HSLU Lucerne University of Applied Sciences and Arts

Introduction & Multiarmed Bandits

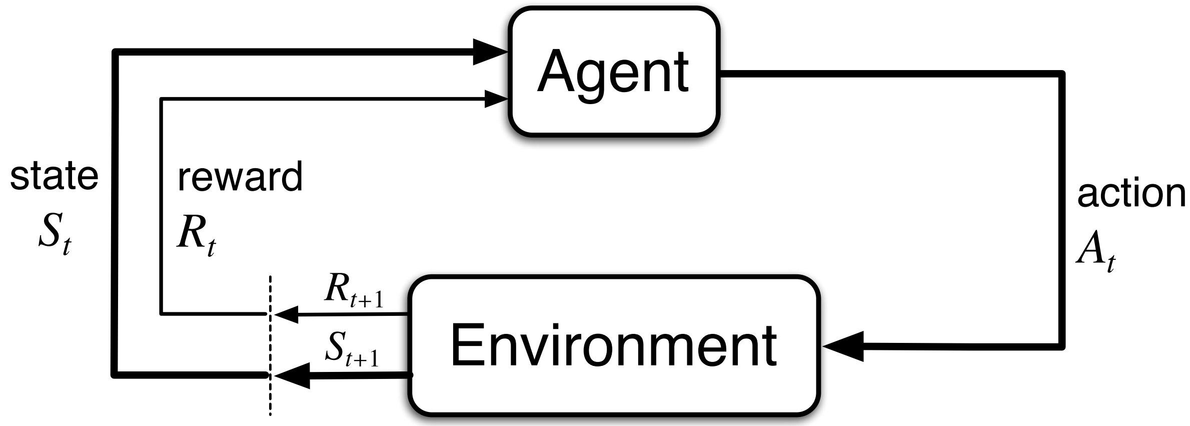
Summary

Reinforcement Learning
September 29, 2022

FH Zentralschweiz

Agent and Environment

In RL: An agent interacts with an environment using actions and gets a reward for each action



The agent's goal is to maximize the expected cumulative reward

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Multi-Armed Bandits

- Solving the multi-armed bandit problem required exploring different actions and exploiting the action which currently seems best
- The expected rewards of each action are estimated by calculating a value function incrementally using

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

$$\text{NewEstimate} \leftarrow \text{OldEstimate} + \text{StepSize}\ [\text{Target} - \text{OldEstimate}]$$

- A epsilon-greedy policy can be defined using this value function and selecting greedy and non-greedy actions with probabilities 1-e, respectively e
- UCB better balances exploration better in the long run

Simple Multi Armed Bandit Agent

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\epsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \epsilon \end{cases}$$

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

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

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

Comparison of exploration