Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 11

Solution by : Behnia Soleymani

Q1

What is the primary objective of a Multi-Armed Bandit (MAB) problem?

A) Minimize the number of actions taken.
B) Maximize cumulative reward over time.
C) Ensure all arms are chosen equally.
D) Predict future rewards accurately.

Correct Answer: B

Explanation: The core objective of MAB is to maximize cumulative rewards through optimal exploration-exploitation balance, not just immediate gains or fairness in arm selection.

Q2

Which best describes the exploration vs. exploitation trade-off in MAB?

A) Choosing arms based only on past rewards.
B) Balancing between gathering information and maximizing immediate rewards.
C) Prioritizing the most explored arm.
D) Ignoring uncertain arms.

Correct Answer: B

Explanation: Balancing between gathering information (exploration) and maximizing immediate rewards (exploitation).

Q3

How do state space and rewards in MAB differ from normal RL?

- **Stateless:** Unlike normal RL, MAB problems do not involve state transitions or a state space. Each decision is independent of previous actions.
- **Immediate Rewards:** Rewards depend only on the chosen arm and are received immediately after the action, unlike normal RL where rewards may depend on state sequences

Q4

Restaurant "Dish of the Day" MAB Scenario:
A restaurant wants to decide its daily "Dish of the Day" to maximize customer satisfaction. Each dish has an unknown average rating (reward). The chef can choose from 5 dishes, but customers only provide feedback after finishing their meal. Explain how a simple MAB strategy could help the chef balance exploration and exploitation. First explain the exploration and exploitation in this scenario, then Propose a basic adjustment to avoid always choosing the same dish too early.

Answer:

- **Exploration:** The chef should occasionally select less-tested dishes to gather feedback (e.g., choosing each dish at least once initially). This avoids missing a potentially high-rated dish that performed poorly early on.
- **Exploitation:** Serve dishes with the highest observed average ratings (using the sample-average method) to maximize immediate satisfaction.
- **strategy:** Introduce a random chance (e.g., 10 percent probability) to pick a non-top dish each day to ensure continued exploration. This prevents premature exploitation of a dish that was initially lucky.

Q5

Clinical Trial Challenges with MAB:
In a clinical trial for a new monoclonal antibody (mAb) therapy, patients are randomly assigned to different treatment arms. However, the trial faces two challenges: (1) limited patient availability and (2) delayed outcomes (e.g., long-term side effects only appear weeks later). How would a basic MAB approach struggle with these challenges?

- **Challenge 1 (Limited Patients):** A basic MAB might waste limited patients on suboptimal arms due to insufficient exploration-exploitation balance.
- **Challenge 2 (Delayed Outcomes):** MAB assumes immediate rewards, but delayed feedback means the agent cannot update estimates quickly, leading to outdated decisions.