

# Multi-Armed Bandits: Exploration vs Exploitation

Fabricio Barth

February 9, 2026

# Outline

- 1 Problem Statement
- 2 Estimating Action Values
- 3 Action Selection

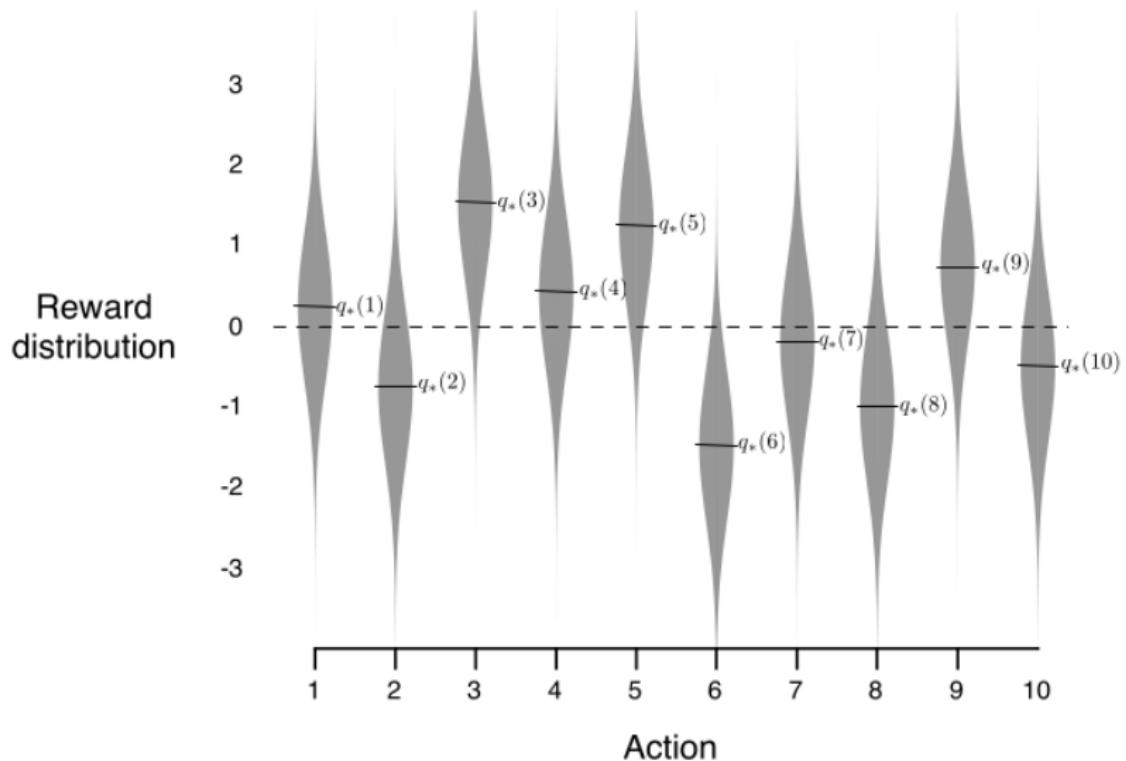
# The Multi-Armed Bandit Problem

- A set of  $k$  actions (arms):  $a \in \{1, \dots, k\}$
- At time  $t$ , choose action  $A_t$
- Observe a stochastic reward  $R_t$
- Each action has an unknown reward distribution

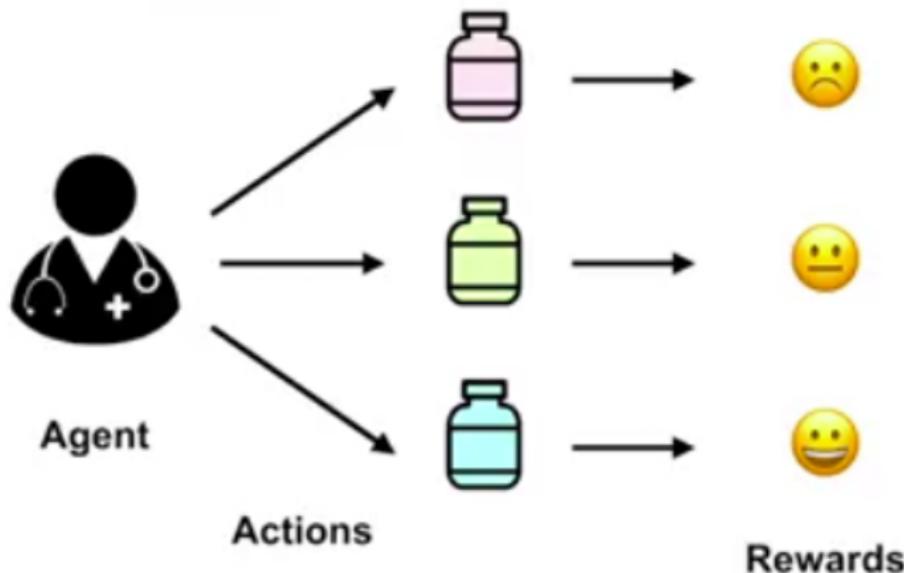
**Goal:** maximize cumulative reward

$$\sum_{t=1}^T R_t$$

# The Multi-Armed Bandit Problem: example



# Clinical Trial Example



It could be you in a restaurant.

# Expected Reward of an Action

The true (unknown) value of action  $a$  is

$$q_*(a) = \mathbb{E}[R_t \mid A_t = a].$$

- $q_*(a)$  is the expected outcome if we always choose  $a$
- The agent never observes  $q_*(a)$  directly

# Action-Value Estimates

Let  $Q_t(a)$  denote the estimate of  $q_*(a)$  at time  $t$ .

Using sample averages:

$$Q_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{N_t(a)} R_i^{(a)}$$

where  $N_t(a)$  is the number of times action  $a$  has been selected.

# Incremental Update Rule

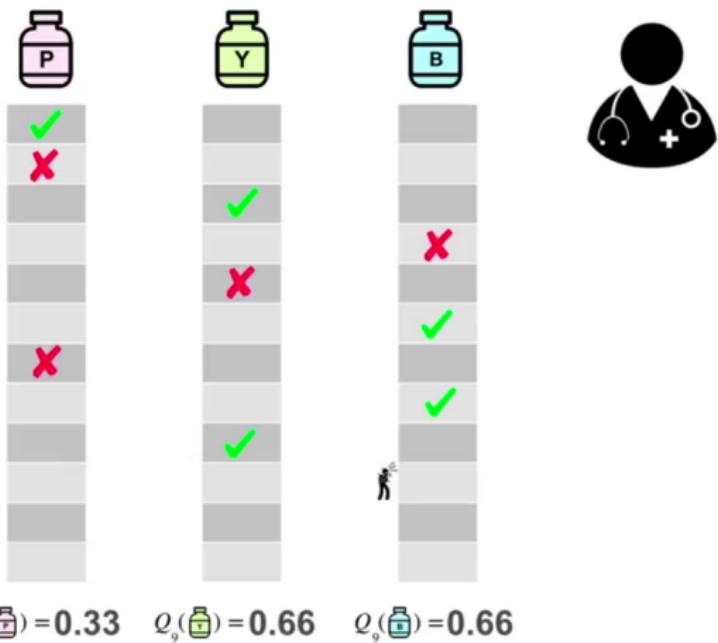
Instead of storing all rewards, we update incrementally:

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

# Estimating the value of $Q(a)$ for the Clinical Trial Example

A reward of 1 if the treatment succeeds otherwise 0

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i}{t - 1}$$



# Greedy Action Selection

If the true values were known:

$$A_t = \arg \max_a q_*(a)$$

In practice, we use estimates:

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

- Exploits current knowledge
- Can get stuck with a suboptimal action

# Explore vs Exploit

- **Exploit:** choose the best estimated action
- **Explore:** try actions with uncertain or lower estimates

Clinical trials must explore to discover better treatments, but also exploit to help current patients.

## $\varepsilon$ -Greedy Action Selection

With probability  $1 - \varepsilon$ : choose greedy action

With probability  $\varepsilon$ : explore uniformly

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

# Clinical Trials: Why $\varepsilon$ -greedy?

- Avoids prematurely committing to a suboptimal treatment
- Ensures all treatments are sampled
- Simple and effective baseline method

Common variants:

- Decaying  $\varepsilon_t$
- Optimistic initialization

# What Did We Learn in This Class?

- The **multi-armed bandit** is the simplest reinforcement learning problem: no states, only actions and rewards
- Real-world problems like **clinical trials** can be modeled as bandits
- Each action has an unknown **expected reward**:

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

- We learn action values using data-driven estimates  $Q_t(a)$  and incremental updates
- Purely greedy decisions can fail due to early randomness
- The core challenge is **exploration vs exploitation**
- $\varepsilon$ -**greedy** balances this trade-off by exploring with probability  $\varepsilon$

# References

- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed., pp. 25–36.