DDA 4230: Reinforcement learning **CUHKSZ**

Assignment 1

TA: Bo Yue, Hengming Zhang

Due Date: Oct. 12th, 11:59 pm
Total points available: 100 pts.

Note: *Please note that external references are allowed only if you give appropriate reference. There is no required format of reference. Please elaborate on your answers as well (do not just give a number, etc).*

Problem 1: Markov Decision Process I [30 points]

Suppose we have an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ with a finite state-action space, $|S \times \mathcal{A}| < \infty$ and $0 \leq \gamma < 1$. For any two arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$,we denote the class of all functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ as $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$. In the questions that follow, let $Q, Q' \in \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ be any two arbitrary action-value functions and consider any fixed state $s \in \mathcal{S}$. Without loss of generality, you may assume that $Q(s,a) \geq Q'(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.

1. Prove that $|\max_{a \in \mathcal{A}} Q(s,a) - \max_{a' \in \mathcal{A}} Q'(s,a')| \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|$. (10 points)
2. Prove that $|\min_{a \in \mathcal{A}} Q(s,a) - \min_{a' \in \mathcal{A}} Q'(s,a')| \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|$. (10 points)
3. Prove that the Bellman operator B ($BV = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s'|s,a)} V(s')\right]$) is a contraction operator for $\gamma \in (0,1)$ with respect to the infinity norm $\|\cdot\|_\infty$ (you may want to search up the definition for this). Specifically, we want to prove that

$$\|BV - BV'\|_\infty \leq \gamma \|V - V'\|_\infty \tag{1}$$

for any two value functions V and V', meaning if we apply it to two different value functions, the distance between value functions (in the ∞ norm) shrinks after application of the operator to each element. (10 points)

Problem 2: A 3-State MDP [40 points]

Consider states s_A, s_B, s_C and actions a_L, a_R. $\gamma = 0.9$. Transitions/rewards:

- s_A: a_L: $(s_A, r = 0)$; a_R: $(s_B, r = 0)$.
- s_B: a_L: $(s_A, r = 0)$; a_R: $(s_C, r = 2)$.
- s_C: terminal/absorbing with 0 reward thereafter (any action leaves you in s_C with $r = 0$).

1. **Evaluate a random policy.** Let π choose a_L/a_R with probability 1/2 in s_A, s_B (action in s_C irrelevant). Compute $V^\pi(s_A), V^\pi(s_B), V^\pi(s_C)$. (10 points).
2. **Bellman optimality equations.** Write $V^*(s)$ for $s \in \{A, B, C\}$ and express $V^*(A), V^*(B)$. (10 points)
3. **Solve for the optimal values and policy.** Find $V^*(A), V^*(B), V^*(C)$ and an optimal deterministic policy. (10 points)
4. **Performance gap.** Compute $V^*(s) - V^\pi(s)$ for each state and briefly explain the difference in words. (10 points)

Problem 3: Bandits Problem [20 points]

Consider a bandit instance with 4 arms and running ϵ-greedy algorithm on this instance. The algorithms maintain an initial estimate of the reward mean of arm i, μ_i, and we denote the estimate at time t as $\hat{\mu}_i(t)$. The algorithm initialize $\hat{\mu}_i(1) = 0$, for all i. Then the algorithm observes the following sequence of actions and rewards, where A_t, R_t denote the action and reward at time t:

- $t = 1$, $A_1 = 1, R_1 = 0$.
- $t = 2$, $A_2 = 2, R_2 = 0$.
- $t = 3$, $A_3 = 3, R_3 = 0$.
- $t = 4$, $A_4 = 4, R_4 = 0$.
- $t = 5$, $A_5 = 1, R_5 = 1$.
- $t = 6$, $A_6 = 2, R_6 = 1$.
- $t = 7$, $A_7 = 2, R_7 = 2$.
- $t = 8$, $A_8 = 2, R_8 = 2$.
- $t = 9$, $A_9 = 3, R_9 = 0$.

Which of the actions was definitely exploratory? (Recall ϵ-greedy explore with probability ϵ). (10 points) Which of the actions was possibly exploratory? (10 points)

Problem 4: Explore and commit [10 points]

Recall the ETC algorithm given in Lecture, and we consider the case where it interacts with a two-armed bandits. We assume that both arms are 1-subgaussian with mean μ_1, μ_2 and $\Delta = |\mu_1 - \mu_2|$. We follow the notations specified in the lecture notes here.

1. Find a new choice of k that is dependent on time horizon T but independent of Δ. Then show that for your choice of k, there exists a constant $C > 0$ and the regret is bounded by

$$\bar{R}_T \leq (\Delta + C)T^{2/3}\,.$$

Hint: Start with the result from Theorem 1 with the two arms case in the lecture note and try to get a bound independent of Δ.