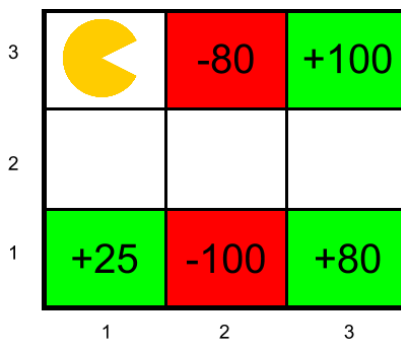


Exercises: Reinforcement learning

- Consider the grid-world given below and Pacman who is trying to learn the optimal policy. If an action results in landing into one of the shaded states the corresponding reward is awarded during that transition. All shaded states are terminal states, i.e., the MDP terminates once arrived in a shaded state. The other states have the *North*, *East*, *South*, *West* actions available, which deterministically move Pacman to the corresponding neighboring state (or have Pacman stay in place if the action tries to move out of the grid). Assume the discount factor $\gamma = 0.5$ and the Q -learning rate $\alpha = 0.5$ for all calculations. Pacman starts in state $(1, 3)$.



- (a) What are the utilities of the following states:

$$U(3, 2) =$$

$$U(2, 2) =$$

$$U(1, 3) =$$

- (b) The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing (s, a, s', r) .

Episode 1	Episode 2	Episode 3
$(1,3), S, (1,2), 0$	$(1,3), S, (1,2), 0$	$(1,3), S, (1,2), 0$
$(1,2), E, (2,2), 0$	$(1,2), E, (2,2), 0$	$(1,2), E, (2,2), 0$
$(2,2), S, (2,1), -100$	$(2,2), E, (3,2), 0$	$(2,2), E, (3,2), 0$
	$(3,2), N, (3,3), +100$	$(3,2), S, (3,1), +80$

Using Q -Learning updates, what are the following Q -values after the above three episodes:

$$Q((3, 2), N) = \qquad \qquad \qquad Q((1, 2), S) = \qquad \qquad \qquad Q((2, 2), E) =$$

(c) Consider a feature based representation of the Q -value function:

$$Q_f(s, a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$$

$f_1(s)$: The x coordinate of the state $f_2(s)$: The y coordinate of the state

$$f_3(N) = 1, \quad f_3(S) = 2, \quad f_3(E) = 3, \quad f_3(W) = 4$$

1. Given that all w_i are initially 0, what are their values after the first episode:

$$w_1 = \qquad \qquad \qquad w_2 = \qquad \qquad \qquad w_3 =$$

2. Assume the weight vector w is equal to $(1, 1, 1)$. What is the action prescribed by the Q -function in state $(2, 2)$?

2. **(Old exam question)** Consider the following MDP: We have infinitely many states $s \in \mathbb{Z}$ and actions $a \in \mathbb{Z}$, each represented as an integer. Taking action a from state s deterministically leads to new state $s' = s + a$ and reward $r = s - a$.

For example, taking action 3 at state 1 results in new state $s' = 1 + 3 = 4$ and reward $r = 1 - 3 = -2$.

We perform approximate Q -Learning, with features and initialized weights defined below.

Feature	Initial weight
$f_1(s, a) = s$	$w_1 = 1$
$f_2(s, a) = -a^2$	$w_2 = 2$

(a) Write down $Q(s, a)$ in terms of w_1 , w_2 , s and a .

(b) Calculate $Q(1, 1)$.

(c) We observe a sample (s, a, r, s') of $(1, 1, 0, 2)$. Assuming a learning rate of $\alpha = 0.5$ and discount factor of $\gamma = 0.5$, compute new weights after a single update of approximate Q -Learning.

(d) Compute the new value for $Q(1, 1)$.

3. **(Old exam question)** Consider a board game where the agent - Pacman moves on a virtual board with the goal to accumulate as many points by eating dots (food). Some ghosts move around the maze unsystematically. Meeting with the ghost in the same square is fatal for Pacman: he loses a life and when all lives have been lost the game is over. Suppose only a part of the maze, as shown in the figures below, is relevant for the current state and possible successor states.

The agent is using feature-based representation to estimate the $Q(s, a)$ value of taking an action a in a state s . The features the agent uses are f_0 and f_1 defined as the inverse of $1 +$ the Manhattan distance to closest food and ghost, respectively.

The four possible successor states from a given state $s = A$ are shown in the figure below, together with their feature representation vectors $f(s, a) = [f_0(s, a), f_1(s, a)]$. For example, $f(A, STOP) = [\frac{1}{4}, \frac{1}{4}]$.

State	$a = STOP$	$a = RIGHT$	$a = LEFT$	$a = DOWN$
A				
$f(s, a)$	$\begin{bmatrix} \frac{1}{4} & \frac{1}{4} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{3} & \frac{1}{5} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{5} & \frac{1}{3} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{3} & \frac{1}{3} \end{bmatrix}$

- (a) The agent picks the action according to

$$\arg \max_a Q(s, a) = \arg \max_a \{w_0 f_0(s, a) + w_1 f_1(s, a)\},$$

where the features $f_i(s, a)$ are as defined above, and w_i are weights, with $i = 0, 1$. Using the weight vector $w = [0.2, 0.5]$, which action, of the ones shown above, would the agent take from state A ? Circle the right answer(s) and show below the corresponding calculation.

- A. STOP B. RIGHT C. LEFT D. DOWN

- (b) Suppose now the situation depicted in the figure below. With the same weights as in (a), the agent goes down aiming to eat the dot, but the ghost moves at the same time to the right. Pacman loses a life, which costs him 100 points. How will the agent adapt its weights based on this unhappy experience? Calculate the resulting values of w_0 and w_1 if the learning rate is 0.0045.

