Prof. Joschka Boedecker, Gabriel Kalweit, Maria Huegle

REINFORCEMENT LEARNING
Solution 5

1 Q-learning: Pen and Paper

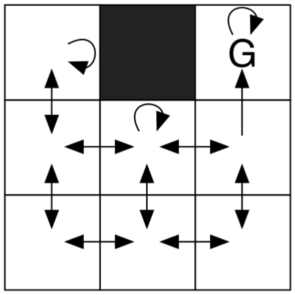

Figure 1: Grid MDP

Consider the deterministic MDP in Figure 1. There exists a terminal state G and a wall that cannot be entered. The agent remains in its current position if it chooses an action that moves it against the wall or off the grid. All transitions have a reward of -1. We discount with 0.5.

(a) How is the Q-learning update defined for a transition from state i to j, if j is a terminal state?

Solution. The Q-learning update was defined as $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$. If there is no next state, i.e. the transition led to a terminal state, then the update boils down to $Q(S, A) \leftarrow Q(S, A) + \alpha[R - Q(S, A)]$.

(b) We initialize all Q-values with 0. The agent starts in the upper left corner. It then moves one cell down, then one cell to the right and tries unsuccessfully to move one cell upwards (i.e. remains in its current cell), then moves one cell to the right and finally moves upwards into the terminal state. Which values of the Q-function change during this episode if we apply Q-learning with a learning rate of 1.0? Calculate the updated Q-function after this first episode. Repeat the calculation for a second identical episode.

Solution. Let $s_{i,j}$ denote the cell in row i and column j. We update the following state-action pairs:

$$\{(s_{0,0}, \text{down}), (s_{1,0}, \text{right}), (s_{1,1}, \text{up}), (s_{1,1}, \text{right}), (s_{1,2}, \text{up})\}.$$

$\alpha = 1.0$, $\gamma = 0.5$. The updates for the first episode are:
$Q(s_{0,0}, \text{down}) = 0 + 1.0 \cdot (-1 + 0.5 \cdot 0 - 0) = -1$

$Q(s_{1,0}, \text{right}) = 0 + 1.0 \cdot (-1 + 0.5 \cdot 0 - 0) = -1$
$Q(s_{1,1}, \text{up}) = 0 + 1.0 \cdot (-1 + 0.5 \cdot 0 - 0) = -1$
$Q(s_{1,1}, \text{right}) = 0 + 1.0 \cdot (-1 + 0.5 \cdot 0 - 0) = -1$
$Q(s_{1,2}, \text{up}) = 0 + 1.0 \cdot (-1 - 0) = -1$

The updates for the second episode are:
$Q(s_{0,0}, \text{down}) = -1 + 1.0 \cdot (-1 + 0.5 \cdot 0 - (-1)) = -1$
$Q(s_{1,0}, \text{right}) = -1 + 1.0 \cdot (-1 + 0.5 \cdot 0 - (-1)) = -1$
$Q(s_{1,1}, \text{up}) = -1 + 1.0 \cdot (-1 + 0.5 \cdot 0 - (-1)) = -1$
$Q(s_{1,1}, \text{right}) = -1 + 1.0 \cdot (-1 + 0.5 \cdot 0 - (-1)) = -1$
$Q(s_{1,2}, \text{up}) = -1 + 1.0 \cdot (-1 - (-1)) = -1$

(c) Calculate the optimal Q-values $Q_*(s, a)$ for all state-action pairs.

Solution. From states that are neighbor of the goal state, an optimal policy would move the agent into the goal state. Therfore $\pi_*(s_{1,2}) = \text{up}$ and $Q_*(s_{1,2}, \text{up}) = -1$. All other actions have to cost more, since we later definitely have to choose action up if we again end up in state $s_{1,2}$ and want to move in the goal state. It follows $Q_*(s_{1,2}, \text{right}) = -1 + 0.5 \cdot \max_a Q(s_{1,2}, a) = -1 + 0.5 \cdot (-1) = -1.5$.

In states that are neighbors of $s_{1,2}$, the optimal policy has to move the agent into $s_{1,2}$. $\pi_*(s_{1,1}) = \text{right}$ and $Q_*(s_{1,1}, \text{right}) = -1 + 0.5 \cdot (-1) = -1.5$. Furthermore, $\pi_*(s_{2,2}) = \text{up}$ and $Q_*(s_{2,2}, \text{up}) = -1 + 0.5 \cdot (-1) = -1.5$. Exemplary, the following actions lead to a worse value: $Q(s_{1,1}, \text{up}) = -1 + 0.5 \cdot (-1.5) = -1.75$ or $Q(s_{1,2}, \text{left}) = -1 + 0.5 \cdot (-1.5) = -1.75$. The same holds for other non-optimal actions analogously.

In state $s_{2,1}$ an optimal policy can move the agent either in $s_{1,1}$ or $s_{2,2}$, since both have the same optimal value. $Q_*(s_{2,1}, \text{up}) = -1 + 0.5 * (-1.5) = -1.75$ and $Q_*(s_{2,1}, \text{right}) = -1 + 0.5 \cdot (-1.5) = -1.75$.

Lastly, $Q_*(s_{1,0}, \text{right}) = -1.75$, $Q_*(s_{0,0}, \text{down}) = -1 + 0.5 \cdot -1.75 = 1\frac{7}{8}$ and $Q_*(s_{2,0}, \text{up}) = Q_*(s_{2,0}, \text{right}) = 1\frac{7}{8}$.