

Reinforcement Learning Fundamentals

Lecture 5: Multi-armed Bandit

Dr Sandeep Manjanna

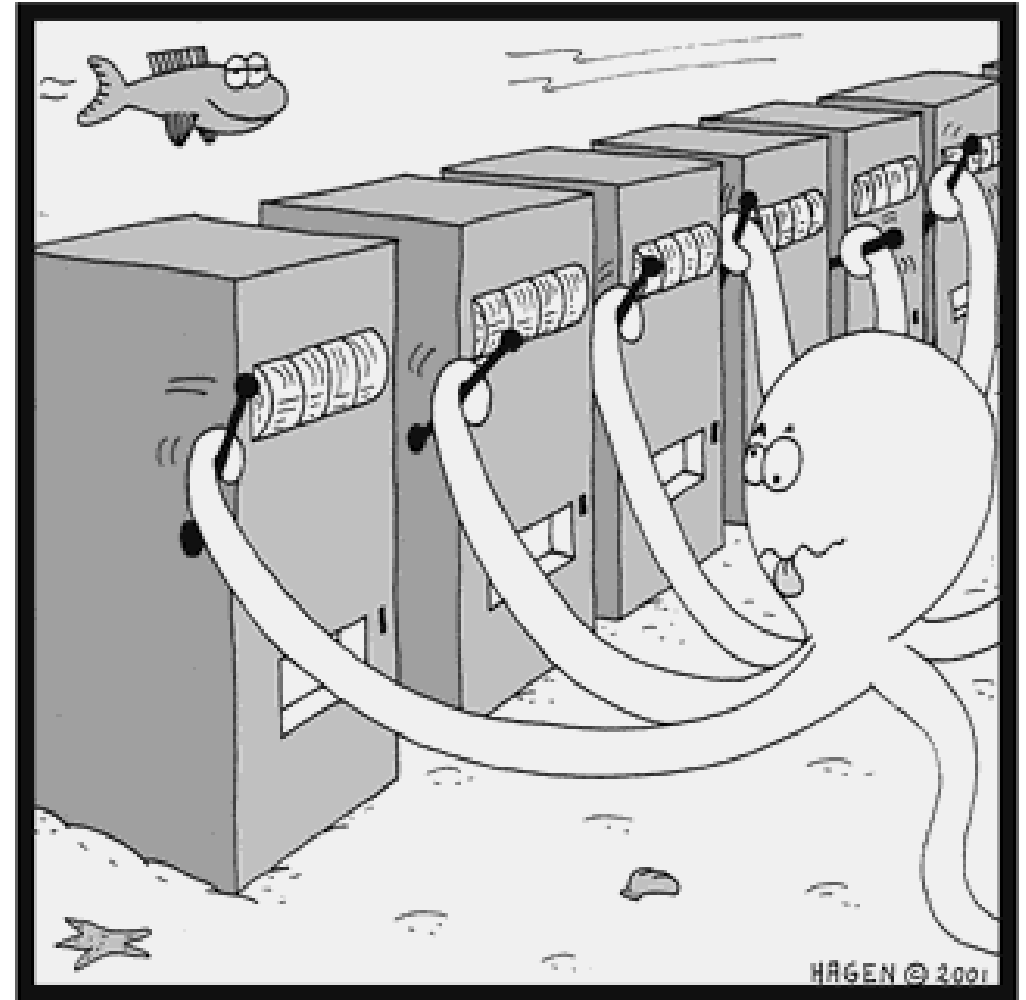
Assistant Professor, Plaksha University

sandeep.manjanna@plaksha.edu.in



In today's class...

- Immediate RL Problems
- Exploration vs. Exploitation
- Multi-armed Bandit
- ϵ -greedy algorithms
- Performance metrics



Immediate RL Problems

- Every time instant t , pick an action a_t and get a reward R_t .
- There is no state!
- Example:
 - Testing a drug for effectiveness
 - Tossing a coin
- 3 Actions: Choose one of the 3 coins

Coin 1



$$\mathbb{P}\{\text{heads}\} = p_1$$

Coin 2



$$\mathbb{P}\{\text{heads}\} = p_2$$

Coin 3



$$\mathbb{P}\{\text{heads}\} = p_3$$

Given 10 trials at tossing, maximize the total number of heads.

If, the probabilities p_1 , p_2 , and p_3 are given, then?

How many heads in 10 tosses?

Exploration and Exploitation

- Reinforcement learning is like trial-and-error learning
- The agent should discover a good policy
 - From its experiences of the environment
 - Without losing too much reward along the way
- **Exploration** finds more information about the environment
- **Exploitation** exploits known information to maximize reward
- It is usually important to explore as well as exploit

Exploration vs. Exploitation

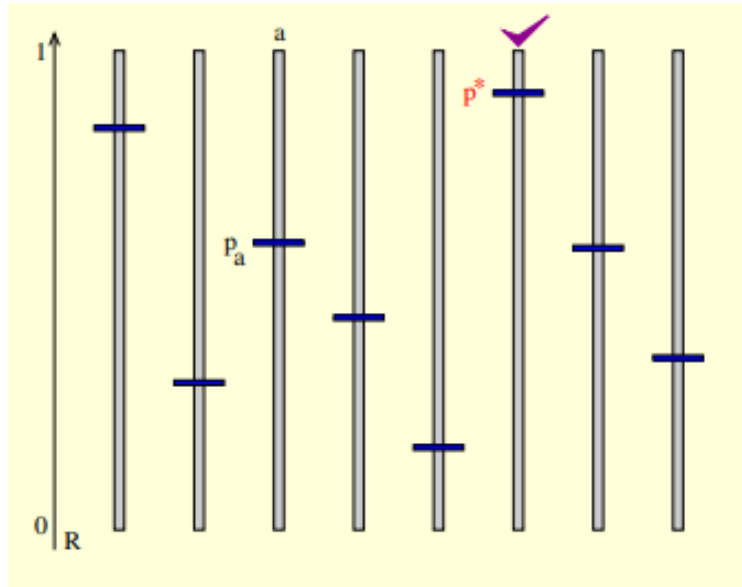
Examples?

What is exploration and exploitation in these examples?

- Restaurant Selection
- Online Advertising: Template optimization
- Clinical trials
- Packet routing in communication networks
- Game playing and reinforcement learning
- Oil Drilling or Mining

Multi-armed Bandit

- A hypothetical experiment where a person must **choose between multiple actions** (i.e., slot machines, the "one-armed bandits"), **each with an unknown payout**.



- The goal is to determine the best or **most profitable outcome** through a series of choices.
- At the beginning of the experiment, when odds and payouts are unknown, the gambler must determine **which machine to pull, in which order and how many times**.

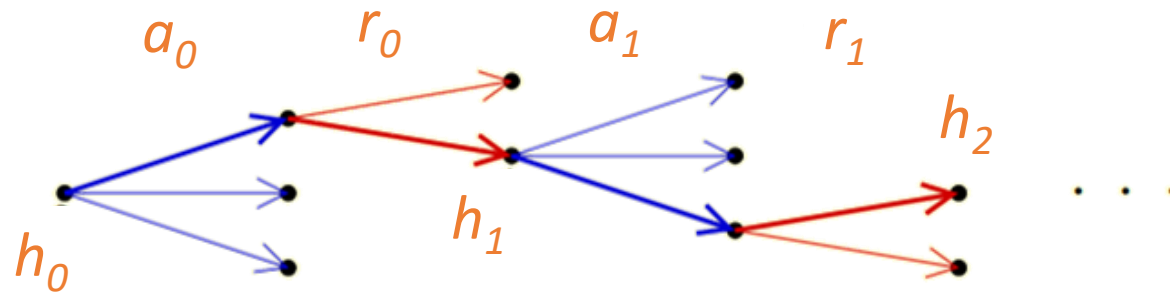
Algorithm to solve Multi-armed Bandit?

- Here is what an algorithm does—

For $t = 0, 1, 2, \dots, T - 1$:

- Given the **history** $h_t = (a_0, r_0, a_1, r_1, a_2, r_2, \dots, a_{t-1}, r_{t-1})$,
 - Pick an **arm** a_t to sample (or “pull”), and
 - Obtain a **reward** r_t drawn from the distribution corresponding to arm a_t .
- T is the total sampling budget, or the **horizon**.
 - Formally: a **deterministic algorithm** is a mapping from the set of all histories to the set of all arms.
 - Formally: a **stochastic** algorithm is a mapping from the set of all histories to the set of all probability distributions over arms.
 - The algorithm picks the arm to pull; the bandit instance returns the reward.

Multi-armed Bandit Tree



- An algorithm, bandit instance pair can generate many possible T -length histories.

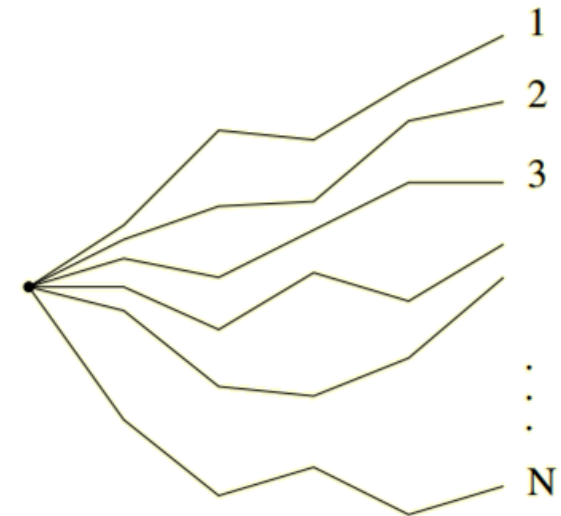
For a complete horizon T ,

$$h_T = (a_0, r_0, a_1, r_1, a_2, r_2, \dots, a_{T-1}, r_{T-1}),$$

$$P(h_T) = \prod_{t=0}^{T-1} \underbrace{P(a_t | h_t)}_{\text{Decided by the Algorithm}} \underbrace{P(r_t | a_t)}_{\text{Decided by the bandit instance}}$$

Decided by
the Algorithm

Decided by the
bandit instance



How many histories possible if the algorithm is deterministic and rewards 0–1?