**Vivekanand Education Society’s Institute of Technology**

**Department of AI&DS Engineering**



**Subject: Reinforcement Learning**

**Class: D16AD**

| ROLL NO: **30** | NAME: [**SUHANEE KANDALKAR**](mailto:2021.suhanee.kandalkar@ves.ac.in) | | |
| --- | --- | --- | --- |
| EXPERIMENT NO:**4** | TITLE:Evaluating Sample-Average Methods in Nonstationary Bandit Problems | | |
| DOP: |  | DOS: **23/02/25** |  |
| GRADES: | LOs MAPPED: | | SIGNATURE: |

## **Aim:**

Evaluating Sample-Average Methods in Nonstationary Bandit Problems

## **Theory:**

## **1. Introduction to the Multi-Armed Bandit Problem**

The **multi-armed bandit (MAB) problem** is a classical reinforcement learning problem where an agent must choose between multiple actions (or "arms") to maximize its total reward over time. The challenge lies in the **exploration-exploitation trade-off**:

* + [Smallpdf—PDF Convert, AI Summarize, Merge, & Sign](https://incompleteideas.net/book/RLbook2020.pdf)
* **Exploration**: Trying different actions to discover potentially better rewards.
* **Exploitation**: Choosing the action that has yielded the highest rewards so far.

This problem is widely applicable in domains such as:

* **Online advertising** (choosing the best ad to display).
* **Clinical trials** (testing new treatments).
* **Recommendation systems** (suggesting personalized content).

## **2. Stationary vs. Nonstationary Bandit Problems**

* **Stationary bandits**: The true value of each action remains constant over time.
* **Nonstationary bandits**: The true value of each action **changes over time** due to external factors.

In **stationary environments**, sample-average methods work well because past data accurately reflects future outcomes.  
In **nonstationary environments**, relying on past data too much can be misleading, requiring adaptive strategies.

## **3. Modeling Nonstationary Environments**

To simulate a **nonstationary bandit environment**, we let each action’s **true value** q∗(a)q^\*(a)q∗(a) change over time using an **independent random walk**:

* + [Latest News from Google Research Blog](https://research.googleblog.com/)

q∗(a)←q∗(a)+N(0,0.01)q^\*(a) \leftarrow q^\*(a) + N(0, 0.01)q∗(a)←q∗(a)+N(0,0.01)

where N(0,0.01)N(0, 0.01)N(0,0.01) is a normal distribution with **mean 0** and **standard deviation 0.01**. This means that:

* The best action today might not be the best action in the future.
* The agent must continuously adapt to changing rewards.

## **4. Action-Value Methods**

### **4.1 Sample-Average Method**

The sample-average method updates the action-value estimate as:

Qn(a)=1n∑i=1nRiQ\_n(a) = \frac{1}{n} \sum\_{i=1}^{n} R\_iQn​(a)=n1​i=1∑n​Ri​

where:

* Qn(a)Q\_n(a)Qn​(a) is the estimated value of action aaa.
* RiR\_iRi​ is the reward received from action aaa at step iii.
* nnn is the number of times action aaa has been chosen.

This method **performs well in stationary environments** but **fails in nonstationary environments** because it gives equal weight to old and new data.

### **4.2 Constant Step-Size Method**

An alternative method is to use a **constant step-size update rule**:

Q(a)←Q(a)+α(R−Q(a))Q(a) \leftarrow Q(a) + \alpha (R - Q(a))Q(a)←Q(a)+α(R−Q(a))

where α\alphaα (e.g., **0.1**) is a fixed learning rate.

* **Advantage**: Gives more weight to **recent rewards**, making it better suited for **nonstationary environments**.
* **Disadvantage**: The estimate never fully converges because older rewards are forgotten.

## **5. Exploration Strategies**

We use the **ε-Greedy algorithm** to balance exploration and exploitation:

* **With probability** ϵ\epsilonϵ (e.g., **0.1**), choose a **random** action (**exploration**).
* **With probability** 1−ϵ1 - \epsilon1−ϵ (**0.9**), choose the action with the highest estimated reward (**exploitation**).

This ensures the agent doesn't get stuck in suboptimal actions.

## **6. Experimental Setup**

We compare:

1. **Sample-Average Method**
2. **Constant Step-Size Method (α = 0.1)**

**Parameters:**

* **10-armed bandit**
* **10,000 time steps**
* **2,000 runs**
* **Nonstationary rewards (random walk with N(0,0.01)N(0, 0.01)N(0,0.01))**
* **ε = 0.1** (ε-greedy exploration)

### **Expected Results**

* **Sample-average methods will struggle** because they rely on outdated rewards.
* **Constant step-size methods will adapt better** and achieve higher long-term rewards.

## **7. Applications of Multi-Armed Bandits**

### **7.1 Online Learning and A/B Testing**

* Websites like Google, Amazon, and Netflix use bandit algorithms to **optimize recommendations** dynamically.
* **Example**: Netflix tests different thumbnails for a movie to see which gets more clicks.

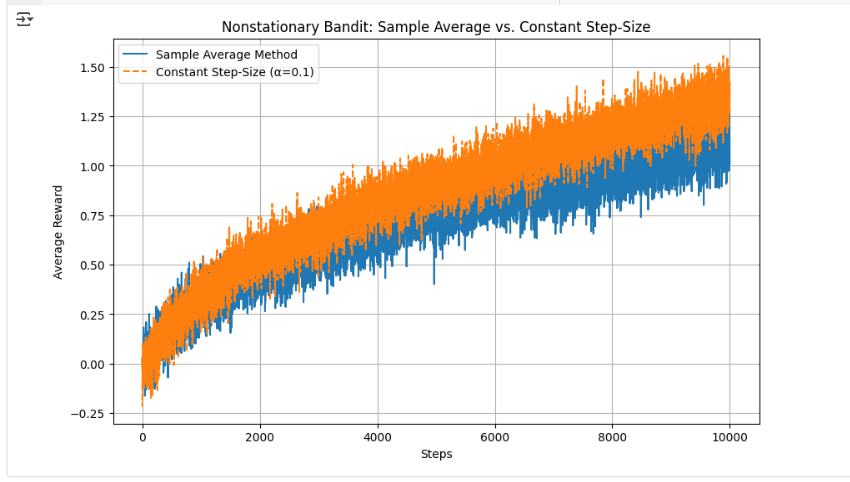
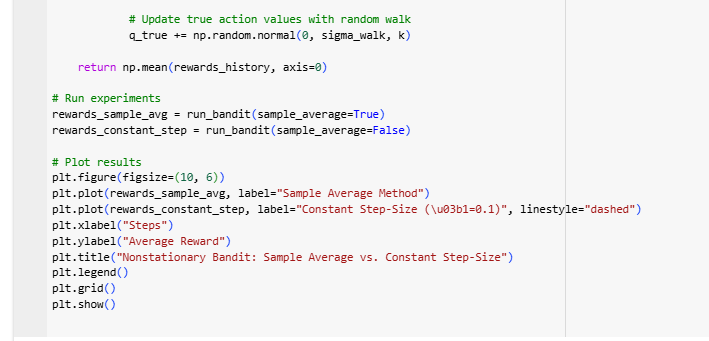
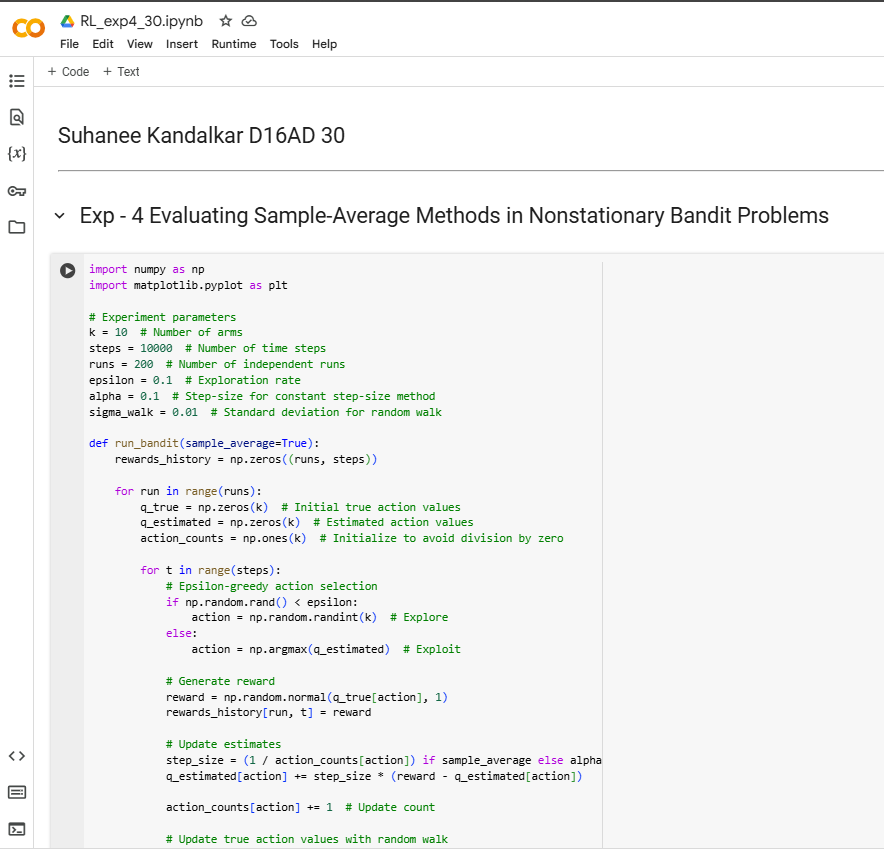
### **7.2 Financial Trading**

* In stock trading, optimal investment choices **change over time**.
* **Bandit models help adapt** to these nonstationary environments.

### **7.3 Clinical Trials**

* Medical researchers must balance **testing new drugs (exploration)** with **using the best-known treatment (exploitation)**.

## **Code:**

[RL\_exp4\_30.ipynb](https://colab.research.google.com/drive/14s5OKX2riZ6BbuAjwMc-QCM3SxWQkdmF?authuser=0#scrollTo=j5xfLgDgv3Gw)  


**Conclusion :**   
Sample-average methods struggle in nonstationary environments because they rely too much on past rewards.  
Constant step-size methods adapt faster because they give more importance to recent rewards.  
The experiment shows that in nonstationary settings, choosing an adaptive learning rate (α) is better than averaging all past rewards.  
Real-world applications include online learning, stock trading, and recommendation systems, where environments change dynamically