RL 期末考

題目

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 $||TU - TV||_\infty \le \gamma ||U - V||_\infty$.

1

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
```

π

— entropy

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?

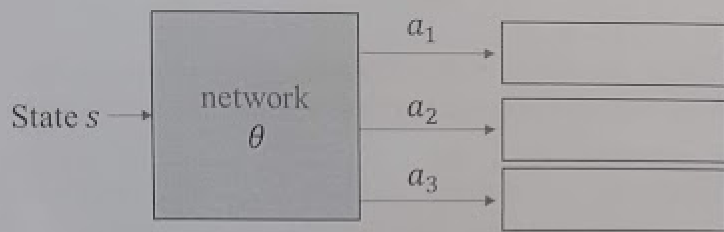
Question 11. (3 points)
What problem can Double-DQN prevent with respect to DQN?

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?

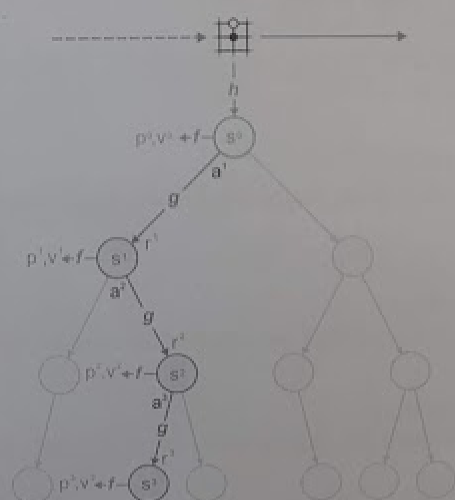
Question 14. (4 points)
What is the projection step in C51? Why does C51 need the projection step?

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?

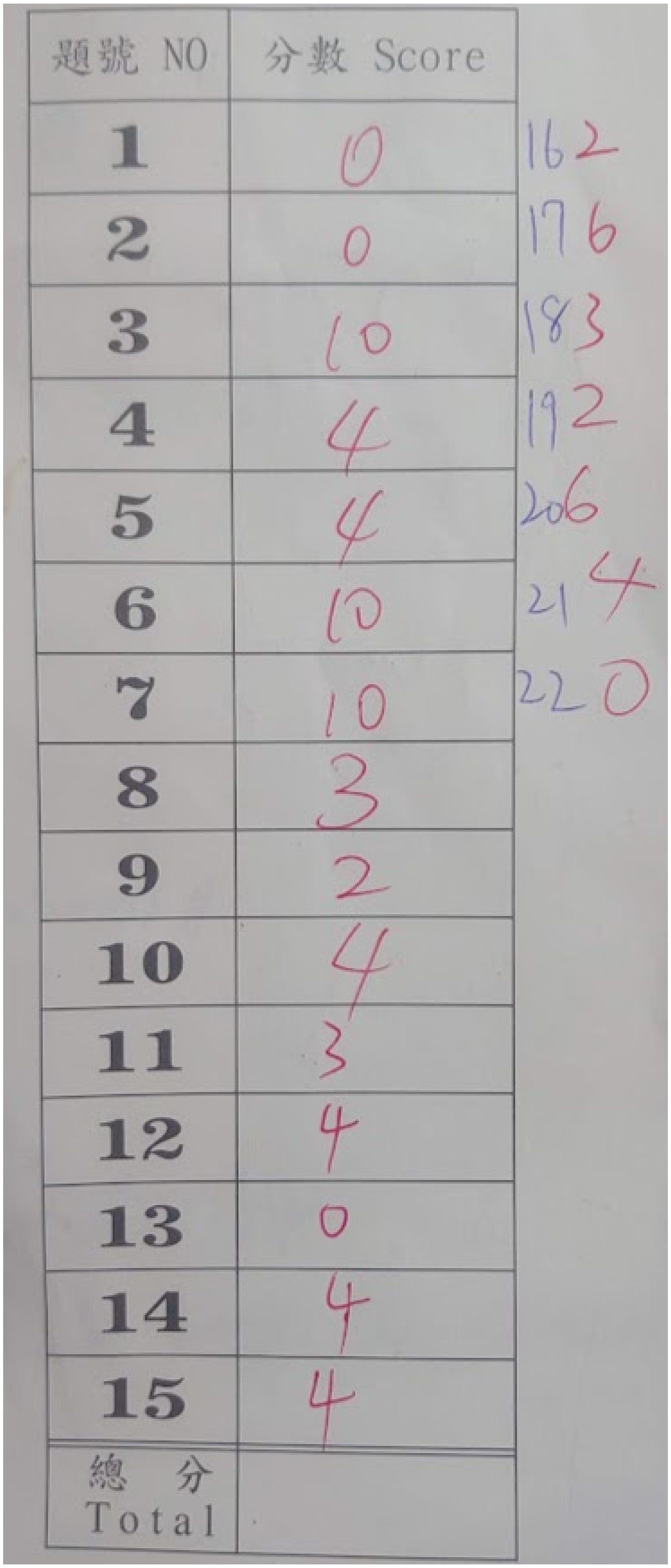
Individual-Global-Maximum (IGM)

– 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

我的答案



題號 NO	分數 Score	
1	0	162
2	0	176
3	10	183
4	4	192
5	4	206
6	10	214
7	10	220
8	3	
9	2	
10	4	
11	3	
12	4	
13	0	
14	4	
15	4	
總分 Total		

Q_1 (a)

$E_{s,a\sim\pi}[V^{\pi}(s)] = E_{s,a\sim\pi}[Q^{\pi}(s,a)]$ ，以π為 policy. state s 的估值

$E_{s_t,a_t\sim\pi}[Q^{\pi}(s_t,a_t)] = E_{s_t,a_t\sim\pi}[R(s_t,a_t)+\gamma V^{\pi}(s_t)]$ ；以π為 policy、after state (s, a) 的估值

(b)

$E_{s,a\sim\pi}[A^{\pi}(s,a)] = E_{s,a\sim\pi}[Q^{\pi}(s,a) - V^{\pi}(s)]$

$E_{s,a\sim\pi}[A^{\pi}(s,a)] = E_{s,a\sim\pi}[Q^{\pi}(s,a) - Q^{\pi}(s,a)] = 0$ #

Q_2

(a) $V_{\pi}(s) = Q_{\pi}(s,\pi(s))$

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

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

Q_3

known $\max_i\{x_i\} - \max_i\{y_i\} \le \max_i\{x_i - y_i\}$

$\|TU(s) - TV(s)\|_\infty = \|\max_{a\in A}\{R_s^a + \gamma\sum_{s'} P_{ss'}^a U(s')\} - \max_{a\in A}\{R_s^a + \gamma\sum_{s'} P_{ss'}^a V(s')\}\|_\infty$

$\le \|\max_{a\in A}\{R_s^a + \gamma\sum_{s'} P_{ss'}^a U(s') - R_s^a - \gamma\sum_{s'} P_{ss'}^a V(s')\}\|_\infty = \|\max_{a\in A}\{\gamma\sum P_{ss'}^a (U(s') - V(s'))\}\|_\infty$

$\le \gamma\|U(s') - V(s')\|_\infty$

$\therefore \|TU - TV\|_\infty \le \gamma\|U - V\|_\infty$ #

Q4 4

off-policy, Q-learning 之更新式 $Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(R_s^a + \gamma \max_{a'} Q(s',a')\right)$

只和 s, a, r, s' 有關, a' 是採最佳策略, 會取代原始策略, 所以訓練之策略目和當前策略不同, 故是 off-policy #

Q5 4

令老虎机 i 之(採樣)勝率 x_i, 已玩次數 n_i

每次玩選擇 $\arg\max_i x_i + \frac{\log \Sigma n_i}{n_i}$

x_i 幫我們找出最大勝率的玩, $\frac{\log \Sigma n_i}{n_i}$ 讓我們玩比較少玩的机会

提高勝率也兼具探索性(可能可找到 x_i 更大的) #

Q6 10

(a)

	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)
up	0	0	0	-0.1	0	0
down	0	-1	0	0	0	0
right	-0.1	0	0	0	0	0
left	0	0	0	0	0	0

① $Q^+((1,0), up) = -1 + 0.9\, Q((0,0), \underset{max}{a'}) = -1$

② $Q^+((0,0), right) = -1 + 0.9 \cdot 0$

Q6 (b)

	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)
up	0	0	0	-0.19	0	0
down	0	-1	0.1	0	0	0
right	-0.19	-0.1	0	0	0	0
left	0	0	0	0	0	0

$Q((1,0),up) = 0.9\cdot(-0.1) + 0.1(-1 + 0.9\cdot 0) = -0.19$

$Q((0,0),right) = 0.9(-0.1) + 0.1(-1 + 0.9\cdot 0) = -0.19$

$Q((0,1),right) = 0.1(-1) = -0.1$

$Q((0,2),down) = 0.1(1)$ #

10

Q7 section 1. 跑 n 个 episode，搜集 $((s_j, a_j, p_j, r_j, s'_j, g_j))$，如下 (大小為 N)

buffer = [ ]

for ep = 1...n

与 env 互动完成一个 episode 得到 V. data = $((s_i, a_i, v_i, p_i, r_i, s'_i))_i$ → $Q_\phi(s_i, a_i)$ → $\pi_\theta(a_i|s_i)$

for $(s_i, a_i, v_i, p_i, r_i, s'_i)$ in data

$\delta_i = r_i + \gamma v_{i+1} - v_i$

for $i = 0 \ldots N_i$

算 GAE: $G_i = \delta_i + \lambda \delta_{i+1} + \lambda^2 \delta_{i+2} + \ldots$

buffer append $(s_i, a_i, v_i, p_i, G_i)$

for $(s_i, a_i, v_i, p_i, G_i)$ in buffer

$$L_Q = \frac{1}{N}\sum_i \left(G_i - Q_\phi(s_i,a_i)\right)^2$$

$$A_i = G_i - v_i,\quad p'_{i,\theta} = \max_a \pi_\theta(a|s_i)$$

$$B_\pi = \frac{1}{N}\sum_i \left(\min\left[A_i \frac{p'_{i,\theta}}{p_i},\ A_i\,\mathrm{clip}\left(\frac{p'_{i,\theta}}{p_i}, 1-\epsilon, 1+\epsilon\right)\right]\right)$$

$$E_\pi = -\frac{1}{N}\sum_i \sum_a \pi_\theta(a|s_i)\log\pi_\theta(a|s_i)$$

code

update θ,ϕ by loss: $L_Q - \tau_a B_\pi - \tau_b E_\pi$

description

希望 $Q_\phi(s_i,a_i)$ 接近 G_i，故最小化 L_Q

希望採取 advantage 大的動作，故 $A_i>0$，提高 $p'_{i,\theta}$，反之減少

故最大化 B_π，clip是避免 π 變化太大

希望 π 可以探索，故最大化 π 的 entropy。

loss $= L_Q - \tau_a B_\pi - \tau_b E_\pi$，$\tau_a . \tau_b$ 為權重

Q8. actor π_θ 一直換策略會導致不穩定，delay actor update 不每回合更新 π_θ 緩和該問題，降低 variance。

Q9. 更新 actor 的參數 ϕ，SAC假設 $\pi_\phi(\cdot|s_t)$ 的distribution 與 $\exp(Q_\theta(s_t,\cdot))$ 很像，用 D_{KL} 比較兩者差距並試圖最小化它。

Q10. ICM

CIM 鼓勵模型探索無法預測 s' 的情況

CIM 處理 noisy TV 會不斷 explore 因為 noisy TV 下一狀態 s' 皆是隨机的

4

RND 則是鼓勵模型探索較少或沒探索过的 state,

這讓它在 noisy TV 中不會过度 exlore #

Q11 避免 Q-network 过度樂觀,

3

Q12

4 DQN 是離散策略, 使用 ε-greedy 讓模型訓練時有固定机率採隨机策略。

DDPG 是連續策略. 在訓練時, 於 action 中加高斯 noise.

Q13

0 use small learning rate α

Q14

4 C51 的 Q vale 的机率分佈是離散的, 而 sample 出來的 Q value 可能介於兩个 Q vale atom 中間, 於是需要 projection 分配到兩个 atom 上。

Q15

4

① 一個分佈，以不同百分點的Q-value代表

② 不同百分點 Q-value的平均

Q16

2

是z，$z = \begin{cases} 1, \text{episode結束後勝利} \\ 0, \text{episode結束後輸了} \end{cases}$ → value

policy？

Q17

6

h：將版面的feature取出

g：預測下一state

f：為state進行價值估算

Q18.

3

可以自己學習遊玩規則

Q19. 2

縮小 $Q(s,a)+I(a_E,a)$ 和 $Q(s,a_E)$ 的差距

這讓 demonstration action a_E 高於其它 action 的 Q value 一個距離 #

$$I(a_E,a) = \begin{cases} 0 & a=a_E \\ \alpha & a\neq a_E \end{cases} \quad (\alpha>0)$$

Q20、每个 agent 會隨著其它 agent 改變动作而改变动作
造成 non-stationarity、這讓模型難以收斂
如剪刀石頭布

agent1	agent2
剪刀	布
⋮	⋮
同上	換石頭
⋮	⋮
換布	同上

6

Q21.
是將不同agent的模型進行集中訓練
Centralized Training，其好處是避免 non-stationarity.
Decentralized Execution，其好處是避免 action space 过大。
是agent分开決策。

4

Q22、這性質讓每个 agent 把自己練好等同於 Q_{joint} 最大
因為 $\frac{\partial Q_{joint}}{\partial Q^i} > 0$

0

期末考答題關鍵

Q1

(a)
V-pi: expected return start from state s to terminal.
Q-pi: expected return start from state s and apply action a to terminal.
(b)
E[A] = E[Q-V] = E[Q] - E[V]
by definition, E[Q] = V and E[V] = V. So, E[A] = V - V = 0

Q4

Off-policy. Because it uses argmax of Q for training target.

Q5

Balance between exploration and exploitation.

Q8

More accurate critic to stabilize training.

Q9

Let the policy get close to softmax of Q, obtaining a multi-modal policy for exploration.

Q10

Forward prediction model visits the states he can't predict. But states are always hard to predict in noisy TV problem. And RND is just like a pseudo-counter, counting how many times he visits the similar states.

Q11

Over-optimistic.

Q12

DQN: epsilon-greedy
DDPG: noise

Q13

Baseline, actor critic, MC->TD

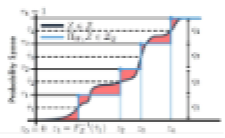
Q14

Distribute probability mass to neighboring atoms. The values of the target distribution may not be on the atoms.

Q15

Quantile values (not probability).

Q-value: $Q(x,a) = \mathbb{E}[Z(x,a)] = \sum_i \frac{1}{N} z_i$
- e.g. $N = 4$
- $Q(x,a) = \frac{1}{4}z_1 + \frac{1}{4}z_2 + \frac{1}{4}z_3 + \frac{1}{4}z_4$



Q16

policy target: the probability from MCTS
value target: the result of the game (win +1, loss -1)
(有寫到這兩個network在做甚麼就可以)

Q17

h: convert observations to embedding states.
g: given the current state and action, get the next state and reward.
f: predict the value and policy for the current state.

Q18

Also learns the dynamics of the environment. Can be applied to Atari games.

Q19

有描述到以下各點分別可得到的分數
- Pretraining phase (2 points)
- Supervised loss (2 points)
- Replay Buffer/PER with Online Data / Online Training (2 points)

Q20

有描述到以下各點分別可得到的分數

- 描述 non-stationary 的概念 (2 points)

- 描述原因 (2 points)
- 用 rock-paper-scissors 的例子是否適切 (2 points)
    - 若此例子能很好的順便表達了前兩者，前兩者也可給分

Q21

有描述到以下各點分別可得到的分數

- Centralized Training 是什麼 (1 point)
- Centralized Training 的優點 (1 point)
- Decentralized Execution 是什麼 (1 point)
- Decentralized Execution 的優點 (1 point)

Q22

描述到以下相關概念：The global optimal action computed during the centralized training phase is consistent with the actions that would be chosen by the agents acting individually based on their local observations during execution (4 points)