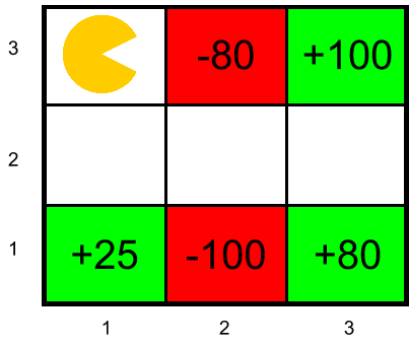
**ARTIFICIAL INTELLIGENCE (E016350A)**
GHENT UNIVERSITY
AY 2024/2025
Aleksandra Pizurica
Asst: Yoann ARHANT
E-mail: ai@lists.ugent.be

# Solutions: Reinforcement learning

1. 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 grad). 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:

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

(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 $\qquad$ $f_2(s)$ : 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:

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

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

**Solution**:

(a) The utility of a state is the expected reward for the next transition plus the discounted utility of the next state, assuming that the agent chooses optimal action. Or briefly: the state utilities are the optimal values for the states found by computing the expected reward for the agent acting optimally from that state onwards. Note that you get a reward when you transition into the shaded states and not out of them. So for example the optimal path starting from $(2,2)$ is to go to the $+100$ square which has a discounted reward of $0 + \gamma \cdot 100 = 50$. For $(1,3)$, going to either of $+25$ or $+100$ has the same discounted reward of 12.5.
$U(3,2) = 100$ $\qquad$ $U(2,2) = 50$ $\qquad$ $U(1,3) = 12.5$

(b) $Q$-values obtained by $Q$-learning updates:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + \gamma \max_{a'} Q(s',a')).$$

- **Episode 1:**

$$Q((1,3),S) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$
$$Q((1,2),E) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$
$$Q((2,2),S) = 0.5 \cdot 0 + 0.5(-100 + 0.5 \cdot 0) = -50$$

Let us explain how we obtain those values. At the beginning the values that are stored inside of the Q-table, are all set to zero. After each episode we update the Q-table and proceed to the next episode. For example:

$$Q((1,3),S) = 0.5 \cdot Q((1,3),S) + 0.5 \left(0 + 0.5 \cdot \max_a Q((1,2),a)\right)$$
$$= 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$

**Solution**:

(b)

- (Continued) **Episode 2:**

$$Q((1,3),S) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$
$$Q((1,2),E) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$
$$Q((2,2),E) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$
$$Q((3,2),N) = 0.5 \cdot 0 + 0.5(100 + 0.5 \cdot 0) = 50$$

- **Episode 3:**

$$Q((1,3),S) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$
$$Q((1,2),E) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 0) = 0$$
$$Q((2,2),E) = 0.5 \cdot 0 + 0.5(0 + 0.5 \cdot 50) = 12.5$$
$$Q((3,2),S) = 0.5 \cdot 0 + 0.5(80 + 0.5 \cdot 0) = 40$$

For clarity, let us show more detailed calculations for $Q((2,2),E)$ and $Q((3,2),S)$. After the previous two episodes we know that all the current Q-values are zero, apart from $Q((2,2),S) = -50$ and $Q((3,2),N) = 50$. Then we obtain that

$$\begin{aligned} Q((2,2),E) &= 0.5 \cdot Q((2,2),E) + 0.5\left(0 + 0.5 \cdot \max_a Q((3,2),a)\right) \\ &= 0.5 \cdot 0 + 0.5\left(0 + 0.5 \cdot Q((3,2),N)\right)) = 0.25 \cdot 50 = 12.5 \end{aligned}$$

Similarly, we obtain:

$$\begin{aligned} Q((3,2),S) &= 0.5 \cdot Q((3,2),S) + 0.5\left(80 + 0.5 \cdot \max_a Q((3,1),a)\right) \\ &= 0.5 \cdot 0 + 0.5\left(80 + 0.5 \cdot 0\right) = 0.5 \cdot 80 = 40. \end{aligned}$$

Now we can fill in the following values:

$Q((3,2),N) = 50$ $\qquad$ $Q((1,2),S) = 0$ $\qquad$ $Q((2,2),E) = 12.5$

(c) 1. Using the approximate $Q$-learning weight updates:

$$w_i \leftarrow w_i + \alpha[(R(s,a,s') + \gamma \max_{a'} Q(s',a')) - Q(s,a)]f_i(s,a).$$

The only time the reward is non zero in the first episode is when it transitions into the $-100$ state.

$w_1 = -100$ $\qquad$ $w_2 = -100$ $\qquad$ $w_3 = -100$

**Solution**:

(b) 1. (Continued) $West$.
Indeed, here we have $Q_f(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$, and since we have that $w_1 = w_2 = w_3 = 1$, then

$$\begin{aligned} Q_f((2,2),a) &= f_1((2,2)) + f_2((2,2)) + f_3(a) = 2 + 2 + f_3(a) \\ &= 4 + \begin{cases} 1, & \text{north} \\ 2, & \text{south} \\ 3, & \text{east} \\ 4, & \text{west} \end{cases} = \begin{cases} 5, & \text{north} \\ 6, & \text{south} \\ 7, & \text{east} \\ 8, & \text{west} \end{cases} \end{aligned}$$

The action prescribed at $(2,2)$ is $\max_a Q((2,2),a)$ where $Q(s,a)$ is computed using the feature representation. In this case, the $Q$-value for $West$ is maximum $(2 + 2 + 4 = 8)$.

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 $Q(1, 1)$.

**Solution**:

(a) $Q(s, a) = w_1 \cdot s + w_2 \cdot (-a^2)$

(b) $Q(1, 1) = w_1 \cdot f_1(1, 1) + w_2 \cdot f_2(1, 1) = 1 \cdot 1 + 2 \cdot (-1^2) = -1$

(c) First we obtain the difference

$$
\begin{aligned}
Difference &= (0 + 0.5 \cdot \max_a Q(2, a)) - (-1) \\
Difference &= (0.5 \cdot \max_a (2 - 2 \cdot a^2)) + 1 \\
Difference &= (0.5 \cdot (2 + 2 \cdot \max_a (-a^2))) + 1 \\
Difference &= (0.5 \cdot (2 + 2 \cdot 0)) + 1 \\
Difference &= 1 + 1 = 2
\end{aligned}
$$

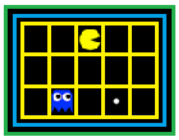
From here: $w_1 = 1 + 0.5 \cdot 2 \cdot 1 = 2$ and $w_2 = 2 + 0.5 \cdot 2 \cdot (-1^2) = 1$.

(d) $Q(1, 1) = w_1 \cdot 1 + w_2 \cdot (-1^2) = 2 \cdot 1 + 1 \cdot (-1) = 1$

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 looses 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, $f(A, STOP) = \left[\frac{1}{4}, \frac{1}{4}\right]$.

| State | $a =$ STOP | $a =$ RIGHT | $a =$ LEFT | $a =$ DOWN |
|---|---|---|---|---|
| $A$ |  |  |  |  |
| $f(s,a)$ | $\left[\frac{1}{4}, \frac{1}{4}\right]$ | $\left[\frac{1}{3}, \frac{1}{5}\right]$ | $\left[\frac{1}{5}, \frac{1}{3}\right]$ | $\left[\frac{1}{3}, \frac{1}{3}\right]$ |

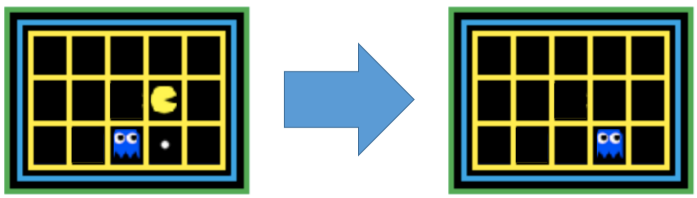
   (a) The agent picks the action according to

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

   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 looses 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.



**Solution**:

(a) Let us denote the actions $\{STOP, RIGHT, LEFT, DOWN\}$ by $\{t, r, l, d\}$, respectively. Then for the state $s = A$ and $w = [w_0, w_1] = [0.2, 0.5]$ we have the following

$$\begin{aligned}\arg\max_{a\in\{t,r,l,d\}} Q(s,a) &= \arg\max_{a\in\{t,r,l,d\}}\{0.2 f_0(s=A,a) + 0.5 f_1(s=A,a)\} \\ &= \arg\max_{a\in\{t,r,l,d\}} \left\{0.2\cdot\frac{1}{4} + 0.5\cdot\frac{1}{4}; 0.2\cdot\frac{1}{3} + 0.5\cdot\frac{1}{5}; 0.2\cdot\frac{1}{5} + 0.5\cdot\frac{1}{3}; 0.2\cdot\frac{1}{3} + 0.5\cdot\frac{1}{3}\right\} \\ &= \arg\max_{a\in\{t,r,l,d\}} \{0.175; 0.1666; 0.206; 0.233\} = 0.2333.\end{aligned}$$

As we can see $Q(s,a)$ for $s = A$ is maximal when the chosen action is $a =$ DOWN.

(b) The estimated Q-value of taking the action $a = DOWN$ from the state $s$ where the agent was before loosing its life was:

$$Q(s, DOWN) = 0.2\, f_0(s, DOWN) + 0.5\, f_1(s, DOWN) = 0.2\cdot 1 + 0.5\cdot\frac{1}{2} = 0.45.$$

However, the agent gets instead a reward or $r = -100$ and lands in a state from which any further state value is 0. Hence, the difference between the actual value of the new state and the estimated value is $Difference = -100 - 0.45 = -100.45$. The new weights are

$$\begin{aligned}w_0 &= 0.2 + \alpha\cdot Difference\cdot f_0(s, DOWN) = 0.2 + 0.0045\cdot(-100.45)\cdot 1.0 = -0.252 \\ w_1 &= 0.5 + \alpha\cdot Difference\cdot f_1(s, DOWN) = 0.5 + 0.0045\cdot(-100.45)\cdot 0.5 = 0.273\end{aligned}$$