學號：r12227118 系級：心理所碩二 姓名：張瑋宸

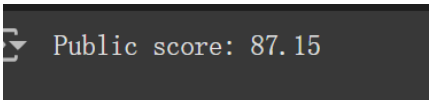
a. (0.5%) Please write down the Bellman consistency equation in terms of V^{π} on both sides.

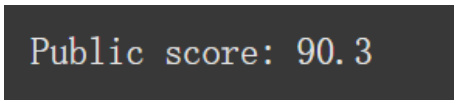
$$V^{\pi}(s) = \mathbb{E}_{a\sim\pi(\cdot|s),s'\sim P(\cdot|s,a)}\left[r(s,a) + \gamma V^{\pi}(s')\right]$$

b. (0.5%) Please implement the epsilon-greedy algorithm or the UCB algorithm. Paste the code and compare the public leaderboard scores of it and the default greedy algorithm (directly choose the state with maximum value).

Implrement UCB

```
def ucb_action_selection(env, state, value_table, count_table, c=2):
  actions = env._knight_moves(state[:2]) # -> List[(knight_x, knight_y)]
  pawn_cur_pos = state[2:]

  total_visits = np.sum(count_table[(*state[:2], *pawn_cur_pos)]) + 1e-6 # Total visits to current state

  max_ucb = -float('inf')
  best_actions = []
  # Calculate UCB for all actions before selecting one

  for a in actions:
    visits = count_table[(*a, *pawn_cur_pos)] + 1e-6
    ucb = value_table[(*a, *pawn_cur_pos)] + c * np.sqrt(np.log(total_visits) / visits)

    if ucb > max_ucb:
      max_ucb = ucb
      best_actions = [a]
    elif ucb == max_ucb:
      best_actions.append(a)

  return random.choice(best_actions) if best_actions else random.choice(actions)
```

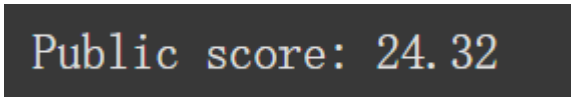
greedy: 

UCB: 

UCB slightly better.

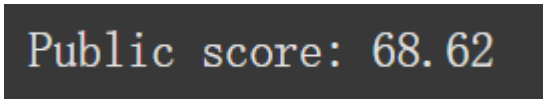
c. (1%) How to encourage the agent to catch the pawn as soon as possible?
Please make two modifications (for example, change the reward function, discount factor, …)
    a. What is your first modification? How does it affect your public score?
    b. What is your second modification? How does it affect your public score?

原始參數設定: GAMMA=1, TAU=1, REWARD_STEP=0 的 public score

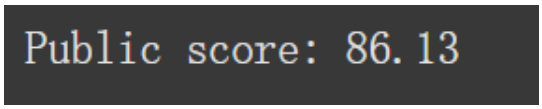


Ans:

a. 增加步數的懲罰，設定REWARD_STEP =-0.05。分數如下



b. 把discount factor (GAMMA)調小成0.3，延續REWARD_STEP =-0.05



與初始給的hyperparameter 相比，兩者都使public score 有所上升。