Final Exam Keypoints

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

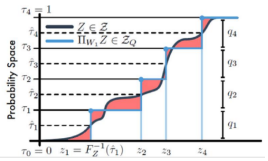
Distribute probability mass to neighboring atoms. The values of the target distribution may not be on the atoms.

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)