Planning, Learning and Decision Making: Homework 5. Reinforcement learning

Mykhaylo Marfeychuk, Roy De Prins

04 December 2018

Excercise 1) a

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t[c_t + \gamma \min_{a' \in A} Q_t(x_{t+1}, a') - Q_t(x_t, a_t)]$$

$$Q_{t+1}(x_t, a_t) = 3.08 + 0.1[1.0 + 0.99 * \min(3.08, 3.25, 3.57, 3.22) - 3.08]$$

$$Q_{t+1}(x_t, a_t) = 3.18$$

Excercise 1) b

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t[c_t + \gamma Q_t(x_{t+1}, a_{t+1}) - Q_t(x_t, a_t)]$$

$$Q_{t+1}(x_t, a_t) = 3.08 + 0.1[1.0 + 0.99 * 3.22 - 3.08]$$

$$Q_{t+1}(x_t, a_t) = 3.19$$

Excercise 1) c

An off-policy algorithm learns the value of one policy while following another policy. An on-policy algorithm learns the value of the policy that it follows. In the case of SARSA it leads to more 'stable' updates than Q-Learning (less chance of a high cost), but it will be less optimal in terms of 'steps'. SARSA is less 'greedy' than Q-Learning. This can also be seen by the higher Q-value produced by the SARSA algorithm in comparison with the Q-Learning algorithm.