ADI TPC 5 - Grupo 3

a) $Q_{t+1}((16,3),D) = Q_t((16,3),D) + 0{,}1\left[0{,}05 + 0{,}95 \min_{a' \in A} Q_t((16,4),a') - Q_t((16,3),D)\right] =$

$= 0{,}32 + 0{,}1\left[0{,}05 + 0{,}95 \times 0{,}29 - 0{,}32\right] =$

$= 0{,}32 + 0{,}1\left[0{,}05 + 0{,}2755 - 0{,}32\right] =$

$= 0{,}32 + 0{,}1\left[0{,}00055\right] = 0{,}320055 \Rightarrow Q^{(t+1)}_{(16,3)} = \left[0{,}25 \;\; 0{,}320055 \;\; 0{,}32 \;\; 0{,}25\right]$

b) $Q_{t+1}((16,3),D) = Q_t((16,3),D) + 0{,}1\left[0{,}05 + 0{,}95\, Q_t((16,4),L) - Q_t((16,3),D)\right] =$

$= 0{,}32 + 0{,}1\left[0{,}05 + 0{,}95 \times 0{,}36 - 0{,}32\right] =$

$= 0{,}32 + 0{,}1\left[0{,}072\right] = 0{,}32 + 0{,}0072 = 0{,}3272$

$\Rightarrow Q^{(t+1)}_{(16,3)} = \left[0{,}25 \;\; 0{,}3272 \;\; 0{,}32 \;\; 0{,}25\right]$

c). Q-learning is an off-policy algorithm since it learns the value of one policy while following another.

- SARSA is an on-policy algorithm since it learns the value of the policy that it follows.

- To illustrate these definitions, we can verify that in exercise 1a) the Q-learning algorithm used a greedy action ("up" with $Q_t(10,6) = 0,29$) to update the Q_t value, even though the original policy chose the action "left" at $t+1$.
- Also, in exercise 1b), since SARSA is an on-policy algorithm it uses the Q_t value based on the current policy and uses the action "left" with $Q_t(10,4) = 0,36$ to compute the update.
- This illustrated distinction disappears if the current policy is a greedy policy.