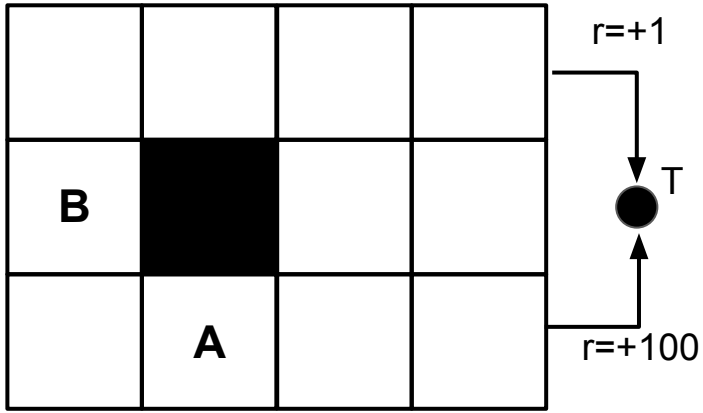


Let's consider a grid-world MDP. Shaded cells represent walls. In all states, the agent has available actions ↑,↓,←,→. Performing an action that would transition to an invalid state (outside the grid or into a wall) results in the agent remaining in its original state. In states with an arrow coming out, the agent has an additional action EXIT. In the event that the EXIT action is taken, the agent receives the labeled reward and ends the game in the terminal state T. All other transitions receive zero reward. All transitions are deterministic. Assume that the value iteration begins with all states initialized to zero.

(1) What are the optimal values for **A** and **B** if the discounting factor $\gamma = 0.5$? How many iterations of value iteration does it take for all states to converge to optimal values?

(2) What if $\gamma = 0.1$? Is the optimal policy same as the case $\gamma = 0.5$

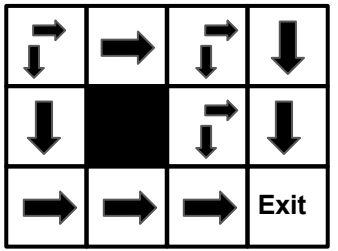
(3) In the case of no discounting factor ($\gamma = 1$), will the optimal policy same as the case $\gamma = 0.5$ or $\gamma = 0.1$?

γ = 0.5
$V^*(A) = 25$
$V^*(B) = 25 / 4$
Takes 6 iterations (shown in the figures)

Initialization

0	0	0	0
0		0	0
0	0	0	0

Policy



(1)

0	0	0	1
0		0	0
0	0	0	100

(2)

0	0	1/2	1
0		0	50
0	0	50	100

(3)

0	1/4	1/2	25
0		25	50
0	25	50	100

(4)

1/8	1/4	25/2	25
0		25	50
25/2	25	50	100

(5)

1/8	25/4	25/2	25
25/4		25	50
25/2	25	50	100

(6)

25/8	25/4	25/2	25
25/4		25	50
25/2	25	50	100

Converged!

γ = 0.1
$V^*(A) = 1$
V*(B) = 1/100
Takes 5 iterations (shown in the figures)

Initialization

0	0	0	0
0		0	0
0	0	0	0

(1)

0	0	0	1
0		0	0
0	0	0	100

(2)

0	0	1/10	1
0		0	10
0	0	10	100

(3)

0	1/100	1/10	1
0		1	10
0	1	10	100

(4)

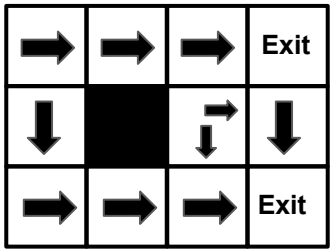
1/1000	1/100	1/10	1
0		1	10
1/10	1	10	100

(5)

1/1000	1/100	1/10	1
1/100		1	10
1/10	1	10	100

Converged!

Policy



Same as the case $\gamma = 0.5$

In the case of no discounting factor, it will consider reward in the future without any discounting. In that case, reward 100 in the terminating state will dominate in value iteration. So it will be the same as the policy found in the case $\gamma = 0.5$.