Q9: MDPs and Reinforcement Learning: Mini-Grids

Problem 9: MDPs and Reinforcement Learning: Mini-Grids

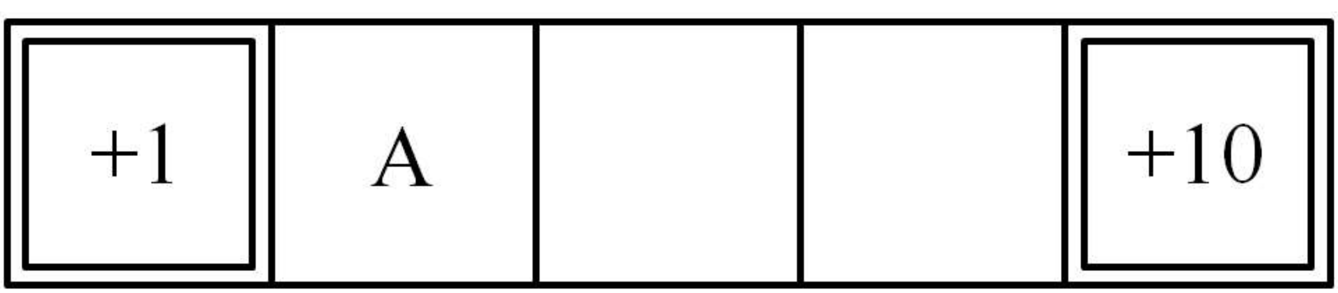
The following problems take place in various scenarios of the gridworld MDP (as in Project 3). In all cases, A is the start state and double-rectangle states are exit states. From an exit state, the only action available is $Exit$, which results in the listed reward and ends the game (by moving into a terminal state X, not shown).

From non-exit states, the agent can choose either $Left$ or $Right$ actions, which move the agent in the corresponding direction. There are no living rewards; the only non-zero rewards come from exiting the grid.

Throughout this problem, assume that value iteration begins with initial values $V_0(s) = 0$ for all states s. Also remember that the reward is only obtained *after* taking the exit action.

Part 1

First, consider the following mini-grid. For now, the discount is $\gamma = 1$ and legal movement actions will always succeed (and so the state transition function is deterministic).



Part 1.1

1/1 point (ungraded)
What is the optimal value $V^*(A)$?

10

Submit

Correct (1/1 point)

Part 1.2

1/1 point (ungraded)
When running value iteration, remember that we start with $V_0(s) = 0$ for all s. What is the first iteration k for which $V_k(A)$ will be non-zero?

2

Submit

Correct (1/1 point)

Part 1.3

1/1 point (ungraded)
What will $V_k(A)$ be when it is first non-zero?

1

Submit

Correct (1/1 point)

Part 1.4

1/1 point (ungraded)

After how many iterations k will we have $V_k(A) = V^*(A)$?

- 2
- 3
- 4 ✔
- 5
- 6
- They will never become equal for any finite value of k.

Submit

✔ Correct (1/1 point)

Part 2

Now the situation is as before, but the discount γ is less than 1.

Part 2.1

2/2 points (ungraded)

If $\gamma = 0.5$, what is the optimal value $V^*(A)$?

1.25 ✔

Submit

✔ Correct (2/2 points)

Part 2.2

2/2 points (ungraded)

For what range of values γ of the discount will it be optimal to go $Right$ from A? Remember that $0 \leq \gamma \leq 1$.

- ○ $0 \leq \gamma \leq 1$
- ○ $\frac{1}{10} \leq \gamma \leq 1$
- ◉ $\frac{1}{\sqrt{10}} \leq \gamma \leq 1$ ✔
- ○ $\gamma = 1$
- ○ $-\infty \leq \gamma \leq +\infty$
- ○ For no values of γ will it be optimal to go $Right$ from A.

Submit

✔ Correct (2/2 points)

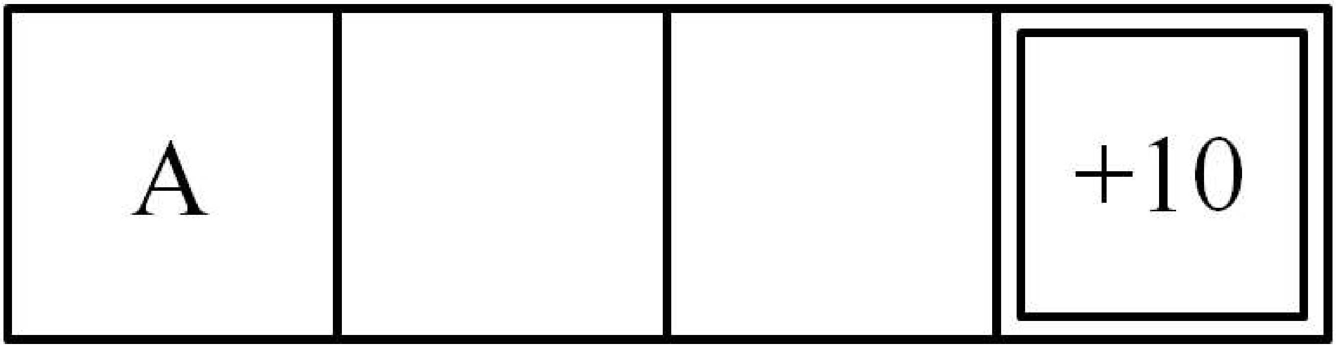
Part 3

Let's kick it up a notch! The $Left$ and $Right$ movement actions are now stochastic and fail with probability f. When an action fails, the agent moves Up or $Down$ with probability $f/2$ each. When there is no square to move Up or $Down$ into (as in the one-dimensional case), the agent stays in place. The $Exit$ action does not fail.

Part 3.1

1/1 point (ungraded)

For the following mini-grid, the failure probability is $f = 0.5$. The discount is back to $\gamma = 1$.



What is the optimal value $V^*(A)$?

10

Submit

Correct (1/1 point)

Part 3.2

1/1 point (ungraded)

When running value iteration, what is the smallest value of k for which $V_k(A)$ will be non-zero?

4

Submit

Correct (1/1 point)

Part 3.3

1/1 point (ungraded)

What will $V_k(A)$ be when it is first non-zero?

1.25

Submit

Correct (1/1 point)

Part 3.4

1/1 point (ungraded)

After how many iterations k will we have $V_k(A) = V^*(A)$?

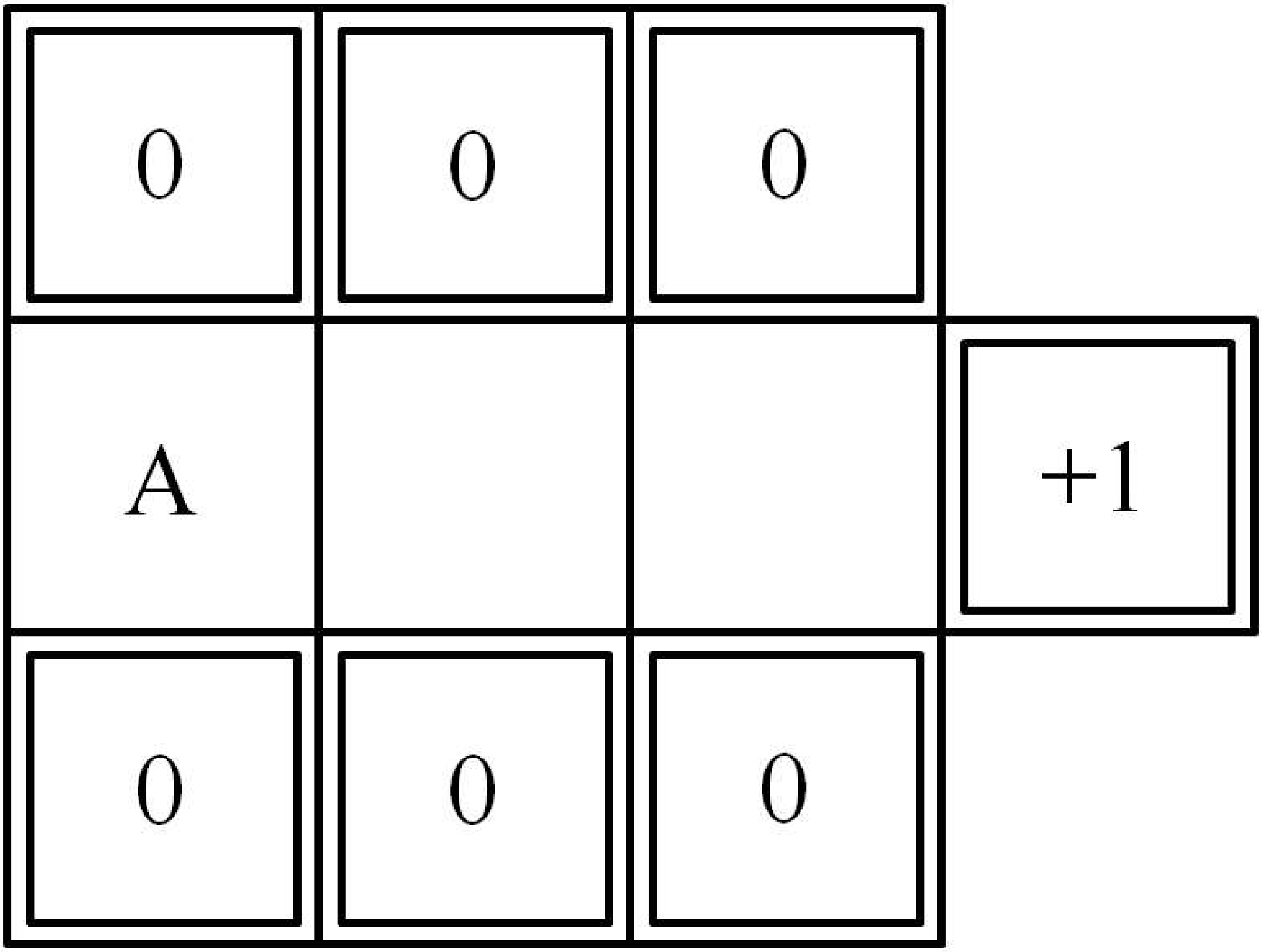
- 2
- 3
- 4
- 5
- 6
- They will never become equal for any finite value of k.

Submit

Correct (1/1 point)

Part 4

Now consider the following mini-grid. Again, the failure probability is $f = 0.5$ and $\gamma = 1$. Remember that failure results in a shift Up or $Down$, and that the only action available from the double-walled exit states is $Exit$.



Part 4.1

1/1 point (ungraded)

What is the optimal value $V^*(A)$?

0.125

Submit

Correct (1/1 point)

Part 4.2

1/1 point (ungraded)

When running value iteration, what is the smallest value of k for which $V_k(A)$ will be non-zero?

4 ✔

Submit

✔ Correct (1/1 point)

Part 4.3

1/1 point (ungraded)

What will $V_k(A)$ be when it is first non-zero?

0.125 ✔

Submit

✔ Correct (1/1 point)

Part 4.4

1/1 point (ungraded)

After how many iterations k will we have $V_k(A) = V^*(A)$?

- ○ 2
- ○ 3
- ◉ 4 ✔
- ○ 5

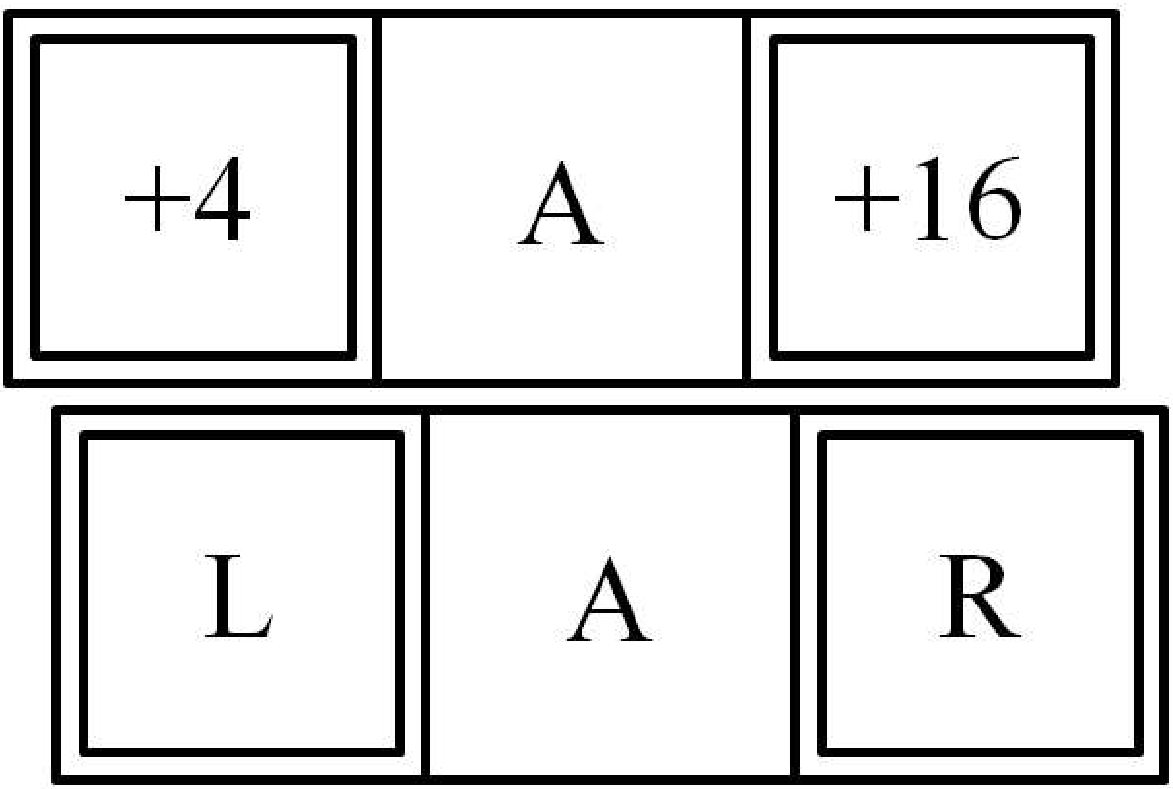
- 6
- They will never become equal for any finite value of k.

Submit

Correct (1/1 point)

Part 5

Finally, consider the following mini-grid (rewards shown on left, state names shown on right).



In this scenario, the discount is $\gamma = 1$. The failure probability is actually $f = 0$, but, now we do not actually know the details of the MDP, so we use reinforcement learning to compute various values. We observe the following transition sequence (recall that state X is the end-of-game absorbing state):

s	a	s'	r
A	$Right$	R	0
R	$Exit$	X	16
A	$Left$	L	0
L	$Exit$	X	4
A	$Right$	R	0
R	$Exit$	X	16
A	$Left$	L	0
L	$Exit$	X	4

Part 5.1

2/2 points (ungraded)

After this sequence of transitions, if we use a learning rate of $\alpha = 0.5$, what would temporal difference learning learn for the value of A? Remember that $V(s)$ is intialized with 0 for all s.

3

Submit

Correct (2/2 points)

Part 5.2

2/2 points (ungraded)

If these transitions repeated many times and learning rates were appropriately small for convergence, what would temporal difference learning converge to for the value of A?

10

Submit

✔ Correct (2/2 points)

Part 5.3

2/2 points (ungraded)

After this sequence of transitions, if we use a learning rate of $\alpha = 0.5$, what would Q-learning learn for the Q-value of (A, $Right$)? Remember that $Q(s, a)$ is initialized with 0 for all (s, a).

4 ✔

Submit

✔ Correct (2/2 points)

Part 5.4

2/2 points (ungraded)

If these transitions repeated many times and learning rates were appropriately small for convergence, what would Q-learning converge to for the Q-value of (A, $Right$)?

16 ✔

Submit

✔ Correct (2/2 points)