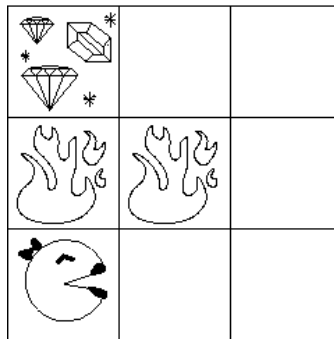
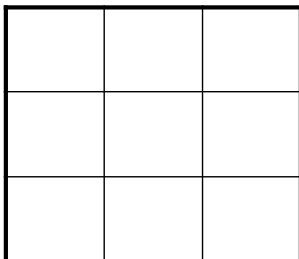


### Q1. MDPs: Treasure Hunting

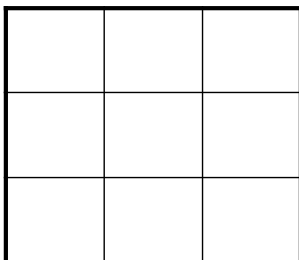
While Pacman is out collecting all the dots from `mediumClassic`, Ms. Pacman takes some time to go treasure hunting in the Gridworld island. Ever prepared, she has a map that shows where all the hazards are, and where the treasure is. From any unmarked square, Ms. Pacman can take the standard actions (N, S, E, W), but she is surefooted enough that her actions always succeed (i.e. there is no movement noise). If she lands in a hazard (H) square or a treasure (T) square, her only action is to call for an airlift (X), which takes her to the terminal ‘Done’ state, receiving a reward of -64 if she’s escaping a hazard, but +128 if she’s running off with the treasure. There is no “living reward.”



(a) What are the optimal values,  $V^*$  of each state in the above grid if  $\gamma = 0.5$ ?



(b) What’s the optimal policy?



Call this policy  $\pi_0$ . Ms. Pacman realizes that her map might be out of date, so she decides to do some Q-

learning to see what the island is really like. Because she thinks  $\pi_0$  is close to correct, she decides to Q-learn while following an  $\epsilon$ -random policy based on (b). Specifically, with probability  $\epsilon$  she chooses amongst the available actions uniformly at random. Otherwise, she does what  $\pi_0$  recommends. Call this policy  $\pi_\epsilon$ .

An  $\epsilon$ -random policy like  $\pi_\epsilon$  is an example of a *stochastic* policy, which assigns probabilities to actions rather than recommending a single one. A stochastic policy can be written as  $\pi(s, a)$ , the probability of taking action  $a$  when the agent is in state  $s$ .

(c) Write out a modified Bellman equation for policy evaluation when the policy  $\pi(s, a)$  is stochastic.

$$V^\pi(s) =$$

(d) If Ms. Pacman's map is correct what relationship will hold for all states?

1.  $V^{\pi_0} \geq V^{\pi_\epsilon}$
2.  $V^{\pi_0} = V^{\pi_\epsilon}$
3.  $V^{\pi_0} \leq V^{\pi_\epsilon}$

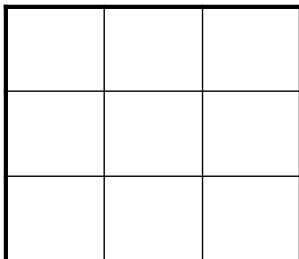
It turns out that Ms. Pacman's map is mostly correct, but some of the fire pits may have fizzled out and become regular squares! Thus, when she starts Q-learning, she observes the following episodes:

[ (0, 0), N, 0, (0, 1), N, 0, (0, 2), X, 128, Done ]  
[ (0, 0), N, 0, (0, 1), N, 0, (0, 2), X, 128, Done ]  
[ (0, 0), N, 0, (0, 1), E, 0, (1, 1), X, -64, Done ]

(e) What are Ms. Pacman's Q-values after observing these episodes? Assume that she initialized her Q-values all to 0 (you only have to write the Q-values that aren't 0) and used a learning rate of 1.0.

(f) In most cases, a learning rate of 1.0 will result in a failure to converge. Why is it safe for Ms. Pacman to use a learning rate of 1.0?

(g) Based on your knowledge about the structure of the maze and the episodes Ms. Pacman observed, what are the *true* optimal values of each state?



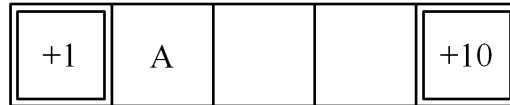
## Q2. MDPs: Mini-Grids

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  $s$ .

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).

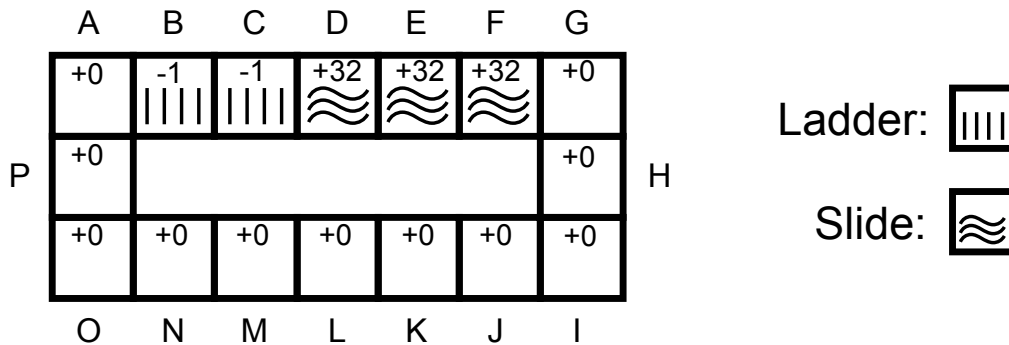


- (a) What is the optimal value  $V^*(A)$ ?
  
  
  
  
  
  
  
  
  
  
- (b) When running value iteration, remember that we start with  $V_0(s) = 0$  for all  $s$ . What is the first iteration  $k$  for which  $V_k(A)$  will be non-zero?
  
  
  
  
  
  
  
  
  
  
- (c) What will  $V_k(A)$  be when it is first non-zero?
  
  
  
  
  
  
  
  
  
  
- (d) After how many iterations  $k$  will we have  $V_k(A) = V^*(A)$ ? If they will never become equal, write *never*.

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

- (e) If  $\gamma = 0.5$ , what is the optimal value  $V^*(A)$ ?
  
  
  
  
  
  
  
  
  
  
- (f) For what range of values  $\gamma$  of the discount will it be optimal to go *Right* from  $A$ ? Remember that  $0 \leq \gamma \leq 1$ . Write *all* or *none* if all or no legal values of  $\gamma$  have this property.

### Q3. MDPs: Water Slide



Consider an MDP representing your experience at a water park, depicted by the figure above. Each state is labeled with a capital letter, A-P. The park has a single water slide which has a ladder that must be climbed (states B and C) before the slide can be ridden (states D, E, F). Apart from the slide states, you have three actions available in every state: stay where you are (Stay) or move to one of your two neighboring states (North, South, East, or West depending on the state's location). In the slides states (D, E, F) you only have one available action: East. Actions are deterministic: you always end up where you intend to.

Of course, you find it fun to ride the slide portion, but you hate exerting yourself while climbing the ladder. Walking around the rest of the park is a neutral experience. Specifically, you experience a reward of +32 when you enter a slide state, a reward of -1 upon entering a ladder state, and zero reward everywhere else. Note, you experience the reward of a state upon entering it (i.e.  $R(s, a, s') = R(s')$ ). Let  $\gamma$  be the future reward discount.

- (a) Suppose we run value iteration to convergence with  $\gamma = 0.5$ . Circle the action(s) from state A that are optimal under the calculated values.

South                      Stay                      East

- (b) Suppose, instead, we run value iteration to convergence with  $\gamma = 0.1$ . Circle the action(s) from state A that are optimal under these calculated values.

South                      Stay                      East

- (c) What happens to  $V^*(A)$  as  $\gamma \rightarrow 1$ ?

(d) Suppose that  $\gamma = 0.5$ . How many iterations of value iteration must be done before the calculated value of state A is positive? In other words, what is the minimum  $n$  such that  $V_n(A) > 0$ ?

(e) Suppose that  $\gamma = 0.5$ . What is the calculated value of A after 13 iterations of value iteration? In other words, what is  $V_{13}(A)$ ?

(f) Suppose that  $\gamma = 0.5$ . Let  $\pi$  be the policy that never leaves the current state unless that is the only available action. For example, under this policy, from state A you will always choose to remain in state A. What value for state A does policy evaluation converge to when you run it on this policy? In other words, what is  $V^\pi(A)$ ?