DDA 4230: Reinforcement learning **CUHKSZ**

Assignment 1

TA: Sheng Xu

Due Date: Oct. 14th, 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)

Consider the an infinite horizon MDP with three states, s_1, s_2, s_3 and two actions a_0, a_1. The transitions are deterministic. The rewards and transitions are summarized as:

- s_1: $P(s_1 \mid s_1, a_0) = 1$, $r(s_1, a_0) = 0$, $P(s_2 \mid s_1, a_1) = 1$, $r(s_1, a_1) = 0$.
- s_2: $P(s_1 \mid s_2, a_0) = 1$, $r(s_2, a_0) = 0$, $P(s_3 \mid s_2, a_1) = 1$, $r(s_2, a_1) = 0$.
- s_3: $P(s_3 \mid s_3, a_0) = 1$, $r(s_3, a_0) = 1$, $P(s_3 \mid s_3, a_1) = 1$, $r(s_3, a_1) = 1$.

Note that taking any action on s_3 will let the agent stay in s_3 and receive a reward of 1. Let $\gamma = 0.5$ for this problem.

1. Consider an agent that follows the random policy (take the two actions with equal probability) in state s_1, s_2. Find the value function for the random policy for all states.
2. Let $V^*(s_1)$ and $V^*(s_2)$ denote the optimal value of state s_1 and s_2, respectively. Write down the Bellman optimality equation for $V^*(s_1)$ and $V^*(s_2)$.
3. Show that $V^*(s_1) < V^*(s_2) < 2$.
4. Based on the previous two parts, find the optimal value function for s_1, s_2 and the optimal policy.

Problem 2: Markov Decision Process II (30 points)

In this question, we consider an infinite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, R, P)$ with finite state and actions.

1. Let $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \gamma, \hat{R}, P)$, where the states and actions spaces are the same with $\mathcal{M}$, but $|\hat{R}(s,a) - R(s,a)| \leq \epsilon, \forall s \in \mathcal{S}, \mathcal{A}$. Let $\hat{V}^*$ denotes the optimal value function on $\hat{\mathcal{M}}$. Show that $V^* - \hat{V}^* \leq \frac{\epsilon}{1-\gamma}$.
2. For the $\hat{\mathcal{M}}$ defined in question 1, will $\mathcal{M}$ and $\hat{\mathcal{M}}$ have the same optimal policy? Please elaborate on your answer.
3. Consider an alternative definition of $\hat{R}$, where $\hat{R}(s,a) - R(s,a) = \epsilon, \forall s \in \mathcal{S}, \mathcal{A}$. Let $\hat{V}^*$ denotes the optimal value function on $\hat{\mathcal{M}}$. Express $\hat{V}^*$ in terms of V^*.
4. For the $\hat{\mathcal{M}}$ defined in question 3, will $\mathcal{M}$ and $\hat{\mathcal{M}}$ have the same optimal policy? Please elaborate on your answer.

Problem 3: Bandits Problem I (20 points)

Let $T \in \mathbb{N}^+$ and $0.5 < p \leq 1$ be constants and assume the following T-round casino game. In the beginning, the gambler has a capital of 1. At each round, the gambler can stack any portion of the capital (when the capital is 0 the bet can only be 0). The gambler wins each round with probability p, independently of other rounds. If the gambler wins a round, the bet doubles up. If the gambler loses a round then the bet is lost. The goal is to maximize the expected capital after all T rounds of play.

1. Formulate this game into a discrete-time stationary or non-stationary Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho_0, \gamma)$ with a finite horizon T.
2. Find an optimal policy.

Problem 4: Bandits Problem II (20 points)

Consider a multi-arm bandit problem with $k = 5$ actions, denoted $1, 2, 3, 4$, and 5. Consider applying to this problem a bandit algorithm using ε-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$ for all a.

1. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = 2, A_2 = 2, R_2 = 2, A_3 = 3, R_3 = 1, A_4 = 2, R_4 = 3, A_5 = 3, R_5 = 0, A_6 = 4, R_6 = 5$. On some of these time steps the ε case may have occurred causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

   (Hint: Write down the sequence of values for each action at each time step. If the selected action's value is not the maximum one, it must be selected at random caused by ϵ. In addition, remember that the ϵ may impact at each time step regardless of the action values)