第二次作业（强化学习）

xxx　PB22000000

2024 年 4 月 30 日

本次作业需独立完成，不允许任何形式的抄袭行为，如被发现会有相应惩罚。在上方修改你的姓名学号，说明你同意本规定。

问题 1：热身（10 分）

a. 计算（5 分）

i	$s=0$	$s=1$	$s=2$	$s=3$
0	0	0	0	0
1	$V^{(1)}(0)$	$V^{(1)}(1)$	$V^{(1)}(2)$	$V^{(1)}(3)$
2	$V^{(2)}(0)$	$V^{(2)}(1)$	$V^{(2)}(2)$	$V^{(2)}(3)$

表 1: Value Iteration for $i \in \{0, 1, 2\}$

b. 计算（5 分）

TODO

问题 2：Q-Learning（15 分）

a. 回答问题（2 分）

TODO

b. 计算（8 分）

TODO

c. 回答问题（5 分）

TODO

问题 3：Gobang Programming（55 分）

a. 回答问题（2 分）

TODO

b. 代码填空（33 分）

TODO

```
class Gobang(UtilGobang):

    def get_next_state(self, action: Tuple[int, int, int], noise: Tuple[int, int, int])
        -> np.array:

        # BEGIN_YOUR_CODE (our solution is 3 line of code, but don't worry if you
            deviate from this)

        # END_YOUR_CODE

        if noise is not None:
            white, x_white, y_white = noise
            next_state[x_white][y_white] = white
        return next_state

    def sample_noise(self) -> Union[Tuple[int, int, int], None]:

        if self.action_space:
            # BEGIN_YOUR_CODE (our solution is 2 line of code, but don't worry if you
                deviate from this)

            # END_YOUR_CODE
            return 2, x, y
        else:
            return None

    def get_connection_and_reward(self, action: Tuple[int, int, int],
                                  noise: Tuple[int, int, int]) -> Tuple[int, int, int,
                                      int, float]:

        # BEGIN_YOUR_CODE (our solution is 4 line of code, but don't worry if you
            deviate from this)

        # END_YOUR_CODE

        return black_1, white_1, black_2, white_2, reward
```

```
    def sample_action_and_noise(self, eps: float) -> Tuple[Tuple[int, int, int], Tuple[
        int, int, int]]:

        # BEGIN_YOUR_CODE (our solution is 8 line of code, but don't worry if you
            deviate from this)

        # END_YOUR_CODE
        return action, self.sample_noise()

    def q_learning_update(self, s0_: np.array, action: Tuple[int, int, int], s1_: np.
        array, reward: float,
                          alpha_0: float = 1):

        s0, s1 = self.array_to_hashable(s0_), self.array_to_hashable(s1_)
        self.s_a_visited[(s0, action)] = 1 if (s0, action) not in self.s_a_visited else
            \
            self.s_a_visited[(s0, action)] + 1
        alpha = alpha_0 / self.s_a_visited[(s0, action)]

        # BEGIN_YOUR_CODE (our solution is 18 line of code, but don't worry if you
            deviate from this)

        # END_YOUR_CODE
```

c. 结果复现（10 分）

TODO 你需要将复现结果的截图粘贴在这里。

图 1: 复现结果

图 1是一张示例图片，请你按照示例插入图片以及文字叙述。

d. 回答问题（10 分）

TODO

问题 4：Deeper Understanding（10 分）

a. 回答问题（5 分）

TODO

b. 回答问题（5 分）

TODO

反馈（10 分）

在每次实验报告的最后欢迎反馈你上这门课的感受，你可以写下任何反馈，包括但不限于以下几个方面：课堂、作业难度和工作量、助教工作等等。

TODO

- .....
- .....