Reinforcement Learning Specialization - Lecture Note

Course 1 - Fundamentals of Reinforcement Learning

Module 2

New terms:

short/long-term reward

policies

planning methods

dynamic programming

reward

time steps

Video: Sequential Decision Making with Evaluative Feedback

Action-Value function

Action-Values

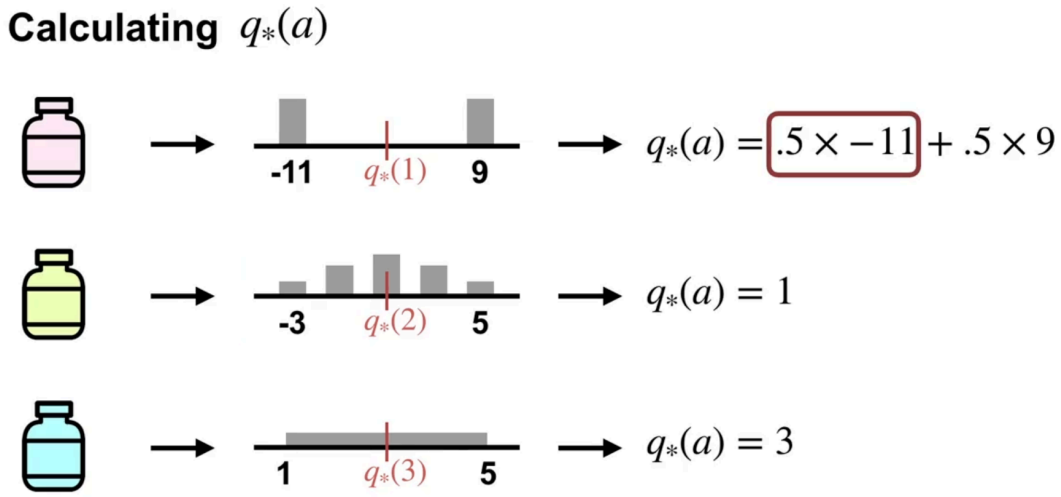
- The value is the expected reward

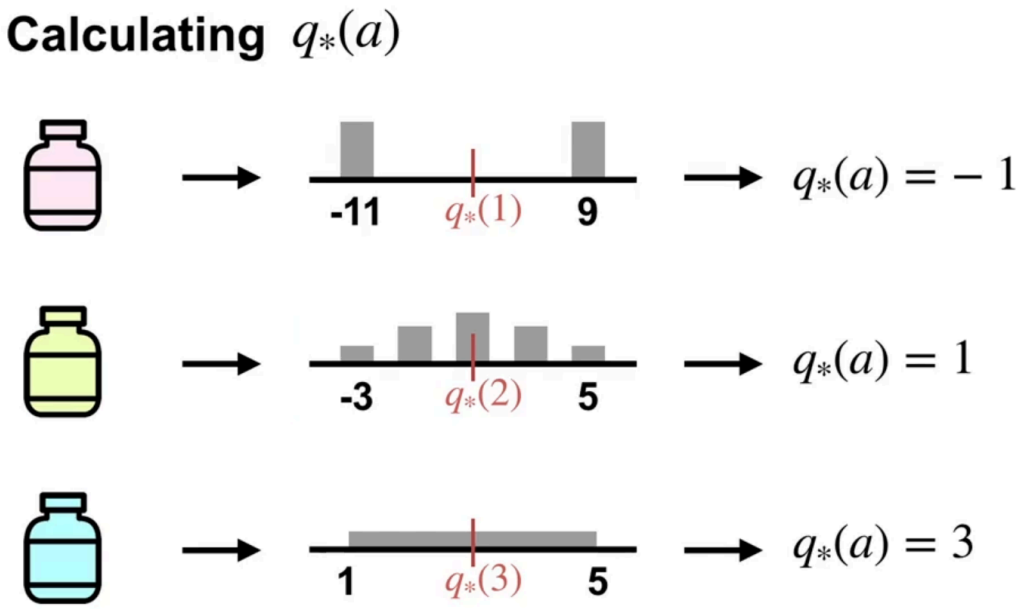
$$q_*(a) \doteq \mathbb{E}[R_t \,|\, A_t = a] \quad \forall a \in \{1, \ldots, k\}$$
$$= \sum_r p(r \,|\, a)\, r$$

- The goal is to maximize the expected reward

$$\underset{a}{\operatorname{argmax}} \; q_*(a)$$

chọn *a* để tối đa hóa phần thưởng kỳ vọng





How is the bandit problem similar or different to the supervised learning problem?

Vietnamese

Giống: cả 2 đều có mục tiêu đạt được kết quả tối ưu được đo lường bởi 1 hàm số (supervised: loss function, bandit problem: q*/reward function), supervised được train trên 1 tập data hữu hạn và cố định, bandit problem thì có số lượng action là 1 tập hữu hạn các action (K).

Khác: supervised learning tối đa hóa hàm mất mát trên 1 tập data cố định, ko thay đổi, label là cố định; bandit problem thì có label là giá trị kỳ vọng

trên 1 phân phối xác suất.

English

Similarities

1. **Optimization Objective**
    - Both aim to optimize a measurable function:
        - *Supervised Learning*: Minimizes a **loss function** (e.g., cross-entropy, MSE).
        - *Bandit Problems*: Maximizes a **reward function** (e.g., Q*-value, expected reward).
2. **Finite Action Space**
    - Supervised learning uses a fixed, finite dataset.
    - Bandit problems assume a finite set of **K actions** (e.g., choosing between K ad variants).

Key Differences

Aspect	Supervised Learning	Bandit Problems
Data Dynamics	Static dataset with fixed labels	Dynamic, stochastic rewards from a distribution
Feedback Type	Full feedback (labels for all inputs)	Partial feedback (reward only for chosen action)
Exploration Strategy	No exploration needed (deterministic training)	Requires **exploration-exploitation trade-off** (e.g., ε-greedy, UCB)
Objective	Generalize to unseen data	Maximize cumulative reward over interactions