Reinforcement Learning

An Introduction

second edition

Richard S. Sutton and Andrew G. Barto

Introduction

- Characteristics: trial-and-error search and delayed reward
- Exploration and exploitation trade-off
- Elements of RL:
    - Policy - mapping from states of the environment to actions (agent’s behavior)
    - Reward - feedback signal every time step that defines the goal
    - Value function - indicates the long-term desirability of states
    - Model – used for planning

Chapter 2 - Multi-armed Bandits

k-armed Bandit

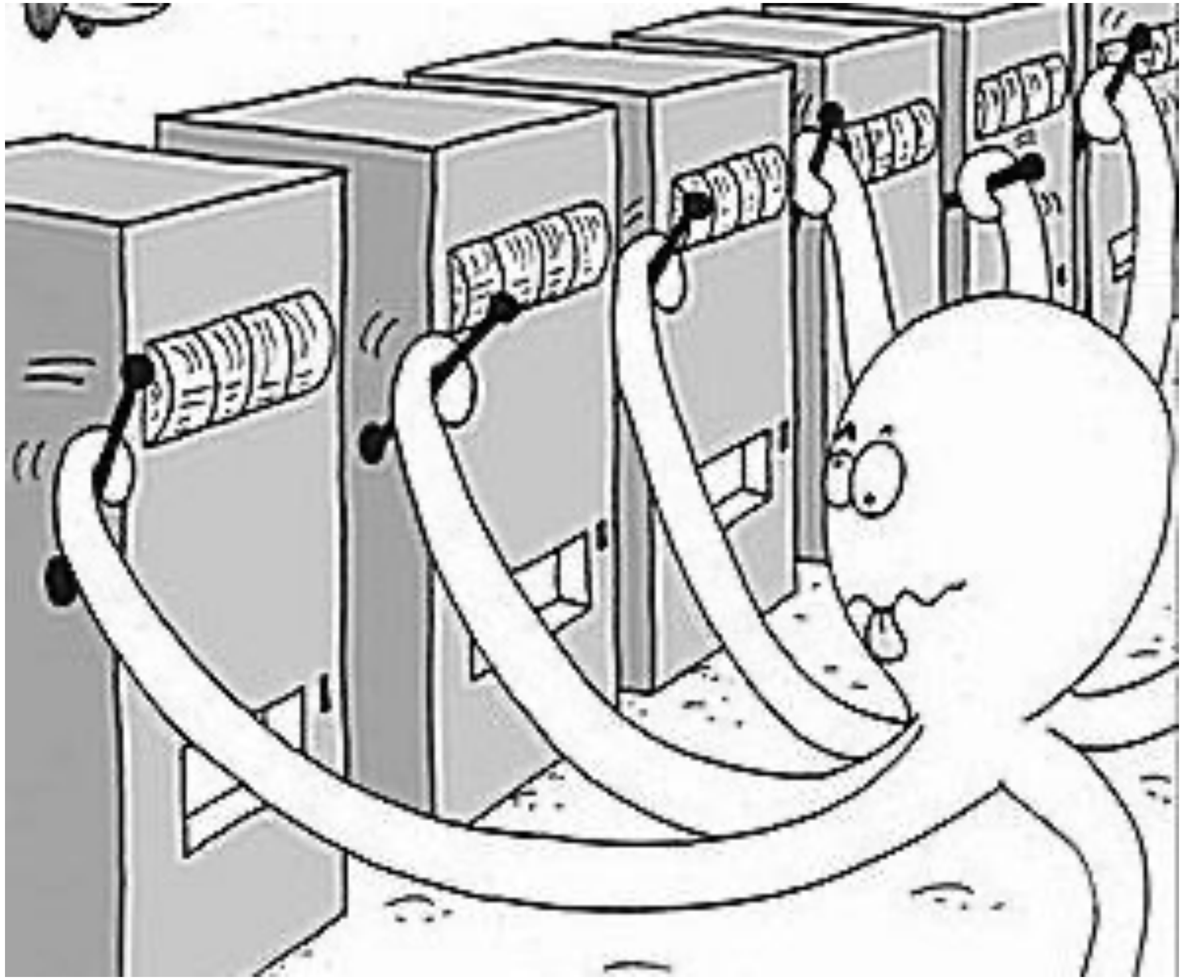
- Value of an action is the expected reward

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a] \,.$$

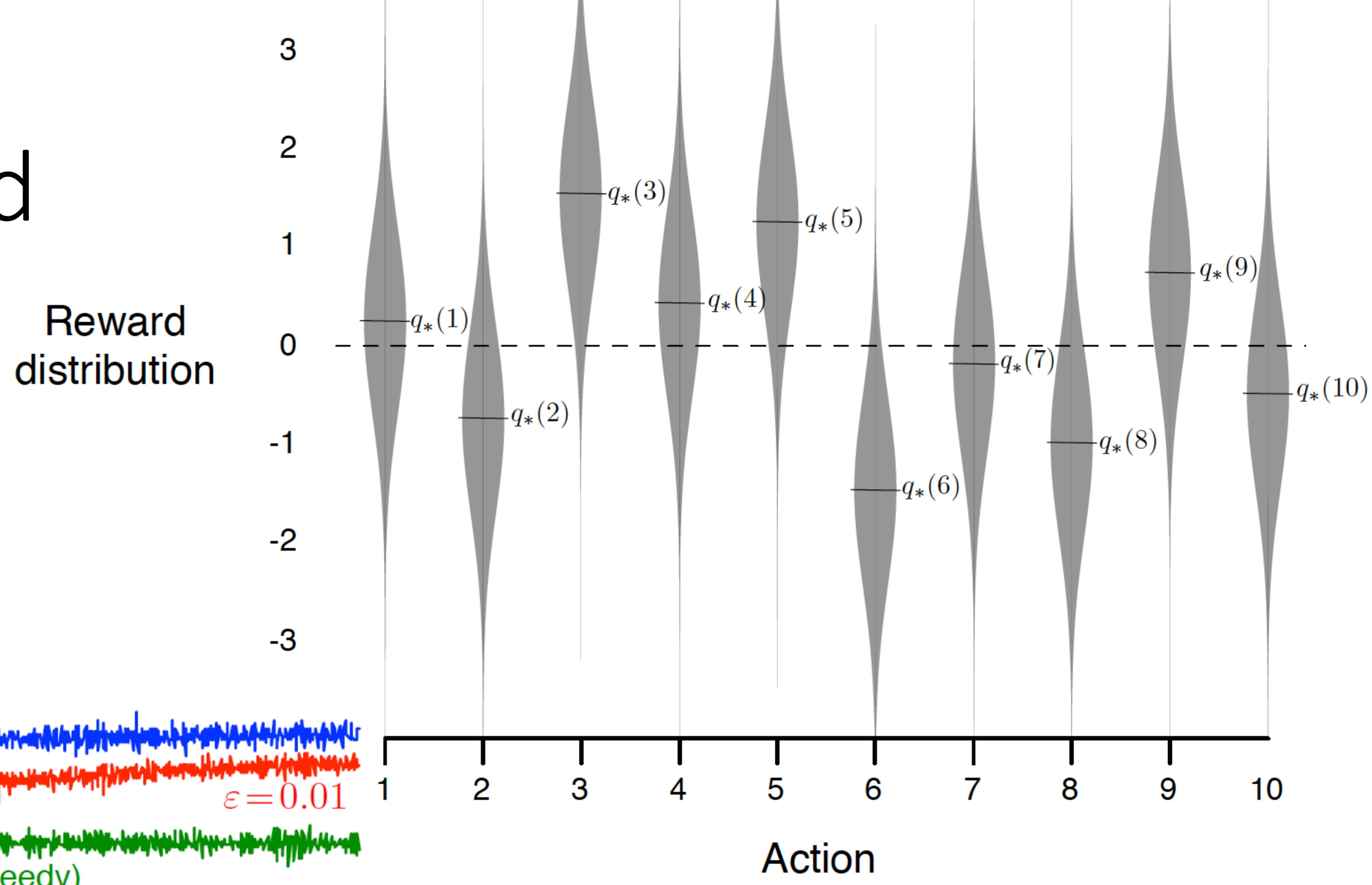
- Sample-average estimation
- Greedy actions

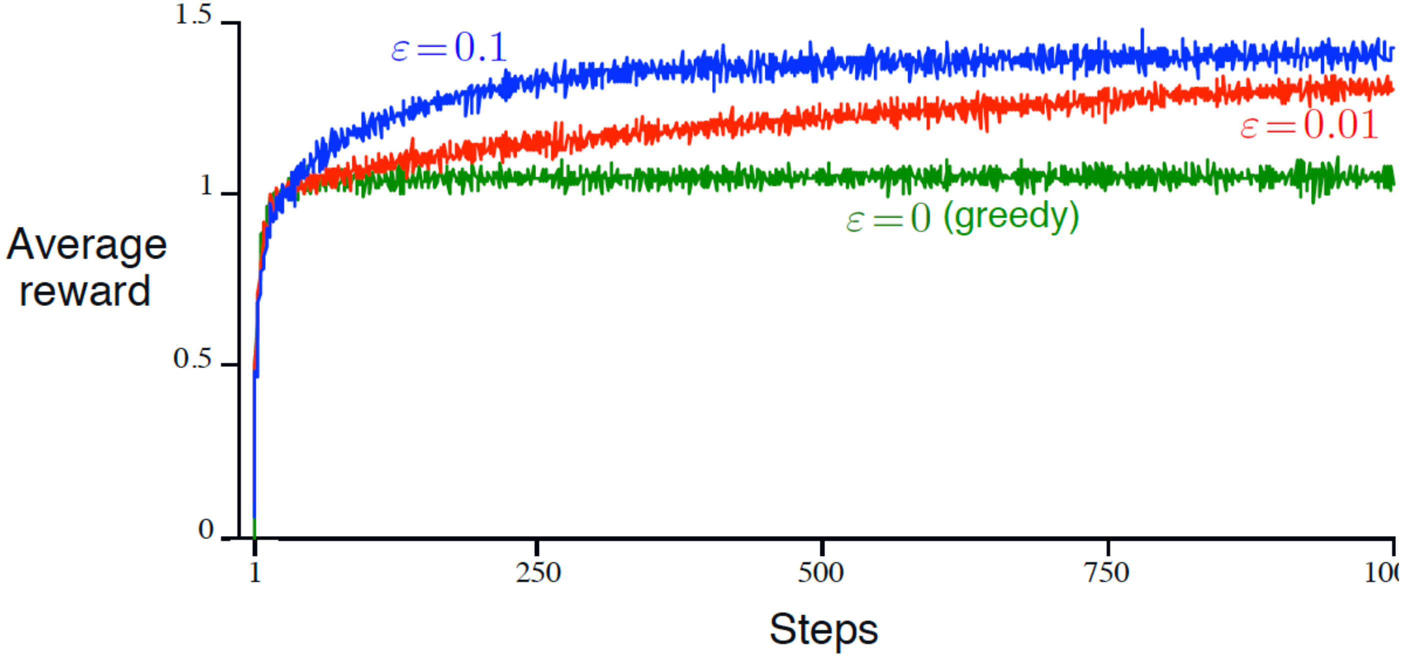
$$A_t \doteq \underset{a}{\arg\max}\, Q_t(a),$$

- Epsilon greedy



10-armed Testbed





Incremental action value

Exponential recency-weighted average for nonstationary problems

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i = Q_n + \frac{1}{n}\Big[R_n - Q_n\Big] = Q_n + \alpha\Big[R_n - Q_n\Big]$$

$$\mathit{NewEstimate} \leftarrow \mathit{OldEstimate} + \mathit{StepSize}\Big[\mathit{Target} - \mathit{OldEstimate}\Big].$$

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow \mathit{bandit}(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$

Balancing exploration and exploitation

Distribution free:

- Epsilon greedy
- Optimistic initial values
- Upper-confidence-bound (UCB) action selection

Bayesian methods:

- Gittins index
- Posterior sampling or Thompson sampling