# Introduction to Artificial Intelligence (INFO8006)

## Exercises 7 – Reinforcement learning

September 16, 2022

## Learning outcomes

At the end of this session you should be able to

- differentiate passive and active RL;
- define and apply direct utility estimation and temporal-difference learning;
- define and apply $Q$-learning.

## Exercise 1 Passive RL

| | | | |
|---|---|---|---|
| 3 | $-6$ | $-1$ | $+5$ |
| 2 | | | |
| 1 | ᗧ | $-8$ | $+3$ |
| | 1 | 2 | 3 |

Consider the grid-world given above and an agent who is trying to learn the optimal policy. The agent starts from the bottom-left corner and can take the actions `north` ($N$), `south` ($S$), `west` ($W$) and `east` ($E$). Rewards are only awarded for reaching the terminal (shaded) states. You observe the following trials, whose trajectories are sequences of tuples $(s_t^i, r_t^i, a_t^i, s_{t+1}^i)$.

| $t$ | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 |
|---|---|---|---|---|---|
| 0 | $(1,1), 0, N, (1,2)$ | $(1,1), 0, N, (1,2)$ | $(1,1), 0, N, (1,2)$ | $(1,1), 0, N, (1,2)$ | $(1,1), 0, N, (1,2)$ |
| 1 | $(1,2), 0, E, (2,2)$ | $(1,2), 0, E, (2,2)$ | $(1,2), 0, E, (2,2)$ | $(1,2), 0, E, (2,2)$ | $(1,2), 0, E, (2,2)$ |
| 2 | $(2,2), 0, N, (2,3)$ | $(2,2), 0, E, (3,2)$ | $(2,2), 0, S, (2,1)$ | $(2,2), 0, E, (3,2)$ | $(2,2), 0, E, (3,2)$ |
| 3 | $(2,3), -1, \varnothing, \varnothing$ | $(3,2), 0, N, (3,3)$ | $(2,1), -8, \varnothing, \varnothing$ | $(3,2), 0, W, (2,2)$ | $(3,2), 0, S, (3,1)$ |
| 4 | | $(3,3), +5, \varnothing, \varnothing$ | | $(2,2), 0, N, (2,3)$ | $(3,1), +3, \varnothing, \varnothing$ |
| 5 | | | | $(2,3), -1, \varnothing, \varnothing$ | |

Assuming a discount factor $\gamma = 1$,

1. Perform direct utility estimation of the expected utilities $V^\pi(s)$, given the four first trials.
2. Update the estimated expected utilities with respect to the fifth trial using temporal-difference learning. Assume a learning rate $\alpha = 0.5$.

## Exercise 2 Q-learning

An agent is in an unknown environment where there are three states $\{A, B, C\}$ and two actions $\{0, 1\}$. We are given the following tuples $(s, a, r, s')$, generated by taking actions in the environment.

| $s$ | $a$ | $r$ | $s'$ |
|---|---|---|---|
| $A$ | 0 | $+2$ | $A$ |
| $C$ | 1 | $-2$ | $A$ |
| $B$ | 1 | $+1$ | $B$ |
| $A$ | 0 | $-1$ | $B$ |
| $B$ | 1 | $-2$ | $C$ |
| $C$ | 0 | $+4$ | $B$ |
| $B$ | 0 | $+1$ | $A$ |

Assuming a discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.75$,

1. Apply the $Q$-learning algorithm to obtain state-action-value $Q(s, a)$ estimates. Estimates are initialized to 0.
2. We now switch to a feature-based estimator $\hat{Q}(s, a) = w_0 + w_1 f_1(s, a)$, with $f_1(s, a) = 2a - 1$. Starting from weights $w_0 = w_1 = 0$, update the weights according to the approximate $Q$-learning algorithm.

## Exercise 3 Football

ULiège's football team is playing against UCL's team next week. With a lot of losses this season, Liège needs to improve their attack strategy to win the game and increase their popularity. Luckily, the team captain follows INFO8006 and knows how to model the attack as a Markov Decision Process. The captain considers four states `close` (`C`), `away` (`A`), `fail` (`F`), and `goal` (`G`), and two actions `pass` (`P`) and `shoot` (`S`). Although the transition probabilities are unsure, the possible transitions $(s, a, s')$ are known. To each transition is associated an increase/decrease of the team's popularity.

| $s$ | $a$ | $s'$ | $R(s, a, s')$ |
|---|---|---|---|
| C | P | C | $+1$ |
| C | P | A | $-1$ |
| C | P | F | $-2$ |
| C | S | C | $+3$ |
| C | S | F | $-5$ |
| C | S | G | $+10$ |
| A | P | C | $+2$ |
| A | P | A | $0$ |
| A | P | F | $-3$ |
| A | S | C | $+3$ |
| A | S | F | $-10$ |
| A | S | G | $+20$ |

The current strategy of the team is to always shoot. Last match, they had several attack opportunities, resulting in the following actions.

| $s$ | C | C | C | C | A | A | A | A |
|---|---|---|---|---|---|---|---|---|
| $a$ | S | S | S | S | S | S | S | S |
| $s'$ | G | C | G | F | F | C | F | F |

Assuming a discount factor $\gamma = 0.75$ and a learning rate $\alpha = 0.25$,

1. Build an estimator of the transition model $P(s'|s, a)$ and, from it, determine the expected utility $V^\pi$ of the team's current policy $\pi$.

The captain found the tapes of the previous season where they had much more success. Together with the team, the captain selects the following instructive actions.

| $s$ | C | C | A | A | C | A | A | C |
|---|---|---|---|---|---|---|---|---|
| $a$ | S | S | S | P | P | P | S | P |
| $s'$ | G | C | F | C | C | C | F | F |

2. Given the selected tuples, apply the $Q$-learning algorithm to obtain state-action-value $Q(s, a)$ estimates. Estimates are initialized to 0.

3. Determine the optimal policy according to the state-action-value estimates.

## Supplementary materials

- Playing Atari with Deep Reinforcement Learning
- Chapter 21 of the reference textbook.