# Homework 5 – Reinforcement learning

Luisa Santo and Filipe Soares, A59

May 18, 2018

## 1 Exercise 1

**(a)** At time step $t$, the Q-values estimated by the for state (16, 3) is:

$$Q^{(t)}_{(16,3)} = \begin{bmatrix}0.25 & 0.32 & 0.32 & 0.25\end{bmatrix} Q^{(t)}_{(16,4)} = \begin{bmatrix}0.29 & 0.36 & 0.36 & 0.29\end{bmatrix} \quad (1)$$

The Q-learning update is given by

$$Q^{(t+1)}(x_t, a_t) = Q^{(t)}(x_t, a_t) + \alpha(c_t + \gamma \min_{a' \in A} Q^{(t)}(x_{t+1}, a') - Q^{(t)}(x_t, a_t)) \quad (2)$$

Therefore, the transitions in time step $t$ thus result in the following updates:

$$Q^{(t+1)}(x_t, a_t) = Q^{(t)}((16,3), D) + 0.1(0.05 + 0.95 \min_{a' \in A} Q^{(t)}_{(16,4)a'}) - Q^{(t)}((16,3), D) =$$

$$= 0.32 + 0.1(0.05 + 0.95 \times 0.29 - 0.32) = 0.32.$$

**(b)** On SARSA, the update rule is given by

$$Q^{(k+1)}(x_t, a_t) = Q^{(k)}(x_t, a_t) + \alpha_t(c_t + \gamma Q^{(k)}(x_{t+1}, a_{t+1}) - Q^{k}(x_t, a_t)) \quad (3)$$

Therefore, the transitions in time step $t$ thus result in the following updates:

$$Q^{(t+1)}(x_t, a_t) = Q^{(t)}((16,3), D) + 0.1(0.05 + 0.95 Q^{(t)}_{((16,4),L)} - Q^{(t)}((16,3), D) =$$
$$= 0.32 + 0.1(0.05 + 0.95 \times 0.36 - 0.32) = 0.32.$$

**(c)** Off-policy refers to reinforcement learning methods that learn the value of a policy while following a different policy. On-policy refers to reinforcement learning methods that learn the value of the policy that the agent is following.

In Q-learning, the update rule is

$$Q^{(k+1)}(x_t, a_t) = Q^{(k)}(x_t, a_t) + \alpha_t(c_t + \gamma \min_{\alpha \in A} Q^{(k)}(x_{t+1}, a) - Q^{(k)}(x_t, a_t)) \quad (4)$$

and we can observe that the update is based on the value of the policy at the next state, independently of the policy that the agent is actually following.

In this case, off-policies is advisable for exploration problems. Conversely, on SARSA, the update rule is

$$Q^{(k+1)}(x_t, a_t) = Q^{(k)}(x_t, a_t) + \alpha_t(c_t + \gamma Q^{(k)}(x_{t+1}, a_{t+1}) - Q^k(x_t, a_t)) \quad (5)$$

and we can observe that the update is based on the value of the actual policy followed by the agent at the next state. Therefore, if a good starting policy is available, on-policies is advisable. For this reason, Q-learning is an off-policy method while SARSA is an on-policy method.