

Lab 9 – Q-Learning

State \Leftrightarrow cell

Action $\in \{left, right, up, down\}$

Reward = $\begin{cases} INF, & \text{for a transition that ends in state } G \\ -1, & \text{for any other transition} \end{cases}$

Parameters:

- $\alpha \in (0,1)$ – learning rate
- $\gamma \in (0,1)$ – discount factor
- number of episodes

Initially, Q values are usually set to 0.

Initialize $Q(s, a)$ arbitrarily

Repeat (for each episode):

Initialize s // $s = \text{initial state}$

Repeat (for each step of episode):

Choose a from s using policy derived from $Q(*)$

Take action a , observe r, s'

Update

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)] (**)$$

$s \leftarrow s'$;

Until s is terminal

$Q(s, a)$ = long term reward if we choose action a from state s and then follow the policy

(*) $\pi(s) = \operatorname{argmax}_a Q(s, a)$ (or alternatively, you can use ϵ -greedy exploration)

