复查测验提交: Homework 5

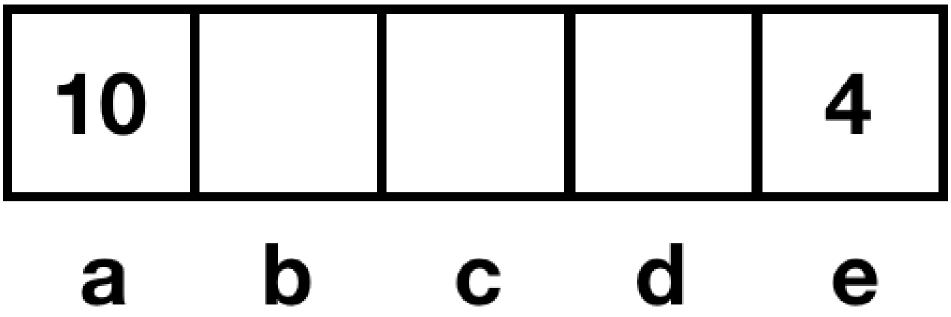
用户	信息科学与技术学院 周守琛
课程	人工智能I
测试	Homework 5
已开始	23-12-15 下午10:23
已提交	23-12-21 上午1:12
截止日期	23-12-22 下午11:59
状态	已完成
尝试分数	得 140 分，满分 140 分
已用时间	122 小时 48 分钟
显示的结果	所有答案, 已提交的答案, 正确答案

问题 1

得 10 分，满分 10 分

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 γ = 0.5. Fill in the following quantities.

$V^*(a) = V_\infty(a) =$ **[x1]**

$V^*(b) = V_\infty(b) =$ **[x2]**

$V^*(c) = V_\infty(c) =$ **[x3]**

$V^*(d) = V_\infty(d) =$ **[x4]**

$V^*(e) = V_\infty(e) =$ **[x5]**

x1 的指定答案： ✓ 10

x2 的指定答案： ✓ 5

x3 的指定答案： ✓ 2.5

x4 的指定答案： ✓ 2

x5 的指定答案： ✓ 4

x1 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	10(.0*)?	

x2 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	5(.0*)?	

x3 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	2.50*	

x4 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	2(.0*)?	

x5 的正确答案：

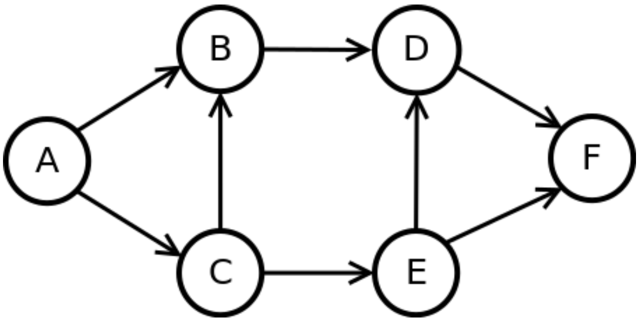
评估方式	正确答案	区分大小写
✓ *模式匹配*	4(.0*)?	

问题 2

得 10 分，满分 10 分

Value Iteration Convergence

We will consider a simple MDP that has six states, A, B, C, D, E, and F. Each state has a single action, *go*. An arrow from a state x to a state y indicates that it is possible to transition from state x to next state y when *go* is taken. If there are multiple arrows leaving a state x, transitioning to each of the next states is equally likely. The state F has no outgoing arrows: once you arrive in F, you stay in F for all future times. The reward is one for all transitions, with one exception: staying in F gets a reward of zero. Assume a discount factor = 0.7. We assume that we initialize the value of each state to 0. (Note: you should not need to explicitly run value iteration to solve this problem.)



After how many iterations of value iteration will the value for state C have become exactly equal to the true optimum? (Enter inf if the values will never become equal to the true optimal but only converge to the true optimal.)**[x1]**

How many iterations of value iteration will it take for the values of all states to converge to the true optimal values? (Enter inf if the values will never become equal to the true optimal but only converge to the true optimal.)**[x2]**

x1 的指定答案： ✓ 3

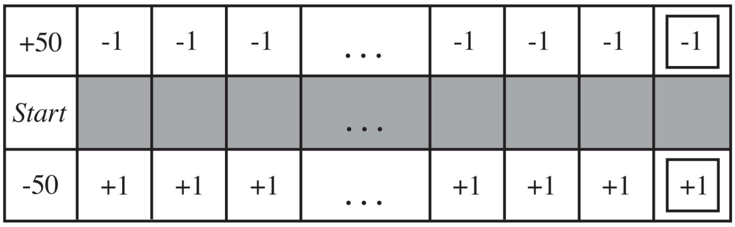
x2 的指定答案： ✓ 4

x1 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	3(.0*)?	
x2 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	4(.0*)?	

问题 3

得 10 分，满分 10 分

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: 101×3 world for a pollution model.

Consider the 101×3 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 γ.

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.

$V_1(Start) =$ **[q1]**

$V^*(Start) = V_\infty(Start) =$ **[q2]**

q1 的指定答案： ✓ 50

q2 的指定答案： ✓ 50

q1 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	50(.0*)?	

q2 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	50(.0*)?	

问题 4

得 10 分，满分 10 分

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:

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

所选答案： ✓ 41

答案： ✓ 41

50

45

49

40

问题 5

得 5 分，满分 5 分

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.

所选答案： ✓ 101

正确答案：

评估方式	正确答案	区分大小写
✓ *完全匹配*	101	

问题 6

得 5 分，满分 5 分

To pollute or not to pollute - Part 4

1.4 Discounted Future

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

所选答案： ✓ $\frac{50}{51} < \gamma \leq 1$

答案：

$0 \leq \gamma < \frac{50}{51}$

✓ $\frac{50}{51} < \gamma \leq 1$

$\frac{1}{51} < \gamma \leq 1$

$0 \leq \gamma < \frac{1}{51}$

问题 7

得 10 分，满分 10 分

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:

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

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?

所选答案: ✓ $\frac{3}{4}$

答案: $\frac{6}{7}$

$\frac{1}{4}$

✓ $\frac{3}{4}$

$\frac{1}{3}$

问题 8

得 10 分，满分 10 分

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.

所选答案: ✓ -0.5

正确答案:

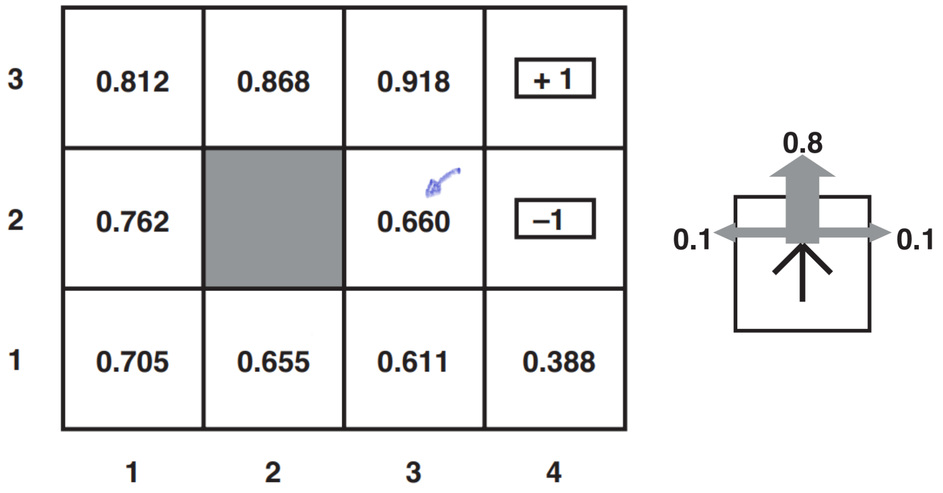
评估方式	正确答案	区分大小写

✓ *模式匹配*	-0?.50*
✓ *完全匹配*	-0.5
✓ *模式匹配*	-1/2

问题 9

得 10 分，满分 10 分

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).



	1	2	3	4
3	0.812	0.868	0.918	+1
2	0.762	▨	0.660	–1
1	0.705	0.655	0.611	0.388

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?

所选答案: ✓ Up

答案: Down

Left

Right

✓ Up

问题 10

得 10 分，满分 10 分

Suppose in reinforcement learning, we want to evaluate a fixed policy π. 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

所选答案: ✓ We can’t rewind time to get samples after sampling from state s.

答案: The transition function $T\left(s,a,s'\right)$ should be considered in the samples.

✓ We can’t rewind time to get samples after sampling from state s.

The average of the samples cannot correctly reflect the value of state s.

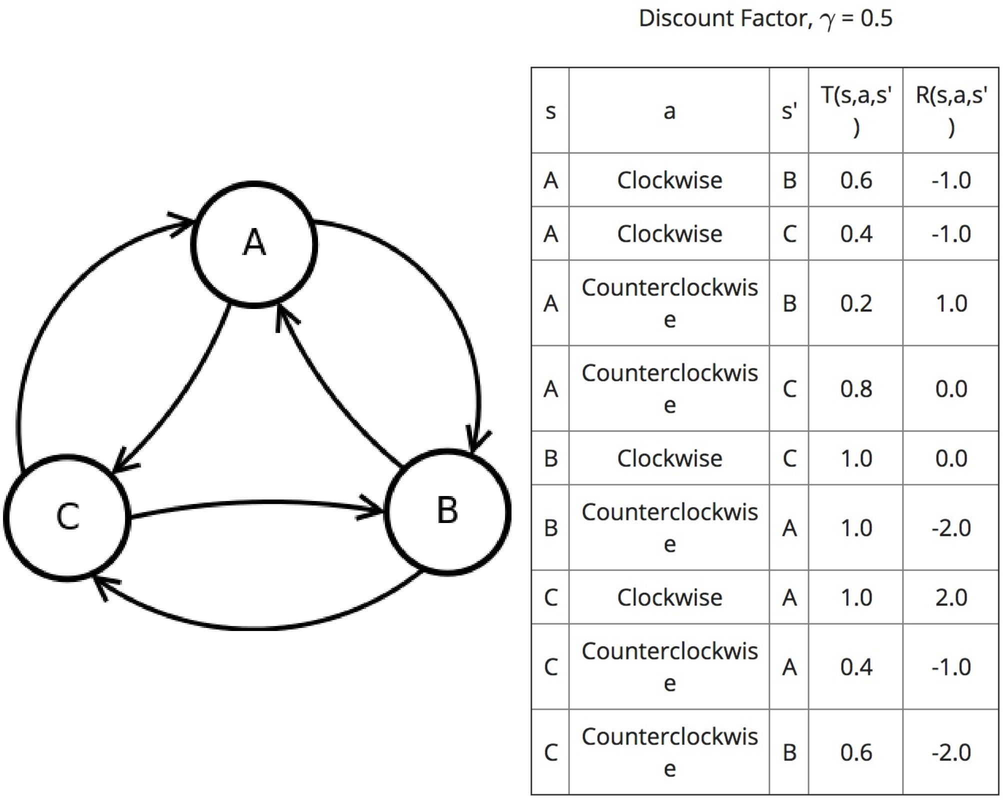
There might be more than one possible actions in state s and the samples only consider one of them $\pi(s)$.

问题 11

得 20 分，满分 20 分

Policy Iteration

Consider the following transition diagram, transition function and reward function for an MDP.



Discount Factor, γ = 0.5

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	0.6	-1.0
A	Clockwise	C	0.4	-1.0
A	Counterclockwise	B	0.2	1.0
A	Counterclockwise	C	0.8	0.0
B	Clockwise	C	1.0	0.0
B	Counterclockwise	A	1.0	-2.0
C	Clockwise	A	1.0	2.0
C	Counterclockwise	A	0.4	-1.0
C	Counterclockwise	B	0.6	-2.0

Suppose we are doing policy evaluation, by following the policy given by the left-hand side table below. Our current estimates (at the end of some iteration of policy evaluation) of the value of states when following the current policy is given in the right-hand side table.

A	B	C
Counterclockwise	Clockwise	Counterclockwise

$V_k^\pi(A)$	$V_k^\pi(B)$	$V_k^\pi(C)$
-0.440	-0.800	-1.560

Answers should keep 3 decimals.

What is $V_{k+1}^\pi(C)$? **[x1]**

Suppose that policy evaluation converges to the following value function, V_∞^π.

$V_\infty^\pi(A)$	$V_\infty^\pi(B)$	$V_\infty^\pi(C)$
-0.724	-1.026	-2.053

What is Q_∞^π(C, clockwise)? **[x2]**

What is Q_∞^π(C, counterclockwise)? **[x3]**

What is the updated action for state C? Enter clockwise or counterclockwise. **[x4]**

x1 的指定答案： ✓ -1.928

x2 的指定答案： ✓ 1.638

x3 的指定答案： ✓ -2.053

x4 的指定答案： ✓ clockwise

x1 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	-1.928	
x2 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	1.638	
x3 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	-2.053	
x4 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	clockwise	

问题 12

得 30 分，满分 30 分

Feature-Based Representation

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

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

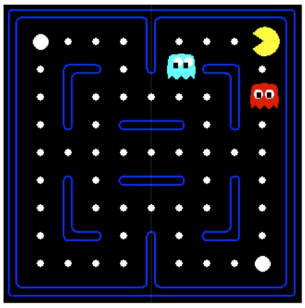
with

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

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

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.



$Q(s, West) =$

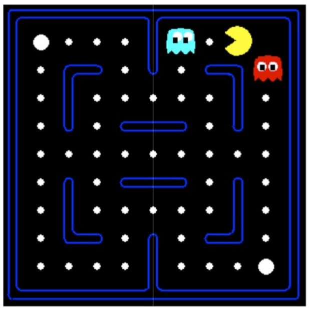
[x1]

$Q(s, South) =$

[x2]

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.

$Q(s', West) =$

[x3]

$Q(s', East) =$

[x4]

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

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

[x5]

Part 3

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

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

[x6]

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

[x7]

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

[x8]

x1 的指定答案： ✓ 31

x2 的指定答案： ✓ 11

x3 的指定答案： ✓ 11

x4 的指定答案： ✓ 11

x5 的指定答案： ✓ 20

x6 的指定答案： ✓ -11

x7 的指定答案： ✓ -4.5

x8 的指定答案： ✓ -6.5

x1 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	31(.0*)?	

x2 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	11(.0*)?	

x3 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	11(.0*)?	

x4 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	11(.0*)?	

x5 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	20(.0*)?	

x6 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	-11(.0*)?	

x7 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	-4.50*	

x8 的正确答案：

评估方式	正确答案	区分大小写
✓ *模式匹配*	-6.50*	

← 确定