AU332 Quiz9

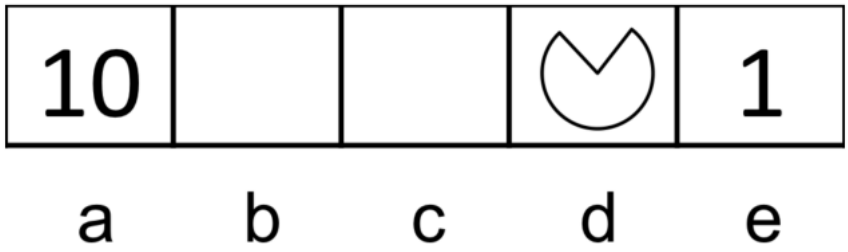
* 基本信息：

姓名： q

学号： q

*1. Solving MDPs

Consider the gridworld MDP for which Left and Right actions are 100% successful. 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 1. Exit actions are successful 100% of the time. There is no living reward.



Let the discount factor γ = 1. Fill in the following blanks.

$V_0(d)$= 0

$V_1(d)$= 0

$V_2(d)$= 1

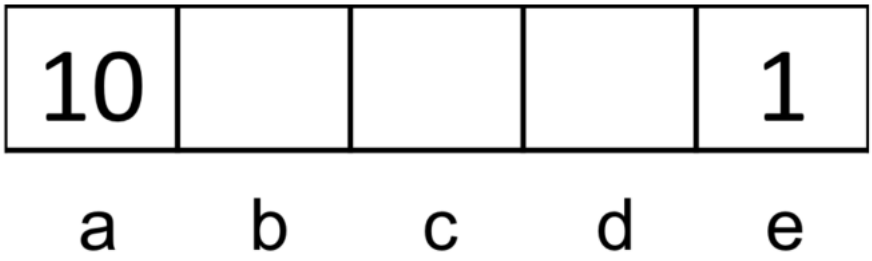
$V_3(d)$= 1

$V_4(d)$= 10

$V_5(d)$= 10

*2. Value Iteration Convergence Values

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 1. Exit actions are successful 100% of the time. There is no living reward.



Let the discount factor γ = 0.2. Fill in the following blanks.

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

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

$V^*(c)=V_\infty(c)=$ 0.4

$V^*(d)=V_\infty(d)=$ 0.2

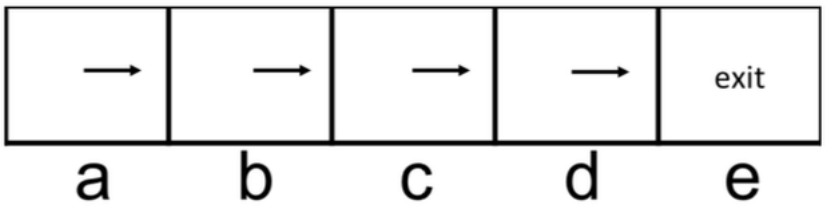
$V^*(e)=V_\infty(e)=$ 1

***3. Policy Evaluation**

Consider the grid world where Left and Right actions are successful 100% of the time. Specifically, th e available actions in each state are to move to the neighboring grid squares. From state a, there is al so an exit action available, which results in going to the terminal state and collecting a reward of 10. S imilarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time. The discount factor (γ) is 1.

Part 1

Consider the policy π_1 shown below, and evaluate the following quantities for this policy.



$V^{\pi 1}(a)=$ 1

$V^{\pi 1}(b)=$ 1

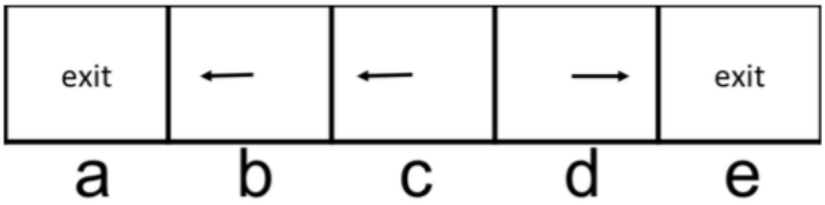
$V^{\pi 1}(c)=$ 1

$V^{\pi 1}(d)=$ 1

$V^{\pi 1}(e)=$ 1

Part 2

Consider the policy π_2 shown below, and evaluate the following quantities for this policy.



$V^{\pi 2}(a)=$ 10

$V^{\pi 2}(b)=$ 10

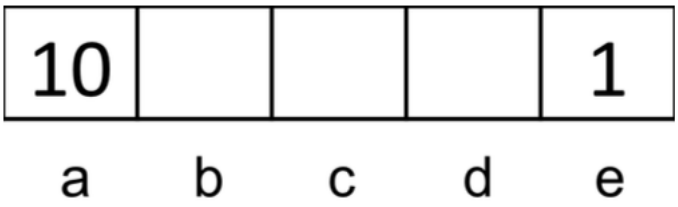
$V^{\pi 2}(c)=$ 10

$V^{\pi 2}(d)=$ 1

$V^{\pi 2}(e)=$ 1

*4. Policy Iteration

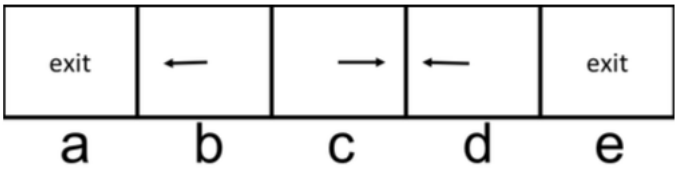
Consider the grid world where Left and Right actions are successful 100% of the time. Specifically, th e available actions in each state are to move to the neighboring grid squares. From state a, there is al so an exit action available, which results in going to the terminal state and collecting a reward of 10. S imilarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time. The discount factor (γ) is 0.9.



We will execute one round of policy iteration.

Part 1 Policy Evaluation

Consider the policy π_i shown below, and evaluate the following quantities for this policy.



$V^{\pi i}(a)=$ 10

$V^{\pi i}(b)=$ 9

$V^{\pi i}(c)=$ 0

$V^{\pi i}(d)=$ 0

$V^{\pi i}(e)=$ 1

Part 2: Policy Improvement

Perform a policy improvement step. The current policy's values are the ones from Part 1.

*5. $\pi_{i+1}(a)=$

- (•) Exit
- ( ) Right

*6. $\pi_{i+1}(b)=$

- (•) Left
- ( ) Right

*7. $\pi_{i+1}(c)=$

- (•) Left
- ( ) Right

*8. $\pi_{i+1}(d)=$

- ( ) Left
- (•) Right

*9. $\pi_{i+1}(e)=$

- ( ) Left
- (•) Exit

提交