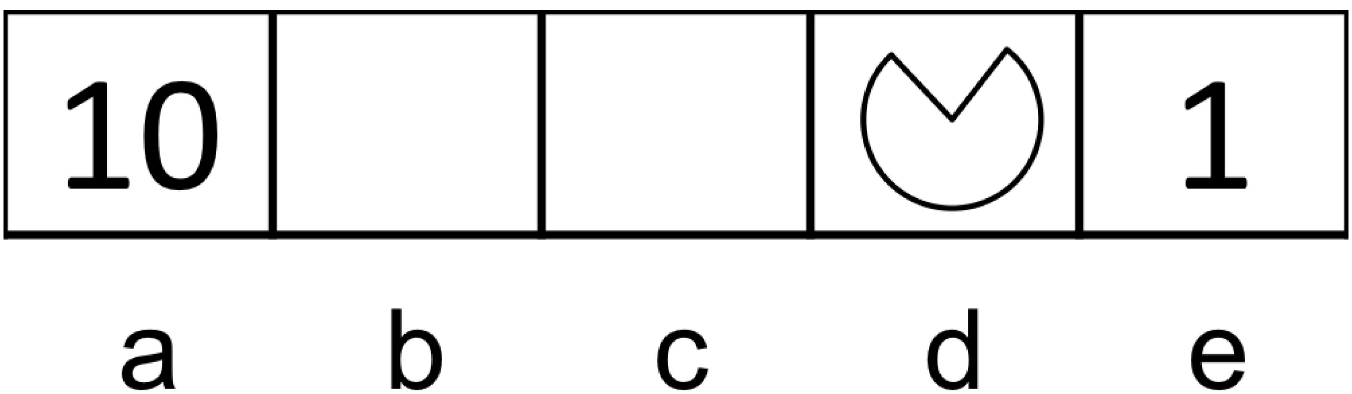
hw4_mdps_q1_solving_mdps

Question 1: Solving MDPs

0.0/6.0 points (graded)

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.



Let the discount factor $\gamma = 1$. Fill in the following quantities.

$V_0(d) =$

0

Answer: 0

When value iteration is initialized, the V_0 value of each state is 0.

$V_1(d) =$

0

Answer: 0

At the first iteration, each state knows about the V_0 values of successor states. Because $V_0(e) = 0$, the state d will not know about the exit reward at state e.

Call t the terminal state. Notice the transition probabilities do not show up in our equation, because the transitions are all deterministic.

$V_1(a) = R(a, exit, t) + V_0(t) = 10$
$V_1(b) = max(V_0(a), V_0(c)) = 0$
$V_1(c) = max(V_0(b), V_0(d)) = 0$
$V_1(d) = max(V_0(c), V_0(e)) = 0$
$V_1(e) = R(e, exit, t) + V_0(t) = 1$

$V_2(d) =$

1

Answer: 1

At the second iteration, each state knows about the V_1 values of successor states. State d will now know about the exit reward of state e, and $V_1(d)$ will be updated to 1.

$V_2(a) = 10$
$V_2(b) = max(V_1(a), V_1(c)) = 10$
$V_2(c) = max(V_1(b), V_1(d)) = 0$
$V_2(d) = max(V_1(c), V_1(e)) = 1$
$V_2(e) = 1$

$V_3(d) =$

1

Answer: 1

$V_3(a) = 10$
$V_3(b) = max(V_2(a), V_2(c)) = 10$
$V_3(c) = max(V_2(b), V_2(d)) = 10$
$V_3(d) = max(V_2(c), V_2(e)) = 1$
$V_3(e) = 1$

$V_4(d) =$

10

Answer: 10

At the fourth iteration, state d will see the exit reward of state a, so $V_4(d)$ will be updated to 10. $V_4(a) = 10$
$V_4(b) = max(V_3(a), V_3(c)) = 10$
$V_4(c) = max(V_3(b), V_3(d)) = 10$
$V_4(d) = max(V_3(c), V_3(e)) = 10$
$V_4(e) = 1$

$V_5(d) =$

10

Answer: 10

Nothing will change between the fourth and fifth iteration.

Alternatively, for a simple MDP like this one, the values could also be computed directly from the meaning of $V_i(d)$, which is the expected discounted sum of rewards if acting optimally for i time steps, starting from state d. With $i = 0$ and $i = 1$, no reward can be obtained from state d, so $V_0(d) = V_1(d) = 0$. With $i = 2$ and $i = 3$, a reward of 1 can be obtained through "Right", "Exit" from state d, so $V_2(d) = V_3(d) = 1$. Because it takes four steps to obtain the reward of 10 ("Left", "Left", "Left", "Exit"), $V_i(d) = 10$ for $i \geq 4$.

Submit

Answers are displayed within the problem