaptiveness in non-stationary environments, especially because this method exhibited episodes of random significant divergence from true values.

5 Extensions

To better understand the limitations of the sample average method in non-stationary environments, the experiment described here could be extended in several ways. First, future research could include experimenting with environments where the degree of non-stationarity varies. This would mean implementing different rates and magnitudes of change within the environment to assess how well the sample average method adapts to fast, slow, minor or major changes. Second, a more extensive comparison of the efficiency and adaptability of the sample average method as compared to other estimation methods in non-stationary environments might help to highlight specific deficiencies in the sample average method. Third, future investigations could examine the method's performance in a diverse range of non-stationary environments with different characteristics. This analysis might include environments with varying degrees of noise, fluctuating reward structures, and/or irregular patterns of change.

6 Secondary Question, Experiment, and Answer

As a secondary investigation, another question was posed: does the performance of the constant step-size method break-down as the environment becomes more non-stationary?

To answer this question, examination of the constant step method's performance within environments characterized by varying degrees of non-stationarity was performed. Specifically, the performance of the constant step method was evaluated in various non-stationarity environments, simulated by manipulating the volatility of the reward distribution. Environments with volatility levels of 0.01, 0.1, and 1 were tested. These varied levels of volatility are representative of different degrees of environmental change and reward unpredictability.

For each volatility level, 100 experiments, each comprising 1000 steps, were executed (additional experiments with a larger number of steps should be conducted in order to generate more robust results). Within each step, the constant step method, with a predetermined step-size $\alpha = 0.1$, was applied, and the obtained rewards were compared against the environment's changing reward distributions. The average rewards were subsequently calculated and compared across different volatility levels.

The results of the experiment exhibit a clear pattern demonstrating the influence of environmental volatility on the performance of the Constant Step Method.

When the environment is relatively stable, denoted by a low volatility of 0.01, the constant step method performs well, yielding a positive average reward of approximately 0.0017. This suggests that in environments with minimal fluctuations in the reward distribution, the method is effective in estimating rewards.

However, as the volatility of the environment increases to 0.1 and 1 a noticeable decline in performance is observed, with the average rewards decreasing to approximately negative 0.3123 and negative 0.3076 respectively. These negative values indicate that in more non-stationary environments, where the reward distributions are subject to more frequent and substantial changes, the constant step method struggles to adapt effectively, leading to sub-optimal reward estimations.

7 References

Sutton, Richard S., and Andrew G. Barto. Reinforcement Learning: An Introduction. The MIT Press, 2018.