Selected Topics in Reinforcement Learning (535519)

Final Exam
DATE: 2023/12/26, 18:30–21:30

INSTRUCTIONS

1. This examination paper includes **22 questions** in **6 pages**.
2. This **IS NOT an OPEN BOOK** exam.
3. This exam has a total of **110 points**.

Question 1. (4 points)

(a) Write the definition of V^π and Q^π. (2 points)

(b) Prove that $\mathbb{E}_{s,a\sim\pi}[A^\pi(s,a)] = 0$. (2 points)

Question 2. (6 points)
Let $Q^\pi(s,a)$ be written in $V^\pi(s)$ as follows.

$$Q^\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V^\pi(s')$$

Write the following three expressions in a similar way with the same notation:

(a) $V^\pi(s)$ written in $Q^\pi(s,a)$. (2 points)

(b) $V^\pi(s)$ written in $V^\pi(s')$. (2 points)

(c) $Q^\pi(s,a)$ written in $Q^\pi(s',a')$. (2 points)

Question 3. (10 points)
(Contraction mapping) Define a Bellman optimality operator T:

$$[TV](s) := \max_{a \in A}(R_s^a + \gamma \sum_{s'} P_{ss'}^a V(s'))$$

Prove that Bellman optimality backup operator T is a γ-contraction.
Hint: for any U and V, show that $||TU - TV||_\infty \leq \gamma ||U - V||_\infty$.

Question 4. (4 points)
Is Q-learning on-policy or off-policy? Explain the reason.

Question 5. (4 points)
Describe the UCB (Upper Confidence Bound) algorithm and its principles and objectives.

Question 6. (10 points)
Consider a small grid world:

(0,0)	(0,1)	(0,2)
(1,0) S	(1,1) Cliff	(1,2) G

The agent always starts at (1,0). Episodes end when the agent falls off the cliff (1,1) or reaches the goal (1,2). The maximal length of each episode is 10 steps. Reward is given in the following rules:

1. $r = -10$ when falling off the cliff.
2. $r = 1$ when reaching the goal.
3. Otherwise, $r = -1$ every step.

Assume that we train an agent using Q-learning with discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.1$. Initialize the Q-table with value 0, and update after every step.

Q-table:

action/state	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)
up	0	0	0	0	0	0
down	0	0	0	0	0	0
right	0	0	0	0	0	0
left	0	0	0	0	0	0

After two episodes:
1: (1,0) to (0,0): $r = -1$; (0,0) to (0,1): $r = -1$; (0,1) to (1,1): $r = -10$
2: (1,0) to (0,0): $r = -1$; (0,0) to (0,1): $r = -1$; (0,1) to (0,2): $r = -1$; (0,2) to (1,2): $r = 1$

(a) What's the value inside the Q-table after finishing episode 1? (4 points)

(b) What's the value inside the Q-table after finishing both episodes 1 and 2? (6 points)

Question 7. (10 points)
Design an PPO (Proximal Policy Optimization) algorithm by writing the pseudocode of following section 1 and section 2. Also explain your design.

```
for iteration=1,2... do
    // Section 1. Using policy π_θold to interact with
    //            the environment to collect data.
    // Section 2. Optimize θ.
    θ_old = θ
end for
```

Question 8. (3 points)
What's the benefit of the delayed actor update in TD3 (Twin Delayed DDPG)?

Question 9. (4 points)
Explain the design purpose of SAC (Soft Actor-Critic) policy objective.

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim D}[D_{KL}(\pi_\phi(.|s_t)||\frac{exp(Q_\theta(s_t,.))}{Z_\theta(s_t)})]$$

Question 10. (4 points)
Why can the RND (Random Network Distillation) algorithm alleviate the noisy TV problem?

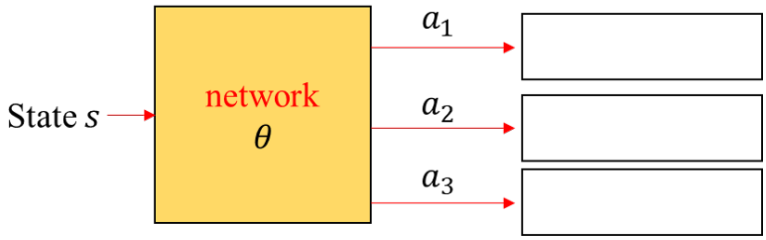
Question 11. (3 points)
What problem can Double-DQN prevent with respect to DQN?

Question 12. (4 points)
What techniques are used for exploration in DQN and DDPG respectively?

Question 13. (3 points)
What can be used to reduce variance for the REINFORCE algorithm?

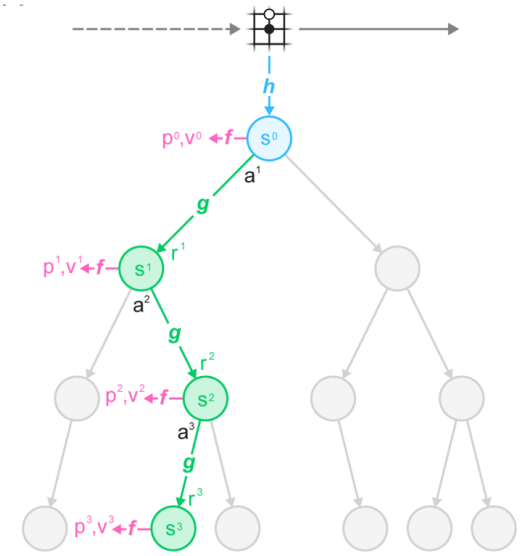
Question 14. (4 points)
What is the projection step in C51? Why does C51 need the projection step?

Question 15. (4 points)
What is the network output of QR-DQN? How are the Q-values calculated for each action?



Question 16. (4 points)
In AlphaZero, what are the training targets for the policy network and value network?

Question 17. (6 points)
In MuZero, explain the roles of the representation network (h), dynamics network (g), and prediction network (f).



Question 18. (3 points)
What is the benefit of MuZero's design compared with AlphaZero?

Question 19. (6 points)
How does DQfD (Deep Q-learning from Demonstrations) utilize demonstrations?

Question 20. (6 points)
What is the problem of non-stationarity in a multi-agent environment, and why does this non-stationarity occur? Use the example of rock-paper-scissors to illustrate your points.

Question 21. (4 points)
What are "Centralized Training" and "Decentralized Execution" in the CTDE approach, and what are the advantages of each?

Question 22. (4 points)
Below is the description of IGM. Why do some cooperative value-based MARL methods that use the CTDE framework often need the IGM property?

Individual-Global-Maximum (IGM)

- For a joint action-value function $Q_{joint}: \mathcal{T}^N \times \mathcal{A}^N \to \mathbb{R}$, where $\boldsymbol{\tau} \in \mathcal{T}$ is a joint action-observation histories, if there exist individual action-value function $\left[Q^i: \mathcal{T} \times \mathcal{A} \to \mathbb{R}\right]_{i=1}^{N}$, such that the following holds:

$$\arg\max_{\boldsymbol{a}} Q_{joint}(\boldsymbol{\tau}, \boldsymbol{a}) = \begin{pmatrix} \arg\max_{a_1} Q^1(\tau^1, a^1) \\ \arg\max_{a_2} Q^2(\tau^2, a^2) \\ \dots \\ \arg\max_{a_N} Q^N(\tau^N, a^N) \end{pmatrix}$$

END OF EXAM