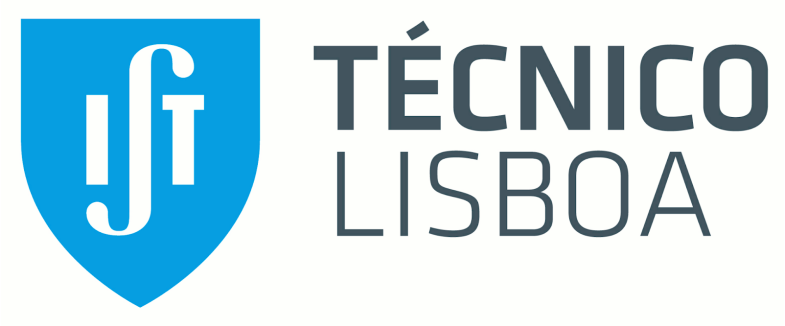

Planning, Learning and Intelligent Decision Making

2022/2023

Homework 4 - Group 25

Guilherme Pereira	105806
Miguel Belbute	96453

Introduction

This project was developed for the course Planning, Learning and Intelligent Decision Making taught at Instituto Superior Técnico under the professor Francisco Saraiva de Melo.

Exercise 1

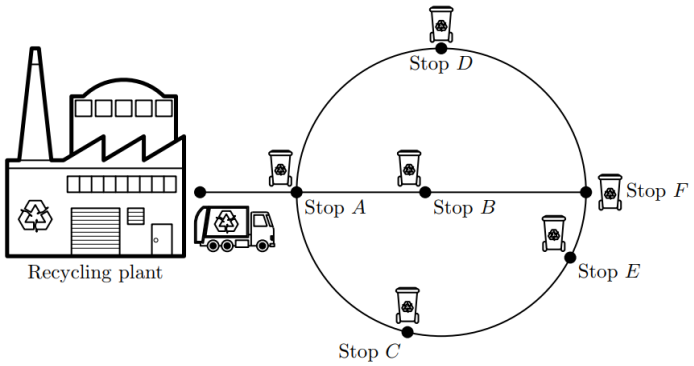

Figure 1: The garbage truck must collect garbage from B, C and D.

a) Given the mentioned trajectory τ, defined by

$$\tau = \{x_{t-1}, \mathrm{a}_{t-1}, \mathrm{c}_{t-1}, x_t, \mathrm{a}_t, \mathrm{c}_t, x_{t+1}, \mathrm{a}_{t+1}, \mathrm{c}_{t+1}\}$$

Where

$$x_{t-1} = (E, 1, 0, 1),$$
$$a_{t-1} = D,$$
$$c_{t-1} = 1.0,$$
$$x_t = (E, 1, 0, 1),$$
$$a_t = R,$$
$$c_t = 0.2,$$
$$x_{t+1} = (F, 1, 0, 1),$$
$$a_{t+1} = R,$$
$$c_{t+1} = 1.0,$$

And knowing that at time step t, the Q-values estimated by the agent for state (E, 1, 0, 1) and for state (F, 1, 0, 1) are, respectively,

$$\mathbf{Q}^{(t)}_{(E,1,0,1),:} = \begin{bmatrix} 2.8 & 2.8 & 2.8 & 2.8 & 2.54 & 2.0 \end{bmatrix}$$

$$\mathbf{Q}^{(t)}_{(F,1,0,1),:} = \begin{bmatrix} 2.8 & 2.8 & 2.95 & 2.0 & 3.14 & 2.8 \end{bmatrix}$$

We can use the following expression from the theoretical classes to calculate the Q-values after a Q-learning update with step-size α = 0.1 and γ = 0.9, resulting from the transition at time step t:

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

From the sample $(x_{t-1}, a_{t-1}, c_{t-1}, x_t)$ we can calculate $Q^{(t+1)}_{(E,1,0,1),D}$:

$$Q^{(t+1)}_{(E,1,0,1),D:} = 2.8 + 0.1\ (1.0 + 0.9\ min([\ 2.8\ 2.8\ 2.8\ 2.8\ 2.54\ 2.0\ ])\ -\ 2.8\ ) = 2.8$$

From the sample (x_t, a_t, c_t, x_{t+1}) we can calculate $Q^{(t+1)}_{(E,1,0,1),R}$:

$$Q^{(t+1)}_{(E,1,0,1),R:} = 2.0 + 0.1\ (0.2 + 0.9\ min([\ 2.8\ 2.8\ 2.95\ 2.0\ 3.14\ 2.8\ ])\ -\ 2.8\ ) = 1.92$$

From these results, we can now update $Q^{(t)}_{(E,1,0,1),:}$ such that it becomes:

$$Q^{(t+1)}_{(E,1,0,1),:} = [\ 2.8\ 2.8\ 2.8\ 2.8\ 2.54\ 1.92\ ]$$

Note: We are unable to calculate the updated Q-values for the state (F, 1, 0, 1) because we don't know which action will be taken at the timestep t+2.

b) Similarly to the first exercise, we can use the following expression from the theoretical classes to calculate the Q-values after a SARSA update with step-size α = 0.1 and γ = 0.9, resulting from the transition at time step t:

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

From the sample $(x_{t-1}, a_{t-1}, c_{t-1}, x_t, a_t)$ we can calculate $Q^{(t+1)}_{(E,1,0,1),D}$:

$$Q^{(t+1)}_{(E,1,0,1),D:} = 2.8 + 0.1\ (1.0 + 0.9 * 2.0\ -\ 2.8\ ) = 2.8$$

From the sample $(x_t, a_t, c_t, x_{t+1}, a_{t+1})$ we can calculate $Q^{(t)}_{(E,1,0,1),R}$:

$$Q^{(t+1)}_{(E,1,0,1),R:} = 2.0 + 0.1\ (0.2 + 0.9 * 2.8\ -\ 2.0) = 2.072$$

From these results, we can now update $Q^{(t)}_{(E,1,0,1),:}$ such that it becomes:

$$Q^{(t+1)}_{(E,1,0,1),:} = [\ 2.8\ 2.8\ 2.8\ 2.8\ 2.54\ 2.072\ ]$$

Note: We are unable to calculate the updated Q-values for the state (F, 1, 0, 1) because we don't know which action will be taken at the timestep t+2.

c) The main difference between **off-policy** and **on-policy** learning, which correspond to the Q-learning (1a) and SARSA (1b) algorithms respectively, is that in the first case the agent is learning the value of a policy (more specifically, the optimal policy) without necessarily following it, whereas in the latter case the agent is learning the value of the policy that it follows (whichever it may be).

This also means SARSA needs policy improvements to converge to an optimal policy (for example, with the usage of ε-greedy with decaying ε), the trade-off being it has **more stable updates** - when following a strictly optimal policy, the agent can get into unfavorable situations due to exploration mechanisms (as an example, see the file regarding lecture 11, slide 82).

This way, we can state that the agent is learning the optimal policy in 1a, despite whatever other policy he might be following to perform actions. In exercise 1b, the agent is learning the policy it is following, not being guaranteed that it is the optimal policy.

In the previous exercises, we can also distinguish a slightly more stable update in SARSA, regarding state (E,1,0,1) and action R: In the Q-learning update, the corresponding Q-value suffers a difference of 0.08, while in the SARSA algorithm the corresponding Q-value suffers a difference of 0.072.

References

[1] - De Melo, F. et al. (2023) Theoretical Lectures, Planning, Learning and Intelligent Decision Making. Instituto Superior Técnico (IST). Available at: https://fenix.tecnico.ulisboa.pt/disciplinas/ADI/2022-2023/2-semestre/material-de-apoio (Accessed: March 24, 2023).