FPT Education
FPT UNIVERSITY

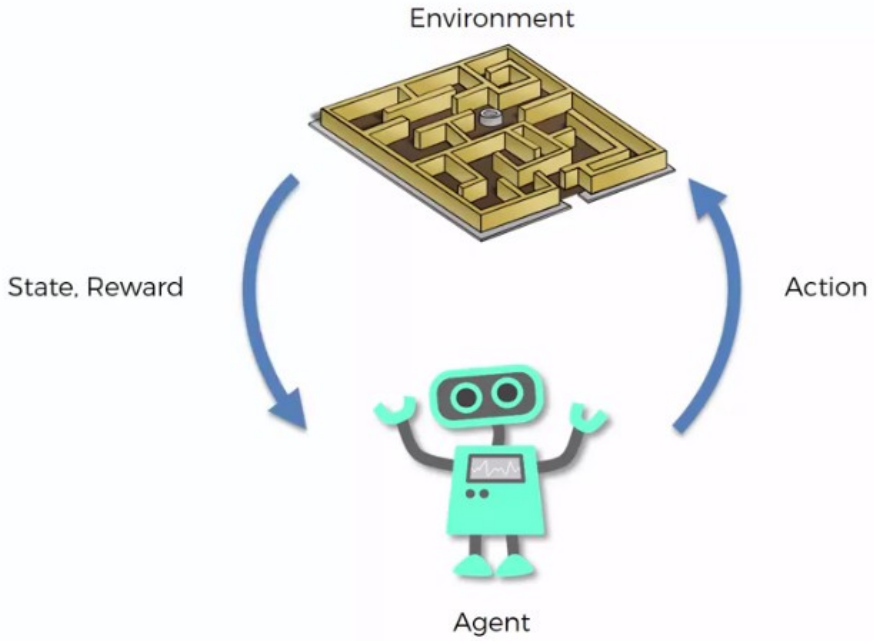
The K-Armed Bandit Problem

Objectives

- Define reward
- Understand the temporal nature of the bandit problem
- Define k-armed bandit
- Define action-values

Define reward

- What is the reward?



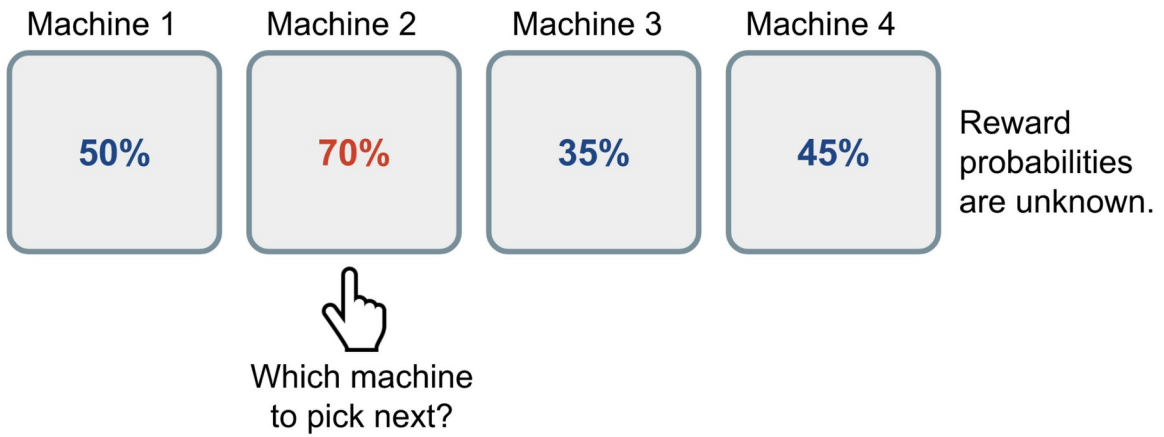
- Give your opinions

Define reward

- Reward is a scalar feedback signal provided by the environment to the learning agent.
- It indicates how well the agent is performing a given task at a particular state and time step.
- The reward signal serves as the primary mechanism for shaping the behavior of the agent, guiding it towards achieving its objectives.

The bandit problem

- The problem involves a gambler facing a row of slot machines (bandits), each with an unknown probability distribution of providing rewards.
- The gambler's objective is to maximize the total reward obtained over a series of plays.



The bandit problem

- The challenge arises from the trade-off between exploration and exploitation.
  - Exploitation involves playing the arm that appears to provide the highest reward based on past experience.
  - Exploration involves trying out different arms to learn more about their reward distributions.
- The bandit problem has applications in various fields, including clinical trials, online advertising, and reinforcement learning.

The bandit problem

- Solving the bandit problem efficiently is essential for optimizing resource allocation and decision-making in these domains.

Define k-armed bandit

- Scenario: Mr HoaDNT pulls machine k sample from unknow reward distribution
- Problem: given a finite number of pulls T, how can Mr HoaDNT optimize his winning? → give your opinions !

Define k-armed bandit

- It is a classic problem in reinforcement learning and decision theory.
- It is a simplified version of sequential decision making under uncertainty.
- In the K-Armed Bandit Problem, an agent is faced with a row of K different slot machines, each with an unknown probability distribution of yielding rewards.

Define k-armed bandit

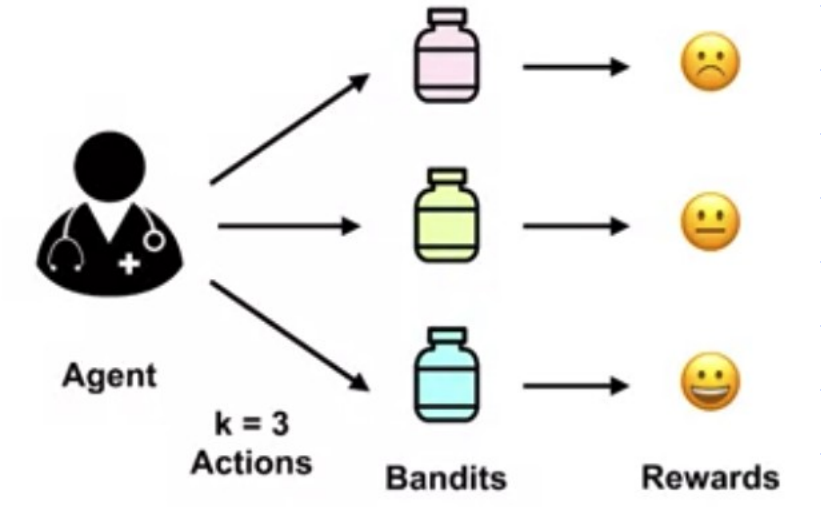
- The agent's objective is to maximize its cumulative reward over a series of trials, where in each trial, it selects one of the K arms (slot machines) to pull and receives a reward based on the outcome.

Define k-armed bandit

- In bandit problems, an agent selects an action given the initial observation, it receives a reward, and the episode terminates. → the agent action does not affect the next observation.
- The environment has no dynamics, so the reward is only influenced by the current action and the current observation

Define action-values

- The medical trial example is a case of the k-armed bandit problem. In the k-armed bandit problem, we have a decision-maker or agent, who chooses between k different actions, and receives a reward based on the actio

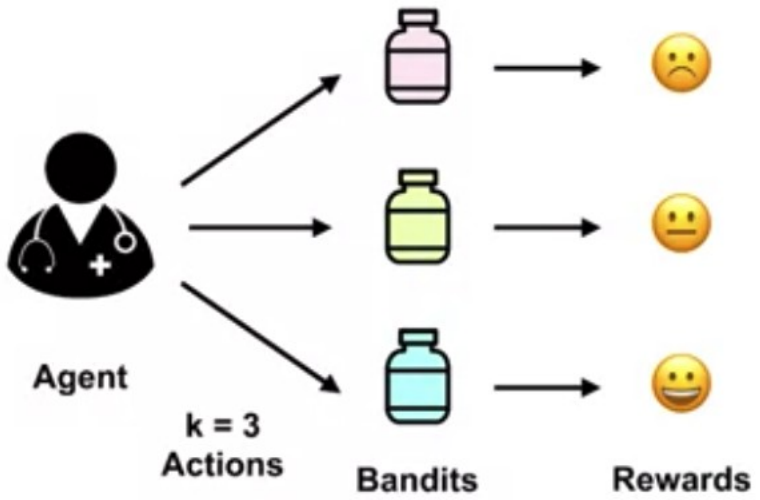


Define action-values

- The role of the agent is played by a doctor.
- The doctor has to choose between three different actions, to prescribe the blue, red, or yellow treatment. Each treatment is an action.
- Choosing that treatment yields some unknown reward.
- The welfare of the patient after the treatment is the reward that the doctor receives.

Define action-values

- For the doctor to decide which action is best, we must define the value of taking each action. We call these values the action values or the action value function.
  - *The expected cumulative reward the agent will receive by taking action A in state S*



Define action-values

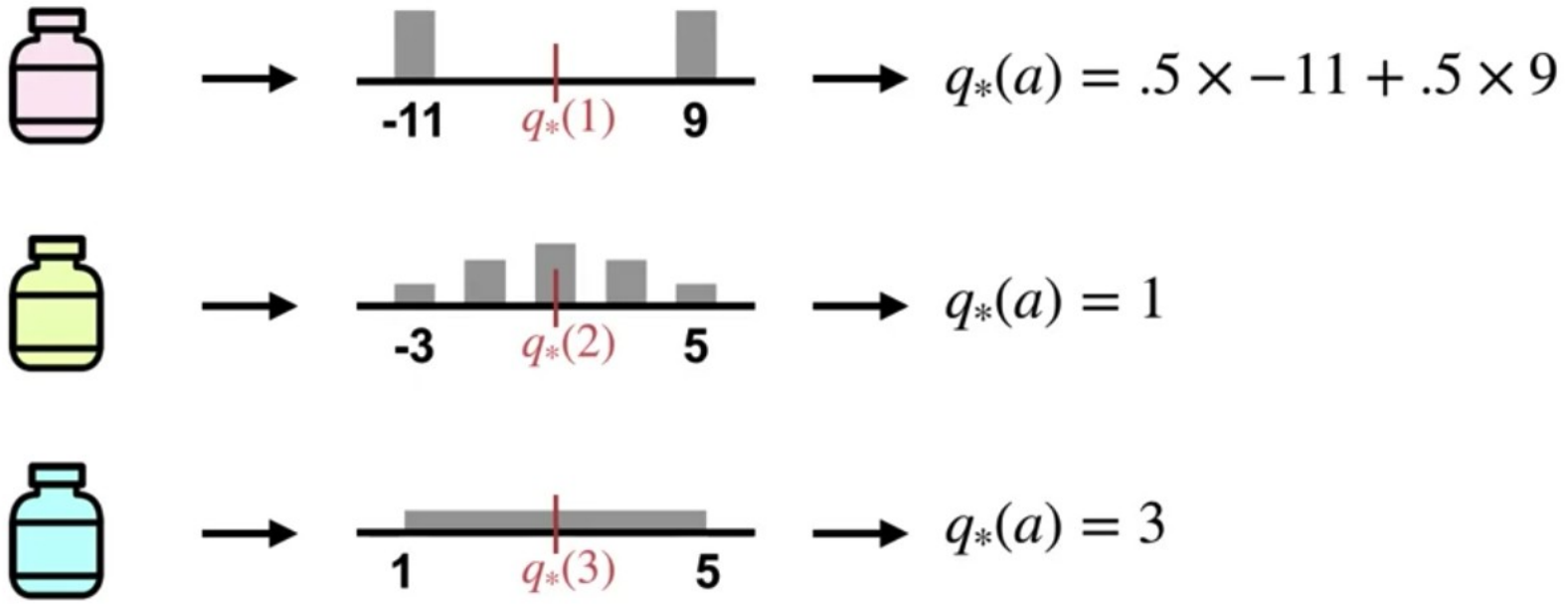
- Mathematically, the action-value function can be defined as the expected reward

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a] \quad \forall a \in \{1, \ldots, k\}$$
$$= \sum_r p(r \mid a)\, r$$

  - q star of a is defined as the expectation of R_t, given we selected action A, for each possible action one through k.
  - This conditional expectation is defined as a sum over all possible rewards.
  - Inside the sum, we have multiplied the possible reward by the probability of observing that reward..

Define action-values

- Example for calculating action value



Summary

- Define reward
- Understand the temporal nature of the bandit problem
- Define k-armed bandit
- Define action-values

Q & A