DDA4230 Reinforcement Learning

Mid-term Examination

Name: ______________________ Student ID: ______________________

Write ALL questions directly on the Question Book.

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

This page is intentionally left blank.

I Regular Questions

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

(1) If the Markov Decision Process (MDP) is stationary, the planning horizon is infinite, and the discount factor $\gamma \in (0, 1)$, a deterministic optimal policy may *not* exist under this MDP.
False. There must exist one that selects a specific action in each state without randomness.

(2) Let $\mathcal{M}_c$ define a constrained MDP where the policy must satisfy a constraint such that $\mathbb{E}_\pi(\sum_{t=0}^{\infty} \gamma^t c_t) \leq \epsilon$ (where c_t denote cost at a time step t) during learning. If $\mathcal{M}_c$ is stationary, the planning horizon is infinite, and the discount factor $\gamma \in (0, 1)$, a deterministic optimal policy may *not* exist under this MDP.
True.

(3) Let Q^π and V^π represent the action-value function and state-value in a stationary MDP. Let π^* define an optimal policy. The expected advantages function $\mathbb{E}_{\pi^*}[A^{\pi^*}(s, a)] = \int_a \pi^*(a|s)[Q^{\pi^*}(s, a) - V^{\pi^*}(s)]\mathrm{d}a$ must be equivalent to 0 (e.g., $\mathbb{E}_{\pi^*}[A^\pi(s, a)] = 0$) for all state s and action a.
True.

(4) Let Q^π and V^π represent the action-value function and state-value in a stationary MDP. Let π and π^* define an arbitrary random and an optimal policy. The expected advantages function $\mathbb{E}_{\pi^*}[A^\pi(s, a)] = \int_a \pi^*(a|s)[Q^\pi(s, a) - V^\pi(s)]\mathrm{d}a$ must be larger or equivalent to 0 (e.g., $\mathbb{E}_{\pi^*}[A^\pi(s, a)] \geq 0$) for all state s and action a.
False. π is not an optimal policy.

(5) In the task of policy evaluation, the Temporal Difference (TD) method tends to exhibit higher variance yet lower bias in the estimation of value functions V^π when compared to the Monte Carlo (MC) method.
False. MC exhibits lower bias and higher variance.

2. (20 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 7 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	6	7
Arm 1	0.3	0.2	0.5	0.3	0.2	0.4	0.6
Arm 2	0.2	0.3	0.5	0.8	0.5	0.3	0.7
Arm 3	0.1	0.05	0.02	0.1	0.03	0.02	0.01

(1) Suppose A applies the ϵ-greedy algorithm with $\epsilon = 0.2$ at round $t = 5$. Compute the chance of each arm being selected.

(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 rounds t in $\{5, 6, 7\}$.

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* $t-1$*.*

Solution

(1)

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:

$P_{best}(Arm1) = P_{best}(Arm2) = (1 - \epsilon)/2 = 0.4$

$P_{random}(Arm1) = P_{random}(Arm2) = P_{random}(Arm3) = 0.2/3 = 0.0667$

Finally we get $P(Arm1) = P(Arm2) = 0.4 + 0.0667 = 0.4667 = 7/15$, $P(Arm3) = 0.0667 = 1/15$.

(2)

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.

At $t = 6$, the UCB for each arm is:

$\text{UCB}_1(t = 6) = 0.3 + \sqrt{\frac{2}{2}} = 0.3 + 1 = 1.3$

$\text{UCB}_2(t = 6) = 0.4 + \sqrt{\frac{2}{2}} = 0.4 + 1 = 1.4$

$\text{UCB}_3(t = 6) = 0.02 + \sqrt{\frac{2}{1}} = 0.02 + 1.414 = 1.434$

So Arm 3 will be selected with a payoff of 0.02.

At $t = 7$, the UCB for each arm is:

$\text{UCB}_1(t = 7) = 0.3 + \sqrt{\frac{2}{2}} = 0.3 + 1 = 1.3$

$\text{UCB}_2(t = 7) = 0.4 + \sqrt{\frac{2}{2}} = 0.4 + 1 = 1.4$

$\text{UCB}_3(t = 7) = 0.02 + \sqrt{\frac{2}{2}} = 0.02 + 1 = 1.02$

So Arm 2 will be selected with a payoff of 0.7.

3. (30 points) **Trajectories, returns, and values.**

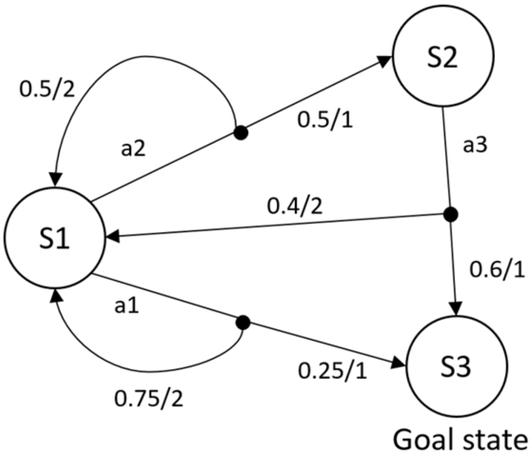

Figure 1: MDP for Question 3.

Consider the MDP above, in which there are three states, $S1, S2$ and $S3$, three actions, $a1, a2$ and $a3$. The transition probabilities and rewards are shown in the line. For example, taking $a1$ at state $S1$ will either transition to $S3$ with probability $p = 0.25$ and reward $r = 1$, or transition to $S1$ with probability $p = 0.75$ and reward $r = 2$. Assume that $V(S3) = 0$, consider two deterministic policies, π_1 and π_2:

$$\pi_1(S1) = a1, \quad \pi_1(S2) = a3$$

$$\pi_2(S1) = a2, \quad \pi_2(S2) = a3$$

(1) Show a trajectory (sequence of states, actions and rewards) from $S1$ for policy π_1.
(2) Show a trajectory (sequence of states, actions and rewards) from $S1$ for policy π_2:
(3) Assuming the discount-rate parameter is $\gamma = 0.5$, what is the return from the initial state for the trajectory in (1) and (2)?
(4) Assuming $\gamma = 0.5$, what is the value of state $S1$ under policy π_1 and policy π_2?
(5) Show the equation representing the optimal value function for state $S1, S2$. *Hint: using representation like:* $V^*(S1) = a + b * V^*(S2)$ *where* a *and* b *are real numbers.*

Solution.

(1)-(3) omitted.

(4)

$V_{\pi_1}(S1) = 0.25 \times (1 + 0.5 \times V_{\pi_1}(S3)) + 0.75 \times (2 + 0.5 \times V_{\pi_1}(S1))$
$V_{\pi_1}(S1) = 2.8$

$V_{\pi_2}(S2) = 0.6 \times (1 + 0.5 \times V_{\pi_2}(S3)) + 0.4 \times (2 + 0.5 \times V_{\pi_2}(S1))$
$V_{\pi_2}(S1) = 0.5 \times (2 + 0.5 \times V_{\pi_2}(S1)) + 0.5 \times (1 + 0.5 \times V_{\pi_2}(S2))$
$V_{\pi_2}(S1) = 2.64$

(5)

$V^*(S1) = \max\Big(1.75 + 0.375 \times V^*(S1), 1.5 + 0.25 \times V^*(S1) + 0.25 \times V^*(S2)\Big)$
$V^*(S2) = 1.4 + 0.2 \times V^*(S1)$

4. (30 points) **Bellman Operator and Its Variant.**

Let the Bellman operator $B : \mathbb{R}^S \longrightarrow \mathbb{R}^S$ defined as:

$$(BV)(s) = \mathbb{E}_{\pi}[\ r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')]$$

Let the state-maximum Bellman operator $B^{max} : \mathbb{R}^S \longrightarrow \mathbb{R}^S$ defined as:

$$(BV)(s) = \mathbb{E}_{\pi}[\ r(s,a) + \gamma \max_{s'} V(s')]$$

(1) Prove that the Bellman operator B is a contraction operator for $\gamma \in (0,1)$ with respect to the infinity norm $\|\cdot\|_\infty$. In other words, please show $\|BV - BV'\|_\infty \leq \gamma \|V - V'\|_\infty$ for any two value functions V and V'. The infinity norm of a value function V can be defined as that $\|V\|_\infty = \max_s \|V(s)\|$.

(2) Please explain whether that the state-maximum Bellman operator B^{max} is a contraction operator for $\gamma \in (0,1)$ with respect to the infinity norm $\|\cdot\|_\infty$. *If yes*, please show $\|B^{max}V - B^{max}V'\|_\infty \leq \gamma \|V - V'\|_\infty$ for any two value functions V and V'. *If no*, please explain why.

Solutions

(1)

For any state $s \in \mathcal{S}$,

$$
\begin{aligned}
&|(BV)(s) - (BV')(s)| \\
&= \left|\mathbb{E}_\pi\left[r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')\right] - \mathbb{E}_\pi\left[r(s,a) + \gamma \sum_{s'} P(s'|s,a)V'(s')\right]\right|. \\
&= \left|\gamma \mathbb{E}_\pi\left[\sum_{s'} P(s'|s,a)\left(V(s') - V'(s')\right)\right]\right|. \\
&= \gamma \left|\mathbb{E}_\pi\left[\sum_{s'} P(s'|s,a)\left(V(s') - V'(s')\right)\right]\right| \\
&= \gamma \left|\sum_a \pi(a|s) \sum_{s'} P(s'|s,a)\left(V(s') - V'(s')\right)\right| \\
&\leq \gamma \sum_a \pi(a|s) \left|\sum_{s'} P(s'|s,a)\left(V(s') - V'(s')\right)\right| \text{(Triangle Inequality)} \\
&\leq \gamma \sum_a \pi(a|s) \sum_{s'} P(s'|s,a)\left|V(s') - V'(s')\right| \text{(Triangle Inequality)} \\
&\leq \gamma \sum_a \pi(a|s) \sum_{s'} P(s'|s,a) \|V - V'\|_\infty \text{(Infinity Norm)} \\
&= \gamma \sum_a \pi(a|s) \|V - V'\|_\infty \text{(Probabilities Sum to One)} \\
&= \gamma \|V - V'\|_\infty \text{(Probabilities Sum to One)}
\end{aligned}
$$

Since the above holds for any state s, it also holds for the state maximizing the LHS, such that:

$$\max_s |BV(s) - BV'(s)| \leq \gamma \|V - V'\|_\infty \,,$$

which means

$$\|BV - BV'\|_\infty \leq \gamma \|V - V'\|_\infty \,.$$

(2)

First we show that for a function g, $|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|$. Assume without loss of generality that $\max_a f(a) \geq \max_a g(a)$, and denote $a^* = \arg\max_a f(a)$. Then,

$$
\begin{aligned}
\left|\max_a f(a) - \max_a g(a)\right| &= \max_a f(a) - \max_a g(a) = f(a^*) - \max_a g(a) \\
&\leq f(a^*) - g(a^*) \\
&\leq \max_a |f(a) - g(a)| \,.
\end{aligned}
$$

For any state $s \in \mathcal{S}$,

$$
\begin{aligned}
|(B^{\max}V)(s) - (B^{\max}V')(s)| \quad &= \left|\left(\mathbb{E}_\pi[r(s,a)] + \gamma \max_{s'} V(s')\right) - \left(\mathbb{E}_\pi[r(s,a)] + \gamma \max_{s'} V'(s')\right)\right| \\
&= \gamma \left|\max_{s'} V(s') - \max_{s'} V'(s')\right| \\
&\leq \gamma \max_{s'} |V(s') - V'(s'))| \\
&= \gamma \|V - V'\|_\infty
\end{aligned}
$$

As in (1), it means that

$$\|B^{\max}V - B^{\max}V'\| \leq \gamma\|V - V'\|_\infty .$$