Planning Lab - Lesson 4
Reinforcement Learning (RL)

Luca Marzari and Alessandro Farinelli

University of Verona
Department of Computer Science

Contact: luca.marzari@univr.it

November 16, 2022

Start Your Working Environment

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

> cd Planning-Lab
> conda activate planning-lab
> jupyter notebook

To open the assignment navigate with your browser to:
lesson_4/lesson_4_problem.ipynb

Assignments

- Your assignments for this lesson are at: *lesson_4/lesson_4_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: $rewards, lengths \leftarrow [0, ..., 0]$ ▷ Null vectors of length $episodes$
3: **for** $i \leftarrow 0$ **to** $episodes$ **do**
4:      Initialize s
5:      **repeat**
6:          $a \leftarrow \text{EXPL\_FUNC}(Q, s, expl\_param)$
7:          $s', r \leftarrow$ take action a from state s ▷ Act and observe
8:          $Q(s, a) \leftarrow Q(s, a) + \alpha(R + \gamma \max\limits_{a' \in A_s} Q(s', a') - Q(s, a))$ ▷ TD
9:          $s \leftarrow s'$
10:      **until** s is terminal
11:      Update $rewards, lengths$
12: $\pi \leftarrow [0, ..., 0]$ ▷ 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: $environment\ [A, S]$, $problem, episodes, \alpha, \gamma, expl\_func, expl\_param$
Output: $policy, rewards, lengths$

1: $\forall a \in A, \forall s \in S$ initialize $Q(s, a)$ arbitrarily
2: $rewards, lengths \leftarrow [0, ..., 0]$ ▷ Null vectors of length $episodes$
3: **for** $i \leftarrow 0$ **to** $episodes$ **do**
4:      Initialize s
5:      $a \leftarrow \text{EXPL\_FUNC}(Q, s, expl\_param)$
6:      **repeat**
7:          $s', r \leftarrow$ take action a from state s ▷ Act and observe
8:          $a' \leftarrow \text{EXPL\_FUNC}(Q, s', expl\_param)$
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: $\pi \leftarrow [0, ..., 0]$ ▷ 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** $\pi, rewards, lengths$