



EL2805 Reinforcement Learning

Homework 2

December 2, 2023

Division of Decision and Control Systems
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Instructions (read carefully):

- Answer the questions of Parts 1 and 2.
- Work in groups of 2 persons.
- **Both** students in the group should upload their scanned report as a .pdf-file to Canvas before December 19, 23:59. The deadline is strict. Please mark your answers directly on this document, and **append** hand-written or typed notes justifying your answers. Reports without justification will not be graded.

Good luck!

1 Part 1. Q-learning and SARSA

Consider a discounted MDP with $\mathcal{S} = \{A, B, C\}$ and $\mathcal{A} = \{a, b, c\}$. We plan to use either the Q-learning or the SARSA algorithm in order to learn to control the system. We initialize the estimated Q-function as all zeros – that is:

$$Q^{(0)} = \begin{matrix} & a & b & c \\ \begin{matrix} A \\ B \\ C \end{matrix} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{matrix} .$$

The observed trajectory is as follows (for these transitions, we are imposed a policy):

$$(? , ? , ?) ; (A , ? , ?) ; (B , a , 100) ; (A , b , 60) ; (B , c , 70) ; (C , b , 40) ; (A , a , 20) ; (C , c , \dots)$$

where each triplet represents the state, the selected action, and the corresponding reward. Some of the information has been corrupted (marked with question marks) in the above sequence.

- a) Before the information became corrupt, we ran the Q-learning algorithm and obtained that

$$Q^{(2)} = \begin{matrix} & a & b & c \\ \begin{matrix} A \\ B \\ C \end{matrix} & \begin{bmatrix} 11 & 0 & 0 \\ 0 & 0 & 60 \\ 0 & 0 & 0 \end{bmatrix} \end{matrix} .$$

The discount factor was $\lambda = 0.5$ and the learning rate was fixed to $\alpha = 0.1$. Can you infer what the corrupt information was (i.e., the first state, the first and second selected actions, and the first and second observed rewards)? **Answer:**

$$(\text{---}, \text{---}, \text{---}) ; (A, \text{---}, \text{---}) ; (B, a, 100) ; (A, b, 60) ; (B, c, 70) ; (C, b, 40) ; (A, a, 20) ; (C, c, \dots)$$

- b) Provide the updated Q-values, using the Q-learning algorithm, at the 7th iteration. Use the same values for λ and α as in a). **Answer:**

$$Q^{(7)} = \begin{matrix} & a & b & c \\ \begin{matrix} A \\ B \\ C \end{matrix} & \begin{bmatrix} \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \end{bmatrix} \end{matrix} .$$

- c) What is the greedy policy w.r.t. the estimated Q function at the 7th iteration? $\pi(A) = \text{---}, \pi(B) = \text{---}, \pi(C) = \text{---}$.
- d) Provide the updated Q-values at the 7th iteration using the SARSA algorithm (initialized with $Q^{(0)}$ as all zeros). Take the first two (state, action, reward)-triplets as those given in your answer to a). Let the discount factor be $\lambda = 0.5$ and the learning rate fixed to $\alpha = 0.1$. **Answer:**

$$Q^{(7)} = \begin{matrix} & a & b & c \\ \begin{matrix} A \\ B \\ C \end{matrix} & \begin{bmatrix} \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \end{bmatrix} \end{matrix} .$$

- e) What is the greedy policy at the 7th iteration? $\pi(A) = \text{---}, \pi(B) = \text{---}, \pi(C) = \text{---}$.
- f) (Tick the correct circle) Are the rewards deterministic? ☐ Yes - ☐ No

2 Part 2: policy gradient and function approximation

Policy gradients. We consider an episodic RL problem with finite state-space \mathcal{S} and action space $\mathcal{A} = \{1, \dots, n+1\}$. For all states s , let $f(s)$ be a real valued function in $[1, 2]$. We parameterize the policy using parameter vector $\theta = (\theta_1, \dots, \theta_n) \in [0, 1]^n$ according to the following recursion: For $i \in \{1, \dots, n\}$, initialize $i = 1$ and draw independent random variable Z_i uniformly from $[0, f(s)]$. If $Z_i \leq \theta_i$, choose action $a = i$, otherwise, set $i \leftarrow i+1$ and repeat. At the last step of the recursion, if $Z_n > \theta_n$, choose $a = n+1$.

- a) Compute in state s , the probability $\pi_\theta(s, i)$ of choosing action i . **Answer:**

$$\pi_\theta(s, 1) =$$

$$\pi_\theta(s, i) = \quad \text{for } i \in \{2, \dots, n\}$$

$$\pi_\theta(s, n+1) =$$

- b) What is the Monte-Carlo REINFORCE update of θ upon observing an episode $\tau = (s_1, a_1, r_1, \dots, s_T, a_T, r_T)$? Provide explicit formulas using the function f , θ and τ only.

$$\frac{\partial \ln \pi_\theta(s, i)}{\partial \theta_i} =$$

$$\frac{\partial \ln \pi_\theta(s, i)}{\partial \theta_k} = \quad \text{for } k < i$$

$$\frac{\partial \ln \pi_\theta(s, i)}{\partial \theta_k} = \quad \text{for } k > i$$

Off-policy control with function approximation. Consider a discounted RL problem, that we wish to solve using approximations of the (state, action) value function (i.e., parametrized by vector θ).

- c) We observe the transition (s_t, a_t, r_t, s_{t+1}) . State the Q update in the Q-learning algorithm with function approximation. Why is it a semi-gradient algorithm? **Answer:**

- d) In the previous updates, the "target" evolves in every step, which could affect the algorithm convergence. What do we mean by target? Can you propose a modification that addresses this problem? **Answer:**