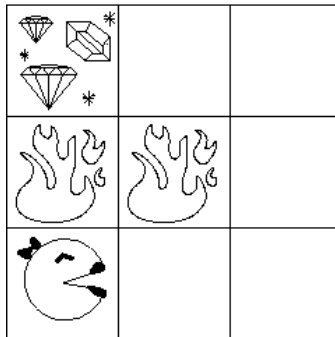


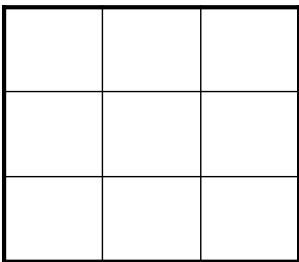
CS188 Fall 2010 Section 5: MDPs and RL

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."

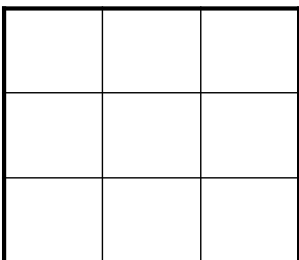


(a) What are the optimal values, V^* of each state in the above grid if $\gamma = 0.5$?



(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 π_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 Pacman realizes that her map might be out of date, so she decides to do some Q-learning to see what the island is really like.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 π_0 is close to correct, she decides to Q-learn while following an ϵ -random policy based on (b). Specifically, with probability ϵ she chooses amongst the available actions uniformly at random. Otherwise, she does what π_0 recommends. Call this policy π_ϵ .

An ϵ -random policy like π_ϵ 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?

