

## Wprowadzenie do RL 9

UWAGA: przed przystąpieniem do zadań zapoznaj się z prezentacją [RL\\_6\\_wykład.pdf](#).

### Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size  $\alpha \in (0, 1]$ , small  $\varepsilon > 0$

Initialize  $Q(s, a)$ , for all  $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$ , arbitrarily except that  $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

    Initialize  $S$

    Loop for each step of episode:

        Choose  $A$  from  $S$  using policy derived from  $Q$  (e.g.,  $\varepsilon$ -greedy)

        Take action  $A$ , observe  $R, S'$

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$

$S \leftarrow S'$

    until  $S$  is terminal

### Zadanie 1

Zaimplementuj algorytm **Q-learning** w środowisku **FrozenLake**.

### Zadanie 2

Zaimplementuj algorytm **Q-learning** w środowisku **CartPole**.

