Reinforcement Learning

Multi-Armed Bandit

Dr. Alireza Aghamohammadi

Multi-Armed Bandit Problem

- Imagine a person wants to maximize their happiness by choosing among three restaurants. Each restaurant provides a different, unpredictable level of happiness on each visit. The goal is to balance:
  - **Exploitation**: Focusing on the restaurant that seems to provide the highest happiness based on past experiences.
  - **Exploration**: Occasionally trying other restaurants to discover if they might offer even greater happiness.
- **Arm**: Each restaurant represents an arm of the bandit. Similar to pulling the lever on a slot machine, selecting a restaurant results in a "reward" (happiness).
- Over time, the person must:
  - Identify the restaurant with the highest average happiness (the best arm).
  - Minimize regret, which is the loss incurred by not always choosing the best restaurant.
- If the reward values for all actions were known beforehand, solving the multi-armed bandit problem would be trivial: always select the action with the highest reward. However, the challenge arises because the reward values are uncertain and must be estimated through repeated trials.

- ❖ A Multi-Armed Bandit (MAB) problem consists of multiple actions (arms) and a single state.
- ❖ While there are multiple episodes, in each episode, you have only one opportunity to select an action. The value of an action is defined as:

  $$q(a) = \mathbb{E}\left[R \mid A = a\right]$$

  where $q(a)$ represents the expected reward when action a is chosen.
- ❖ **Regret** quantifies the difference between the cumulative reward you could have achieved by always selecting the optimal action and the actual cumulative reward received using your strategy:

  $$T = \sum_{e=1}^{E} \mathbb{E}\left[v^\star - q(A_e)\right]$$

  - ❑ T: Total regret over E episodes.
  - ❑ E: Total number of episodes.
  - ❑ $v^\star$: The reward you would receive if you always chose the optimal action (best arm).
  - ❑ $q(A_e)$: The expected reward based on the action taken in episode e.