# 执行测验: Homework 5

## 测试信息

描述

我们将在本次作业中允许多次尝试，不限制提交次数。请注意：

- 作业将使用最后一次尝试的成绩作为最终成绩；
- 未提交的尝试将被记为0分；
- 当开始新的尝试时，所填入的答案将被完全清除。

因此，当决定提交作业时，请在其他设备上妥善保存已经完成的答案；否则，请保存答案但不要提交。在截止日期之前，请确保作业的最后一次尝试已经提交。在截止日期之后，如果发现作业成绩有任何问题，可以随时联系助教处理。

**FAQ**

1. 作业有grace day吗?

BB作业没有grace day，Autolab编程作业有5个grace day。

2. 我忘记提交作业了，可以请助教帮忙提交吗?

在同时满足以下条件时，你可以联系助教在ddl之后为你提交作业：
a. 你的当前作业没有成绩，没有提交记录；
b. 你的作业完成记录显示你的所有操作在ddl之前完成。

注意，BB会记录助教的所有操作，这些操作也都将需要归档。

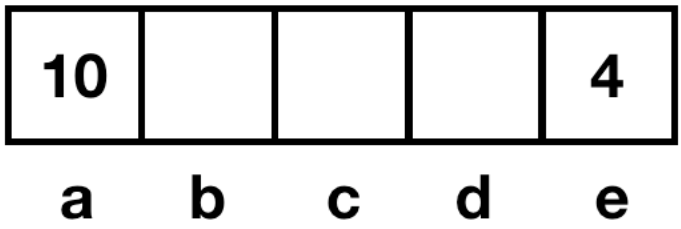
| | |
|---|---|
| 说明 | 注意：本作业不会自动提交。请在完成作业检查无误后，单击右下角“保存并提交”按钮提交作业。逾期未提交的作业不会被保存或计分。 |
| 多次尝试 | 此测试允许进行多次尝试。 |
| 强制完成 | 本测试可保存并可稍后继续。 |

问题完成状态：

问题 1 10 分 已保存

**Value Iteration**

Consider the gridworld where Left and Right actions are successful 100% of the time. Specifically, the available actions in each state are to move to the neighboring grid squares. From state $a$, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state $e$, the reward for the exit action is 4. Exit actions are successful 100% of the time.




Let the discount factor $\gamma$ = 0.5. Fill in the following quantities.

$V^*(a) = V_\infty(a) =$ 10

$V^*(b) = V_\infty(b) =$ 5

$V^*(c) = V_\infty(c) =$ 2.5

$V^*(d) = V_\infty(d) =$ 2

$V^*(e) = V_\infty(e) =$ 4

问题 2 10 分 已保存

## To pollute or not to pollute - Part 1

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| +50 | -1 | -1 | -1 | ... | -1 | -1 | -1 | -1 |
| *Start* | | | | ... | | | | |
| -50 | +1 | +1 | +1 | ... | +1 | +1 | +1 | +1 |

Figure 1: $101 \times 3$ world for a pollution model.

Consider the $101 \times 3$ grid world shown in Figure 1 (omitting 93 identical columns in the middle). The start state has reward 0. In the start state, the agent has a choice of two deterministic actions, Up or Down, but in the other states the agent has only one deterministic action, Right. The game ends when the agent reaches a right-most state. The agent receives the reward in a state when it transits to that state, including the right-most ones. The discount factor is $\gamma$.

This simple example actually reflects many real-world situations in which one must weigh the value of an immediate action versus potential long-term consequences, such as choosing to dump pollutants into a lake.

### 1 Markov Decision Process

#### 1.1 Value Iteration

Assume $\gamma = 1$. Recall the update function in value iteration:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T\left(s, a, s'\right)\left[R\left(s, a, s'\right) + \gamma V_k\left(s'\right)\right]$$

Calculate the following. Any decimal points will be acceptable.

$V_1(Start) =$ 50

$V^*(Start) = V_\infty(Start) =$ 50

问题 3 10 分 已保存

**To pollute or not to pollute - Part 2**

**1.2 The Discount Factor**

Similar to the previous problem, but now we assume $\gamma = 0.9$. Then what is the value of $V^*(Start)$?

You may use $0.9^{100} \approx 0$ in your calculation.

HINT:

$$x + x^2 + x^3 + \ldots + x^n = \frac{x^{n+1} - x}{x - 1}$$

- (●) 41
- ( ) 50
- ( ) 45
- ( ) 49
- ( ) 40

问题 4 — 5 分 — 已保存

**To pollute or not to pollute - Part 3**

**1.3 Stay Stable**

How many iterations does value iteration need to get a stable result? That is if $\forall t \geq t_0$

$$V_t(s) = V^*(s) = V_\infty(s)$$

then what is the minimum value of $t_0$? Assume there is no approximation in the calculation. You answer should be a positive integer.

101

问题 5 — 5 分 — 已保存

**To pollute or not to pollute - Part 4**

**1.4 Discounted Future**

For what values of the discount factor $\gamma$ should the agent choose Down (i.e. $\pi^*(Start) = Down$)? Assume that $\forall 0 < x < 1, x^{100} \approx 0$.

- ( ) $0 \leq \gamma < \frac{1}{51}$
- ( ) $\frac{1}{51} < \gamma \leq 1$
- ( ) $0 \leq \gamma < \frac{50}{51}$
- (●) $\frac{50}{51} < \gamma \leq 1$

问题 6 — 10 分 — 已保存

**To pollute or not to pollute - Part 5**

**2 Undetermined Transitions**

Assume the transition of the bottom-left grid is no longer deterministic. After taking the action `Right`, it is possible to transit to

1. the bottom grid on the second column as before, or
2. the top grid on the second column.

**2.1 Model-Based Learning**

After running the game several times, we observe the following state sequences:

After running the game several times, we observe the following state sequences:

- $Start, +50, -1, -1, \ldots, -1$
- $Start, -50, -1, -1, \ldots, -1$
- $Start, +50, -1, -1, \ldots, -1$
- $Start, -50, -1, -1, \ldots, -1$
- $Start, -50, -1, -1, \ldots, -1$
- $Start, -50, +1, +1, \ldots, +1$
- $Start, +50, -1, -1, \ldots, -1$

What is the estimated $T(s_{-50}, Right, s_{-1})$, where $s_{-50}$ is the bottom-left grid and $s_{-1}$ is the top grid on the second column?

- ○ $\frac{1}{3}$
- ○ $\frac{6}{7}$
- ◉ $\frac{3}{4}$
- ○ $\frac{1}{4}$

问题 7 10 分 已保存

**To pollute or not to pollute - Part 6**

**2.2 Temporal Difference Learning**

Assume $\gamma = 1$, learning rate $\alpha = 0.5$. The policy is to go down from the start state, then keep going right. Initially, $\forall s, V^{\pi}(s) = 0$. Recall that in TD learning,

$$V^{\pi}(s) \leftarrow (1-\alpha)V^{\pi}(s) + \alpha\left[R\left(s, \pi(s), s'\right) + \gamma V^{\pi}\left(s'\right)\right]$$

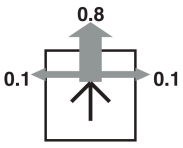
After observing the state sequence $Start, -50, -1, -1, \ldots, -1$, what is the updated $V^{\pi}(s_{-50})$, where $s_{-50}$ is the bottom-left grid? Any decimal points will be acceptable.

-0.5

问题 8 10 分 已保存

In the grid world below, an agent moves with noise: each action achieves the intended effect with probability 0.8, but for the rest of the time, the action moves the agent at a perpendicular angle to the intended direction (0.1 for the left and 0.1 for the right, see the figure below). Furthermore, if the agent bumps into a wall, it stays in the same square. The living reward is -0.04 and the discount factor is 1 (i.e., no discount).

| | | | | |
|---|---|---|---|---|
| 3 | 0.812 | 0.868 | 0.918 | +1 |
| 2 | 0.762 | | 0.660 | −1 |
| 1 | 0.705 | 0.655 | 0.611 | 0.388 |
| | 1 | 2 | 3 | 4 |




The value of each state has been shown in the figure. How should the agent in grid (2, 3) (with value 0.660) act based on these values?

- ○ Right
- ◉ Up
- ○ Left
- ○ Down

问题 9

10 分 已保存

Suppose in reinforcement learning, we want to evaluate a fixed policy $\pi$. One possible way to compute the value of state $s$ is to take samples of outcomes $s'$ and average:

$$\text{sample}_i = R\left(s, \pi(s), s_i'\right) + \gamma V_k^{\pi}\left(s_i'\right)$$
$$V_{k+1}^{\pi}(s) \leftarrow \frac{1}{n}\sum_i \text{sample}_i$$

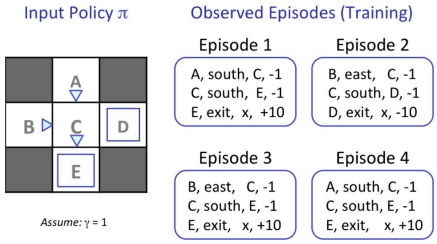
It doesn't work in practice because

- (●) We can't rewind time to get samples after sampling from state $s$.
- ( ) The transition function $T\left(s, a, s'\right)$ should be considered in the samples.
- ( ) The average of the samples cannot correctly reflect the value of state $s$.
- ( ) There might be more than one possible actions in state $s$ and the samples only consider one of them $\pi(s)$.

问题 10

10 分 已保存

**Model Based RL**




What model would be learned from the above observed episodes?

T(A, south, C) = 1

T(B, east, C) = 1

T(C, south, E) = 0.75

T(C, south, D) = 0.25

问题 11

30 分 已保存

**Feature-Based Representation**

Consider the following feature based representation of the Q-function:

$$Q\left(s, a\right) = w_1 f_1\left(s, a\right) + w_2 f_2\left(s, a\right)$$

with

$$f_1\left(s, a\right) = 1/\left(\text{Manhattan distance to nearest dot after having executed action } a \text{ in state } s\right)$$
$$f_2\left(s, a\right) = \left(\text{Manhattan distance to nearest ghost after having executed action } a \text{ in state } s\right)$$

Part 1

Assume $w_1 = 1, w_2 = 10$. For the state $s$ shown below, find the following quantities. Assume that the red and blue ghosts are both sitting on top of a dot.

$Q\left(s, West\right) =$

31

$Q\left(s, South\right)$

$Q\left(s, South\right) =$

11

Part 2

Assume Pac-Man moves West. This results in the state $s'$ shown below.

The reward for this transition is $r = +10 - 1 = 9$ (+10: for food pellet eating, -1 for time passed). Fill in the following quantities. Assume that the red and blue ghosts are both sitting on top of a dot.

$Q\left(s', West\right) =$

11

$Q\left(s', East\right) =$

11

What is the sample value (assuming $\gamma = 1$)?

$\text{sample} = \left[r + \gamma \max_{a'} Q\left(s', a'\right)\right] =$

20

Part 3

Now let's compute the update to the weights. Let $\alpha = 0.5$.

$\text{difference} = \left[r + \gamma \max_{a'} Q\left(s', a'\right)\right] - Q\left(s, a\right) =$

-11

$w_1 \leftarrow w_1 + \alpha\left(\text{difference}\right) f_1\left(s, a\right) =$

-4.5

$w_2 \leftarrow w_2 + \alpha\left(\text{difference}\right) f_2\left(s, a\right) =$

-6.5

*单击"保存并提交"以保存并提交。单击"保存所有答案"以保存所有答案。*

保存所有答案 保存并提交