- Assignment 3 -
Reinforcement Learning Spring 2025

1. We consider a corridor environment of 2 rooms. The environment is represented with 2 non-terminal states $\mathcal{S} = \{1, 2\}$ and one terminating state 3. The state labels correspond to the room numbers; for instance, state 2 corresponds to Room 2. The leftmost room is Room 1 and the rightmost room is Room 3. The actions are LEFT (ActL) and RIGHT (ActR) with their usual behaviour on the left edge.

   Initialization: $Q(s, ActL) = 10$**, and** $Q(s, ActR) = 3$ $s \in \mathcal{S}$ **and for the terminal state** $Q(3, a) = 0$ $\forall a$.

   Applying Dyna-Q with $n = 2$, we obtain the following episode:

t	S_t	A_t	R_{t+1}
0	Room 1	ActR	-1
1	Room 2	ActR	+10
2	Room 3 (terminate)	-	-

   The line references (a) – (f) below, refer to the line references in Dyna-Q pseudo-code on page 164 at the textbook. For the planning step (f), as a convention we use the last n state, action pairs that we have encountered during our real interaction with the environment. The last encountered pair is used first. (These conventions are only to create predictable output for this exercise not a property of DynaQ!) If there is only one real interaction available, it is used repeatedly as much as needed.

   The discount rate is $\gamma = 1$. The step size is $\alpha = 0.5$. We will keep track of $Q(s, a)$. In the quiz you will be asked the below quantities for different state action pairs. To refer to these quantities in the quiz, we will use the notation Q(s, a, time, line-reference) as explained below.

   "After line (x)" means just after executing line (x) and before executing the next line. Note that for line (f), one needs to calculate the result of the whole loop under line (f).

   (a) Time t=0: After line (c), find $Q(s, a)$, i.e. $Q(s, a, 0, c)$.
   (b) Time t=0: After line (d), find $Q(s, a)$, i.e. $Q(s, a, 0, d)$.
   (c) Time t=0: After line (f), find $Q(s, a)$, i.e. $Q(s, a, 0, f)$.
   (d) Time t=1: After line (c), find $Q(s, a)$, i.e. $Q(s, a, 1, c)$.
   (e) Time t=1: After line (d), find $Q(s, a)$, i.e. $Q(s, a, 1, d)$.
   (f) Time t=1: After line (f), find $Q(s, a)$, i.e. $Q(s, a, 1, f)$.

2. The environment is represented with 2 non-terminal states $\mathcal{S} = \{1, 2\}$ and one terminating state 3. For some policy π we already know that $v_\pi(1) = 3$ and $v_\pi(2) = 5$. However, we now want to find a linear function approximator. We decide on the features

$$x_1(s) = s$$
$$x_2(s) = (s - 1)^2$$

   and thus use the approximator

$$\hat{v}(s, \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{x}(s) = \sum_{i=1}^{2} w_i x_i(s).$$

Show that it is possible to make $\hat{v}(s, \boldsymbol{w}) = v_\pi(s)$ in this case. What should w_1 and w_2 be?

Note: This is a conceptual question. You do not need to use any RL-methods to solve it.

3. We consider a corridor environment of 100 rooms. The environment is represented with 100 non-terminal states $\mathcal{S} = \{1, 2, \ldots, 100\}$ and one terminating state 101. The state labels correspond to the room numbers; for instance, state 55 corresponds to Room 55.

   For a given policy π, we want to predict $v_\pi(s)$ for $s \in \mathcal{S}$ using value function approximation $\hat{v}(s, \boldsymbol{w}) = \sum_{i=1}^{2} w_i x_i(s)$, where $\boldsymbol{w} \in \mathbb{R}^2$. Note that $v_\pi(101) = 0$ since it is a terminating state. The features are given by

   $$x_1(s) = 0.5$$
   $$x_2(s) = s/100$$

   The policy π is used to generate the following episode (given in the format of $S_t, R_{t+1} \ldots$):

   Room 1, +1, Room 35, -1, Room 25, +5, Room 85, +10, Room 101(terminate)

   The discount rate is $\gamma = 1$. The step size is $\alpha = 0.9$. The weight vector is initialized as $\boldsymbol{w} = \boldsymbol{0}$. Using the above episode data and gradient Monte-Carlo with function approximation determine $\boldsymbol{w}_{final}$ after the updates.

   In the quiz, you will be asked for $\boldsymbol{w}_{final}$ and the associated quantities, such as $\hat{v}(s, \boldsymbol{w}_{final})$, $\nabla \hat{v}(s, \boldsymbol{w}_{final})$ or $\boldsymbol{x}(s)$ for a given s.

   Note: In Studium, $\hat{v}(\cdot)$ is denoted by $v_{app}(\cdot)$ where the subscript "app" stands for approximation.

4. We again consider the corridor environment of 2 rooms (similar to the DynaQ problem). The environment is represented with 2 non-terminal states $\mathcal{S} = \{1, 2\}$ and one terminating state 3. The state labels correspond to the room numbers; for instance, state 2 corresponds to Room 2.The leftmost room is Room 1 and the rightmost room is Room 3. The actions are LEFT (ActL) and RIGHT (ActR) with their usual behaviour on the left edge.

   We want to find an approximation of $q_*(s, a)$ for $s \in \mathcal{S}, \forall a$ using the approximation $\hat{q}(s, a, \boldsymbol{w}) = \sum_{i=1}^{2} w_i x_i(s, a)$, where $\boldsymbol{w} \in \mathbb{R}^2$. The features are given by $x_i(s, ActL) = s^{i-1}$ and $x_i(s, ActR) = s^i$ for i=1, 2.

   The discount rate is $\gamma = 1$. The step size is $\alpha = 0.9$. The weight vector is initialized as $\boldsymbol{w} = \boldsymbol{0}$. Using Episodic Semi-gradient Sarsa, we obtain the following episode:

t	S_t	A_t	R_{t+1}
0	Room 1	ActR	-1
1	Room 2	ActR	+10
2	Room 3 (terminate)	-	-

   In the quiz, you will be asked for $\boldsymbol{w}_{final}$ after the end of the episode and the associated quantities, such as $\hat{q}(s, a, \boldsymbol{w}_{final})$, $\nabla \hat{q}(s, a, \boldsymbol{w}_{final})$ or $\boldsymbol{x}(s, a)$ for a given s and a.

   Note: In Studium, $\hat{q}(\cdot)$ is denoted by $q_{app}(\cdot)$ where the subscript "app" stands for approximation.