

AI Lab - Lesson 5

Reinforcement Learning

Davide Corsi
Alessandro Farinelli

University of Verona
Department of Computer Science

January 17th 2022



UNIVERSITÀ
di **VERONA**

Dipartimento
di **INFORMATICA**

Start Your Working Environment

Start the previously installed (lesson 1) conda environment *ai-lab*

Listing 1: Update Environment

```
cd AI-Lab
git stash (NB: remember to backup the previous lessons before this step!)
git pull
git stash pop
conda activate ai-lab
jupyter notebook
```

Listing 2: Open Lesson

To open the tutorial navigate with your browser to:
`lesson_5/lesson_5_problem.ipynb`

- Your assignments for this lesson are at:
lesson_5/lesson_5_problem.ipynb. You will be required to implement Q-Learning and SARSA algorithms
- In the following you can find the pseudocode

Q-Learning

Input: *environment* $[A, S]$, *problem*, *episodes*, α, γ , *expl_func*, *expl_param*

Output: *policy*, *rewards*, *lengths*

```
1:  $\forall a \in A, \forall s \in S$  initialize  $Q(s, a)$  arbitrarily
2:  $rewards, lengths \leftarrow [0, \dots, 0]$  ▷ Null vectors of length episodes
3: for  $i \leftarrow 0$  to episodes do
4:   Initialize  $s$ 
5:   repeat
6:      $a \leftarrow \text{EXPL\_FUNC}(Q, s, \text{expl\_param})$ 
7:      $s', r \leftarrow \text{take action } a \text{ from state } s$  ▷ Act and observe
8:      $Q(s, a) \leftarrow Q(s, a) + \alpha(R + \gamma \max_{a' \in A_s} Q(s', a') - Q(s, a))$  ▷ TD
9:      $s \leftarrow s'$ 
10:  until  $s$  is terminal
11:  Update rewards, lengths
12:   $\pi \leftarrow [0, \dots, 0]$  ▷ Null vector of length  $|S|$ 
13:  for each  $s$  in  $S$  do ▷ Extract policy
14:     $\pi_s \leftarrow \underset{a \in A_s}{\operatorname{argmax}} Q(s, a)$ 
15: return  $\pi, rewards, lengths$ 
```

SARSA

Input: *environment* $[A, S]$, *problem*, *episodes*, α , γ , *expl_func*, *expl_param*

Output: *policy*, *rewards*, *lengths*

```
1:  $\forall a \in A, \forall s \in S$  initialize  $Q(s, a)$  arbitrarily
2:  $rewards, lengths \leftarrow [0, \dots, 0]$  ▷ Null vectors of length episodes
3: for  $i \leftarrow 0$  to episodes do
4:   Initialize  $s$ 
5:    $a \leftarrow \text{EXPL\_FUNC}(Q, s, \text{expl\_param})$ 
6:   repeat
7:      $s', r \leftarrow \text{take action } a \text{ from state } s$  ▷ Act and observe
8:      $a' \leftarrow \text{EXPL\_FUNC}(Q, s', \text{expl\_param})$ 
9:      $Q(s, a) \leftarrow Q(s, a) + \alpha(R + \gamma Q(s', a') - Q(s, a))$  ▷ TD
10:     $s \leftarrow s'$ 
11:     $a \leftarrow a'$ 
12:   until  $s$  is terminal
13:   Update rewards, lengths
14:  $\pi \leftarrow [0, \dots, 0]$  ▷ Null vector of length  $|S|$ 
15: for each  $s$  in  $S$  do ▷ Extract policy
16:    $\pi_s \leftarrow \underset{a \in A_s}{\text{argmax}} Q(s, a)$ 
17: return  $\pi, rewards, lengths$ 
```