DDA4230 Reinforcement Learning

Mid-term Examination

Name: ______________________ Student ID: ______________________

Answer ALL questions in this paper.

Question	Points	Score
1	20	
2	30	
3	30	
4	20	
Total:	100	

This page is intentionally left blank.

1. (20 points) **True or False. If your answer is "False"**, please explain the reason.

(1) In a Markov Decision Process (MDP) with finite state and action spaces, suppose a policy π satisfies the following condition for all states s: $V^\pi(s) = \max_a Q^\pi(s, a)$, then π must be an optimal policy.
False. The condition $V^\pi(s) = \max_a Q^\pi(s, a)$ only ensures that π is greedy with respect to its own value function, but not necessarily that it achieves the optimal value function V^*.

(2) Starting from the same initial value function V_0, a single application of the Bellman Optimality operator T^* will always produce a value function T^*V_0 that is greater than or equal to the result of applying the Bellman operator of any policy $T^\pi V_0$, for all states s.
True.

(3) The trade-off between exploration and exploitation is a fundamental issue in reinforcement learning. In both the ϵ-greedy algorithm ($\epsilon > 0$) and the Upper Confidence Bound (UCB) algorithm, every action retains a non-zero probability of being selected in every iteration for exploration purposes.
False, because the UCB algorithm selects the action that maximizes its upper confidence bound value, giving other suboptimal actions zero probability at that iteration.

(4) Within a model-free policy iteration algorithm for an episodic MDP (assuming the MDP is discrete and stationary), let π denote an ϵ-greedy policy defined as:

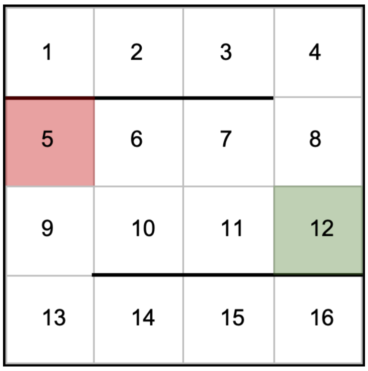
$$\pi(a \mid s) = \begin{cases} \text{Uniform}(\mathcal{A}) & \text{with probability } \epsilon_k \\ \arg\max_a Q^\pi(s, a) & \text{with probability } 1 - \epsilon_k \, . \end{cases}$$

If $\epsilon_k \in (0, 1)$ and the number of running episodes $k \to \infty$, then the policy π converges to the optimal policy, satisfying $\forall s, a, Q^\pi(s, a) = Q^*(s, a)$.
False. It requires $\epsilon_k \to 0$ as $k \to \infty$.

(5) In Q-Learning, the update rule uses the maximum estimated future reward (off-policy), whereas in SARSA, the update rule uses the actual action taken by the agent (on-policy).
True.

2. (30 points) **Gridworld & MDP.**

Consider the following grid environment. Starting from any unshaded square, you can move up, down, left, or right. Actions are deterministic and always succeed (e.g. going left from state 16 goes to state 15) unless they will cause the agent to run into a wall. The thicker edges indicate walls, and attempting to move in the direction of a wall results in staying in the same square (e.g. going in any direction other than left from state 16 stays in 16). Taking any action from the green target square (no. 12) earns a reward of r_g (so $r(12, a) = r_g \ \forall a$) and ends the episode . Taking any action from the red square of death (no. 5) earns a reward of r_r (so $r(5, a) = r_r \ \forall a$) and ends the episode. Otherwise, from every other square, taking any action is associated with a reward $r_s \in \{-1, 0, +1\}$ (even if the action results in the agent staying in the same square). Assume the discount factor $\gamma = 1$, $r_g = +5$, and $r_r = -5$ unless otherwise specified.



(1) ⓐ Define the value of r_s that would cause the optimal policy to return the shortest path to the green target square (no. 12); ⓑ Using this r_s, find the optimal value for each square.

$$r_s = -1, \quad \begin{bmatrix} 0 & 1 & 2 & 3 \\ -5 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 1 & 0 & -1 & -2 \end{bmatrix}$$

(2) Let's refer to the value function derived in (1) as $V_{old}^{\pi_g}$ and the policy as π_g. Suppose we are now in a new gridworld where all the rewards (r_s, r_g, and r_r) have +2 added to them. Consider still following π_g of the original gridworld, what will the new values $V_{new}^{\pi_g}$ be in this second gridworld?

$$\begin{bmatrix} 12 & 11 & 10 & 9 \\ -3 & 10 & 9 & 8 \\ 10 & 9 & 8 & 7 \\ 11 & 12 & 13 & 14 \end{bmatrix}$$

(3) Consider a general MDP with rewards and transitions. Consider a discount factor of γ. For this case, assume that the horizon is infinite (so there is no termination). A policy π in this MDP induces a value function V^π (let's refer to this as V^π_{old}). Now, suppose we have a new MDP where the only difference is that all rewards have a constant c added to them. Can you come up with an expression for the new value function V^π_{new} induced by π in this second MDP in terms of V^π_{old}, c, and γ?

$V^\pi_{\text{old}}(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s]\,,$
$V^\pi_{\text{new}}(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t (r_t + c) \mid s_0 = s] = V^\pi_{\text{old}}(s) + \frac{c}{1-\gamma}.$

(4) Let's go back to our gridworld from (1) with the default values for r_g, r_r, γ, and with the value you specified for r_s. Suppose we now derive a second gridworld by adding a constant c to all rewards (r_s, r_g, and r_r) such that $r_s = +2$. ⓐ How does the optimal policy change (Just give a one or two-sentence description)? ⓑ What do the values of the unshaded squares become?

It will become looping at states other than the terminal states, because each step will gain a positive reward value.
Unshaded squares will become value $+\infty$.

(5) Now take the second gridworld from part (4) and change γ such that $0 < \gamma < 1$. Can the optimal policy change, and does it depend on your choice of gamma? (A brief description is sufficient; no formal proof or mathematical analysis required).

If γ is close to 1, the result is the same, all unshaded squares will have value $+\infty$, and the optimal policy is to keep looping.
If γ is equal to some value close to 0, at some point optimal policy will be able to find the shortest path to the green square.

3. (30 points) **Multi-Armed Bandit (MAB).**

Consider the stochastic bandit problem with 3 arms, where the (random) reward associated with the 3 arms for the first 6 rounds are shown in Table 1. Note that these numbers are *unknown* to the bandit algorithm. A bandit algorithm A has respectively selected Arms $1, 2, 3$, and 1 in the first 4 rounds (for t in $\{1, 2, 3, 4\}$).

Table 1: Arm rewards over time.

Time (t)	1	2	3	4	5
Arm 1	0.3	0.2	0.5	0.3	0.2
Arm 2	0.2	0.3	0.5	0.8	0.5
Arm 3	0.1	0.05	0.02	0.1	0.03

(1) Suppose A applies the ϵ-greedy algorithm with $\epsilon = 0.3$ at round $t = 5$. ⓐ Compute the chance of each arm being selected. ⓑ If the algorithm chooses Arm 3 at round $t = 5$, is this action an instance of exploitation or exploration?

We first calculate the average reward for each arm up to round 4.
Arm 1 is selected at $t = 1, 4$, so the average reward for Arm 1 is $(0.3 + 0.3)/2 = 0.3$.
Arm 2 is selected at $t = 2$, so the average reward for Arm 2 is 0.3.
Arm 3 is selected at $t = 3$, so the average reward for Arm 3 is 0.02.
We know that the ϵ-greedy algorithm will select the best arm with probability $1 - \epsilon$, and a random arm with probability ϵ. According to the above average reward, the best arm for $t = 5$ is Arm 1 and Arm 2. So we can calculate the probability of selecting each arm as follows:
$P_{best}(Arm1) = P_{best}(Arm2) = (1 - \epsilon)/2 = 0.35$
$P_{random}(Arm1) = P_{random}(Arm2) = P_{random}(Arm3) = 0.3/3 = 0.1$
Finally we get $P(Arm1) = P(Arm2) = 0.35 + 0.1 = 0.45$, $P(Arm3) = 0.1$.

Choosing Arm 3 at round $t = 5$ is (definitely) exploration.

(2) Suppose A intends to apply the UCB algorithm (with confidence level $\delta = 0.5$) in the following rounds after $t = 4$. We want to trace the algorithm for these rounds. ⓐ Please show how the algorithm works at round $t = 5$. ⓑ Based on this, compute the relative advantage (regarding the upper confidence bound) of choosing Arm 2 over Arm 3 in round $t = 5$. ⓒ Is it larger or smaller than the empirical advantage of choosing Arm 2 *over* Arm 3 in round $t = 5$?
Hint: The UCB algorithm follows

$$UCB_i(t-1, \delta) = \begin{cases} \infty \,, & N_{i,t-1} = 0 \,, \\ \dfrac{1}{N_{i,t-1}} \displaystyle\sum_{t' \le t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} + \sqrt{\dfrac{2 \log_2(1/\delta)}{N_{i,t-1}}} \,, & N_{i,t-1} > 0 \,; \end{cases}$$

where $\frac{1}{N_{i,t-1}} \sum_{t' \le t-1} r_{t'} \mathbb{1}\{a_{t'} = i\}$ *is the average reward of arm* i *up to time* $t-1$*, and* $N_{i,t-1}$ *is the number of times arm* i *has been selected up to time* $t-1$*.*

At $t = 5$, the UCB for each arm is:
$\text{UCB}_1(t=5) = 0.3 + \sqrt{\frac{2}{2}} = 0.3 + 1 = 1.3$
$\text{UCB}_2(t=5) = 0.3 + \sqrt{\frac{2}{1}} = 0.3 + 1.414 = 1.714$
$\text{UCB}_3(t=5) = 0.02 + \sqrt{\frac{2}{1}} = 0.02 + 1.414 = 1.434$
So Arm 2 will be selected with a payoff of 0.5.

The estimated advantage of choosing Arm 2 over Arm 3 at round $t = 5$ is $1.714 - 1.434 = 0.28$. The empirical advantage of choosing Arm 2 over Arm 3 at round $t = 5$ is $0.5 - 0.03 = 0.47 > 0.28$.

Therefore, the estimated advantage is smaller than the empirical advantage of choosing Arm 2 over Arm 3 at round $t = 5$.

4. (20 points) **Finite-Horizon MDP**

Consider a finite-horizon MDP with horizon $H = 3$ and discount factor $\gamma = 0.9$. In the initial state s_0, two actions are available: a_1 and a_2. The immediate rewards are given by:

$$r(s_0, a_1) = 2, \quad r(s_0, a_2) = 0.$$

At time step t, the Q-value under policy π is defined as

$$Q_t^\pi(s', a') = \mathbb{E}\left[\sum_{l=0}^{H-t-1} \gamma^l r(s_l', a_l')\right]$$

At timestep $t = 1$, if a_1 is taken, the next-state value is $V_{t+1}(s') = 0$; if a_2 is taken, the next-state value is $V_{t+1}(s') = 10$. The episode ends after step $H = 3$.

(1) Compute $Q_1^\pi(s_0, a_1)$ and $Q_1^\pi(s_0, a_2)$.

At $t = 1$: $Q_1^\pi(s_0, a_1) = 2 + 0.9 \times 0 = 2, \qquad Q_1^\pi(s_0, a_2) = 0 + 0.9 \times 10 = 9.$

(2) Compute $Q_2^\pi(s_0, a_1)$ and $Q_2^\pi(s_0, a_2)$.

At $t = 2$: $Q_2^\pi(s_0, a_1) = 2, \qquad Q_2^\pi(s_0, a_2) = 0.$

(3) Determine which action is optimal at $t = 1$ and at $t = 2$.

Optimal actions: At $t = 1 : a_2$ is optimal since $9 > 2$; At $t = 2 : a_1$ is optimal since $2 > 0$.

(4) Explain why this implies that the optimal policy π^* must be non-stationary.

The optimal action changes with time — a_2 is preferred early ($t = 1$) in the episode when there is still time to receive future rewards, while a_1 is preferred near the end when only the immediate reward matters. Hence, the optimal policy π^* depends explicitly on time t, and is therefore **non-stationary** in a finite-horizon MDP.