# Q9: MDPs and Reinforcement Learning: Mini-Grids

## Problem 9: MDPs and Reinforcement Learning: Mini-Grids

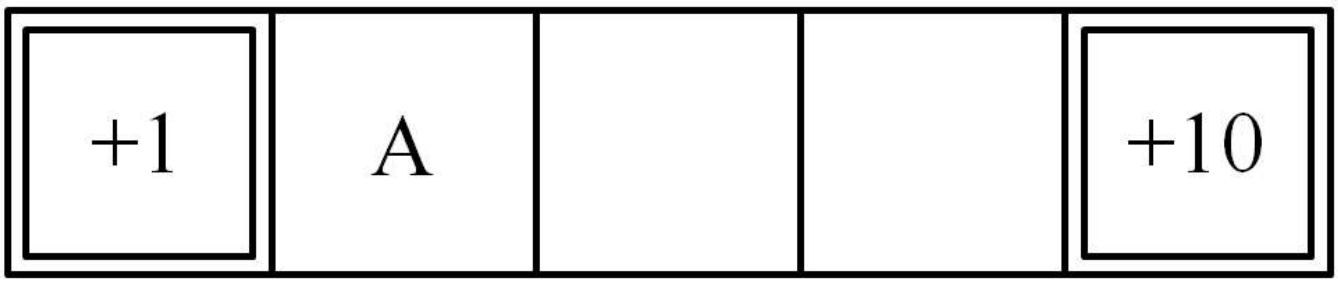
The following problems take place in various scenarios of the gridworld MDP (as in Project 3). In all cases, $\boldsymbol{A}$ is the start state and double-rectangle states are exit states. From an exit state, the only action available is $\boldsymbol{Exit}$, which results in the listed reward and ends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

From non-exit states, the agent can choose either $\boldsymbol{Left}$ or $\boldsymbol{Right}$ 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 $V_0(\boldsymbol{s}) = 0$ for all states $\boldsymbol{s}$. 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

0.0/1.0 point (ungraded)

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

10

**Answer:** 10

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

## Part 1.2

0.0/1.0 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

**Answer:** 2

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

## Part 1.3

0.0/1.0 point (ungraded)

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

1

**Answer:** 1

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

## Part 1.4

0.0/1.0 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 You have used 0 of 2 attempts

Answers are displayed within the problem

## Part 2

Now the situation is as before, but the discount $\gamma$ is less than $1$.

## Part 2.1

0.0/2.0 points (ungraded)

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

1.25 **Answer:** 1.25

Submit You have used 0 of 2 attempts

## Part 2.2

0.0/2.0 points (ungraded)

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

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

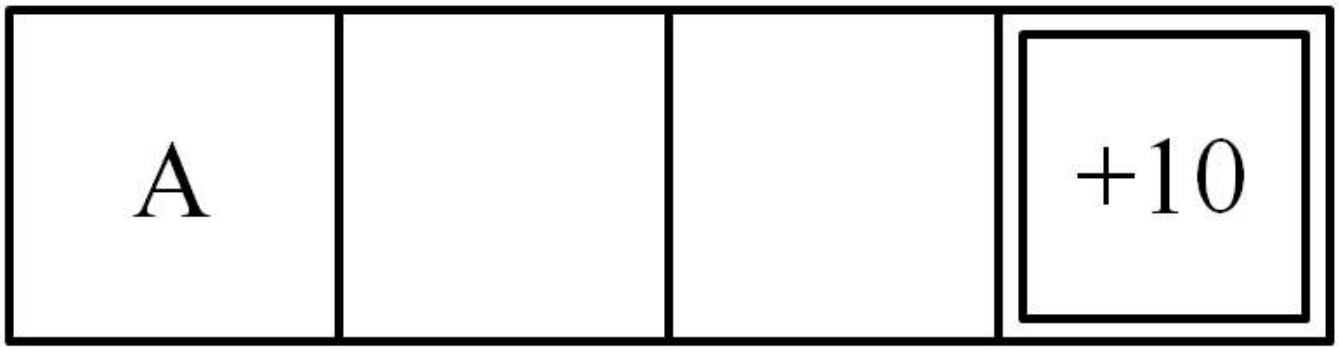
Submit You have used 0 of 2 attempts

## Part 3

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

## Part 3.1

0.0/1.0 point (ungraded)
For the following mini-grid, the failure probability is $\boldsymbol{f = 0.5}$. The discount is back to $\gamma = \mathbf{1}$.




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

10

**Answer:** 10

Submit

You have used 0 of 2 attempts

## Part 3.2

0.0/1.0 point (ungraded)

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

4

**Answer:** 4

Submit

You have used 0 of 2 attempts

## Part 3.3

0.0/1.0 point (ungraded)

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

1.25

**Answer:** 1.25

Submit

You have used 0 of 2 attempts

## Part 3.4

0.0/1.0 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 $k$. ✔

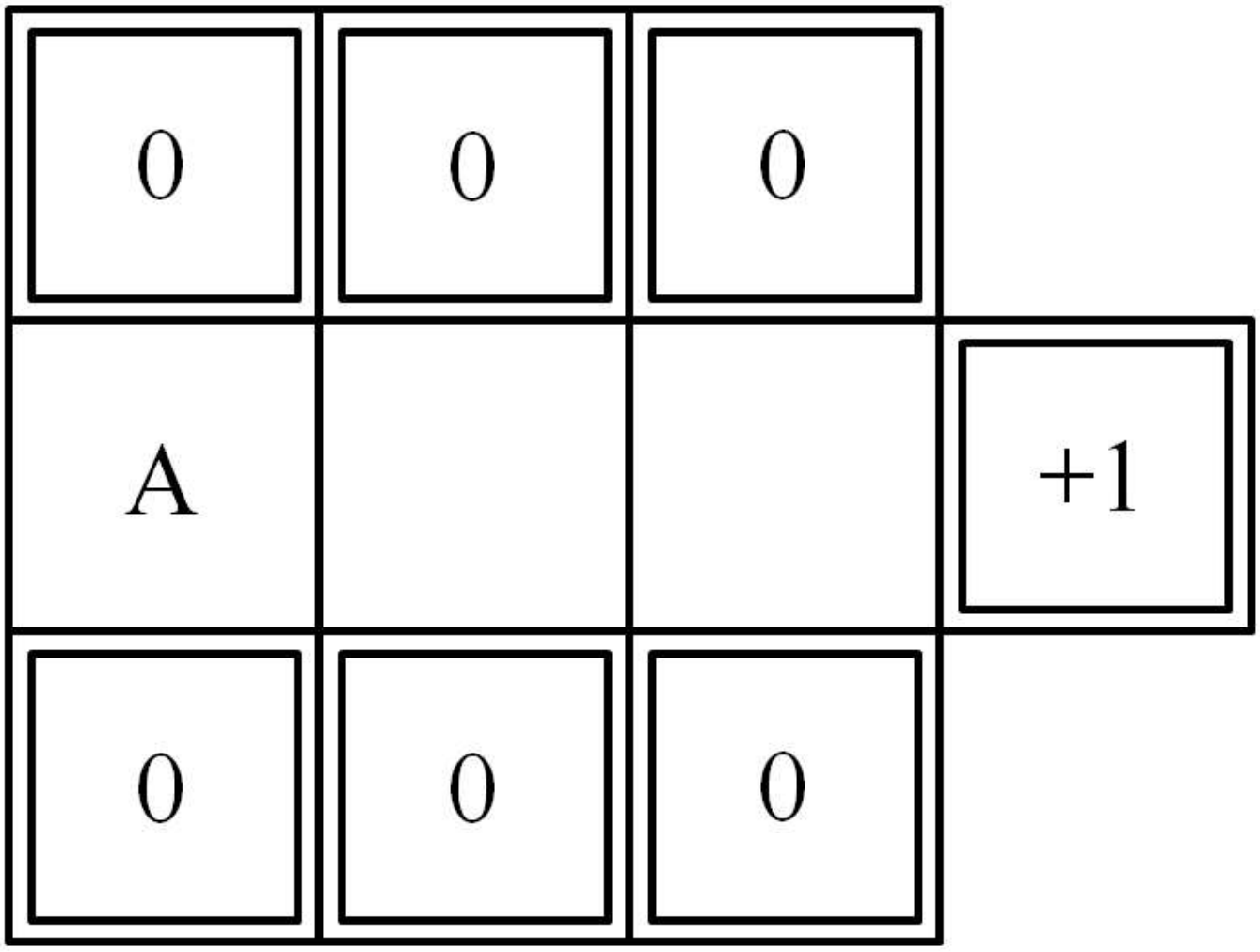
Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

## 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

0.0/1.0 point (ungraded)

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

0.125

**Answer:** 0.125

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

## Part 4.2

0.0/1.0 point (ungraded)

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

4

**Answer:** 4

Submit

You have used 0 of 2 attempts

## Part 4.3

0.0/1.0 point (ungraded)

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

0.125

**Answer:** 0.125

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

## Part 4.4

0.0/1.0 point (ungraded)

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

- 2
- 3
- 4 ✔
- 5
- 6

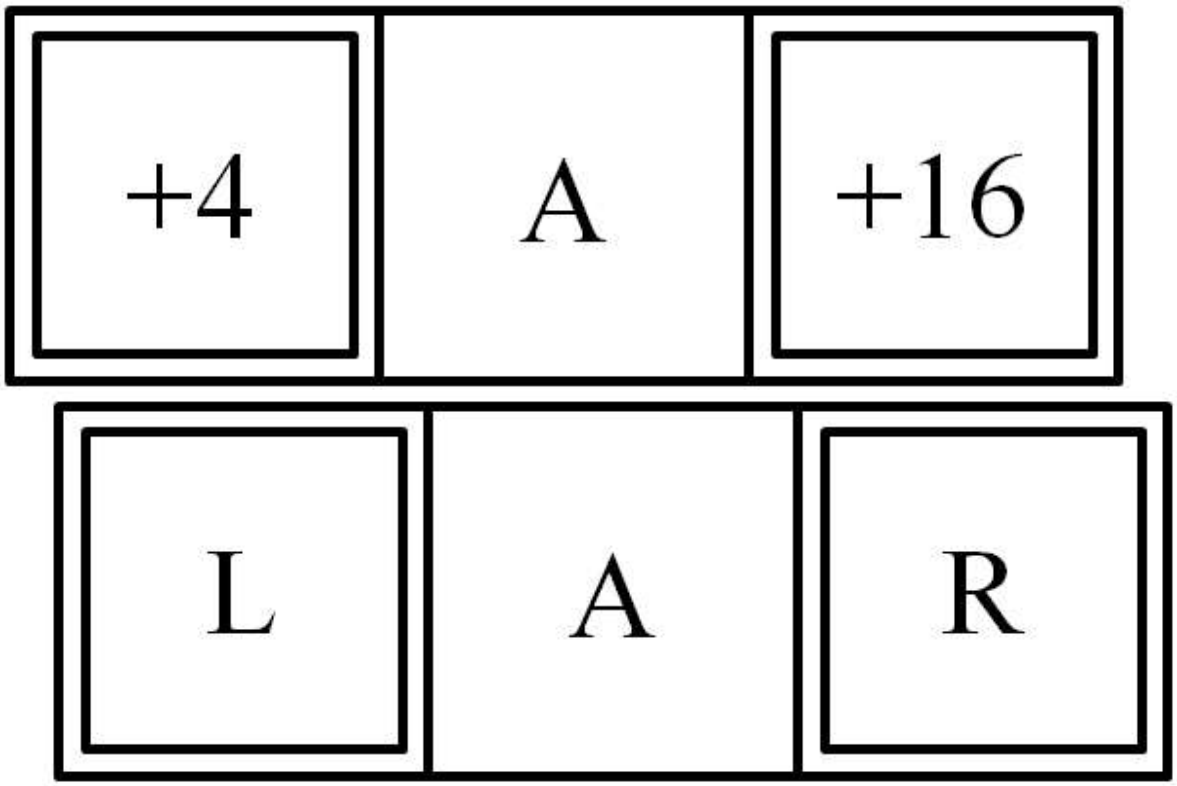
- They will never become equal for any finite value of $k$.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

## 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

0.0/2.0 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

**Answer:** 3

Submit You have used 0 of 2 attempts

## Part 5.2

0.0/2.0 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

**Answer:** 10

Submit You have used 0 of 2 attempts

## Part 5.3

0.0/2.0 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

**Answer:** 4

Submit

You have used 0 of 2 attempts

## Part 5.4

0.0/2.0 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

**Answer:** 16

Submit

You have used 0 of 2 attempts