

Tracking Non-Stationary Problems: Bandits with Changing Rewards

Fabricio Barth

February 18, 2026

Outline

- ① Stationary vs Non-Stationary
- ② Constant Step-Size Update
- ③ Exploration for Non-Stationary Problems
- ④ Optimistic Initial Values

Stationary vs Non-Stationary

- **Stationary bandit:** action values are constant

$$q_*(a) = \mathbb{E}[R_t \mid A_t = a] \quad (\text{independent of } t)$$

- **Non-stationary bandit:** action values can change over time

$$q_t(a) = \mathbb{E}[R_t \mid A_t = a]$$

- Examples: user preferences drift, ads saturate, treatment efficacy changes, markets shift

Sample-Average Estimates Are Too Slow to Adapt

For the selected action A_t :

$$Q_{t+1}(A_t) = Q_t(A_t) + \frac{1}{N_t(A_t)}(R_t - Q_t(A_t))$$

- Step-size is $\alpha_t = 1/N_t(A_t)$
- As $N_t(A_t)$ grows, α_t shrinks \Rightarrow updates become tiny
- In a changing environment, this **locks in old information**

Constant Step-Size Update

To track change, use a **constant** step-size α :

$$Q_{t+1}(A_t) = Q_t(A_t) + \alpha(R_t - Q_t(A_t)), \quad 0 < \alpha \leq 1$$

- Recent rewards influence the estimate more than older rewards
- Faster adaptation for larger α (but more variance)

Choosing α : Memory vs Noise

- Smaller α :
 - smoother estimates (less variance)
 - slower tracking (more lag)
- Larger α :
 - faster tracking
 - noisier estimates

Key Point: Exploration Must Continue

In a non-stationary bandit:

- The identity of the best arm can change
- If exploration stops, the agent may never notice the change

A simple baseline is **constant ε -greedy**:

$$A = \begin{cases} \arg \max_a Q_t(a) & \text{with probability } 1 - \varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

Decaying ε_t : Use with Care

Sometimes we decrease exploration over time:

$$\varepsilon_t \downarrow$$

- Good intuition in **stationary** tasks: explore early, exploit later
- Risk in **non-stationary** tasks: exploration may vanish while the world keeps changing

Practical fix: decay to a **floor**:

$$\varepsilon_t = \max\left(\varepsilon_{\min}, \frac{\varepsilon_0}{1 + \beta t}\right)$$

Optimistic Initial Values

- Initialize action-value estimates high:

$$Q_1(a) = Q_0 \quad \text{with } Q_0 \text{ above typical rewards}$$

- Then act greedily:

$$A_t = \arg \max_a Q_t(a)$$

- Effect: encourages **initial exploration** because untried arms look best

References

- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed., Section 2.5.