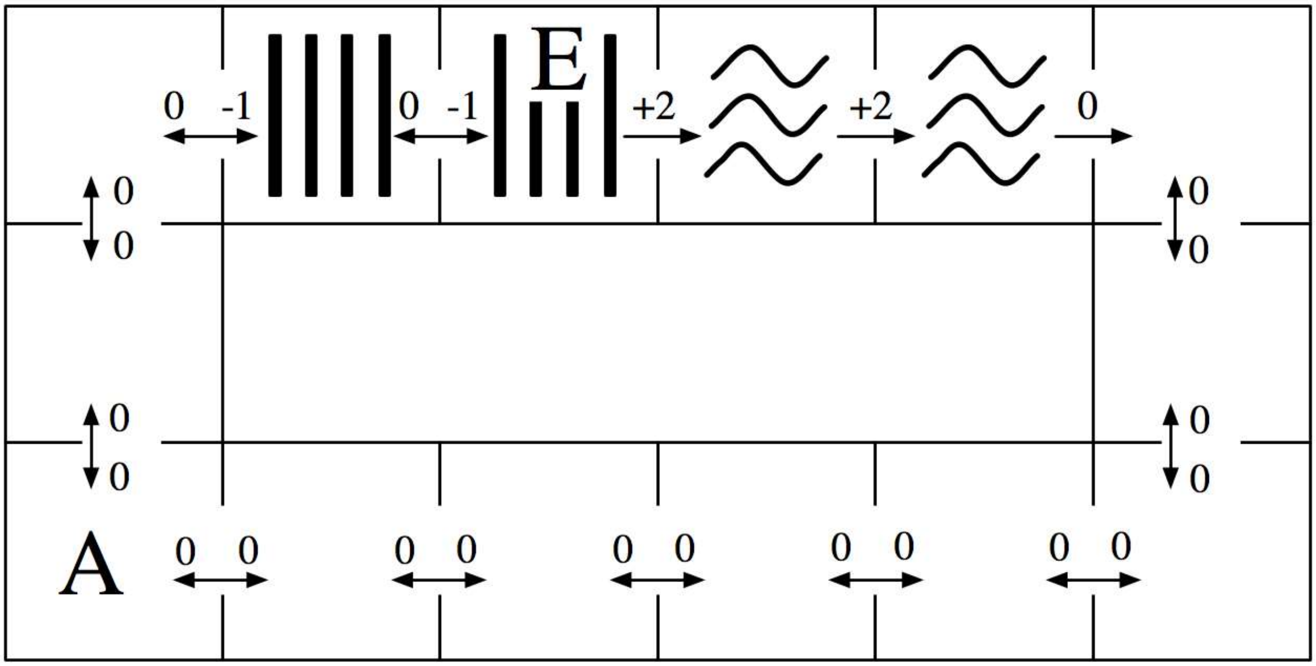
# Q5: MDPs: Grid-World Water Park

## Problem 5: MDPs: Grid-World Water Park

Consider the MDP drawn below. The state space consists of all squares in a grid-world water park. There is a single waterslide that is composed of two ladder squares and two slide squares (marked with vertical bars and squiggly lines respectively).

An agent in this water park can move from any square to any neighboring square, unless either the current square is a slide, in which case it must move forward one square along the slide, or the current square is just after the bottom of the slide, in which case the agent cannot go back into the slide square without going all the way back around the grid and up the ladder.

The actions are denoted by arrows between squares on the map and all deterministically move the agent in the given direction. The agent cannot stand still: it must move on each time step. Rewards are also shown below: the agent feels great pleasure as it slides down the water slide (+2), a certain amount of discomfort as it climbs the rungs of the ladder (-1), and receives rewards of 0 otherwise. The time horizon is infinite; this MDP goes on forever.



## Part 1

1/1 point (ungraded)

How many (deterministic) policies $\pi$ are possible for this MDP?

○ 22

○ 13

○ $11^2$

● $2^{11}$ ✓

○ $11$

Submit

✓ Correct (1/1 point)

## Part 2

6/6 points (ungraded)

Fill in the blank cells of this table with values that are correct for the corresponding function, discount, and state. If one of your values is $\infty$, enter "inf" (without the quotes) in the cell. For $-\infty$, enter "-inf". *Hint: You should not need to do substantial calculations here*.

| | $\gamma$ | $s = A$ | $s = E$ |
|---|---|---|---|
| $V_3^*(s)$ | 1.0 | 0 ✓ | 4 ✓ |
| $V_{10}^*(s)$ | 1.0 | 2 ✓ | 4 ✓ |
| $V_{10}^*(s)$ | 0.1 | 0 ✓ | 2.2 ✓ |
| $Q_1^*(s, \text{west})$ | 1.0 | — | 0 ✓ |
| $Q_{10}^*(s, \text{west})$ | 1.0 | — | 3 ✓ |
| $V^*(s)$ | 1.0 | inf ✓ | inf ✓ |
| $V^*(s)$ | 0.1 | 0 ✓ | 2.2 ✓ |

Submit

✓ Correct (6/6 points)

## Part 3

8/8 points (ungraded)

Fill in the blank cells of this table with the Q-values that result from applying the Q-update for the transition specified on each row. **Leave Q-values that are unaffected by the current update blank.** Use discount $\gamma = 1.0$ and learning rate $\alpha = 0.5$. Assume all Q-values are initialized to 0. (Note: the specified transitions would not arise from a single episode.)

| | $Q(D, \text{west})$ | $Q(D, \text{east})$ | $Q(E, \text{west})$ | $Q(E, \text{east})$ |
|---|---|---|---|---|

| Initial: | 0 | 0 | 0 | 0 |
|---|---|---|---|---|
| Transition 1:<br>$(s = D, a = \text{east}, r = -1, s' = E)$ | ✓ | -0.5<br>✓ | ✓ | ✓ |
| Transition 2:<br>$(s = E, a = \text{east}, r = +2, s' = F)$ | ✓ | ✓ | ✓ | 1<br>✓ |
| Transition 3:<br>$(s = E, a = \text{west}, r = 0, s' = D)$ | ✓ | ✓ | 0<br>✓ | ✓ |
| Transition 4:<br>$(s = D, a = \text{east}, r = -1, s' = E)$ | ✓ | -0.25<br>✓ | ✓ | ✓ |

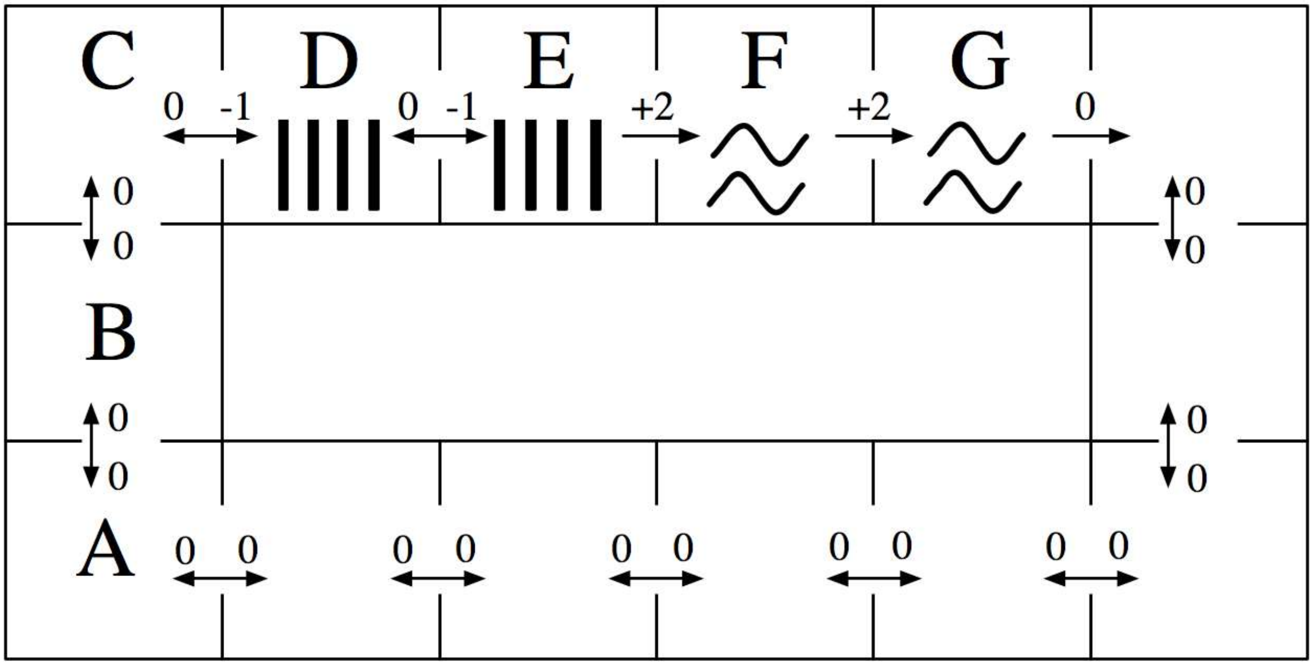
Submit

✓ Correct (8/8 points)

## Part 4

2/2 points (ungraded)

The agent is still at the water park MDP, but now we're going to use function approximation to represent Q-values. Recall that a policy $\pi$ is *greedy* with respect to a set of Q-values as long as $\forall a, s\ Q(s, \pi(s)) \geq Q(s, a)$ (so ties may be broken in any way).

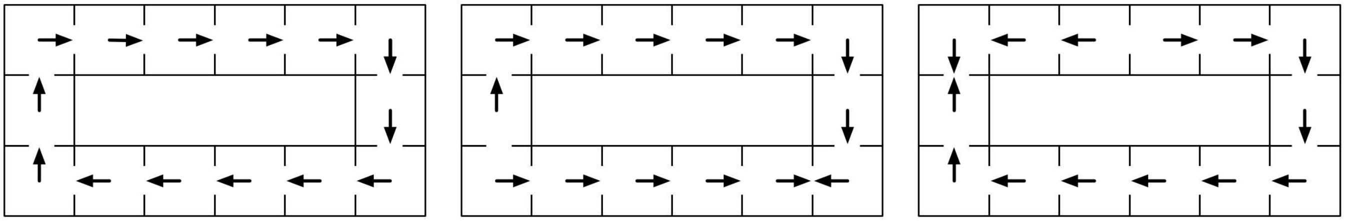


For the next subproblem, consider the following feature functions:

$f(s,a) = \begin{cases} 1 & \text{if } a = \text{east}, \\ 0 & \text{otherwise.} \end{cases}$ $f'(s,a) = \begin{cases} 1 & \text{if } (a = \text{east} \wedge \text{isSlide}(s)), \\ 0 & \text{otherwise.} \end{cases}$

(Note: isSlide($s$) is true if and only if the state $s$ is a slide square, i.e. either $F$ or $G$.)

Also consider the following policies:



$\pi_1$ $\pi_2$ $\pi_3$

Which are greedy policies with respect to the Q-value approximation function obtained by running the single Q-update for the transition $(s = F, a = \text{east}, r = +2, s' = G)$ while using the feature function $f$? You may assume that all feature weights are zero before the update. Use discount $\gamma = 1.0$ and learning rate $\alpha = 1.0$. Check all that apply.

☐ $\pi_1$

☑ $\pi_2$

☐ $\pi_3$

✓

Submit

✓ Correct (2/2 points)

## Part 5

2/2 points (ungraded)

Which are greedy policies with respect to the Q-value approximation function obtained by running the single Q-update for the transition $(s = F, a = \text{east}, r = +2, s' = G)$ while using the feature function $f'$? You may assume that all feature weights are zero before the update. Use discount $\gamma = 1.0$ and learning rate $\alpha = 1.0$. Check all that apply.

☑ $\pi_1$

☑ $\pi_2$

☑ $\pi_3$

✓

Submit

✓ Correct (2/2 points)