An introduction to Reinforcement Learning

14th of June 2022

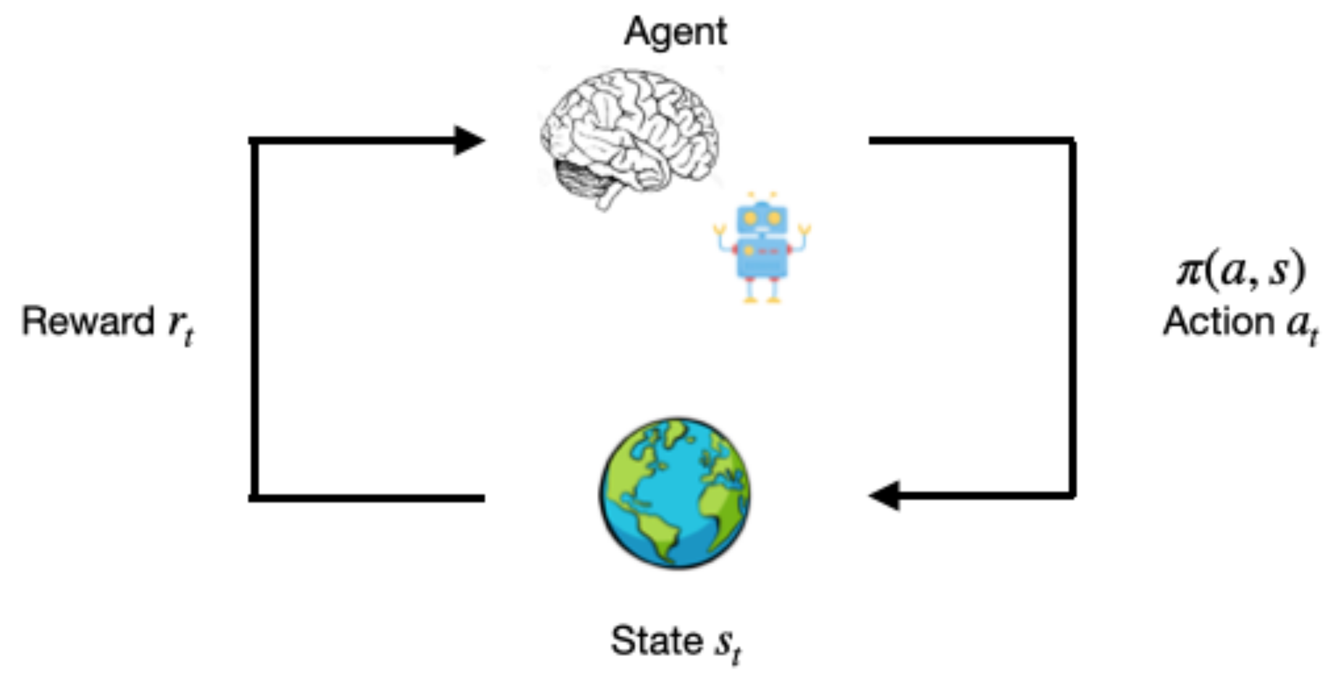
Recap: Temporal Difference Learning

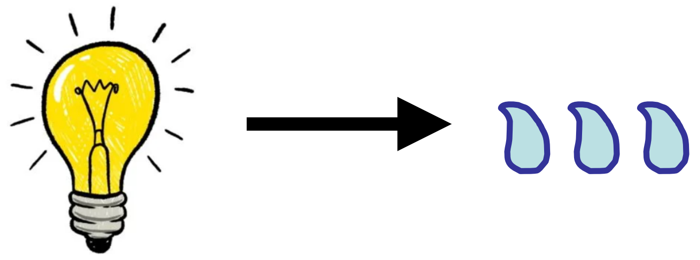
TD Learning:

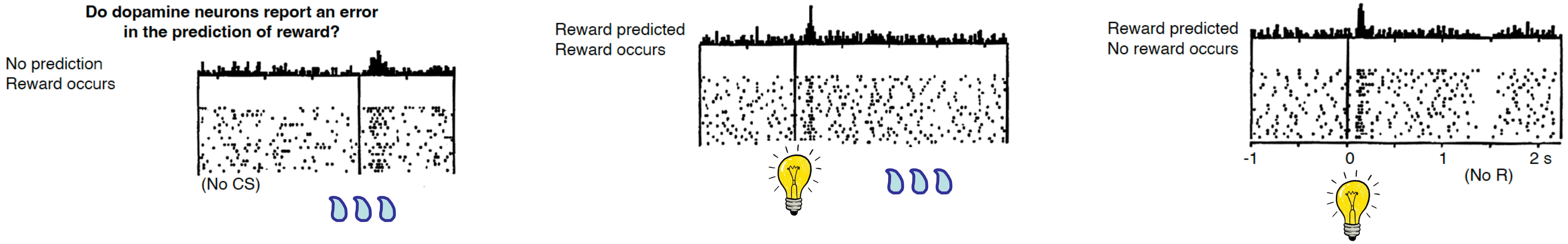
Prediction error

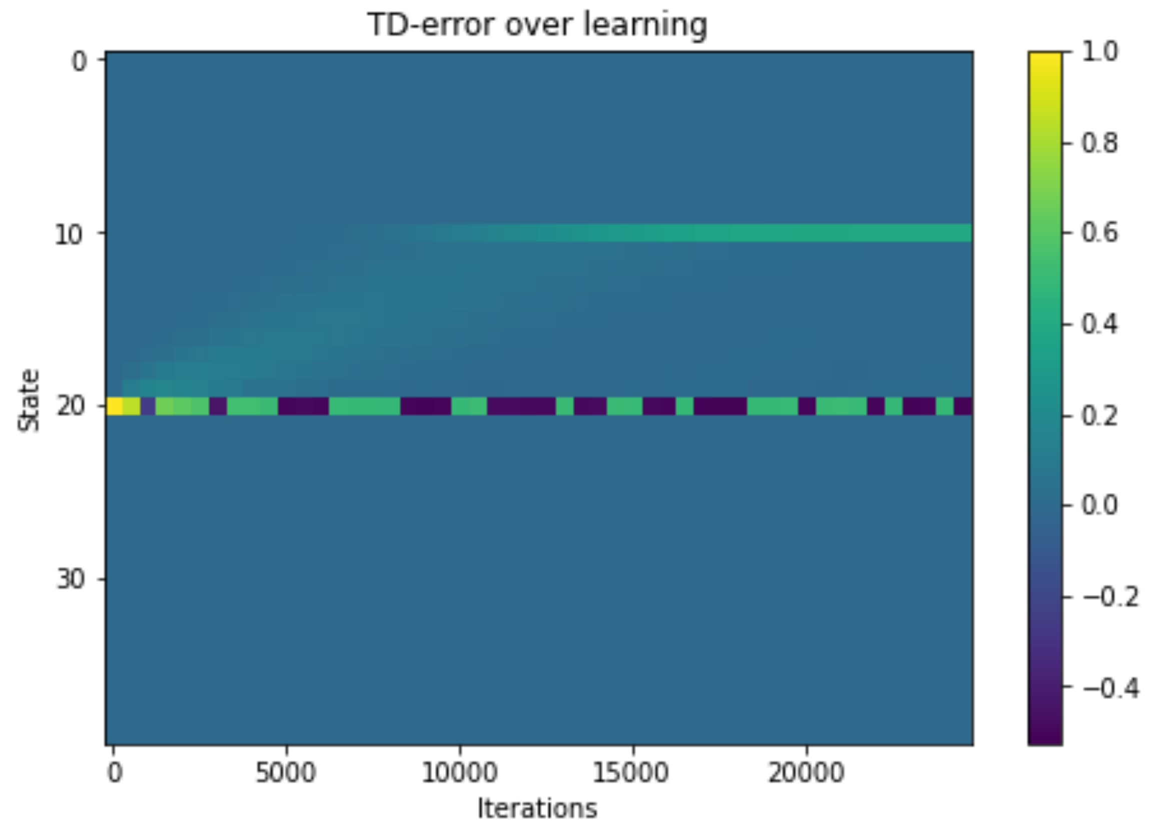
$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r + \gamma \cdot V(s_{t+1}) - V(s_t))$$

Learning rate   Discount rate









But what about actions?

Dates and topics

14.06.2022
- Models of Action Selection, Exploration

21.06.2022
- Combine Learning and action selection: Q-learning, SARSA

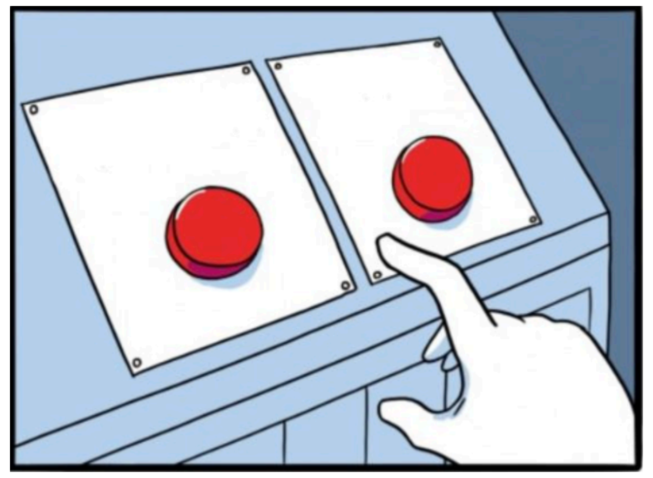
28.06.2022
- Model-based RL

05.07.2022

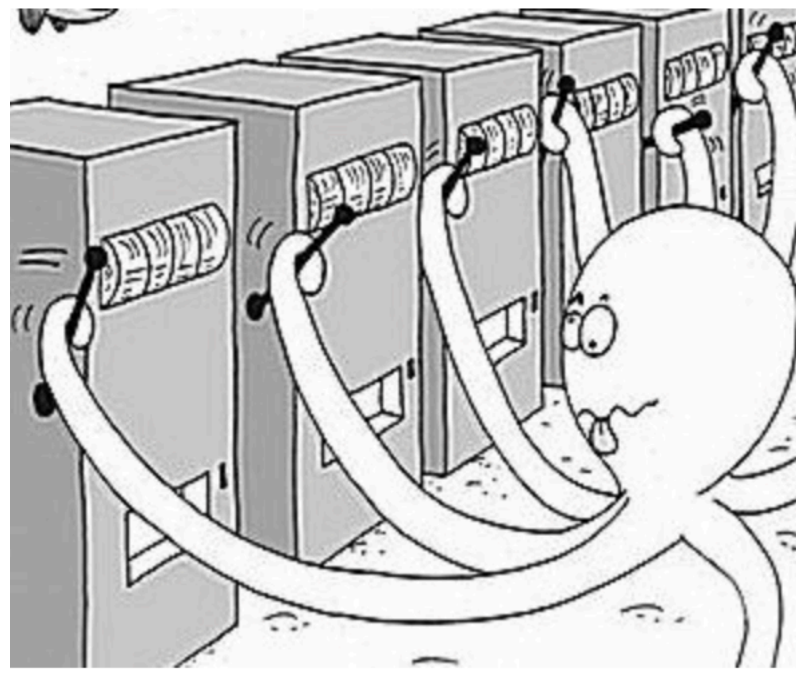
12.07.2022

19.07.2022
- Applications
  - Model fitting, testing psych hypotheses
  - Deep RL (maybe)
  - Current research (maybe)

26.07.2022
- Recap and talk about essay/project ideas



Multi-armed bandits



- Problems where agents are faced with different options
  - Have to find out which of these are good or bad via trial-and-error
- Key problem: **exploitation vs. exploration**
  - **Random** vs. **goal-directed exploration**
- At the heart of many modern RL studies
  - Ideal testbed for different **models of action selection**
- Still in simplified RL setting
  - *Stationary* environment
  - Only consider *immediate reward* (for now)
  - Non-*contextual*
  - *Tabular*

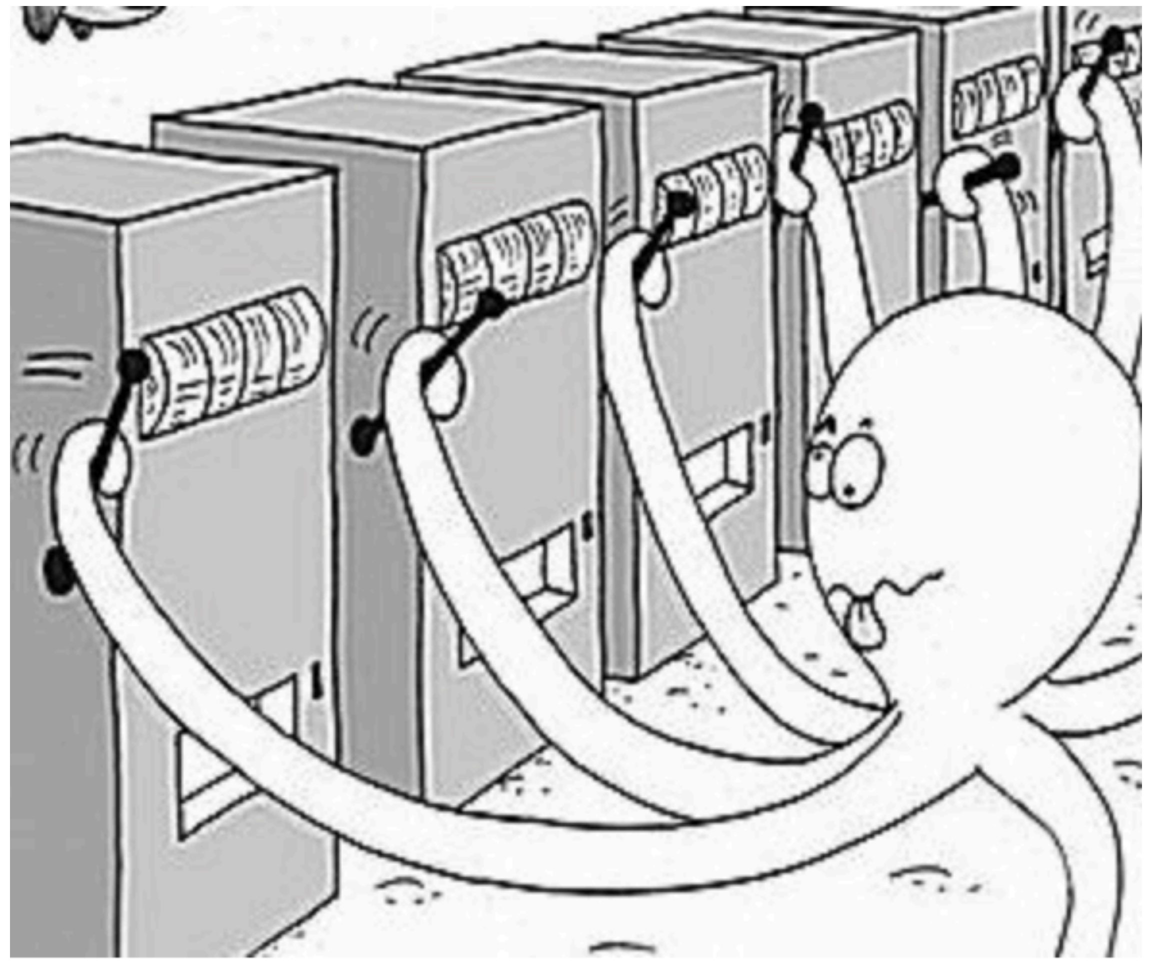
Multi-armed bandits

Greedy action selection:

$$P(a_t = a) = \begin{cases} 1 & \text{if } a_t = \text{argmax}_a\ V_t(a) \\ 0 & \text{otherwise} \end{cases}$$

Epsilon-greedy action selection:

$$P(a_t = a) = \begin{cases} 1 - \epsilon & \text{if } a_t = \text{argmax}_a\ V_t(a) \\ \epsilon/N & \text{otherwise} \end{cases}$$



Softmax action selection:

$$P(a_t = a) = \frac{e^{V_t(a) \cdot \beta}}{\sum_{i=1}^{N} e^{V_t(a_i) \cdot \beta}}$$

Action is governed by a **policy**:

$$\pi(a, s) = P(a_t = a \,|\, s_t = s)$$

Upper-confidence-bound (UCB) action selection:

$$P(a_t = a) = \text{argmax}_a[V_t(a) + c \cdot \sqrt{\frac{\ln t}{N_t(a)}}]$$

Coding: Multi-Armed Bandits

https://github.com/schwartenbeckph/RL-Course/tree/main/2022_06_14