Exercise on Reinforcement Learning

Alberto Maria Metelli

October 21, 2021

Consider the following sequential decision making-problem. An agent in a 3×3 grid can move in the four directions or stay still, provided that it does not crush against a border. Whenever performing a valid action, the agent reaches **deterministically** to the corresponding cell. The interaction starts in the lower left cell (blue) and the upper right cell (green) is a terminal state. The immediate reward is represented in the following grid:

0	0	2
-1	-10	0
0	-1	0

1. Formalize the problem as a Markov decision process (MDP);
2. For which values of the discount factor $\gamma \in [0, 1]$ the optimal policy consists in staying in the initial state forever?
3. Simulate the execution of Q-learning, starting with a Q-table initialized with the immediate reward, supposing to have observed the following trajectories:

$$
\begin{aligned}
&(0,0) \xrightarrow{\rightarrow} (1,0) \xrightarrow{\uparrow} (1,1) \xrightarrow{\rightarrow} (2,1) \xrightarrow{\uparrow} (2,2) \\
&(2,1) \xrightarrow{\downarrow} (2,0) \xrightarrow{\uparrow} (2,1) \\
&(1,1) \xrightarrow{\downarrow} (1,0) \xrightarrow{\rightarrow} (2,0) \\
&(0,0) \xrightarrow{\rightarrow} (1,0) \xrightarrow{\uparrow} (1,1)
\end{aligned}
$$

   Use discount factor $\gamma = 0.9$ and learning rate $\alpha = 1$.

4. Say which is the greedy policy once completed the updates of the Q-table.

Formalization We numerate rows and columns from 0 starting from the lower left cell.

$$
\mathcal{S} = \{(i,j) \,:\, i,j \in \{0,1,2\}\},
$$

$$
\begin{aligned}
\mathcal{A} &= \{(\Delta i, \Delta j) \,:\, \Delta i, \Delta j \in \{-1, 0, +1\} \wedge |\Delta i| + |\Delta j| \leq 1\} \\
&= \{(-1,0), (0,+1), (0,-1), (0,+1), (0,0)\}.
\end{aligned}
$$

An action $(\Delta i, \Delta j)$ is admissible in a state (i,j) if $i + \Delta i, j + \Delta j \in \{0,1,2\}$. In such a case, the next state is given by:

$$
\mathcal{P}((i',j')|(i,j),(\Delta i, \Delta j)) = \mathbb{1}\{(i',j') = (i + \Delta i, j + \Delta j)\}.
$$

The initial state distribution is deterministic on $(0,0)$, i.e, $\mu_0((i,j)) = \mathbb{1}\{(i,j) = (0,0)\}$. The reward function is a function of the state only and is defined as represented in the grid.

Optimal Policy varying γ It is not hard to prove that this problem, depending on the value of γ can admit two possible optimal policies: either staying still in the initial state or moving to the terminal state, with the minimum number of steps, avoiding passing through the -10 cell (two possible paths, leading to the same reward are possible). Let us compute the value function of these two policies:

$$V^{\pi_{\text{still}}}((0,0)) = 0,$$
$$V^{\pi_{\text{go}}}((0,0)) = 0 + \gamma \cdot (-1) + \gamma^2 \cdot 0 + \gamma^3 \cdot 2 = -\gamma + 2\gamma^3.$$

Requiring that $V^{\pi_{\text{still}}}((0,0)) > V^{\pi_{\text{go}}}((0,0))$ leads to $\gamma < \frac{1}{\sqrt{2}}$.

Q-learning Simulation In gray, the Q-table cells of the actions that are not allowed.

	$r(s)$	$Q(s,a)$					$V(s) =$
		$(-1,0)$	$(+1,0)$	$(0,-1)$	$(0,+1)$	$(0,0)$	$\max_{a \in \mathcal{A}} Q(s,a)$
$(0,0)$	0		0, $\mathbf{-0.9}$[1], $\mathbf{0.4122}$[9]		0	0	0, $\mathbf{0.4122}$[9]
$(0,1)$	-1		-1	-1	-1	-1	-1
$(0,2)$	0		0	0		0	0
$(1,0)$	-1	-1	-1, $\mathbf{0.458}$[8]		-1, $\mathbf{-10}$[2], $\mathbf{-10}$[10]	-1	-1, $\mathbf{0.458}$[8]
$(1,1)$	-10	-10	-10, $\mathbf{-10}$[3], $\mathbf{-10.9}$[7]	-10	-10	-10	-10
$(1,2)$	0	0	0	0		0	0
$(2,0)$	0	0			0, $\mathbf{1.62}$[6]	0	0, $\mathbf{1.62}$[6]
$(2,1)$	0	0		0, $\mathbf{0}$[5]	0, $\mathbf{1.8}$[4]	0	0, $\mathbf{1.8}$[4]
$(2,2)$	2						2

We apply the update rule for each transition in order:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \left(r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s',a') \right)$$

[1] $(0,0) \xrightarrow{\rightarrow} (1,0) \implies Q((0,0),(+1,0)) \leftarrow r((0,0)) + \gamma \max_{a \in \mathcal{A}} Q((1,0),a) = 0 + 0.9 \cdot (-1) = -0.9$

[2] $(1,0) \xrightarrow{\uparrow} (1,1) \implies Q((1,0),(0,+1)) \leftarrow r((1,0)) + \gamma \max_{a \in \mathcal{A}} Q((1,1),a) = -1 + 0.9 \cdot (-10) = -10$

[3] $(1,1) \xrightarrow{\rightarrow} (2,1) \implies Q((1,1),(+1,0)) \leftarrow r((1,1)) + \gamma \max_{a \in \mathcal{A}} Q((2,1),a) = -10 + 0.9 \cdot 0 = -10$

[4] $(2,1) \xrightarrow{\rightarrow} (2,2) \implies Q((2,1),(0,+1)) = r((2,1)) + \gamma \max_{a \in \mathcal{A}} Q((2,2),a) = 0 + 0.9 \cdot 2 = 1.8$

[5] $(2,1) \xrightarrow{\downarrow} (2,0) \implies Q((2,1),(0,-1)) = r((2,1)) + \gamma \max_{a \in \mathcal{A}} Q((2,0),a) = 0 + 0.9 \cdot 0 = 0$

[6] $(2,0) \xrightarrow{\uparrow} (2,1) \implies Q((2,0),(0,+1)) = r((2,0)) + \gamma \max_{a \in \mathcal{A}} Q((2,1),a) = 0 + 0.9 \cdot 1.8 = 1.62$

[7] $(1,1) \xrightarrow{\downarrow} (1,0) \implies Q((1,1),(0,-1)) = r((1,1)) + \gamma \max_{a \in \mathcal{A}} Q((1,0),a) = -10 + 0.9 \cdot (-1) = -10.9$

[8] $(1,0) \xrightarrow{\rightarrow} (2,0) \implies Q((1,0),(+1,0)) = r((1,0)) + \gamma \max_{a \in \mathcal{A}} Q((2,0),a) = -1 + 0.9 \cdot 1.62 = 0.458$

[9] $(0,0) \xrightarrow{\rightarrow} (1,0) \implies Q((0,0),(+1,0)) = r((0,0)) + \gamma \max_{a \in \mathcal{A}} Q((1,0),a) = 0 + 0.9 \cdot 0.458 = 0.4122$

[10] $(1,0) \xrightarrow{\uparrow} (1,1) \implies Q((1,0),(0,+1)) = r((1,0)) + \gamma \max_{a \in \mathcal{A}} Q((1,1),a) = -1 + 0.9 \cdot (-10) = -10$

Greedy Policy The greedy policy is represented in the following:

all	all	
all	all except $(+1,0)$	$(0,+1)$
$(+1,0)$	$(+1,0)$	$(0,+1)$