Tags: #Exercises

Exercise 3.1

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

1. Consider a Roomba.
    - States: Battery sufficient, battery low, cleaning, in-transit, collision.
    - Actions: Return to charger, leave charger, move to location, avoid obstacle, start clean, stop clean.
    - Rewards: Pick up trash (positive reward), hit an obstacle (negative reward), run out of batteries (very negative reward), run over pet feces (extremely negative reward)
2. Preparing to go outside.
    - States: Sunny, rainy, overcast, hot, cold.
    - Actions: Take sunglasses, take umbrella, take jacket.
    - Rewards: If sunny reward sunglasses, if rainy reward umbrella, if cold reward jacket.
3. Bartending.
    - States: customer needs drink, customer received drink, good on liquor, running low on liquor, out of liquor, glasses available, glasses dirty.
    - Actions: Make drink, give drink to customer, get more liquor, clean glasses.
    - Rewards: Discounted positive reward through time for getting a drink to customer, negative reward for glasses dirty, positive reward for glasses dirty to clean glasses, negative reward for running out of liquor or trying to make a drink when out of liquor.

Exercise 3.2:

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

Answer

Maybe not always. Perhaps when you need more information about things you cannot glean from just the prior state (like if you need the history of the previous states) to be able to make an action. Vectors of rewards make the framework break down.

Exercise 3.3

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

Depends on where you want to get your rewards from, because the rewards need to be external to the learning agent.

Exercise 3.4

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

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1 - \alpha$
low	search	high	-3	$1 - \beta$
low	search	low	r_{search}	β
high	wait	high	r_{wait}	1
low	wait	low	r_{wait}	1

s	a	s'	r	$p(s', r \mid s, a)$
low	recharge	high	0	1

Exercise 3.5

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).

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s).$$

Answer

$$\sum_{s' \in \mathcal{S}^+} \sum_{r \in \mathbb{R}} p(s', r|s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

with $\mathcal{S}^+$ being all states (including the terminal state) and $\mathcal{S}$ excluding the terminal states.

Exercise 3.6:

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

The return would still be $-\gamma^k$ with k being the time step in the episode. Discounting would not matter in this case, since the goal is to balance the pole as long as possible. The return in discounted continuing formulation of the task accumulates because the agent can continue to fail.

Exercise 3.7:

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

It would accumulate rewards for staying in the maze. Giving a negative reward at all other times that it stays in the maze would be better. You need to communicate to the agent that it wants to not remain in the maze and escape ASAP.

Exercise 3.8:

Suppose γ = 0.5 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 $G_0, G_1, \ldots, G_5$? Hint: Work backwards.

Answer

$G_0 = 2$
$G_1 = 6$
$G_2 = 8$
$G_3 = 4$
$G_4 = 2$
$G_5 = 0$

Exercise 3.9

Suppose γ = 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

$G_0 = 65$
$G_1 = 70$

Exercise 3.10

Prove the second equality in (3.10).

Answer

$$
\begin{aligned}
\text{Let } S &= \sum_{k=0}^{\infty} \gamma^k \text{ with } \gamma < 1 \\
\text{then } S &= 1 + \gamma + \gamma^2 + \gamma^3 + \gamma^4 + \cdots + \gamma^{\infty} \\
\text{with } \gamma S &= \gamma + \gamma^2 + \gamma^3 + \gamma^4 + \cdots + \gamma^{\infty} \\
\text{if we take } S - \gamma S &= 1 + (\gamma + \gamma^2 + \gamma^3 + \gamma^4 + \cdots + \gamma^{\infty}) - (\gamma + \gamma^2 + \gamma^3 + \gamma^4 + \cdots + \gamma^{\infty}) \\
S - \gamma S &= 1 \\
(1 - \gamma) S &= 1 \\
S &= \frac{1}{1 - \gamma}
\end{aligned}
$$

Exercise 3.11

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

$$
\mathbb{E}[R_{t+1} | S_t = s] = \sum_a \pi(a|s) \sum_{s'} \sum_r r \cdot p(s', r | s, a)
$$

Exercise 3.12

Give an equation for v_π in terms of q_π and π.

Answer

$$
v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) q_\pi(s, a)
$$

Exercise 3.13

Give an equation for q_π in terms of v_π and the four-argument p.

Answer

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

Exercise 3.14

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.)

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Answer

$$\begin{aligned} r &= 0 \\ v_\pi(s_{center}) &= [0.25 \times [0.9(2.7) + r]] + [0.25 \times [0.9(0.4) + r]] + [0.25 \times [0.9(0.4) + r]] + [0.25 \times [0.9(0.7) + \\ &= 0.25 \times 0.9(2.3 + 0.4 - 0.4 + 0.7) \\ &= 0.675 \\ &\approx 0.7 \end{aligned}$$

Exercise 3.15

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

Signs should matter, since we want things to be a strict penalty.

$$
\begin{aligned}
G'_t &= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \\
&= \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1} + \gamma^k c) \\
&= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c \\
&= G_t + \sum_{k=0}^{\infty} \gamma^k c \\
&= G_t + \frac{c}{1-\gamma}
\end{aligned}
$$

For v_c:

$$
\begin{aligned}
v'(s) &= \mathbb{E}[G'_t | S_t = s] \\
&= \mathbb{E}[G_t + \frac{c}{1-\gamma} | S_t = s] \\
&= \mathbb{E}[G_t | S_t = s] + \frac{c}{1-\gamma} \\
&= v(s) + \frac{c}{1-\gamma}
\end{aligned}
$$

Exercise 3.16

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.

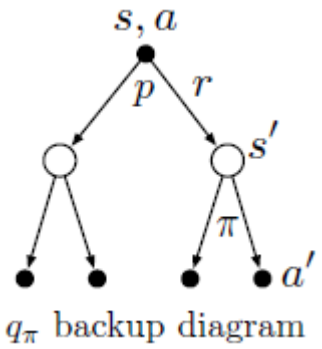
Answer

Adding a c that would make all rewards positive would cause the agent to spend too much time in the maze, as it is not incentivized to exit since rewards would accumulate.

Exercise 3.17

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.



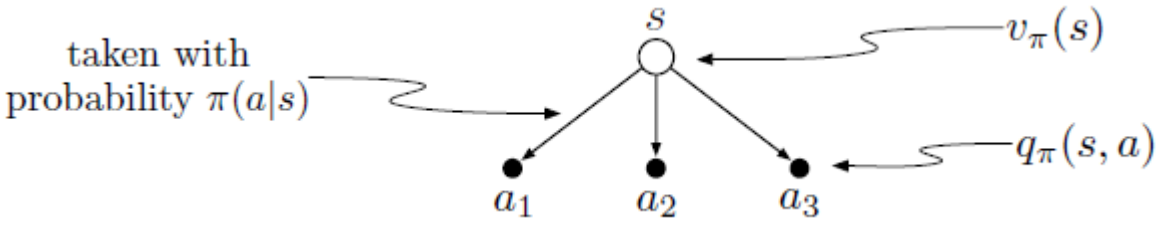
q_π backup diagram

Answer

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[G_t|S_t=s,A_t=a] \\
&= \mathbb{E}_\pi[R_{t+1}+\gamma G_{t+1}|S_t=s,A_t=a] \\
&= \mathbb{E}[R_{t+1}|S_t=s,A_t=a]+\gamma\mathbb{E}_\pi[G_{t+1}|S_t=s,A_t=a] \\
&= r(s,a)+\gamma\mathbb{E}_\pi[G_{t+1}|S_t=s,A_t=a] \\
&= r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)\mathbb{E}_\pi[G_{t+1}|S_{t+1}=s',S_t=s,A_t=a] \\
&= r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)\mathbb{E}_\pi[G_{t+1}|S_{t+1}=s'] \\
&= r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)\mathbb{E}_\pi[q_\pi(s',A_{t+1})|S_{t+1}=s'] \\
&= r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)\sum_{a'\in\mathcal{A}}\pi(a'|s')q_\pi(s',a')
\end{aligned}
$$

Exercise 3.18

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 $S_t=s$. This equation should include an expectation conditioned on following the policy, π. Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ such that no expected value notation appears in the equation.

Answer

First equation:

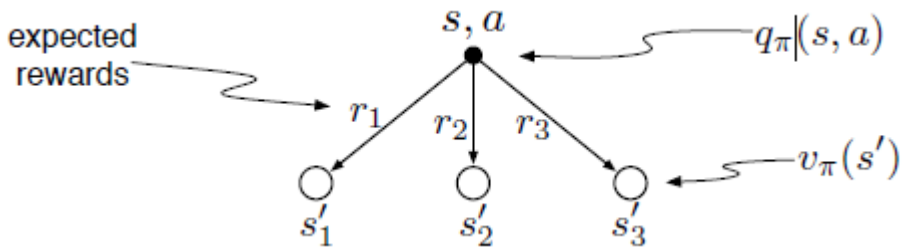
$$v_\pi(s) = \mathbb{E}_\pi[q_\pi(s,a)|S_t = s, a \in \mathcal{A}(s)]$$

Second equation:

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) q_\pi(s,a)$$

Exercise 3.19

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 $A_t = a$. This equation should include an expectation but not one conditioned on following the policy. Then give a second equation, writing out the expected value explicitly in terms of $p(s', r|s, a)$ defined by (3.2), such that no expected value notation appears in the equation.

Answer

First equation:

$$q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s, A_t = a]$$

Second equation:

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

Exercise 3.20

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

Answer

The hole (terminal state) has a value of 0. Everywhere in the green has a value of -1 (in the green). Outside of the green, if there is one stroke to get to the green, it is -2. We would use the driver until we get to the green, and then the putter to get to the hole.

Exercise 3.21

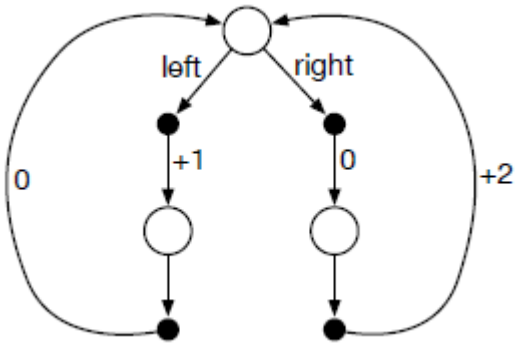
Draw or describe the contours of the optimal action-value function for putting, $q_*(s, putter)$, for the golf example.

Answer

Based on the range of the putter, we would see contours from the green. Anywhere on the green gets you to the hole with the putter. Then the contours just show you how many putts it takes you to get to the green starting at -2.

Exercise 3.22

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, π_{left} and π_{right}. What policy is optimal if γ = 0? If γ = 0.9?
If γ = 0.5?



Answer

For γ = 0:
π_{left} is optimal, since only the immediate reward is considered (1 > 0).

For γ = 0.9:
π_{right} is optimal, since the delayed reward is taken into account heavily (1 < 1.8).

For γ = 0.5:
Either policy is optimal, this is the point where the values are both equal (1).

Exercise 3.23

Give the Bellman equation for q_* for the recycling robot.

Answer

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

Exercise 3.24

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

Optimal policy when we hit point A is to go right back to A from A'.

$$
\begin{aligned}
v_*(A) &= R_{A \to A'} + \gamma R_{A' \to north_1} + \gamma^2 R_{north_2} + \gamma^3 R_{north_3} + \gamma^4 R_{north_4} + \gamma^5 R_{A \to A'} + \cdots \\
&= 10 + \gamma(0) + \gamma^2(0) + \gamma^3(0) + \gamma^4(0) + \gamma^5(10) + \cdots \\
&= \sum_{k=0} \gamma^{5k}(10 + \gamma(0) + \gamma^2(0) + \gamma^3(0) + \gamma^4(0)) \\
&= \sum_{k=0} \gamma^{5k} 10 \\
&= \frac{10}{1 - \gamma^5} \\
&= \frac{10}{1 - 0.9^5} \\
&= 24.419
\end{aligned}
$$

Exercise 3.25

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

Answer

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

Exercise 3.26

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

Answer

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

Exercise 3.27

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

Answer

$$
\pi_*(a|s) = \begin{cases} 1, & \text{if } a = \arg\max_{a'} q_*(s, a') \\ 0, & \text{else} \end{cases}
$$

Exercise 3.28

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

Answer

$$\pi_*(a|s) = \begin{cases} 1, & \text{if } a = \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')] \\ 0, & \text{else} \end{cases}$$

Exercise 3.29

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

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