CS 747 (Autumn 2025) Week 8 Test (Batch 1)

5.35 p.m. – 6.00 p.m., October 7, 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. Suppose Q-learning is run on a continuing MDP, in which discount factor $\gamma < 1$, and each reward is strictly positive. ϵ-greedy action-selection is performed at each step, and updates made using a learning rate α. Both ϵ and α are constants in $(0, 1)$; there is no annealing with time. All entries in the Q-table are initialised to 0.

1a. Is it guaranteed that the entries in the Q-table will remain non-negative for the entire run (that is, after any number of updates have been performed)? Justify your answer. [2 marks]

1b. Suppose the agent uses Sarsa for performing learning updates instead of Q-learning, but otherwise all aspects of the setup (such as the initialisation, the exploration and learning rates) remain as described above. Is it guaranteed that the entries in the Q-table will remain non-negative for the entire run (that is, after any number of updates have been performed)? Justify your answer. [1 mark]

Answer 1. Let Q^t be the Q-table at time step $t \geq 0$. Suppose the transition (s^t, a^t, r^t, s^{t+1}) happens at this time step. Then the next Q-table, Q^{t+1}, satisfies

$$Q^{t+1}(s^t, a^t) = Q^t(s^t, a^t)(1 - \alpha) + \alpha(r + \gamma \max_{a'} Q^t(s^{t+1}, a')),$$
$$Q^{t+1}(s, a) = Q^t(s, a) \text{ for } (s, a) \neq (s^t, a^t).$$

As induction hypothesis, assume that all entries in Q^t are non-negative. The base case of Q^0 being initialised to 0 satisfies this requirement. From the relationship between Q^{t+1} and Q^t, it is apparent that if Q^t is non-negative, so is Q^{t+1} (notice that we have been given that the rewards are positive). Hence, we conclude that the claim in 1a is correct.

The same argument continues to apply when Sarsa is used instead of Q-learning. The only change is to the target in the update, which remains a sum of a reward and a discounted Q-value. So the answer to 1b is also yes.

CS 747 (Spring 2025) Week 8 Test (Batch 2)

6.15 p.m. – 6.40 p.m., October 7, 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 the single-state, single-action continuing MDP shown below. The reward is $r > 0$ and the discount factor is $\gamma \in (0, 1)$.

Suppose the TD(0) algorithm is run to estimate the value of the state under the only policy that exists. The initial estimate is V^0, and for $t \geq 1$, the value estimate after t transitions (and hence t updates) have been performed is V^t. Learning updates all use a constant learning rate $\alpha \in (0, 1)$, with no annealing.

1a. Provide an expression for V^t in terms of r, γ, and α. [2 marks]

1b. Does $\lim_{t \to \infty} V^t$ exist? If yes, what is it? If not, explain why it does not exist. [1 mark]

Answer 1. For $t \geq 1$, the TD(0) update yields

$$\begin{aligned} V^t &= V^{t-1}(1-\alpha) + \alpha(r + \gamma V^{t-1}) \\ &= xV^{t-1} + y \text{ where} \\ x &= 1 - \alpha + \alpha\gamma \text{ and} \\ y &= \alpha r. \end{aligned}$$

Expanding this recurrence, we notice the pattern

$$\begin{aligned} V^0 &= 0, \\ V^1 &= y, \\ V^2 &= xy + y, \\ V^3 &= x^2y + xy + y, \\ &\vdots \\ V^t &= x^{t-1}y + x^{t-2}y + \cdots + xy + y. \end{aligned}$$

Hence,

$$V^t = y\frac{1-x^t}{1-x}.$$

Clearly $\lim_{t\to\infty} V^t$ exists, and is equal to

$$\lim_{t\to\infty} V^t = \frac{y}{1-x} = \frac{r}{1-\gamma},$$

which is as expected since TD(0) should converge to the value of the given state, which is clearly

$$r + \gamma r + \gamma^2 r + \cdots = \frac{r}{1-\gamma}.$$