DDA4230 Reinforcement Learning

Mid-term Examination

Name: ______________________ Student ID: ______________________

Answer ALL questions in the Answer 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) Q-learning operates **strictly** as (can only be) an off-policy algorithm while Monte-Carlo Control operates **strictly** as (can only be) an on-policy algorithm.
   False, Monte-Carlo Control can be off-policy.

   (2) Let Q^π represent the action-value function. The optimal policy π^* in the MDP can be represented as $\arg\max_a Q^\pi(s,a)$, $\forall \pi \in \Pi$ (for any policy).
   False, since $\pi^*(a|s) = \arg\max Q^*(s,a) = \arg\max Q^{\pi^*}(s,a)$, not for any policy π.

   (3) Let Q^π and V^π represent the action-value function and state-value in the stationary Markov Decision Process (MDP). Let π^* define the optimal policy. The advantages function $A^{\pi^*}(s,a) = Q^{\pi^*}(s,a) - V^{\pi^*}(s) \leq 0$ for all state s and action a.
   True

   (4) Let Q^π and V^π represent the action-value function and state-value in the stationary Markov Decision Process (MDP). Then we have both $Q^\pi \leq \frac{r_{\max}}{1-\gamma}$ and $V^\pi \leq \frac{r_{\max}}{1-\gamma}$ where $r_{\max} = \max_{s,a} r(s,a)$ and γ is the discounted factor.
   True

   (5) In the bandit problems, we choose the action i with the largest UCB values, $\text{UCB}_i(t-1,\delta) = \frac{1}{N_{i,t-1}} \sum_{t' \leq t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} + \sqrt{\frac{2\log(1/\delta)}{N_{i,t-1}}}$ for $N_{i,t-1} > 0$. When the confidence level δ becomes smaller, the exploratory level becomes larger, and the action i with less visitation number $N_{i,t-1}$ will be more likely to be visited.
   True

2. (20 points) **Multi-Armed Bandit (MAB).**

Consider a multi-armed bandit with four arms, 1, 2, 3, and 4, each of which returns a positive-valued reward (i.e., reward $r \geq 0$). Imagine there have been 7 prior arm pulls – 2 pulls for each of arms 1, 2, and 3, and 1 pull for arm 4.

Arms	Pulls
1	2
2	2
3	2
4	1

(1) Given the total reward for arms 1, 2, 3, and 4 are 1, 2, 3, 2 and consider a ϵ-greedy algorithm with $\epsilon = 0.4$. What's the probability of selecting each arm (1,2,3,4) in the next time step?

0.1, 0.1, 0.1, 0.7

(2) Given the values of the rewards received up through that point, the UCB heuristic (with $\delta = 0.5$) says to pull arm 4 as the 8th arm pull. After the N=8 arm pulls the relevant statistics are shown in the following table. What is the smallest and largest values of the reward that arm 4 could ever have returned for its first pull?

Arms	Pulls	Total rewards
1	2	1
2	2	2
3	2	3
4	2	4

Hint: the UCB heuristic is:

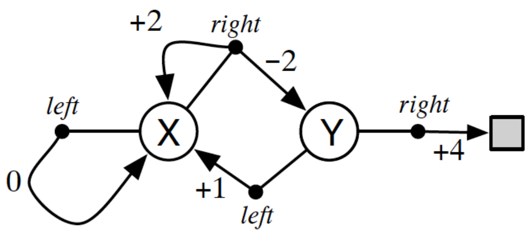
$$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} \infty\,, & N_{i,t-1} = 0\,, \\ \dfrac{1}{N_{i,t-1}} \displaystyle\sum_{t' \leq t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} + \sqrt{\dfrac{2\log_2(1/\delta)}{N_{i,t-1}}}\,, & N_{i,t-1} > 0\,; \end{cases}$$

2.5-$\sqrt{2}$, 4

3. (30 points) **Trajectories, returns, and values.** Consider the MDP below, in which there are two states, X and Y, two actions, right and left, and the deterministic rewards on each transition are as indicated by the numbers. Note that if action right is taken in state X, then the transition may be either to X with a reward of +2 or to Y with a reward of -2. These two possibilities occur with probabilities 2/3 (for the transition to X) and 1/3 (for the transition to state Y).

Consider two deterministic policies:

$$\pi_1(X) = \text{left}, \pi_1(Y) = \text{right}, \tag{1}$$
$$\pi_2(X) = \text{right}, \pi_2(Y) = \text{right}, \tag{2}$$



(1) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_1 (Maximum length is 5):
X, left, 0, X, left, 0, X, left, 0, X, left, 0, X, left, 0

(2) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_2:
X, right, +2, X, right, -2, Y, right, 4, End

(3) Assuming $\gamma = 0.5$, what is the value of state Y under policy π_1 (what is $V^{\pi_1}(Y)$):
4

(4) Assuming $\gamma = 0.5$, what is the action-value of X, left under policy π_1 (what is $Q^{\pi_1}(X, \text{left})$):
0

4. (30 points) **Soft Bellman-Equation.**

Let the soft Q-function be defined by:

$$Q^*_{\text{soft}}(s_t, a_t) = r_t + \mathbb{E}_{\pi^*, p_{\mathcal{T}}}\Big[\sum_{h=1}^{\infty} \gamma^h \Big(r_{t+h} + \mathcal{H}[\pi^*(a_{t+h}|s_{t+h})]\Big)\Big]$$

where $p_{\mathcal{T}}$ dnotes the transition probability and $\mathcal{H}(\pi(a|s)) = -\sum_a \pi(a|s)\log(\pi(a|s))$ denotes the causal entropy.

Let the soft Value function be defined by:

$$V^*_{\text{soft}}(s_t) = \log \sum_{a'} \exp\Big(Q^*_{\text{soft}}(s_t, a')\Big)$$

Let the policy be defined by:

$$\pi^*(a|s) = \exp\Big(Q^*_{\text{soft}}(s, a) - V^*_{\text{soft}}(s)\Big)$$

Show that the above soft Q-function satisfies the soft Bellman equation:

$$Q^*_{\text{soft}}(s_t, a_t) = r_t + \gamma \mathbb{E}_{p_{\mathcal{T}}}[V^*_{\text{soft}}(s_{t+1})]$$

Hint: drive a representation of V^*_{soft} from the first formula.

Proof.

$$\begin{aligned}
Q^*_{\text{soft}}(s_t, a_t) &= r(s_t, a_t) + \mathbb{E}_{\pi^*, p_{\mathcal{T}}}\Big[\sum_{h=1}^{\infty} \gamma^h \Big(r_{t+h} + \mathcal{H}[\pi^*(a_{t+h}|s_{t+h})]\Big)\Big] \\
&= r(s_t, a_t) + \gamma \mathbb{E}_{p_{\mathcal{T}}}\Big[\mathcal{H}[\pi^*(a_{t+1}|s_{t+1})] + \mathbb{E}_{\pi^*}[Q^*_{\text{soft}}(s_{t+1}, a_{t+1})]\Big] \qquad (3)
\end{aligned}$$

Since the entropy $\mathcal{H}(\pi(a|s)) = -\sum_a \pi(a|s)\log(\pi(a|s))$, we have:

$$\begin{aligned}
&\mathcal{H}(\pi(a|s)) + \mathbb{E}_{\pi^*}[Q^*_{\text{soft}}(s, a)] \\
&= \sum_a \pi(a|s)\Big[Q^*_{\text{soft}}(s, a) - \log(\pi(a|s))\Big] \\
&= \sum_a \pi(a|s)\Big(Q^*_{\text{soft}}(s, a) - Q^*_{\text{soft}}(s, a) + V^*_{\text{soft}}(s)\Big) \\
&= \Big(\sum_a \pi(a|s)\Big) V^*_{\text{soft}}(s) \\
&= 1 \cdot V^*_{\text{soft}}(s) \qquad (4)
\end{aligned}$$

By plugging equation 4 to equation 3, we have:

$$Q^*_{\text{soft}}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{p_{\mathcal{T}}}[V^*_{\text{soft}}(s_{t+1})]$$

□