COMP 138 RL: Programming Assignment 1

Catherine Ding

September 30, 2020

1 Introduction

The k-armed bandit problem is a popular reinforcement learning problem where one is faced with k possible actions, each with a expected or mean reward. The bandits in this problem are analogous to slot machines with with k number of levers. Each action selection is a pull on a lever arm and each reward is the profit gained from the slot machine pull. The objective of the reinforcement learning agent in this scenario is to maximize the expected total reward or payoff gained over a number of action selections or time steps.

A main idea related to the k-armed bandit problem is the issue of exploration vs. exploitation. Bandits each have an action-value denoting the expected reward given an action is selected, q_*. Assuming we do not know these action-values, we can have estimates that are hopefully close to q_*, denoted as $Q_t(a)$, where t is the time step and a is the action. To exploit the knowledge of current estimates and obtain the largest immediate reward, an agent would select a greedy action, or the action with the greatest estimated value. On the other hand, if the agent chooses a non-greedy action, the agent is said to be exploring. In this assignment, we use the ϵ-greedy method to balance this conflict between exploration and exploitation. In this method, the agent chooses a non-greedy action with probability ϵ, and a greedy action with probability $1 - \epsilon$.

Another main idea of this exercise is that of stationary vs. non-stationary reinforcement learning problems. With stationary problems, the reward probabilities do not change over time. Often, there are non-stationary problems where we may need to consider giving more weight to more recent rewards than earlier ones. An example of a non-stationary problem is in a game of chess, if one action is repeatedly chosen, your opponent may adapt or change their playing style. In this exercise we compare action-value methods using sample averages and using a constant step size parameter.

2 Problem

In short, this problem asks for us to design and conduct an experiment to demonstrate the difficulties that sample-average methods have for non-stationary problems. Since sample-average methods will update an arm's estimates for the action-value by summing up total reward received from the arm and dividing by the total number of times the arm was pulled. We will use another action-value method using a constant step-size parameter in comparison to the sample-average methods. In the method using the constant step-size parameter, more weight is given to recent rewards than to earlier rewards. The method to go about solving this problem is outlined in the following section.

3 Experiment

In order to demonstrate the short-comings of these methods, it is suggested within the exercise statement that we are to use a 10-armed testbed (10 bandits) where all the $q_*(a)$ start out equal and take independent random walks. For one experiment, I will use the method suggested in the problem: adding a normally distributed increment with mean zero and standard deviation 0.01 to all the $q_*(a)$ on each step. Note that this is a very small, gradual change to the true reward distribution each step. In my second experiment, I will use the same method, but with standard deviation of 0.1. This will change the action-values much more drastically than before and hopefully illustrate the results in a consistent but a more exaggerated manner. For both these experiments, I will use use ϵ-greedy action selection, where $\epsilon = 0.1$. I will also perform 2000 runs each with 10,000 steps.

Both action-value methods using sample-averages and a constant step-size parameter will be tested and compared in the two experiments. For methods using sample-averages, I will be updating the estimates incrementally using $Q_{n+1} = Q_n + \frac{1}{n}[R_n - Q_n]$. For methods using a constant step-size parameter, I will be updating using $Q_{n+1} = Q_n + \alpha[R_n - Q_n]$, where the step-size parameter, $\alpha = 0.1$.

At the end of each experiment, I will plot two plots similar to the plots pictured in Figure 2.2 from the text [1]. One plot will display average reward on the y-axis and the other will display percent optimal action taken on the y-axis. Both plots will have number of steps on the x-axis and contain values that are averaged over the 2000 runs. When evaluating my plots, the action-value method with higher average reward and higher percent optimal action will mostly likely be the action-value method which acts the most favorably in non-stationary problems.

4 Hypothesis

Given that sample-average methods will only take the average reward by dividing total reward from a bandit by number of times the bandit was selected, it does not distinguish between whether the reward came from a more recent action or a much earlier action. In a non-stationary problem, it is especially important that later rewards have a larger weight on the estimates than earlier rewards since the true action-values are constantly changing. If each reward is equally weighted when calculating the estimate, when the current true action-values are much different than what they started out to be, the rewards obtained in the beginning are still a heavy factor in calculating the estimate, thus will not result in an accurate estimate and will impact the greedy action selections in the future negatively. On the other hand, for the action-value method using constant step-size parameter, more weight is given to recent rewards, thus as the reward distribution changes, it is able to adapt far better by counting the later rewards as more relevant to the estimate than earlier rewards.

In terms of the plots, for the Average Reward vs. Time-Step plot, I believe the methods using a constant step-size parameter will perform much better in both the cases where the random walk had a standard deviation of 0.01 and 0.1. That is, it will receive a greater average reward as the time-steps increase. However, since in the second experiment the random walks have a much more sudden change to the mean, the overall rewards for the methods tested will be much more, thus average reward of the gradual vs. sudden random walks for each method should not be directly compared. Similarly for the Percent Optimal Action vs Time-Step plot, I predict that the methods using a constant step-size parameter will have a much higher percentage of taking the optimal action regardless of the standard deviation of the random walks.

5 Results

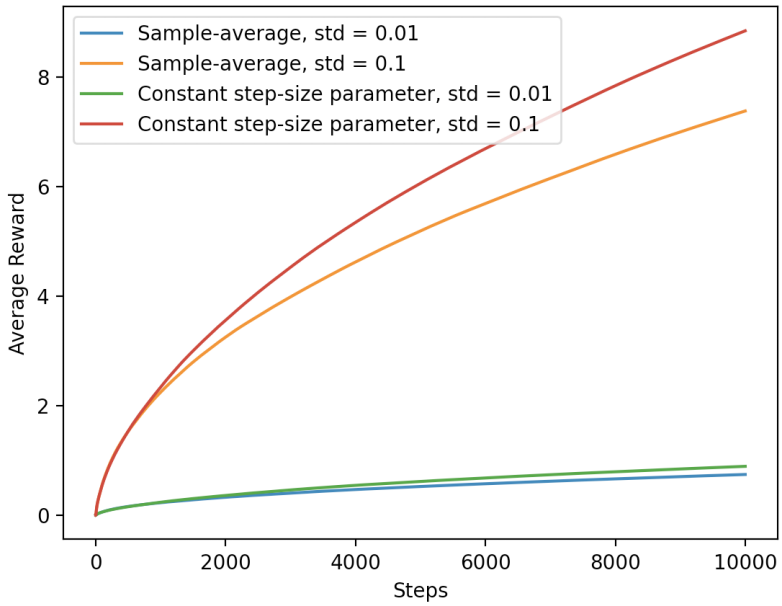


Figure 1: Average rewards over 2000 runs for 10-arm non-stationary bandit problem with ϵ-greedy action-value methods using sample-averages and a constant step-size parameter as their action-value estimates. Independent random walks with a normally distributed value with mean zero and standard deviation 0.01 and 0.1 taken after each step.

In Figure 1, we must look at the two different random walk standard deviations separately. In the gradual random walk case, where standard deviation is 0.01, we see that over 2000 runs of 10,000 steps, the action-value method using a constant step-size parameter (blue) has a greater average reward than the method using sample-averages (green). This is in accordance to my hypothesis. Also in accordance with my hypothesis were the results with the more sudden random walk with standard deviation of 0.1 as well. Looking at the upper two lines, the method using a constant step-size parameter (red) in this case also is shows to have greater average reward over the method using sample averages (yellow). As mentioned in my hypothesis, the two experiments with two different random walks should not be directly compared with each other, since with the random walk with greater standard deviation, the overall reward was greater, which is to be expected with greater variance.

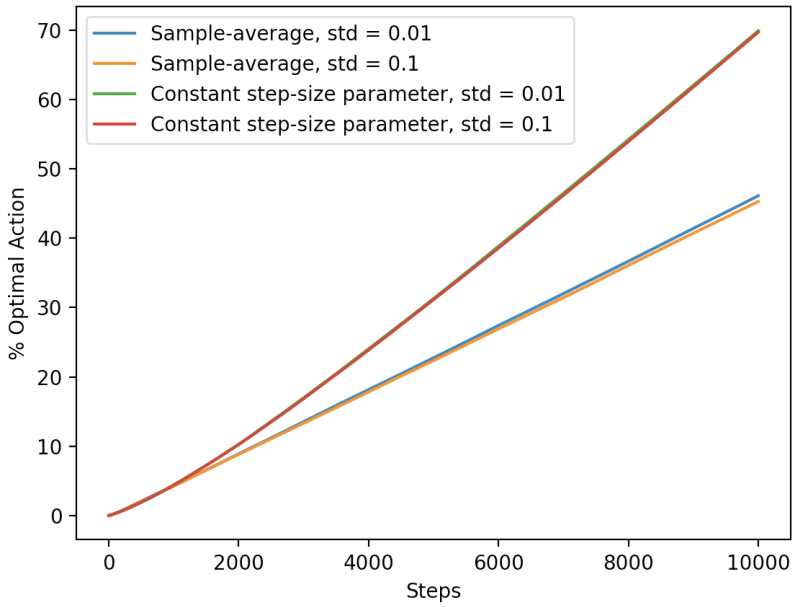


Figure 2: Average percent optimal actions taken over 2000 runs for 10-arm non-stationary bandit problem with ϵ-greedy action-value methods using sample-averages and a constant step-size parameter as their action-value estimates. Independent random walks with a normally distributed value with mean zero and standard deviation 0.01 and 0.1 taken after each step.

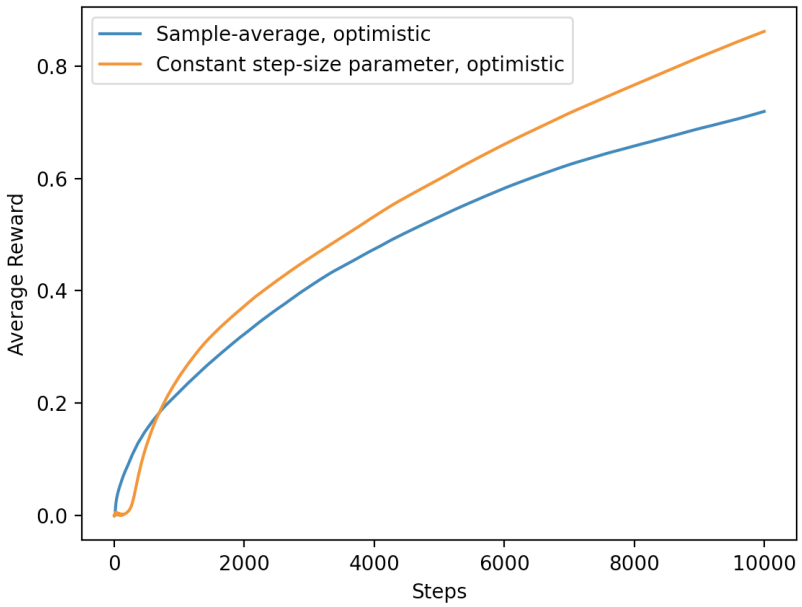
Unlike with Figure 2, the four lines in figure two can be compared with one another, since the variance of random walks only impact the reward, not the percent of time the agent picks the optimal action. This is shown in the plot as there is nearly no distinguishable difference between the same methods with different random walk variances, as the lines clearly overlap significantly. We can see from Figure 2 that both methods using a constant step-size parameter have percent optimal actions that are significantly higher than the methods using sample averages. While at the end of the 10,000 steps, the methods using s constant step-size parameter was choosing the optimal action around 70% of the time, the methods using sample averages were around the 40% to 50% range. These results are also in accordance to my hypothesis.

6 Conclusion

As we can see from the results, in both types of random walks, one with gradual change and one with sudden change, the action-value methods using a constant step-size parameter always had greater average reward and a greater percent optimal action. These results agreed with my hypothesis and makes sense intuitively, since if the reward distributions change, more weight should surely be put on the later rewards and less on those closer to the beginning. Due to the methods using a constant step-size parameter weighting the more recent rewards, it is much better at adapting to the changing distribution and thus are able to maximize the rewards much more effectively than the methods using sample averages.

7 Extensions

I briefly experimented with optimistic initial values. Instead of initializing the bandits' estimates to 0, I initialized them to 5.



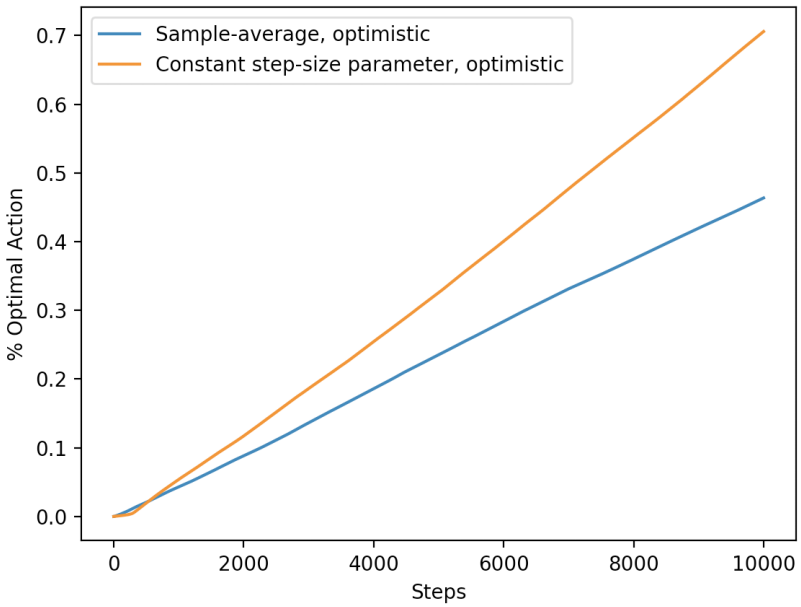


Figure 3: Average rewards and percent optimal action over 200 runs for 10-arm non-stationary bandit problem with ϵ-greedy action-value methods using sample-averages and a constant step-size parameter as their action-value estimates. Optimistic action-values were were initialized to 5.

As you can see, the results of the data in these plots are consistent with the conclusions formed based on previous plots, that is, the methods with a constant step-size parameter is still better at adapting and maximizing rewards in a non-stationary problem. In the very beginning, a dip can be noticed in the line for constant step-size parameter (yellow), this shows that the effect of the optimistic estimate caused the initial reward to be lower than the methods with sample averages, but it quickly adjusts and results in our earlier conclusion.

References

[1] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2017.