CS 747 (Autumn 2025) Week 11 Test (Batch 1)

5.35 p.m. – 6.00 p.m., October 28, 2025, LA 001

Name: _____ Roll number: _____

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. Consider episodic MDP (S, A, T, R, γ), notation as usual, in which $S = \{s_1, s_2, s_\top\}$, $A = \{a_1, a_2\}$, and $\gamma = 1$. States s_1 and s_2 are non-terminal, whereas $s_\top$ is terminal.

An agent follows a policy π_θ with 2-dimensional parameter-vector $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$. The parameters influence decision making as follows:

$$\pi_\theta(s_1, a_1) = \frac{1}{1 + \exp(-\theta_1)}; \qquad \pi_\theta(s_1, a_2) = 1 - \pi_\theta(s_1, a_1);$$
$$\pi_\theta(s_2, a_1) = \frac{1}{1 + \exp(-\theta_2)}; \qquad \pi_\theta(s_2, a_2) = 1 - \pi_\theta(s_2, a_1).$$

The agent is initialised with parameter vector $\theta = (0, 0)$. It executes the policy π_θ and encounters the following state-action-reward trajectory.

$$s_1, a_1, 4, s_2, a_2, 1, s_\top.$$

The agent uses REINFORCE to update its policy parameters at the end of this episode, with learning rate $\alpha = \frac{1}{5}$. The objective function is the value of the starting state s_1. No baseline subtraction is performed. Work out the parameter vector θ' obtained after making the REINFORCE update. Show the sequence of steps used to arrive at your answer. [3 marks]

Answer 1. We first work out $\nabla_\theta \ln \pi_\theta(s_1, a_1)$ and $\nabla_\theta \ln \pi_\theta(s_2, a_2)$.

$$\frac{\partial}{\partial \theta_1} \ln \pi_\theta(s_1, a_1) = \frac{\exp(-\theta_1)}{1 + \exp(-\theta_1)}; \quad \frac{\partial}{\partial \theta_2} \ln \pi_\theta(s_1, a_1) = 0.$$

$$\frac{\partial}{\partial \theta_1} \ln \pi_\theta(s_2, a_2) = 0; \quad \frac{\partial}{\partial \theta_2} \ln \pi_\theta(s_2, a_2) = -\frac{\exp(-\theta_2)}{1 + \exp(-\theta_2)}.$$

We execute the REINFORCE update rule to obtain our answer for the update from $\theta = (0, 0)$ to θ':

$$\begin{aligned}
\theta' &= \theta + \alpha \left(\nabla_\theta \ln \pi_\theta(s_1, a_1) \times (4 + 1) + \nabla_\theta \ln \pi_\theta(s_2, a_2) \times 1 \right) \\
&= (0, 0) + \frac{1}{5} \left(\left(\frac{1}{2}, 0 \right) \times 5 + \left(0, -\frac{1}{2} \right) \times 1 \right) \\
&= \left(\frac{1}{2}, -\frac{1}{10} \right).
\end{aligned}$$

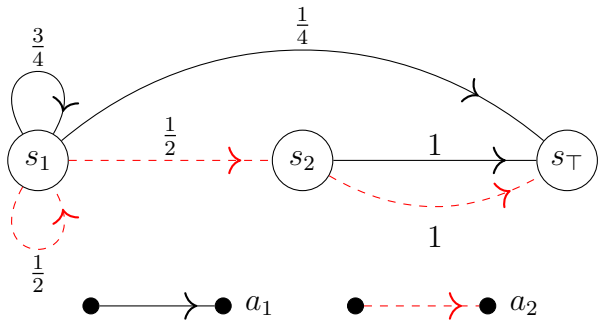
CS 747 (Autumn 2025) Week 11 Test (Batch 2)

6.15 p.m. – 6.40 p.m., October 28, 2025, LA 001

Name: _____ Roll number: _____

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. Consider episodic MDP (S, A, T, R, γ), with notation as usual, in which $S = \{s_1, s_2, s_\top\}$, $A = \{a_1, a_2\}$, and $\gamma = 1$. States s_1 and s_2 are non-terminal, whereas $s_\top$ is terminal. The figure on the right shows the transitions for all state-action pairs, with each transition annotated with its probability. Every reward is 1. Hence, the value of a state under a policy corresponds to the expected number of time steps to terminate if starting at that state and following that policy.



An agent follows a policy π_θ with a single parameter $\theta \in \mathbb{R}$. Action probabilities under π_θ are:

$$\pi_\theta(s_1, a_1) = \frac{\exp(\theta)}{1 + \exp(\theta)}; \quad \pi_\theta(s_1, a_2) = \frac{1}{1 + \exp(\theta)}; \qquad \pi_\theta(s_2, a_1) = \frac{1}{2}; \quad \pi_\theta(s_2, a_2) = \frac{1}{2}.$$

Define the objective $J(\theta) \stackrel{\text{def}}{=} V^{\pi_\theta}(s_1)$. Work out $J(\theta)$ as a function of θ, and then calculate its derivative with respect to θ—that is, $\frac{d}{d\theta}(J(\theta))$—at $\theta = 0$. Show the sequence of steps used to arrive at your answer. For your convenience, here are the partial derivatives of the action probabilities under π_θ.

$$\frac{d}{d\theta}(\pi_\theta(s_1, a_1)) = \pi_\theta(s_1, a_1)\pi_\theta(s_1, a_2);$$
$$\frac{d}{d\theta}(\pi_\theta(s_1, a_2)) = -\pi_\theta(s_1, a_1)\pi_\theta(s_1, a_2);$$
$$\frac{d}{d\theta}(\pi_\theta(s_2, a_1)) = 0;$$
$$\frac{d}{d\theta}(\pi_\theta(s_2, a_2)) = 0.$$

[3 marks]

Answer 1. We write down and solve the Bellman equations to obtain V^{π_θ}. First, it is easy to see that

$$V^{\pi_\theta}(s_2) = \pi_\theta(s_2, a_1) \times 1 + \pi_\theta(s_2, a_2) \times 1 = 1.$$

For s_1, we get

$$V^{\pi_\theta}(s_1) = 1 + \pi_\theta(s_1, a_1) \left(\frac{3}{4} V^{\pi_\theta}(s_1) + \frac{1}{4}(0) \right) + \pi_\theta(s_1, a_2) \left(\frac{1}{2} V^{\pi_\theta}(s_1) + \frac{1}{2}(1) \right),$$

which, upon solving, yields

$$V^{\pi_\theta}(s_1) = J(\theta) = \frac{6 - 2\pi_\theta(s_1, a_1)}{2 - \pi_\theta(s_1, a_1)}.$$

The derivative of this function is

$$\begin{aligned} \frac{d}{d\theta}\left(J(\theta)\right) &= \frac{-2}{2 - \pi_\theta(s_1, a_1)} \frac{d}{d\theta}\left(\pi_\theta(s_1, a_1)\right) + \frac{6 - 2\pi_\theta(s_1, a_1)}{(2 - \pi_\theta(s_1, a_1))^2} \frac{d}{d\theta}\left(\pi_\theta(s_1, a_1)\right) \\ &= \frac{2}{(2 - \pi_\theta(s_1, a_1))^2} \pi_\theta(s_1, a_1) \pi_\theta(s_1, a_2). \end{aligned}$$

For $\theta = 0$, we have $\pi_\theta(s_1, a_1) = \pi_\theta(s_1, a_2) = \frac{1}{2}$ and $\exp(\theta) = 1$, and hence we get

$$\frac{d}{d\theta}\left(J(\theta)\right)\Big|_{\theta=0} = \frac{2}{9}.$$