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

Exercises: Markov decision processes

1. [Adapted from R&N, 3rd Ed] Sometimes MDPs are formulated with a reward function $R(s, a)$ that depends on the action taken or with a reward function $R(s, a, s')$ that also depends on the outcome state. In some formulations, rewards are assigned to the states alone: $R(s)$.

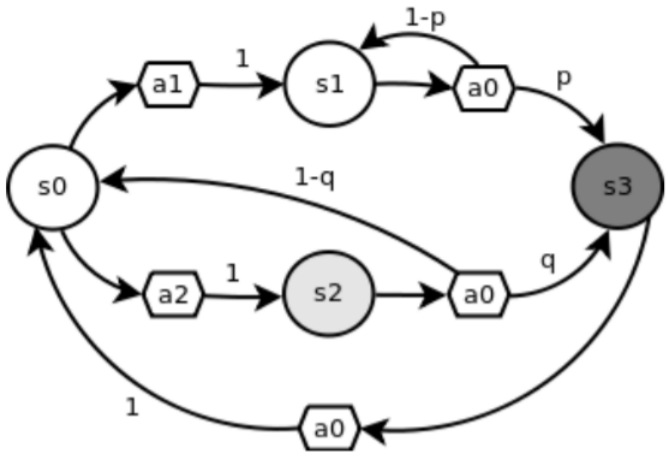
   (a) Write the Bellman equations for these formulations.

   (b) Explain compactly in words the meaning of the Bellman equation for the formulation with $R(s)$.

   (c) Show how an MDP with reward function $R(s, a, s')$ can be transformed into a different MDP with reward function $R(s, a)$, such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP.

   (d) Now do the same to convert MDPs with $R(s, a)$ into MDPs with $R(s)$.

2. Consider the infinite-horizon Markov decision process $\mathcal{M}$ represented by figure below with discount factor $\gamma \in [0, 1)$. States are represented by circles and actions by hexagons.



The numbers on the arrows from actions to states represent the transition probability, for instance, $T(s_2, a_0, s_3) = P(s_3 \mid s_2, a_0) = q$. Each of the parameters p and q are in the interval $[0, 1]$. The reward is 10 for state s_3, 1 for state s_2 and 0 otherwise.

(a) List all the possible policies for $\mathcal{M}$.

(b) Show the equation representing the optimal value function for each state of $\mathcal{M}$, i.e., $U^*(s_0)$, $U^*(s_1)$, $U^*(s_2)$ and $U^*(s_3)$.

(c) Is there a value for p such that for all $\gamma \in [0, 1)$ and $q \in [0, 1]$, $\pi^*(s_0) = a_2$? Explain.

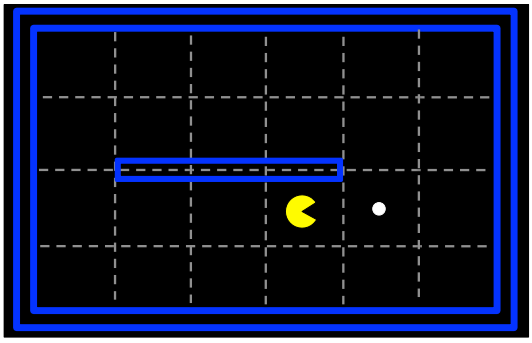
(d) Is there a value for q such that for all $\gamma \in [0, 1)$ and $p > 0$, $\pi^*(s_0) = a_1$? Explain.

(e) Using $p = q = 0.25$ and $\gamma = 0.9$, compute π^* and U^* for all states of $\mathcal{M}$.

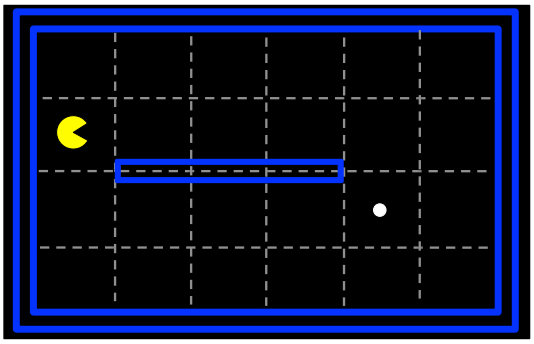
3. Let us consider a Pacman agent in a deterministic environment. A goal state is reached when there are no remaining food pellets on the board. Pacman's available actions are $\{N, S, E, W\}$, but Pacman **can not** move into a wall. Whenever Pacman eats a food pellet he receives a reward of $+1$.

   Assume that pacman eats a food pellet as soon as he occupies the location of the food pellet - i.e., the reward is received for the transition into the square with the food pellet.

   Consider the particular Pacman board states shown below. Throughout this problem assume that $U_0(s) = 0$ for all states s. Let the discount factor be $\gamma = 1$.



State A



State B

(a) What is the optimal value of state A, $U^*(A)$?

(b) What is the optimal value of state B, $U^*(B)$?

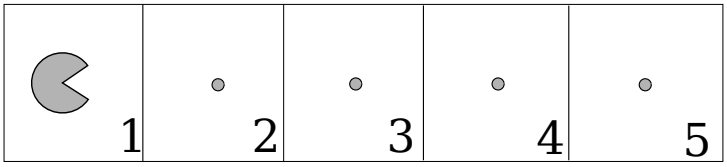
(c) At what iteration, k, will $U_k(B)$ first be non-zero?

(d) How do the optimal q-state values of moving W and E from state A compare? (*choose one*)

◯ $Q^*(A, W) > Q^*(A, E)$ ◯ $Q^*(A, W) < Q^*(A, E)$ ◯ $Q^*(A, W) = Q^*(A, E)$

(e) If we use this MDP formulation, is the policy found guaranteed to produce the shortest path from pacman's starting position to the food pellet? If not, how could you modify the MDP formulation to guarantee that the optimal policy found will produce the shortest path from pacman's starting position to the food pellet?

4. Pacman is in a bonus level! With no ghosts around, he can eat as many dots as he wants. He is in the 5×1 grid shown. The cells are numbered from left to right as $1, \ldots, 5$. In cells 1 through 4, the actions available to him are to move *Right* (R) or to *Fly* (F) out of the bonus level. The action *Right* deterministically lands Pacman in the cell to the right (and he eats the dot there), while the *Fly* action deterministically lands him in a terminal state and ends the game. From cell 5, *Fly* is the only action. Eating a dot gives a reward of 10, while flying out gives a reward of 20. Pacman starts in the leftmost cell (cell 1).



We write this as an MDP where the state is the cell that Pacman is in. The discount is γ.

Consider the following 3 policies:

$$\pi_0(s) = F \text{ for all } s$$
$$\pi_1(s) = R \text{ if } s \leq 3,\ F \text{ otherwise}$$
$$\pi_2(s) = R \text{ if } s \leq 4,\ F \text{ otherwise}$$

(a) Assume $\gamma = 1.0$. Calculate:

1. $U^{\pi_0}(1) =$ 2. $U^{\pi_1}(1) =$ 3. $U^{\pi_2}(1) =$ 4. $U^*(1) =$

(b) Now consider an arbitrary value for γ.

1. Does there exist a value for γ such that π_0 is strictly better than both π_1 and π_2? If yes, give a value for γ. If no, write *none*.

2. Does there exist a value for γ such that π_1 is strictly better than both π_0 and π_2? If yes, give a value for γ. If no, write *none*.

3. Does there exist a value for γ such that π_2 is strictly better than both π_0 and π_1? If yes, give a value for γ. If no, write *none*.

5. It is useful to know that partially observable MDP (POMDP) can be solved by defining an *observable* MDP on *belief states*. Furthermore, it can be shown that an optimal policy for this MDP $\pi^*(b)$ is also an optimal policy for the original MDP. In other words, solving a POMDP on a physical state space can be reduced to solving an MDP on the corresponding belief-state space. To specify the needed MDP on the belief space, we need to define the corresponding transition model and a reward function.

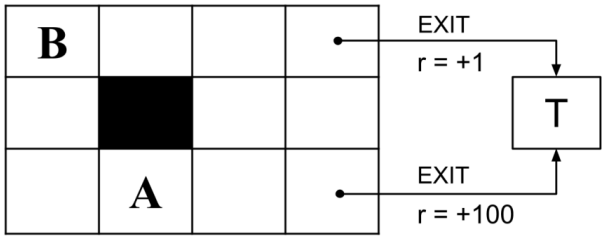
(a) How can we calculate the transition model for the belief states $P(b' \mid b, a)$ from the available information, which includes the transition model on the physical states and the the sensor model? Write the concrete expression.

(b) Write the expression for the expected reward if the agent does a in belief state b.

6. Let us consider the following problem which takes place in various instances of a grid world MDP. Shaded cells represent walls. In all states, the agent has available actions $\{\uparrow, \downarrow, \rightarrow, \leftarrow\}$. Performing an action that would transition to an invalid state (outside the grid or into a wall) results in the agent remaining in its original state. In states with an arrow coming out, the agent has an additional action $EXIT$. In the event that the $EXIT$ action is taken, the agent receives the labeled reward and ends the game in the terminal state T. Unless otherwise stated, all other transitions receive no reward, and all transitions are deterministic.

   For all parts of the problem, assume that value iteration begins with all states initialized to zero, i.e., $U_0(s) = 0$, for every s. Let the discount factor be $\gamma = \frac{1}{2}$.

   (a) Suppose that we are performing value iteration on the grid world MDP below.



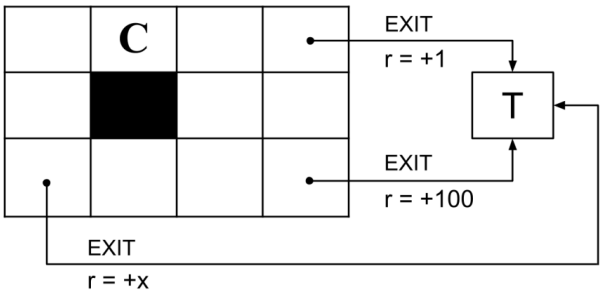
   (i) Find in the optimal values for A and B, i.e., determine $U^*(A)$ and $U^*(B)$.

   (ii) After how many iterations k will we have $U_k(s) = U^*(s)$ for all states s? If it never occurs, write "never". Explain your reasoning.

   (iii) Suppose that we wanted to re-design the original reward function $R(s, a, s')$. For which of the following new reward functions would the optimal policy remain unchanged?

   ◯ $R_1(s, a, s') = 10\, R(s, a, s')$
   ◯ $R_2(s, a, s') = 1 + R(s, a, s')$
   ◯ $R_3(s, a, s') = (R(s, a, s'))^2$
   ◯ $R_4(s, a, s') = -1$
   ◯ None

   (b) For the following problem, we add a new state in which we can take the $EXIT$ action with a reward of $+x$.



   (i) For what values of x is it guaranteed that our optimal policy π^* has $\pi^*(C) = \leftarrow$? Write $+\infty$ and $-\infty$ if there is no upper or lower bound, respectively. Write the upper and lower bounds on the following lines:

   _____ $< x <$ _____

(ii) For what values of x does value iteration take the minimum number of iterations k to converge to U^* for all states? Write $+\infty$ and $-\infty$ if there is no upper or lower bound, respectively.Write the upper and lower bounds on the following lines:

$$\_\_\_\_\_ \leq x \leq \_\_\_\_\_$$

(iii) What is the minimum number of iterations k until U_k has converged to U^* for all states.