University of Stuttgart

Reinforcement Learning

Multi-armed Bandits

Daniel Hennes

17.04.2017

University Stuttgart - IPVS - Machine Learning & Robotics

Tabular solution methods

- *State* and *action spaces* are small enough to be represented as arrays, or *tables*
- Methods can often find *exact* solutions, i.e., **optimal value function** or **optimal policy**
- Later: function approximation & policy search

Multi-armed bandits: RL problems with only one state

- *Nonassociative** setting: only the action space is relevant
- Finite bandit (k-armed bandit) $\rightarrow$ finite action space
- Training information:
  - *evaluate* actions taken
  - **not** *instructs* of "correct" action

k-armed bandit

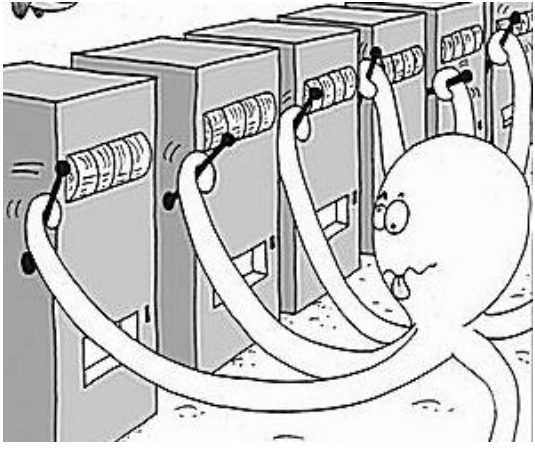

image credits: Microsoft Research

- There are k actions (machines)
- Each machine returns a reward from a stationary probability distribution
- Objective is to maximize the expected total reward, collected over the first T trials

Value

- Each action a has an expected or mean reward, the **value**:

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

- If you would know the true value q_* the next choice would be trivial
- Estimate of the action-value at time step t: $Q_t(a)$

Exploration vs. exploitation

- At each time step t there is (at least) one action that maximizes Q_t, called the *greedy* action:

$$A_t = \arg\max_a Q_t(a)$$

- Exploitation: selecting *greedy* action
- Exploration: selecting *nongreedy* action
  - improving estimate of the nongreedy action's value
  - reward lower in the short run
  - potentially much higher in the long run
- What is better? What does it depend on?
  - current action-value estimates
  - uncertainties
  - number of remaining steps

Estimating action-values

- *Sample average* method:

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

- Next action is greedy:

$$A_t = \arg\max_a Q_t(a)$$

- $\mathbb{1}$ is the indicator function,
  i.e., 1 if predicate is true, else 0
- If denominator is zero, set $Q_t(a) = 0$

ϵ-greedy action selection

- Simple idea to force continued exploration
- With probability $1 - \epsilon$ take the *greedy* action
- With probability ϵ take a random action
- All actions are chosen with non-zero probability