10-Armed Bandit Testbed Report

Naman Goel

February 18, 2026

1 Introduction

The objective of this project was to implement the classical 10-armed bandit testbed and compare different exploration–exploitation strategies.

Each action $a \in \{1, ..., 10\}$ has a true value defined as:

$$q^*(a) \sim \mathcal{N}(0, 1)$$

When an action is selected at time t, the observed reward is:

$$R_t \sim \mathcal{N}(q^*(a), 1)$$

Thus, rewards are stochastic and centered around the true value of the selected arm.

Experiments were conducted for:

- 1000 time steps
- 2000 independent runs

2 The Exploration–Exploitation Dilemma

At each time step, the agent must choose between:

- **Exploitation:** Select the action with the highest estimated value $Q_t(a)$.
- **Exploration:** Select uncertain actions to improve estimates.

The dilemma arises because the true values $q^*(a)$ are unknown and must be estimated through interaction.

3 Action-Value Estimation

The agent maintains estimates:

$$Q_t(a)$$

We use the sample-average update rule:

$$Q_{n+1}(a) = Q_n(a) + \frac{1}{N(a)}(R - Q_n(a))$$

where:

- $N(a)$ is the number of times action a has been selected
- R is the observed reward

This update is equivalent to the sample mean:

$$Q_n(a) = \frac{1}{n}\sum_{i=1}^{n} R_i$$

4 Strategies Compared

4.1 Greedy ($\epsilon = 0$)

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

This strategy performs no exploration and may converge prematurely.

4.2 ϵ-Greedy

With probability ϵ:

$$A_t \sim \text{Uniform}(1, ..., k)$$

Otherwise:

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

Tested values:

- $\epsilon = 0.01$
- $\epsilon = 0.1$

4.3 Optimistic Initialization

Initial estimates are set as:

$$Q_1(a) = 5$$

Since true means are near zero, this forces systematic early exploration.

4.4 Upper Confidence Bound (UCB)

$$A_t = \arg\max_a \left[Q(a) + c\sqrt{\frac{\ln t}{N(a)}} \right]$$

where c controls exploration strength.

This method balances exploitation and uncertainty and achieves logarithmic regret:

$$\mathcal{O}(\log T)$$

5 Evaluation Metrics

5.1 Average Reward

$$\mathbb{E}[R_t]$$

Estimated empirically as:

$$\frac{1}{\text{runs}} \sum_{i=1}^{\text{runs}} R_t^{(i)}$$

5.2 Average Regret

Let the optimal action be:

$$a^* = \arg\max_a q^*(a)$$

Instantaneous regret is defined as:

$$\text{Regret}_t = q^*(a^*) - R_t$$

5.3 Percentage of Optimal Action

$$P(A_t = a^*)$$

This measures how frequently the optimal arm is selected.

6 Hyperparameter Effects

6.1 Effect of ϵ

Higher ϵ:

- Faster early learning
- Lower asymptotic performance

6.2 Effect of UCB Parameter c

Higher c:

- Stronger exploration
- Slower early reward growth

6.3 Effect of Optimistic Initial Value

Higher initial values:

- Strong early exploration
- Slower convergence

7 Conclusion

The 10-armed bandit experiment illustrates the exploration–exploitation dilemma clearly.

Greedy strategies fail due to insufficient exploration. ϵ-greedy ensures exploration but does so inefficiently. Optimistic initialization induces structured early exploration. UCB provides principled uncertainty-driven exploration and achieves superior long-term performance with logarithmic regret.