Reinforcement Learning

Homework 1

Reykjavik University

Teacher: Stephan Schiffel

Fall 2023

Þórir Hrafn Harðarson – thorirhh21@ru.is

Q1

An environment was implemented that uses a 10-armed bandit testbed where the true values of the actions (i.e., the mean of the reward distribution) changes slowly over time. To observe the process of learning in a non-stationary environment two agents were used, one that uses sample averages, incrementally computed, and another that uses an action-value method with a constant step-size parameter, $\alpha = 0.1$. The results can be seen in the figure 1 below:

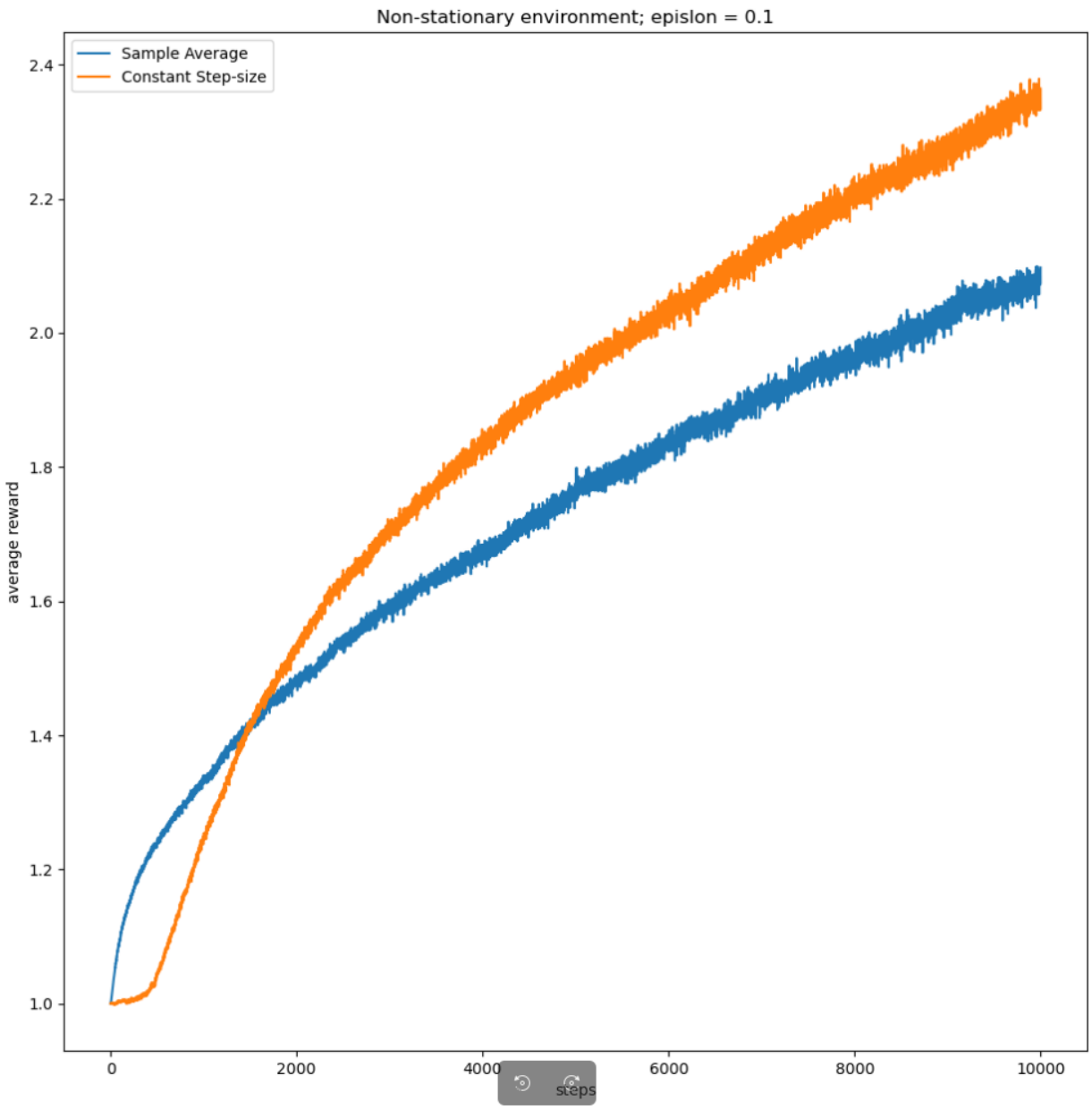

Figure 1 - Average rewards, one agent using sample average and another using constant step-size of α = 0.1

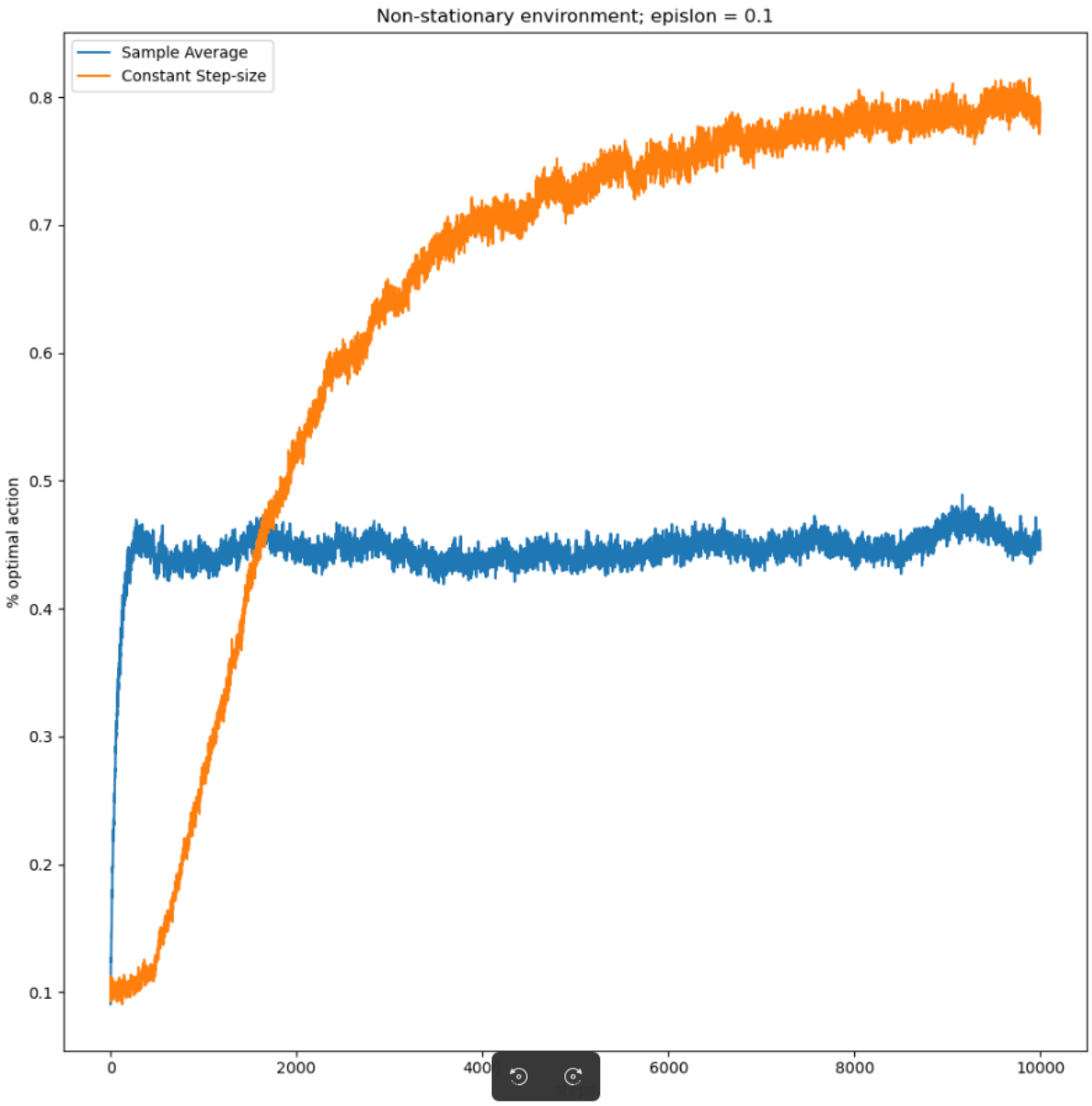

Figure 2 - Percentage of optimal actions, one agent using sample average and another using a constant step-size of α = 0.1

Looking at the results we can see that after a certain number of steps the agent with a constant step-size is able to adapt better to the changes in the environment and thereby learn better which action is most likely to result in the highest expected reward. The reason for this is that the constant step-size allows the agent to put more weight on recent rewards while the weighted average is forced to equally consider all past rewards and is therefore slower to adapt to recent changes in the environment. This can be seen in figure 2 where the weighted average agent quickly plateaus around 50% when looking for optimal actions while the step-size agent is able

to climb upwards and approach 80%.

Looking at the difference between the estimated values Qt(a) and the true values q*(a) over time it can be seen that with constant step-size the differences are greater in the beginning but as the agent adapts the difference begins to decrease quickly and then slowly approaches a value close to 0. On the other hand the agent that uses sample averages begins with a low difference between Qt(a) and q*(a) but as the environment changes it is slow to adapt and the difference starts to increase over time.

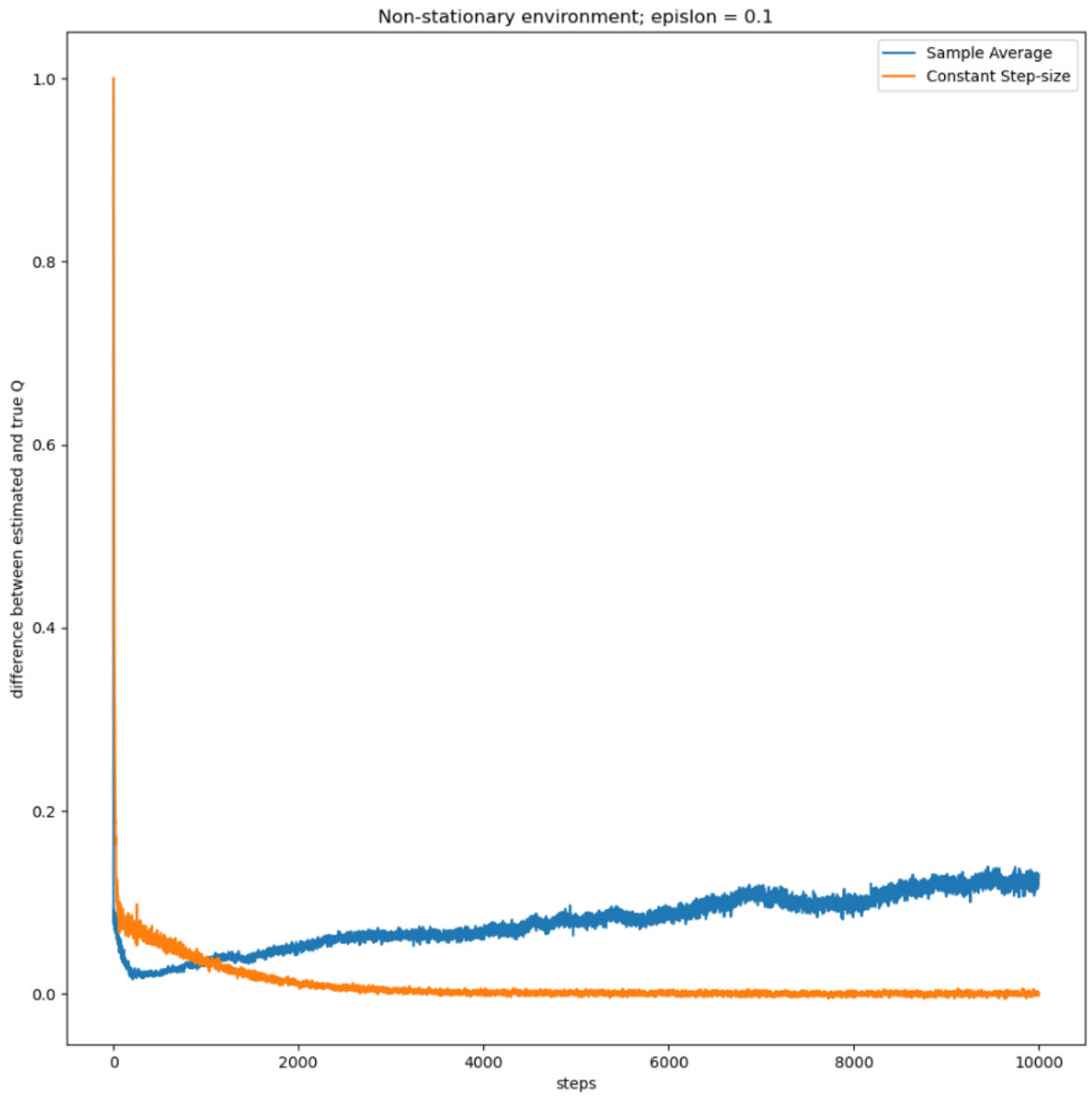

Figure 3 - Difference between estimated and true Q-values, one agent using sample average and another using a constant step-size of α = 0.1

To further investigate the relationship of the step-size parameter with how quickly a non-stationary environment changes, another experiment was set up comparing agents with step-sizes of $\alpha = 0.01$ and $\alpha = 0.3$. The experiment uses two non-stationary environments, one that is slowly changing (std-dev of 0.01) and one that is quickly changing (std-dev of 0.1). The results of this can be seen in figure 4 and 5 below:

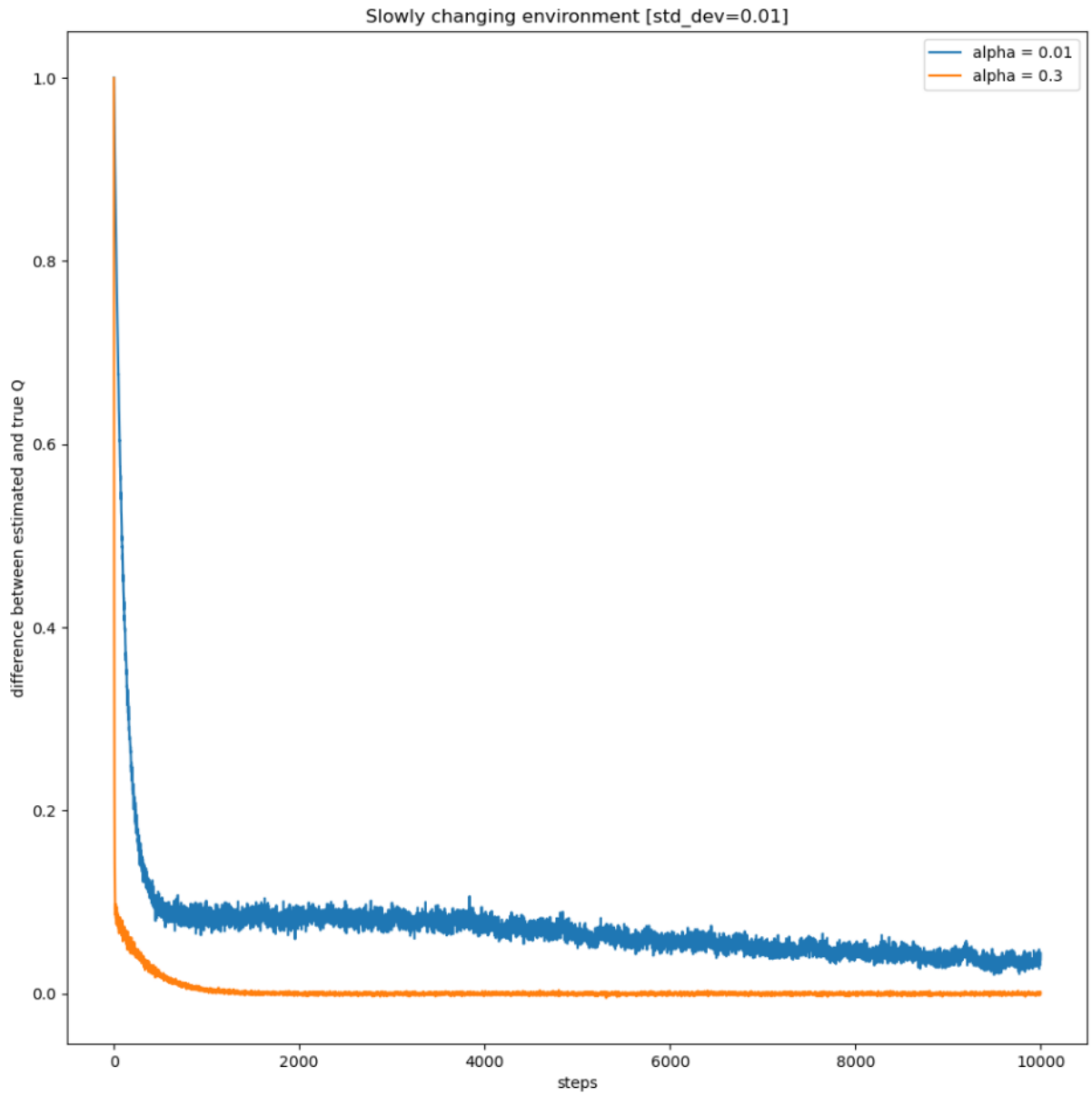

Figure 4 - Difference between estimated and true Q-values using constant step-sizes of $\alpha = 0.01$ and $\alpha = 0.3$ in a slowly changing environment.

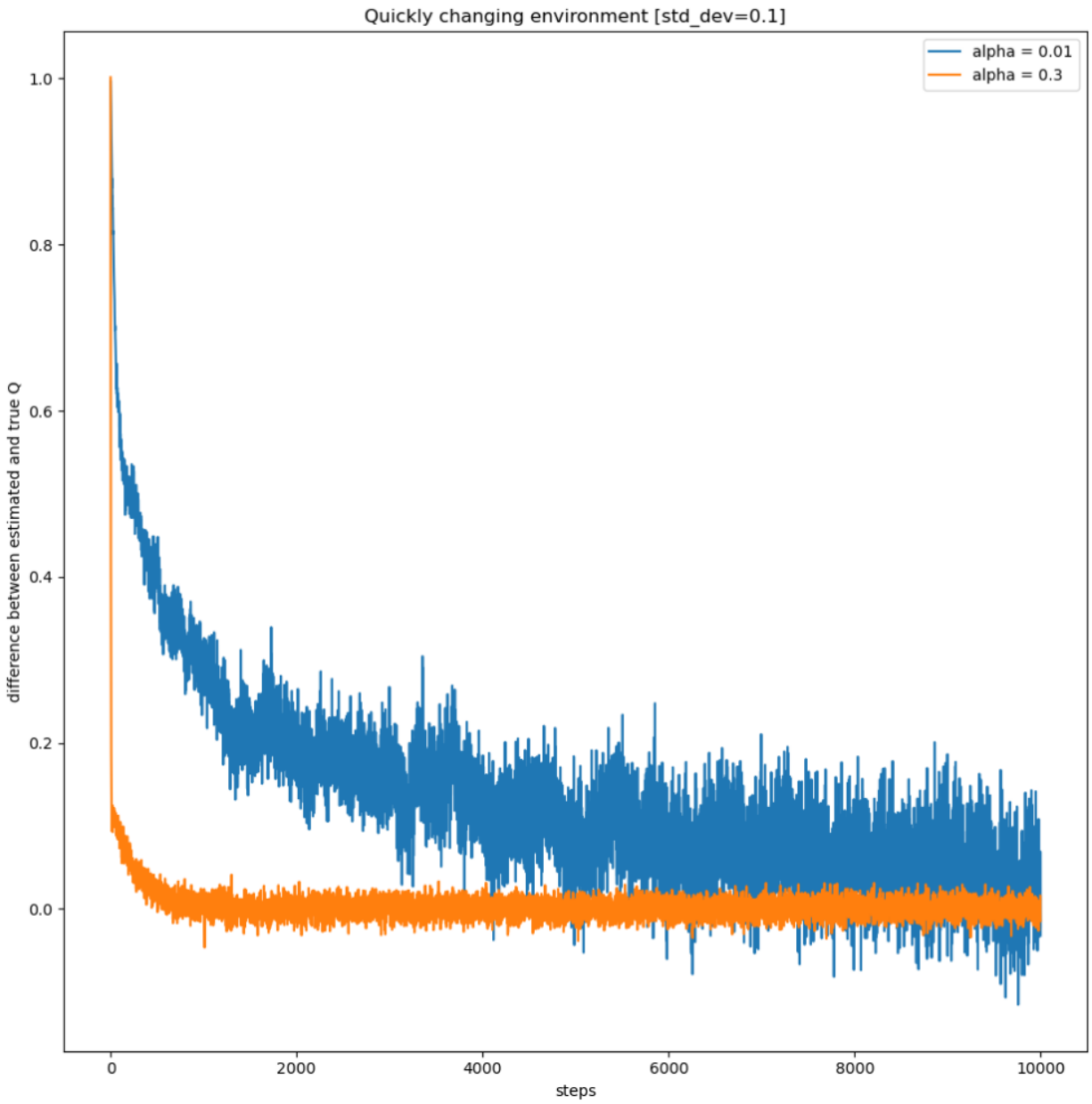

Figure 5 - Difference between estimated and true Q-values using constant step-sizes of α = 0.01 and α = 0.3 in a quickly changing environment.

Looking at the results we can see that both agents are able to adapt and learn, though the agent with the lower step-size is slower to learn and has a more difficult time when trying to adapt to changes in the environment, causing it to have more noise in its results. This is more pronounced when the environment is changing quickly.

Q2

Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 10 and 20 with probability 0.5 (case A), and 90 and 80 with probability 0.5 (case B).

1. If we cannot tell which case we are facing when choosing an action then choosing action 1 will result in an expected reward of:

   0.5 * 10 + 0.5 * 90 = 0.5 * 100 = 50.

   Choosing action 2 will result in:

   0.5 * 20 + 0.5 * 80 = 0.5 * 100 = 50.

   In this case we can just randomly pick an action since they will both result in the same expected reward.

2. If we are now told which case we are facing we should be able to learn what the true values of each action is for both cases. In that case we should choose to perform action 2 in case A and action 1 in case B, resulting in an expected reward of:

   0.5 * 20 + 0.5 * 90 = 0.5 * 110 = 55.