

June 11, 2022

3 Exercises

3.1 Question

Devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. Make the three examples as different from each other as possible. The framework is abstract and flexible and can be applied in many different ways. Stretch its limits in some way in at least one of your examples

Answer

Stock Trading

An agent managing the given amount of cash in a stock. The agent buys and sells in order to maximize the cash it is given.

States: Cash amount, stock amount, buy/sell prices, market indicators, global indices (e.g. risk index)

Actions: Buy, Sell, Do nothing

Rewards: Cash amount deviation from the initial amount.

Landing an UAV

An agent that can land an UAV in any weather conditions. The agent locates the landing point, considers the distance and the altitude, and controls engine thrust and control planes to land the aircraft.

States: Distance to the landing point, altitude, azimuth, velocity, wind info, engine thrust etc.

Actions: Increase/Decrease elevation, change heading, change engine thrust, control landing gears.

Rewards: Positive reward upon landing, possibly increasing with proportionally to proximity to desired landing location. A very low reward upon crash.

Playing ATARI Games

An agent that can play atari games by looking at screen pixel images.

States: Screen pixels.

Actions: Game controls, e.g. left, right, shoot

Rewards: Positive reward upon winning. Negative reward upon losing.

3.2 Question

Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions?

Answer

Agent needs to learn the reward at the end of the episode. If the environment does not provide a reward or does provide a vector of rewards the agent cannot learn.

3.3 Question

Consider the problem of driving. You could define the actions in terms of the accelerator, steering wheel, and brake, that is, where your body meets the machine. Or you could define them farther out—say, where the rubber meets the road, considering your actions to be tire torques. Or you could define them farther in—say, where your brain meets your body, the actions being muscle twitches to control your limbs. Or you could go to a really high level and say that your actions are your choices of where to drive. What is the right level, the right place to draw the line between agent and environment? On what basis is one location of the line to be preferred over another? Is there any fundamental reason for preferring one location over another, or is it a free choice?

Answer

It depends on the problem we are trying to solve.

Car driving is designed to be a function of acceleration, steering wheel and the brake. This interface is already present in any car. Natural approach would be to exploit that interface.

Now let's consider using a lower level control and use brain signals. In order to drive the brain process and produces many signals. Much of the

processed signals are even may not always be related to the driving. Produced signals are do not directly control the car but the body. More work is needed to translate the signals. The brain level control introduces a lot of extra work.

Now let's move to a higher level and use choices of locations as actions. In this case it will be necessary to relate all the locations, road conditions e.g. pedestrians, lights, and car state e.g. turning, acceleration which produces an infinite MDP. It is very hard to learn in this setting.

3.4 Question

Give a table analogous to that in Example 3.3, but for $p(s', r | s, a)$. It should have columns for s, a, s', r , and $p(s', r | s, a)$, and a row for every 4-tuple for which $p(s', r | s, a) > 0$

Answer

Rewards are deterministic. $p(s', r | s, a)$ column will be similar to $p(s' = s, a)$ column. Answer will be similar to the table in example 3.3 excluding rows with 0 probability.

3.5 Question

The equations in Section 3.1 are for the continuing case and need to be modified (very slightly) to apply to episodic tasks. Show that you know the modifications needed by giving the modified version of (3.3).

Answer

$$\sum_{s' \in S^+} \sum_{r \in R} p(s', r | s, a) = 1 \text{ for all } s \in S, a \in A(s), S^+ \text{ being all states, } S \text{ being non-terminal states.} \quad (1)$$

3.6 Question

Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for 1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task?

Answer

Similar to the continuing case, the return will be related to $-K$.

The difference is that in continuing case the return will keep accumulating.

3.7 Question

Imagine that you are designing a robot to run a maze. You decide to give it a reward of +1 for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes—the successive runs through the maze—so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

Answer

The robot should be incentivized to exit the maze as soon as possible.

3.8 Question

Suppose $\gamma = 0.50$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are G_0 , G_1, \dots , G_5 ? Hint: Work backwards.

Answer

$$G_t = R_{t+1} + \gamma G_{t+1} \quad (2)$$

$$\begin{aligned} G_5 &= 0 \\ G_4 &= R_5 + \gamma G_5 = 2 + 0.5 * 0 = 2 \\ G_3 &= R_4 + \gamma G_4 = 3 + 0.5 * 2 = 4 \\ G_2 &= R_3 + \gamma G_3 = 6 + 0.5 * 4 = 8 \\ G_1 &= R_2 + \gamma G_2 = 2 + 0.5 * 8 = 6 \\ G_0 &= R_1 + \gamma G_1 = -1 + 0.5 * 6 = 2 \end{aligned}$$

3.9 Question

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0 ?

Answer

$$G_t = R_{t+1} + \gamma G_{t+1} \quad (3)$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma} \text{ given } \gamma < 1 \quad (4)$$

$$G_1 = 7 * \frac{1}{1-0.9} = 70$$
$$G_0 = R_1 + \gamma G_1 = 2 + 0.9 * 70 = 65$$

3.10 Question

Prove the second equality in (3.10).

Answer

Equation $\sum_{k=0}^{\infty} ar^k$ is sum of geometric series which equals to $\frac{a}{1-r}$ if $r < 1$.
In equation (3.10) $a = 1$ and $r = \gamma = 0.9$

3.11 Question

If the current state is S_t , and actions are selected according to stochastic policy π , then what is the expectation of R_{t+1} in terms of π and the four-argument function p (3.2)?

Answer

$$E_{\pi}[R_{t+1}|S_t = s] = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) * r \quad (5)$$

3.12 Question

Give an equation for V_{π} in terms of q_{π} and π .

Answer

$$v_{\pi}(s) = E_{\pi} \left[\sum_{a \in A(s)} \pi(a|s) q_{\pi}(a, s) \right] \quad (6)$$

3.13 Question

Give an equation for q_{π} in terms of v_{π} and the four-argument p .

Answer

$$q_{\pi}(s, a) = \sum_{s', r} p(s', r | s, a) [v_{\pi}(s') * \gamma + r] \quad (7)$$

3.14 Question

The Bellman equation (3.14) must hold for each state for the value function v_{π} shown in Figure 3.2 (right) of Example 3.5. Show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, 0.4, and +0.7. (These numbers are accurate only to one decimal place.)

Answer

$$v_{\pi}(s_{center}) = 0.25 * 0.9(2.7 + 0.3 + 0.4 - 0.4) = 0.675 \approx 0.7 \quad (8)$$

3.15 Question

In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, using (3.8), that adding a constant c to all the rewards adds a constant, $v c$, to the values of all states, and thus does not affect the relative values of any states under any policies. What is $v c$ in terms of c and γ ?

Answer

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (9)$$

Adding constant c :

$$G'_t = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \gamma^k c = G_t + \sum_{k=0}^{\infty} \gamma^k c \quad (10)$$

$$G'_t = G_t + \frac{c}{1 - \gamma} \quad (11)$$

Showing v' in terms of c and γ

$$v'(s) = E[G'_t | S_t = s] = E[G_t + \frac{c}{1-\gamma} | S_t = s] = E[G_t | S_t = s] + \frac{c}{1-\gamma} = v(s) + \frac{c}{1-\gamma} \quad (12)$$

3.16 Question

Now consider adding a constant c to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.

Answer

In an episodic task adding a constant may change the behaviour. If c is big enough it may assign a positive reward to taking time in the maze which makes the agent prefer not to exit the maze.

3.17 Question

What is the Bellman equation for action values, that is, for q_π ? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state-action pair (s, a) . Hint: the backup diagram to the right corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values.

Answer

$$q(s, a) = E_\pi[G_t | S_t = s, A_t = a] = E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$

$$q(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma \sum_{a' \in A(s')} \pi(a' | s') q(s', a')]$$

3.18 Question

The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action. Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s, a)$, given $S_t = s$.

Answer

$$v_{\pi}(s) = E_{\pi}[G_t | S_t = s] = E_{\pi}[q_{\pi}(s, a) | S_t = s, a \in A(s)]$$
$$v_{\pi}(s) = \sum_{a \in A(s)} \pi(s, a) q_{\pi}(s, a)$$

3.19 Question

The value of an action, $q_{\pi}(s, a)$, depends on the expected next reward and the expected sum of the remaining rewards. Again we can think of this in terms of a small backup diagram, this one rooted at an action (state-action pair) and branching to the possible next states. Give the equation corresponding to this intuition and diagram for the action value, $q_{\pi}(s, a)$, in terms of the expected next reward, R_{t+1} , and the expected next state value, $v_{\pi}(S_{t+1})$, given that $S_t = s$ and $A_t = a$.

Answer

$$q_{\pi}(s, a) = E[G_t | S_t = s, A_t = a] = E[R_{t+1} + \gamma v_{\pi}(s') | S_t = s, A_t = a, s' \in A(s)]$$
$$q_{\pi}(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma v_{\pi}(s')]$$

3.20 Question

Draw or describe the optimal state-value function for the golf example.

Answer

It is optimal to use putter inside the green region, elsewhere driver is the optimal club.

State value for the green region is -1; State value for the contour where green region can be reached in one shot using a driver is -2; State value for the contour where region with state value -2 can be reached in one shot using a driver is -3;

3.21 Question

Draw or describe the contours of the optimal action-value function for putting, $q^*(s, \text{putter})$, for the golf example.

Answer

Action value for the putter inside the green region is -1; Action value for the putter for the contour where green region can be reached in one shot

using a putter is -2; Action value for the putter for the outer contour 1 is -3 because after first putter two more shots are needed in any case. Action value for the putter for the outer contour 2 is -3 because after first putter using one driver and one putter may be optimal. As we go out regions will be larger than first two regions. Overall, since using putter outside green region is sub-optimal, action values will be lower compared to $q^*(s, \text{driver})$ illustrated in the book.

3.22 Question

Consider the continuing MDP shown on to the right. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, policy left and policy right. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

Answer

$$\begin{aligned}
 v_\pi(s_0) &= E_\pi[G_0] = E_\pi[R_1 + \gamma G_1] \\
 v_\pi(s_{top}) &= \sum_{k=0} \gamma^{2k} a + \sum_{k=0} \gamma^{2k} \gamma * b \\
 v_\pi(s_{top}) &= \frac{a}{1-\gamma^2} + \frac{\gamma * b}{1-\gamma^2} \\
 \text{let } d &= \frac{1}{1-\gamma^2} \\
 v_{left}(s_{top}) &= \frac{1}{1-\gamma^2} + \frac{\gamma * 0}{1-\gamma^2} = \frac{1}{1-\gamma^2} = d \\
 v_{right}(s_{top}) &= \frac{0}{1-\gamma^2} + \frac{\gamma * 2}{1-\gamma^2} = \frac{\gamma * 2}{1-\gamma^2} = 2\gamma d
 \end{aligned}$$

Case $\gamma = 0$

$$v_{left}(s_{top}|\gamma = 0) = d \quad v_{right}(s_{top}|\gamma = 0) = 0$$

Optimal action is left.

Case $\gamma = 0.5$

$$v_{left}(s_{top}|\gamma = 0.5) = d \quad v_{right}(s_{top}|\gamma = 0.5) = d$$

Optimal action is left or right.

Case $\gamma = 0.9$

$$v_{left}(s_{top}|\gamma = 0.9) = d \quad v_{right}(s_{top}|\gamma = 0.9) = 1.8d =$$

Optimal action is right.

3.23 Question

Give the Bellman equation for q^* for the recycling robot.

Answer

$$\begin{aligned}
 q_*(s, a) &= E[R + \max_{a'} q(s', a')] \\
 q(h, s) &= \alpha[r_{search} + \max_{a'} q(h, a')] + (1 - \alpha)[r_{search} + \max_{a'} q(l, a')] \\
 q(l, s) &= \beta[r_{search} + \max_{a'} q(l, a')] + (1 - \beta)[-3 + \max_{a'} q(h, a')] \\
 q(h, w) &= r_{wait} + \max_{a'} q(h, a') \\
 q(l, w) &= r_{wait} + \max_{a'} q(l, a') \\
 q(l, w) &= \max_{a'} q(h, a')
 \end{aligned}$$

3.24 Question

Figure 3.5 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places.

Answer

$$\begin{aligned}
 G_t &= \sum_{k=0} \gamma^k R_{t+k+1} \\
 \text{Optimal policy is to go from A' to A.} \\
 V_*(A) &= 10 + \gamma * 0 + \gamma^2 * 0 + \gamma^3 * 0 + \gamma^4 * 0 + \dots \\
 V_*(A) &= \sum_{k=0} \gamma^{5k} (10 + \gamma * 0 + \gamma^2 * 0 + \gamma^3 * 0 + \gamma^4 * 0) = \sum_{k=0} \gamma^{5k} 10 = \frac{10}{1 - \gamma^5} \\
 V_*(A) &= \frac{10}{1 - 0.9^5} = 24.419
 \end{aligned}$$

3.25 Question

Give an equation for v^* in terms of q^* .

Answer

$$v_*(s) = \max_a q_*(s, a)$$

3.26 Question

Give an equation for q^* in terms of v^* and the four-argument p .

Answer

$$q_*(s, a) = \sum_{s', r} p(s', r | s, a) (r + \gamma v_*(s'))$$

3.27 Question

Give an equation for π_* in terms of q_* .

Answer

$$\pi_*(a | s) = \begin{cases} 1, & \text{if } a = \operatorname{argmax}_{a'} q_*(a' | s) \\ 0, & \text{else} \end{cases} \quad (13)$$

3.28 Question

Give an equation for π_* in terms of v_* and the four-argument p .

Answer

$$\pi_*(a | s) = \begin{cases} 1, & \text{if } a = \operatorname{argmax}_{a'} \sum_{r, s'} p(r, s' | a', s) (r + \gamma v_*(s')) \\ 0, & \text{else} \end{cases} \quad (14)$$

3.29 Question

Rewrite the four Bellman equations for the four value functions (v_π, v_*, q_π and q_*) in terms of the three argument function p (3.4) and the two-argument function r (3.5).

Answer

$$G_t = R_{t+1} + \gamma G_{t+1} \quad (15)$$

$$p(s' | s, a) = \Pr\{S_t = s' | S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in R} p(s', r | s, a) \quad (16)$$

$$r(s, a) = E[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R, s' \in S} p(s', r | s, a) * r \quad (17)$$

$$v_*(s) = \max_a \sum_{s', r} p(s', r | s, a) [r + \gamma v_*(s')] \quad (18)$$

$$q_*(s, a) = \sum_{s', r} p(s', r | s, a) [r + \gamma \max_{a'} q_*(s', a')] \quad (19)$$

Derivations

$$\begin{aligned}
v_\pi(s) &= E_\pi[G_t | S_t = s] = E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] \\
v_\pi(s) &= \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a) [r + \gamma v_\pi(s')] \\
v_\pi(s) &= \sum_a \pi(a|s) \left[\sum_{s',r} p(s', r | s, a) r + \sum_{s'} \sum_r p(s', r | s, a) \gamma v_\pi(s') \right] \quad (20) \\
v_\pi(s) &= \sum_a \pi(a|s) [r(s, a) + \sum_{s'} p(s' | s, a) \gamma v_\pi(s')]
\end{aligned}$$

$$\begin{aligned}
q_\pi(s, a) &= E_\pi[G_t | S_t = s, A_t = a] = E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
q_\pi(s, a) &= \sum_{s',r} p(s', r | s, a) [r + \gamma v_\pi(s')] \\
q_\pi(s, a) &= \sum_{s',r} p(s', r | s, a) r + \sum_{s'} \sum_r p(s', r | s, a) \gamma v_\pi(s') \quad (21) \\
q_\pi(s, a) &= \sum_a \pi(a|s) [r(s, a) + \sum_{s'} p(s' | s, a) \gamma v_\pi(s')]
\end{aligned}$$

$$\begin{aligned}
v_*(s) &= \max_a \left[\sum_{s',r} p(s', r | s, a) r + \sum_{s'} \sum_r p(s', r | s, a) \gamma v_*(s') \right] \\
v_*(s) &= \max_a [r(s, a) + \sum_{s'} p(s' | s, a) \gamma v_*(s')] \quad (22)
\end{aligned}$$

$$\begin{aligned}
q_*(s, a) &= \sum_{s',r} p(s', r | s, a) * r + \sum_{s'} \sum_r p(s', r | s, a) * \gamma \max_{a'} q_*(s', a') \\
v_*(s) &= r(s, a) + \sum_{s'} p(s' | s, a) * \gamma \max_{a'} q_*(s', a') \quad (23)
\end{aligned}$$