Fraunhofer IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Reinforcement Learning

—

Exercise 1: Markov Decision Processes (MDPs)

Nico Meyer

Opening Remarks

Hello There!

Christopher Mutschler
(course instructor)

Alexander Mattick
(teaching assistant / exercises)

Nico Meyer
(teaching assistant / exercises)

- Alex and myself will take turns in holding the exercise session
- For specific question, please write us an E-Mail or (better) **ask directly in the StudOn forum**
  - Will try to answer your questions there asap -> please try to be as concise as possible
- You will find exercise sheets and code skeletons here: **https://cmutschler.de/rl**

Overview

Exercise Content

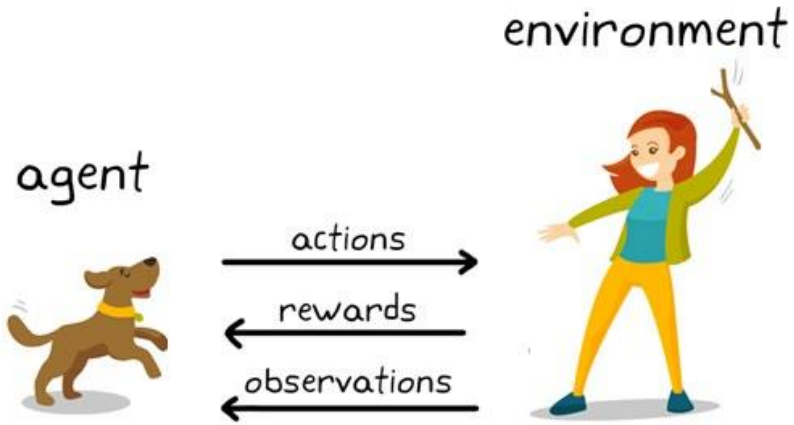
Week	*Date*	*Topic*	*Material*	*Who?*
0			*no exercises*	
1	23.04.	MDPs		Nico
2	30.04.	Dynamic Programming		Alex
3	07.05.	OpenAI Gym, PyTorch-Intro		Alex
4	14.05.	TD-Learning		Nico
5	22.05.	Practical Session (zoom@home)	**Attention: Lecture Slot!**	Nico + Alex
6	28.05.	TD-Control		Nico
7	04.06.	DQN		Nico
8	11.06.	VPG		Alex
9	18.06.	A2C		Nico
10	25.06.	Multi-armed Bandits		Alex
11	02.07.	RND/ICM		Alex
12	09.07.	MCTS		Alex
13	16.07.	BCQ		Nico

Fraunhofer IIS

Recap:
Reinforcement Learning

Recap
The RL paradigm

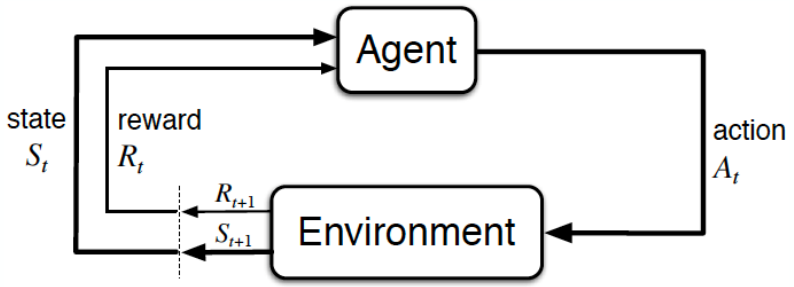
”All goals can be described by the maximization of expected cumulative reward.”



Recap

Markov Decision Processes

- Agent learns by interacting with an environment over many time-steps:
- Markov Decision Process (MDP) is a tool to formulate RL problems
  - Description of an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$:



Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Note:
If the interaction does stop at some point in time (T) then we have an *episodic RL problem.*

- At each step t, the agent:
  - is at state $\mathrm{S_t}$,
  - performs action $\mathrm{A_t}$,
  - receives reward $\mathrm{R_t}$.

- At each step t, the environment:
  - receives action $\mathrm{A_t}$ from the agent,
  - provides reward $\mathrm{R_t}$,
  - moves at state $\mathrm{S_{t+1}}$,
  - increments time $\mathrm{t} \leftarrow \mathrm{t} + 1$.

Fraunhofer IIS

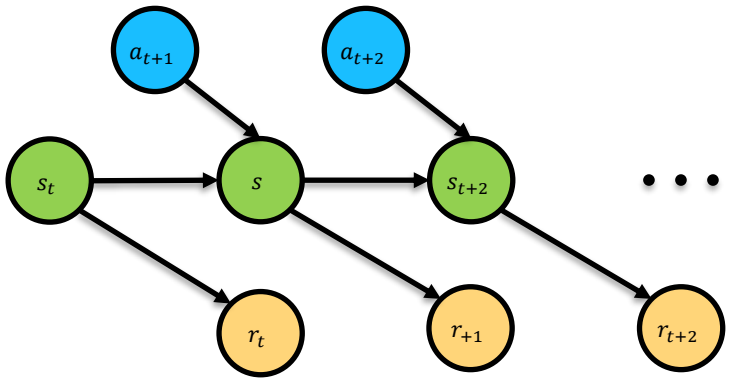
Recap
Markov Property

- Assumption of MDPs: Markov Property
  - A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_1, \dots, S_{t-1}, S_t] = \mathbb{P}[S_{t+1} \mid S_t]$$

  - Past states $S_1, \dots, S_{t-1}$ do not change the outcome for the next state S_{t+1}.
  - The current state S_t captures all relevant information from the history.
  - **"The future is independent of the past given the present"**

$$H_{1:t} \rightarrow S_t \rightarrow H_{t+1:\infty}$$



- State is the information used to determine what happens next
  - Direct (fully observable): $O_t = S_t^e$
  - Indirect (partially observable): $O_t = f(S_t^e)$

Fraunhofer IIS

Recap

Reward and Discounted Return

- We want to "solve" the MDP, by maximizing future rewards.
  - We see the episodes in the form of

$$S_0 \xrightarrow{(A_0, R_0)} S_1 \xrightarrow{(A_1, R_1)} S_2 \xrightarrow{(A_2, R_2)} S_3 \dots \; S_{t-1} \xrightarrow{(A_{t-1}, R_{t-1})} S_t$$

  - **Question:** what happens if our problem never stops (i.e., $T = \infty$)?
    - Examples: data center cooling, recommender systems, etc.

  - Total discounted (γ) reward (**return**) (of one sample)

$$\mathrm{G} = R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \cdots = \sum_{t=0}^{\infty} \gamma^t R_t$$

Fraunhofer IIS

Recap

The Policy π

- Expected long-term value of state s:

$$v(s) = \mathbb{E}(\mathrm{G}) = \mathbb{E}(R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \ldots + \gamma^t R_t)$$

- **Goal: maximize the expected return $\mathbb{E}(\mathrm{G})$.**

- We need a controller that helps us select the actions to maximize $\mathbb{E}(G)$!
- A policy π represents this controller:
  - π determines the agent's behavior, i.e., its way of acting
  - π is a mapping from state space $\mathcal{S}$ to action space $\mathcal{A}$

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

  - Two types of policies:
    - Deterministic policy: $a = \pi(s)$.
    - Stochastic policy: $\pi(a \mid s) = \mathbb{P}[A_t = a \mid S_t = s]$.

- **New goal: find a policy that maximizes the expected return!**

Fraunhofer IIS

Exercise Sheet 1

Markov Decision Processes

Fraunhofer IIS

Fraunhofer
IIS

Fraunhofer-Institut für Integrierte Schaltungen IIS

Thank you for your attention!