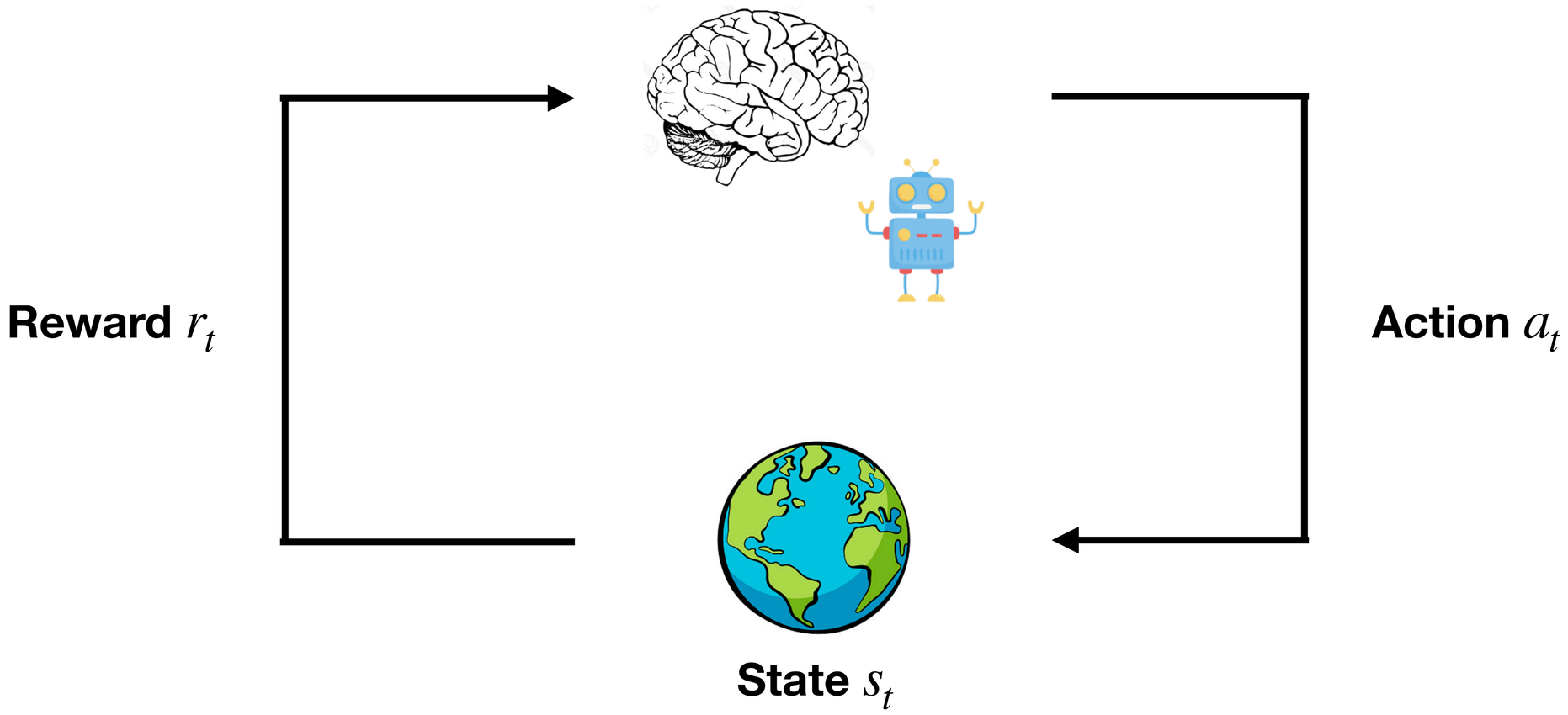
An introduction to Reinforcement Learning

26nd of April 2022

Philipp Schwartenbeck

AI Center, University of Tübingen

Recap: Basic setup of RL



Recap: What is reinforcement learning (RL)?

- RL is a **computational approach** to learning from **interactions** with the **environment**
  - Trial-and-error
  - Delayed reward
- Considers whole problem of **goal-directed** agent interacting with an **uncertain** environment
- Three main machine learning approaches
  - Supervised
  - Unsupervised
  - RL
- Very general account

Where are we?

Intro

Intro (cont)
Theories of Learning
- **Psychology, behaviour**
- **Rescorla-Wagner Learning**

Theories of Learning
- Neuroscience
- TD learning

Markov Decision Processes

Theories of control, action selection

Model-free and model-based RL
Exploitation vs. Exploration

Some coding
- Role of different parameters
- Model-fitting
- If possible: parameter recovery, model comparison

- ‘Advanced’ topics and current applications
  - Planning, Dyna, replay
  - Clever ways of planning, tree-search etc
  - Deep RL
  - Other current fancy developments

Recap This seminar: components

- Most of this is first time material - tell me if something doesn't work, open for suggestions
  - Especially for second half of the seminar
- Structure
  - Theory (key reference: Sutton & Barto, 1998)
  - Research (key papers)
  - Coding (Python)
- Missing May date (24th of May): coding
- Grading: essay
  - Tell me if you would like to have additional grading during the term (coding exercise/s, presentation)

Basics of (Reinforcement) Learning

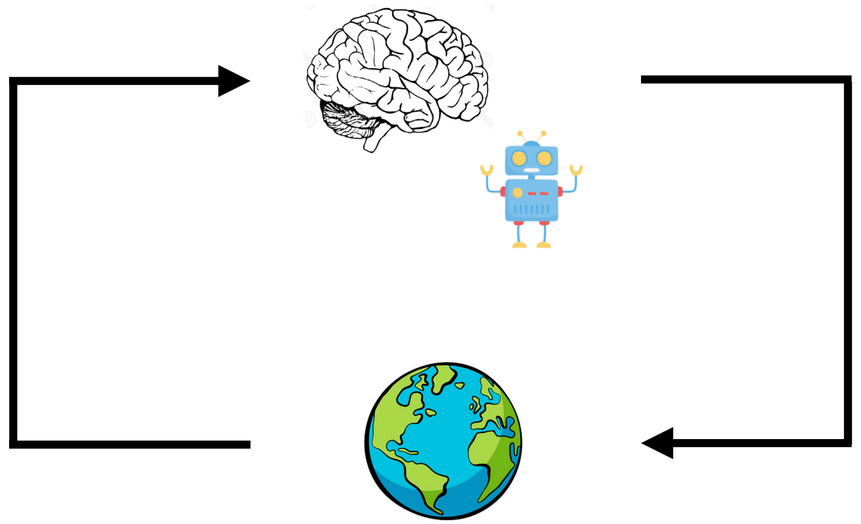
Basic setup: how to agents learn to act?

Based on a reward signal, agents learn **values of actions/states:**

$$V_\pi(s) = \mathbb{E}_\pi[R \mid s_0 = s]$$

Agents can learn a **model of the environment** to make smarter decisions, e.g.:

$$P(s_{t+1} = s \mid s_t = s, a_t = a)$$



Values can be **learnt** (simplified!!):

$$V(s) \leftarrow V(s) + \alpha \cdot (r - V(s))$$

Learning rate

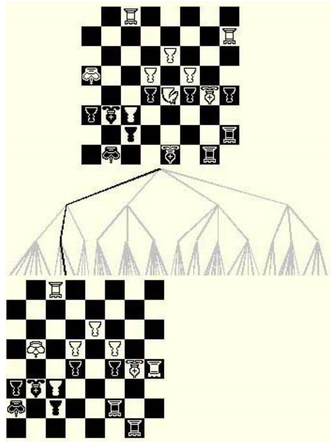
Prediction error

Action is governed by a **policy**:

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

(More) Examples

- **Chess**: what is…
  - The state?
  - An action?
  - A reward?



- How can values be learned over time?
- How could a model of the environment be useful?

Other relevant components:

- tree search
- position evaluation
- situation memory

Taken from Peter Dayan

(More) Examples

- **Learn how to walk**: what is…
  - The state?
  - An action?
  - A reward?

- How can values be learned over time?
- How could a model of the environment be useful?

Examples extended..

- Other examples (see Sutton & Barto, pp 4-5):
- **Adaptive controller** adjusts parameters of a petroleum refinery's operation in real time
  - Optimise yield/cost/quality trade-off
  - Objective: specified marginal costs
  - Without sticking strictly to pre-defined set points
- **Mobile robot** decides to search for trash to collect or find its way back to battery recharging station
  - Decision based on
    - Current charge level of battery
    - How quickly recharger has been found in the past.
- **Prepare breakfast**
  - Subgoals, hierarchies
  - Conditional behaviour
  - Sense/access bodily states

Key features of all these examples

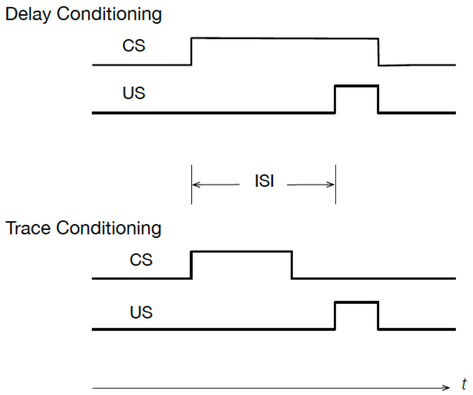
- (Danger of repeating myself): **Interaction** between active decision-making agent and its environment
  - Agent seeks to achieve a **goal**
  - **Uncertainty** about its environment
- Take into account **indirect, delayed consequences** of actions
  - Requires foresight or planning
- Need to **monitor** environment frequently
- **Judge progress** toward goal based on what can be sensed directly
- Use experience to **improve performance** over time (online vs. offline learning)
  - Basis for adjusting behaviour to exploit specific features of the task

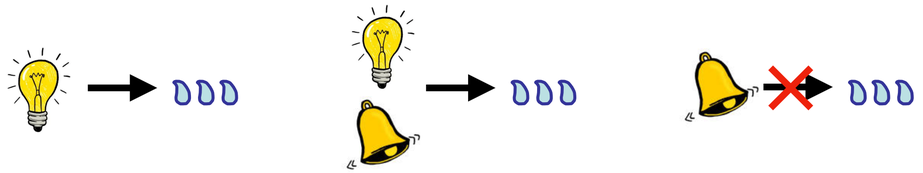
History and Theories of (Reinforcement) Learning

“Three” historical branches of RL

- Association learning, prediction (early 1900s)
- Optimal control (1950 onward)
- Learning *and* control (1980 onward)

History: Psychology

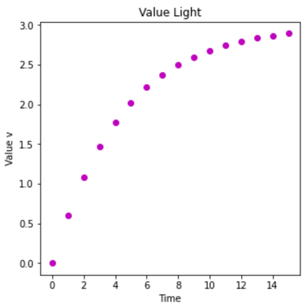


- **Classical** (Pavlovian) **conditioning** (roughly) in domain of algorithms for **prediction**
  - Algorithms for **control**: **instrumental** (operant) **conditioning**
  - You probably know all this..
- At least two interesting phenomena in classical conditioning from algorithmic perspective:
  - **Higher-order conditioning**  Temporal Difference (TD) Learning
  - **Blocking**  Rescorla-Wagner Learning

Basics of Learning: Blocking and Rescorla-Wagner Learning

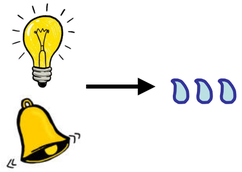
Learn associative strength between a CS and US

$$V(s) \leftarrow V(s) + \alpha \cdot (r - V(s))$$



$V($  $) \leftarrow V($  $) + \alpha \cdot V($ $)$

Introduce a second CS:



$V($  $) \leftarrow V($  $) + \alpha \cdot V($ $)$

$V($  $) = V($  $) :=$ $V($  $)$ + $V($ $)$

$V($  $) \leftarrow V($  $) + \alpha \cdot V($  $))$

What does the value of the sound CS look like at different stages of learning?

Coding: Python, Google Collab

https://github.com/schwartenbeckph/RL-Course/tree/main/2022_04_26