# COMP 138 RL: Homework 1

Mattia Danese

September 28, 2022

## 1 Motivation

A reinforcement learning agent has several components that it uses in order to pick its next action. One of these components is the agent's estimates of how optimal choosing each available action will be, known as its *action-value estimates*. In this report, a 10-Armed Bandit is used as the agent and the two *action-value estimate* methods under question are the *Sample Average* method and the *Constant Step-Size* method. The purpose of this experiment is to see how each *action-value estimate* method affects the performance of the agent in a non-stationary environment, that is an environment where the true values of each arm (i.e. action) changes in every time step.

## 2 Introduction

### 2.1 Sample Average Method

The Sample Average Method is a way for the Bandit to update its current estimate of the arm it chose by following the equation below.

$$Q_{n+1} = Q_n + \frac{1}{n}[R_n - Q_n]$$

Basically, the next estimate of some arm, $Q_{n+1}$, is based on the current estimate of that arm, $Q_n$, plus some discounted error. This error is composed of the reward the agent got from the arm, $R_n$, minus the current estimate of the arm, $Q_n$, scaled by the inverse of the amount of times that arm has been chosen by the Bandit, $\frac{1}{n}$. It is important to note that, due to the how the error term is being scaled, there is an inherent bias towards the earlier rewards of each arm. That is, later arm rewards are going to count less in the update of the Bandit's estimate for that arm. Thus, the purpose of the error is to initially have more drastic updates of the Bandit's estimate for that arm in the beginning, and then gradually have more and more subtle updates as the arm is chosen more.

### 2.2 Constant Step-Size Method

The Constant Step-Size Method is derived from the Sample Average Method and follows the equation below.

$$Q_{n+1} = Q_n + \alpha[R_n - Q_n]$$

In this method, the Bandit again updates its estimate of some arm, $Q_{n+1}$, based on the sum of the current estimate of that arm, $Q_n$, and some error. However, the Constant Step-Size method differs from the Sample Average method in the way in which the error term is calculated. In this method, the error term is calculated by scaling the difference between the reward of the arm, $R_n$, and the Bandit's current estimate of the arm, $Q_n$, by some constant value, $\alpha$. This makes it so the number of times an arm is chosen has no affect on the arm's new estimate. In other words, all arm rewards are valued equally, no matter *when* they occurred, and thus there is no bias, or preference, towards earlier rewards.

### 2.3 Experiment Parameters

In this experiment, each action-value method will be used by 2000 10-Armed Bandits where each Bandit will run for 10,000 time steps. In order to track the performance of both action-value methods, the reward per time step and percentage of optimal arm chosen per time step will be tracked and then averaged across all 2000 10-Armed Bandits for the respective action-value method. Additionally, the environment of every Bandit will be non-stationary, thus a normally distributed increment, with mean 0 and standard deviation 0.01, will be added to the true value of each arm in every time step. The results are showcased and analyzed below.

## 3 Results

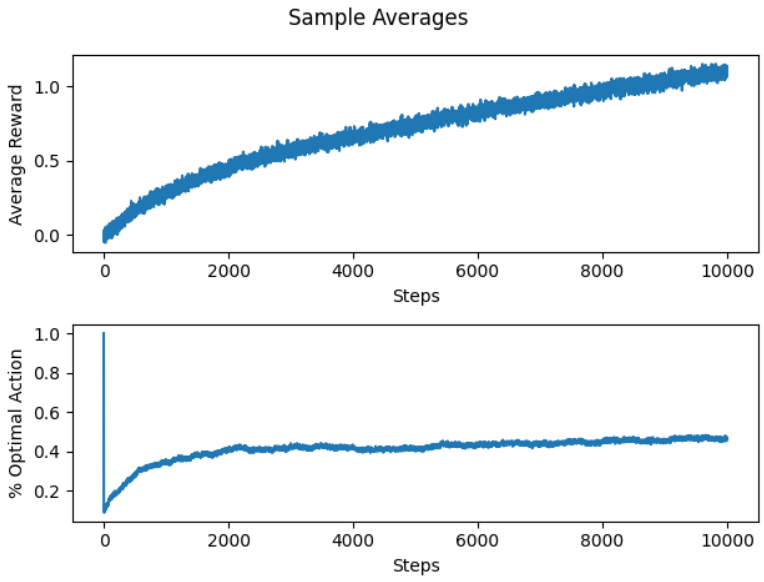
### 3.1 Sample Average

The figures below showcase the results of this experiment when the 10-Armed Bandits had the Sample Average method as their action-value estimate method.

The top figure shows the average reward at every time step, across all 2000 Bandits. It is clear that the Bandits did learn which arms to choose as more time steps passed; this is shown by a constant increase in the average reward as the time steps increase. One thing to note is that the beginning of the graph, say between time step 0 and time step 2000, has a much steeper slope than the rest of the graph. This could be attributed to drastic early arm estimate updates, relative to later arm estimate updates, since all arm estimates started at 0 and gradually got closer to the true arm values. Similarly, the slope at the end of the graph does seem relatively close to a slope of 1 (i.e. a horizontal line) though not close enough to deem that the graph has converged. However,

the overall slope of the graph is getting less positive, especially relative to the beginning of the graph. Thus, given more time steps, it seems like the average reward will converge somewhere between 1.0 and 2.0. Lastly, this graph is fairly thick relative to the respective graphs of stationary environments, though this is most likely due to the non-stationary nature of the environment and the variance introduced by changing the true values of each arm at every time step.

The bottom figure shows the percentage, at each time step, of times when the optimal arm was chosen, across all 2000 Bandits. It is clear that, starting at around the 2000th time step, the percentage of times when the optimal arm was chosen converges to roughly 40%. This might be considered fairly low, however this could be attributed to the environment being non-stationary. It very well may be difficult for the Bandits to frequently choose the optimal arm when, at every time step, the true value of each arm changes as this, in turn, could change which arm is the optimal one. Leading up to the 2000th time step (and eventual convergence), it is clear that the rate at which the Bandits chose the optimal arm is increasing. This is most likely due to the initial conditions of the arm estimates not being close to the true arm values but gradually get closer to them. It should also be noted that the great spike at the very beginning of this figure is due to the true values of each arm starting out at 0. Therefore, whichever arm the Bandit chooses on the first time step, it will always be one of the "optimal" arms since all arms are tied for being optimal.

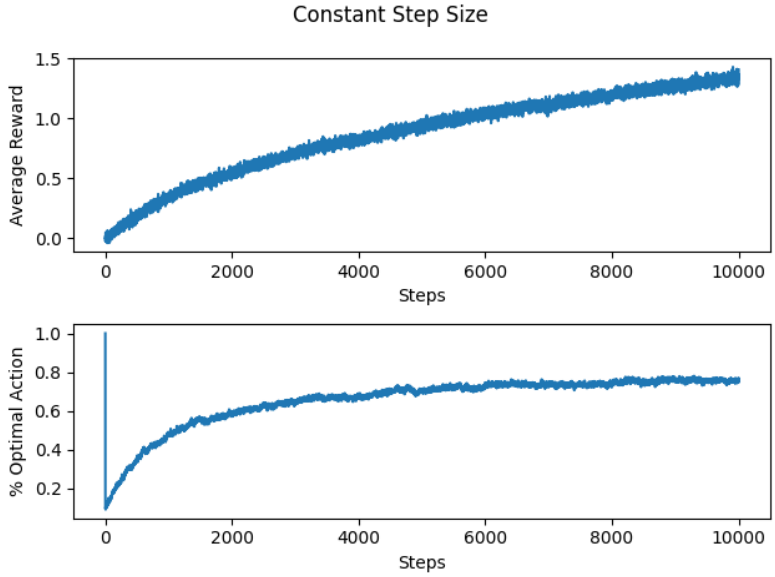


### 3.2 Constant Step-Size

The figures below showcase the results of this experiment when the 10-Armed Bandits had the Constant Step-Size method as their action-value estimate method.

The top figure shows the average reward at every time step, across all the 2000 Bandits. Much like the Bandits that implemented the Sample Average action-value estimate method, it is clear that these Bandits also learned which arms to choose as the average reward is constantly increasing as the time steps increase. By 10000 steps, the average reward seems to be approaching 1.5 and the slope at the end of the graph appears to be approaching a horizontal line. Thus, it can be inferred that the average reward will converge to 1.5, or to some value relatively close to 1.5. The graph also exhibits an initial steep slope similar to that of the average reward graph of the Sample Average Bandits; this is most likely due to the same reasoning as previously stated. Lastly, this graph is also fairly thick, though this is once again attributed to the variance introduced by changing the true values of each arm at every time step, just as it previously was with the Sample Average Bandits.

The bottom figure shows the percentage, at each time step, of times when the optimal arm was chosen, across all 2000 Bandits. It is evident that, around the 6000 time steps, the percentage of times when the optimal arm was chosen converges to just under 80%. Statistically speaking, this percentage should converge to 90%, since the probability of the Bandit exploring instead of exploiting is 0.10. This discrepancy may be due to the non-stationary environment; changing the true value of each arm at every time step may be throwing off the Bandit if the current optimal arm changes to no longer being optimal because the Bandit's arm estimates will not immediately reflect this change. Additionally, it is evident that the Constant Step-Size Bandits experienced the same setbacks as the Sample Average Bandits due to nonoptimal initial conditions of the arm value estimates. This can be seen by the sharp increase in the rate of which the optimal arm was chosen leading up to 2000 time steps. Lastly, there is, once again, a great spike at the very beginning of the graph but this is due to the same reasoning previously stated for the Sample Average Bandits.




### 3.3 Comparison

Based on the two figures above, it can be concluded that the Constant Step-Size Bandits performed better, in this environment, than the Sample Average Bandits. The Constant Step-Size Bandits reached a greater average reward after 10000 time steps and had double the percentage of choosing the optimal arm at every time step relative to the Sample Average Bandits. The reasoning for this may be that the Constant Step-Size Bandits are more adaptive to their environment, as they weigh each reward equally, and therefore are better suited for this non-stationary environment. In other words, the Sample Average Bandits value earlier rewards more than later rewards (on a per arm basis), thus if the true values of the arms change and the earlier rewards are weighed more than current and future rewards, then the Bandits' arm estimates will be focused more towards the older true arm values and will have increasing trouble adjusting its estimates as the time steps increase since each update's error term will be weighed less and less. Similarly, the thickness of the average reward graph for the Constant Step-Size Bandits is less than that of the Sample Average Bandits. This may be another indication that the Constant Step-Size method is better at mitigating the variance introduced by the non-stationary environment than the Sample Average method.

# 4 Further Experiment: Varying $\epsilon$

## 4.1 Introduction

Another component that is crucial in an agent's process of choosing a next action is its $\epsilon$, the probability of the agent *exploring* (picking an available action at random) as apposed to *exploiting* (picking what it thinks is the optimal action based on its action estimates). The value for $\epsilon$ must be chosen wisely because a value too low will result in the agent possibly not finding the optimal action or taking more time than is preferred to do so and a value too large will result in the agent exploring unnecessarily, not choosing the optimal action as much as is preferred, and therefore prolonging its convergence to the optimal policy. In the previous experiment, the Bandits had their $\epsilon = 0.1$, so the Bandits always had a 10% chance of exploring, and picking a random arm, no matter the time step. Intuitively, the longer a Bandit has been running, the less it should have to explore and the more it should exploit what it thinks is the optimal arm, since the Bandit's estimate should theoretically become more accurate the more time steps pass. As such, in this next experiment the $\epsilon$ of each Bandit will again start at 0.1 but will decrease by 50% every 10% of the time steps (i.e at the 1000th, 2000th, 3000th, etc. time step).

## 4.2 Results

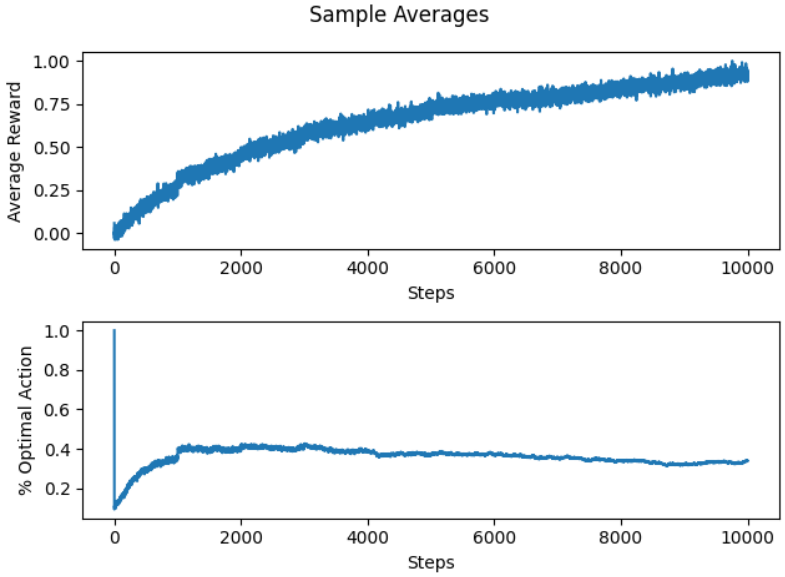
## 4.3 Sample Average

The figures below showcase the results of this experiment when the 10-Armed Bandits had the Sample Average method as their action-value estimate method and a varying $\epsilon$ that decreased by 50% every 1000 time steps.

The top figure shows the average reward at every time step, across all the 2000 Bandits. The graph clearly shows that the Bandits did learn which arms were optimal and which weren't as the average reward increases with the number of time steps. The graph also appears to have the same initial steep slope, relative to the rest of the graph, which has been discussed in the prior experiment. Additionally, the end of the graph does not seem to be relatively flat, therefore the average reward did not converge. Surprisingly, the overall shape of the graph looks similar to that of $f(x) = x^3$, where there is a steeper slope in the beginning, at flatter part in the middle (time steps 5000-7000), and steeper slope at the end. Lastly, the thickness of the graph is fairly similar to that of the graphs in the previous experiment and is most likely caused by the same reasons.

The bottom figure shows the percentage, at each time step, of times when the optimal arm was chosen, across all 2000 Bandits. The graph seems to converge at around 40% between time steps 1000 and 3000, but then gradually declines from time step 4000 onwards. This is most likely due to the fact that $\epsilon$ is inversely proportional to the number of time steps and the effect of the non-

stationary environment. In other words, at around the 5000th time step the probability of the Bandits exploring was too small to keep up with the changing true arm values, thus the Bandits' kept choosing what they thought was the optimal arm but no longer was. Additionally, the same initial spike and initial steep slope are again present and occur for the same reasons previously stated in the prior experiment.



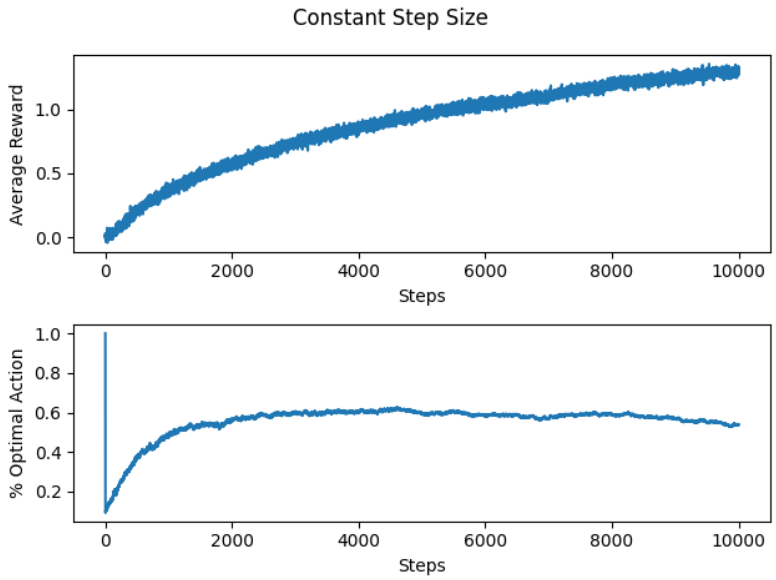

## 4.4 Constant Step-Size

The figures below showcase the results of this experiment when the 10-Armed Bandits had the Constant Step-Size method as their action-value estimate method and a varying $\epsilon$ that decreased by 50% every 1000 time steps.

The top figure shows the average reward at every time step, across all the 2000 Bandits. It is clear that the Bandits did learn which arms were optimal and which weren't as the average reward increases with the number of time steps. The graph also appears to have the same initial steep slope, relative to the rest of the graph, which has been discussed in the prior experiment. Additionally, the end of the graph does not seem to be relatively flat, which indicates that the average reward has not converged yet. That being said, it seems reasonable to assume that average reward would converge somewhere between 1.0 and 2.0. A final thing to note is the thickness of the graph is fairly similar to that of the graphs in the previous experiment and is most likely due to the same reasoning.

The bottom figure shows the percentage, at each time step, of times when the optimal arm was chosen, across all 2000 Bandits. At around the 4000th

time step, the graph appears to converge at roughly 60%, but then, as the time steps continue to increase, the percentage starts to decline. This decline can be attributed to the same reasoning stated for the decline in the corresponding graph of the Sample Average Bandits in this experiment. On another note, the same initial spike and initial steep slope are again present and occur for the same reasons previously stated in the prior experiment.




## 4.5 Comparison

Based on the two figures above, it can once again be concluded that the Constant Step-Size Bandits performed better than the Sample Average Bandits. The Constant Step-Size Bandits reached a greater average reward after 10000 time steps and consistently had better percentages of choosing the optimal arm throughout all time steps. That being said, varying $\epsilon$ in this manner most definitely impaired both the Sample Average and the Constant Step-Size Bandits. For both types of bandits, a nominal decrease (i.e. a down shift) in average reward and percentage of choosing the optimal arm occurred. Even worse, after roughly 4000-6000 time steps, the percentage of choosing the optimal arm started to decrease for both types of bandits. As stated in the analysis of the results for the Average Sample Bandits, the active decrease in percentage is most likely caused by an $\epsilon$ that is too small which causes the Bandit to keep choosing the arm it thinks is optimal but is no longer optimal due to the non-stationary environment. That being said, varying $\epsilon$ did not have other effects; the thickness of the average rewards graphs and initial spike and steep slope of the percentage graphs remained relatively unchanged.

# 5 Conclusion

In this report, two experiments were explored. In the first experiment, the performance of Bandits with the Sample Average action-value estimate method and Bandits with the Constant Step-Size action-value estimate method in a non-stationary environment were compared. It was evident that the Constant Step-Size Bandits had a greater average reward and percentage of choosing the optimal arm. It was discussed that having an action-value estimate method scale the error term by the number of times an arm was chosen was the ultimate downfall of the Sample Average Bandits. This caused the arm estimates of the Sample Average Bandits to be more tailored toward the earlier true values of the arms; as opposed to the Constant Step-Size action-value estimate method which values each error term equally and is therefore better suited to adapt to the changing environment.

In the second experiment, the performance of both types of Bandits in a non-stationary environment were again compared, but this time their $\epsilon$, the probability of a Bandit exploring a random arm instead of exploiting what it thinks is the optimal arm, varied as the time steps increased. More specifically, $\epsilon$ decreased by 50% every 1000 time steps since, intuitively, a Bandit should explore less the more time steps pass. However, it was observed that varying $\epsilon$ in this manner was detrimental to the performance of both types of bandits. Both types of bandits had a lesser average reward and percentage of choosing the optimal arm than when they had a constant $\epsilon = 0.1$. What was most surprising, though, was that the percentage of both types of bandits started to decline roughly half way through the time steps. It was then discussed that this can be attributed to the non-stationary environment and a $\epsilon$ too small for the agent to explore and find the new optimal arm.