1 Problem Statement

Design and conduct an experiment to demonstrate the difficulties that sample-average methods have for non-stationary problems. Use a modified version of the 10-armed testbed in which all the q(a) start out equal and then take independent random walks (say by adding a normally distributed increment with mean zero and standard deviation 0.01 to all the q(a) on each step). Prepare plots like Figure 2.2 for an action-value method using sample averages, incrementally computed, and another action-value method using a constant step-size parameter, = 0.1. Use = 0.1 and longer runs, say of 10,000 steps.

Goal: design an experiment to demonstrate which action-value method has better performance for non-stationary 10-armed testbed problem.

2 Experiment Design

2.1 Introduction

Throughout the whole experimental process, two sets of comparative experiments were conducted and total six algorithms were implemented.

In the first part of the experiment, I combined the incremental sample average method with three different action selection methods, which are the greedy method, ϵ-greedy method, and greedy optimal initial method to run the non-stationary 10-arm testbed.

In the second part of the experiment, we choose the exponential recency-weighted(ERW) average method to combine with the same action selection methods used in the first part to run the same non-stationary 10-arm testbed.

2.2 Parameter Setting

All six algorithms were tested over 1000 runs, and each run has 10000 steps. I chose five as the initial estimated reward $Q_1(a)$ in the optimal initial greedy method. Besides, for the other algorithms, the initial $Q_1(a)$ was set to 0. The initial q*(a) for each arm was assigned to the same value 0.5. The ϵ value for ϵ-greedy method was set to 0.1. The step-size parameter α for the

ERW average was set to 0.1.

3 Related Knowledge

3.1 Sample-average method

The sample-average method is a way to estimate action values using an average sample of relevant rewards. To improve computational efficiency, the incremental formulas of the sample-average method was devised. Given Qn and the n_{th} reward, Rn, the new average of all n rewards can be computed by:

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

3.2 Exponential recency-weighted average method

Tracking non-stationary problems in which the reward probabilities change over time makes sense to give more weight to recent rewards than long-past rewards. One of the most popular ways of doing this is to use a constant step-size parameter α. We call it the exponential recency-weighted average method.

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \tag{2}$$

3.3 Greedy method

The greedy method is the most straightforward action selection rule to select one of the highest estimated value actions.

$$A_t \doteq \arg\ \max_a Q_t(a) \tag{3}$$

3.4 ϵ-Greedy method

The ϵ-greedy method is to behave greedily most of the time, but with small probability ϵ, instead, select randomly from among all the actions with equal probability.

3.5 Optimistic initial value method

Optimistic Initial Values is a sample way to encourage exploration. By selecting a large initial estimated reward $Q_1(a)$, the reward is less than the starting estimates; the learner switches to other actions, being "disappointed" with the rewards it is receiving. The result is that all actions are tried several times before the value estimates converge. The system does a fair amount of exploration, even if greedy actions are selected all the time.

4 Results Analysis

4.1 Part I: Sample Average

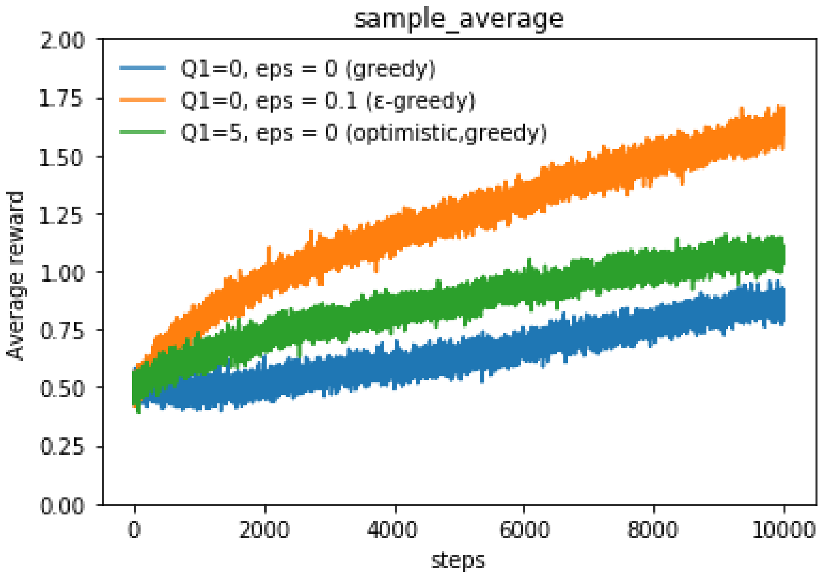

Figure 1: Average reward of three action selection methods on the 10-armed non-stationary testbed. All methods used sample averages as their action-value estimates. These data are averages over 1000 runs with 10-arm bandit non-stationary problem.

Figure 1 shows that under the premise of using the simple average method, the ϵ greedy method has better performance than the other two methods. After 10,000 steps of learning, its average reward can reach 1.75. Next is the optimistic greedy method; after learning, its average reward is around 1.0. The worst performance is method three, and its average reward is about 0.75.

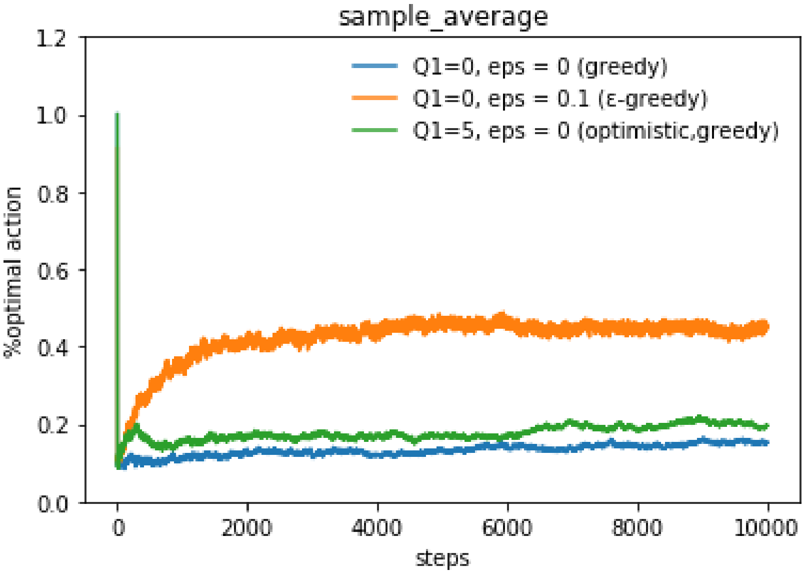

Figure 2: The percentage of optimal action of three action selection methods on the 10-armed non-stationary testbed. All methods used sample averages as their action-value estimates. These data are averages over 1000 runs with 10-arm bandit non-stationary problem.

Figure 2 shows the percentage of optimal action was selected among 1000 runs on each step. Since all q*(a) and Q(a) start equal, the first action for all the three methods should be the optimal action. That's why there is a spike on the first step for each of the three lines. Figure 2 also shows that no matter which action selection method we use, the percentage of optimal action selected is very low. The *epsilon*-greedy method got the highest value, which is only around 40%. For the other two, they only have about 15% and 20%.

4.2 Part II: ERW Average

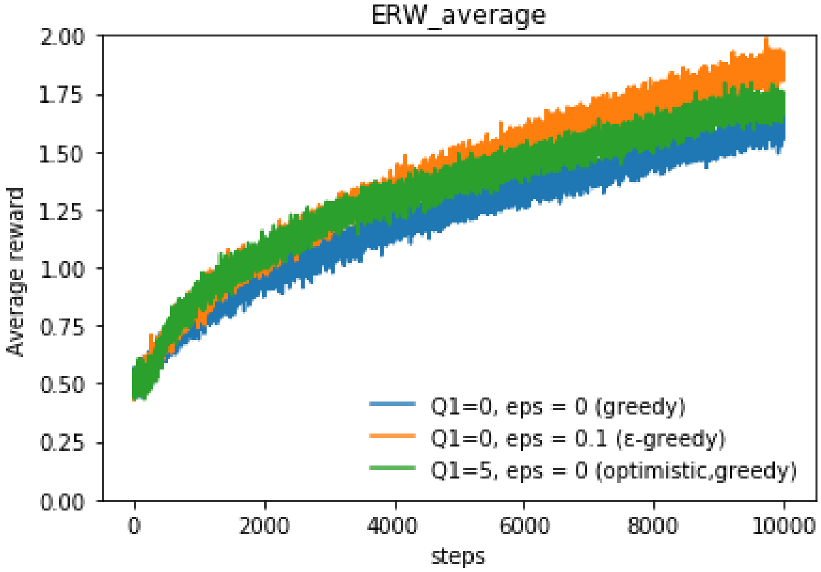

Figure 3: Average reward of three action selection methods on the 10-armed non-stationary testbed. All methods used ERW average as their action-value estimates. These data are averages over 1000 runs with 10-arm bandit non-stationary problem.

Figure 3 shows that, at first, the optimistic greedy method performed best among the three methods. When the learning progressed to about 4000 steps, the average reward of ϵ-greedy method starts to higher than the other two methods. After 10000 steps learning, the average reward of ϵ-greedy method was as high as 2.0. The second-ranked algorithm is the optimistic greedy method; its average reward is around 1.75. The worst performer is still the greedy method, with an average reward between 1.25 and 1.5.

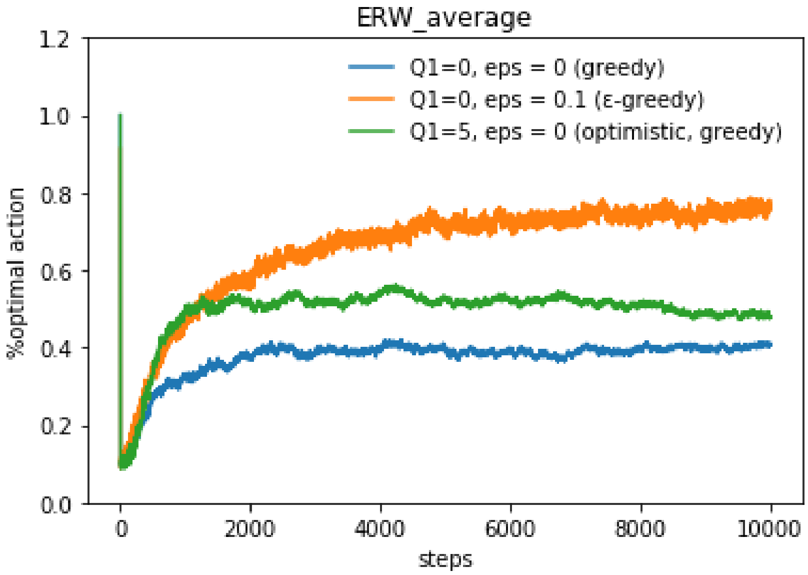

Figure 4: The percentage of optimal action of three action selection methods on the 10-armed non-stationary testbed. All methods used ERW average as their action-value estimates. These data are averages over 1000 runs with 10-arm bandit non-stationary problem.

Like Figure 2, each of the three lines in figure 4 has a distinct spike on the first step, which means most of the runs selected optimal action on that step. By using ERW average method, the percentage of optimal action chosen for all these three is much higher than using the sample-average method. Now, the ϵ-greedy method got the highest percentage value, which is as high as 80%. For the other two, they got about 50% and 40%.

4.3 Part III: Conclusion

Table 1. Average Reward(approximation) after 10000 steps using different algorithms

	ϵ-Greedy	Optimal Initial Greedy	Greedy
Sample average	1.75	1.0	0.75
ERW average	2.0	1.75	1.0

Table 2. The percentage(approximation) of optimal action after 10000

steps using different algorithms

	ϵ-Greedy	Optimal Initial Greedy	Greedy
Sample average	0.4	0.2	0.15
ERW average	0.8	0.5	0.4

From these two tables, we can see that for all these three action selection methods we choose, the ERW average method had better performance than the sample-average method. So we achieved the goal of our experiments.

Besides, we also find that the performance of the greedy method has always been the worst. It's because this selection method often got stuck performing suboptimal action. In contrast, the ϵ-greedy method eventually performed better because it continued to explore and improve its chances of recognizing the optimal action.