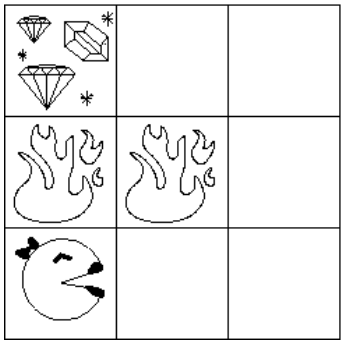
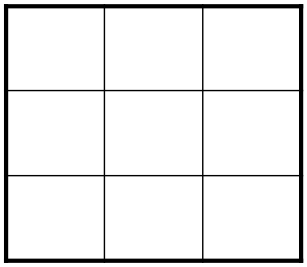
# CS188 Fall 2012 Section 6: MDPs and RL

# 1 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; this results in 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."

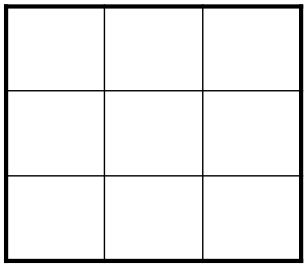


(a) What are the optimal values, $V^*$ of each state in the above grid if $\gamma = 0.5$?



(b) What are the Q-values for the last square on the second row (i.e., the one without fire)?

(c) 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 Pacman realizes that her map might be out of date, so she decides to do some Q-learning to see what the island is really like.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 $s$.

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

$V^\pi(s) =$

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

(i) $V^{\pi_0} \geq V^{\pi_\epsilon}$

(ii) $V^{\pi_0} = V^{\pi_\epsilon}$

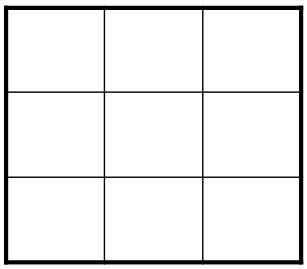
(iii) $V^{\pi_0} \leq V^{\pi_\epsilon}$

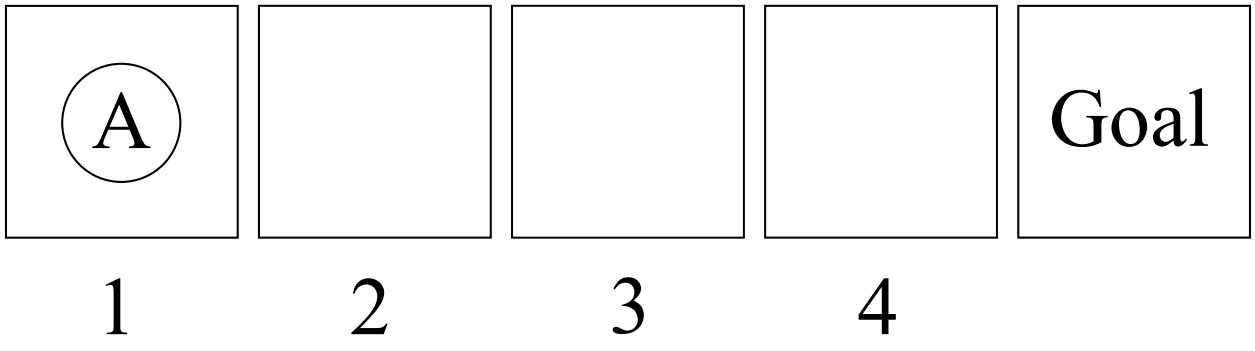
It turns out that Ms. Pacman's map is mostly correct, but some of the fire pits may have fizzled out and become regular squares! Thus, when she starts Q-learning, she observes the following episodes:
[ (0, 0), N, 0, (0, 1), N, 0, (0, 2), X, 128, Done ]
[ (0, 0), N, 0, (0, 1), N, 0, (0, 2), X, 128, Done ]
[ (0, 0), N, 0, (0, 1), E, 0, (1, 1), X, -64, Done ]

(f) 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.

(g) 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?

(h) 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?






# 2 Soccer

A soccer robot $A$ is on a fast break toward the goal, starting in position 1. From positions 1 through 3, it can either shoot (S) or dribble the ball forward (D). From 4 it can only shoot. If it shoots, it either scores a goal (state G) or misses (state M). If it dribbles, it either advances a square or loses the ball, ending up in $M$. When shooting, the robot is more likely to score a goal from states closer to the goal; when dribbling, the likelihood of missing is independent of the current state.

In this MDP, the states are 1,2,3,4,G and M, where G and M are terminal states. The transition model depends on the parameter $y$, which is the probability of dribbling success. Assume a discount of $\gamma = 1$.

$$\begin{aligned} T(k, S, G) &= \frac{k}{6} \\ T(k, S, M) &= 1 - \frac{k}{6} \\ T(k, D, k+1) &= y \quad \text{for } k \in \{1, 2, 3\} \\ T(k, D, M) &= 1 - y \quad \text{for } k \in \{1, 2, 3\} \\ R(k, S, G) &= 1 \end{aligned}$$

Rewards are 0 for all other transitions.

(a) What is $V^{\pi}(1)$ for the policy $\pi$ that always shoots?

(b) What is $Q^{*}(3, D)$ in terms of $y$?

(c) Using $y = \frac{3}{4}$, complete the first two iterations of value iteration.

| $i$ | $Q_i(1,S)$ | $Q_i(2,S)$ | $Q_i(3,S)$ | $Q_i(4,S)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 1 | | | | |
| 2 | | | | |

| $i$ | $Q_i(1,D)$ | $Q_i(2,D)$ | $Q_i(3,D)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | | | |
| 2 | | | |

| $i$ | $V_i^*(1)$ | $V_i^*(2)$ | $V_i^*(3)$ | $V_i^*(4)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 1 | | | | |
| 2 | | | | |

(d) After how many iterations will value iteration compute the optimal values for all states?

(e) For what range of values of $y$ is $Q^*(3,S) \geq Q^*(3,D)$?

(f) Now consider Q-learning, in which we do not have a model (T, y and k above), and instead learn from a series of experienced transitions. Using a learning rate of $\alpha = \frac{1}{2}$, execute Q-learning on these episodes:

- 1-D, 2-D, 3-D, 4-S, G
- 1-D, 2-D, 3-D, 4-S, M
- 1-D, 2-D, 3-S, G
- 1-D, 2-S, M
- 1-D, 2-D, 3-D, M

| $i$ | $Q_i(1,S)$ | $Q_i(2,S)$ | $Q_i(3,S)$ | $Q_i(4,S)$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 1 | | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |

| $i$ | $Q_i(1,D)$ | $Q_i(2,D)$ | $Q_i(3,D)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |