Reinforcement Learning: An Introduction
Attempted Solutions
Chapter 3

Scott Brownlie & Rafael Rui

1 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.

An e-commerce sit could use reinforcement learning to control daily pricing of products. The actions would be the the prices set for each product on each day. The states might include the month of the year, the day of the week and the proximity to special days such as Christmas and Valentine's Day. The reward would be the profit at the end of each day.

The manager of a football team could use reinforcement learning to pick the 11 players to play each game. The actions would be the team selection. The states could be the opponent, whether the game is home or away and the fitness of the players. The reward would be 0, 1 or 3 depending on whether the team lost, drew or won the game.

A company could use reinforcement learning to control the air temperature in its office. The actions would be the specific settings of the heating/air-conditioning system. The states would include the current outdoor temperature and the indoor temperature in each room in the building. The reward would be the satisfaction of the employees, which could be measured by selecting 10 employees at random every hour and asking them to rate the their comfort on a scale of 1 to 10 and then averaging the ratings.

2 Exercise 3.2

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

The MDP framework "proposes that whatever the details of the sensory, memory, and control apparatus, and whatever objective one is trying to achieve, any problem of learning goal-directed behavior can be reduced to three signals passing back and forth between an agent and its environment: one signal to represent the choices made by the agent (the actions), one signal to represent the basis on which the choices are made (the states), and one signal to define the agent's goal (the rewards)."

One exception could be tasks for which actions are regularly taken but we only know if the goal has ultimately been achieved at some distant point in the future.

Consider the example of a government that would like to learn to set political policies with the goal of ensuring that 90% of all children born in 2020 live beyond 2100. Each year the government would review health statistics (the states) and decide which political policies to implement (the

actions). In the MDP framework, at the end of each year the government would receive a reward, given the state and action selected at the end of the previous year. However, how do we define this yearly reward given that we will only know if the government has ultimately achieved its goal in 2100?

One idea might be to define the reward at the end of year n as the life expectancy in year n. However, maximising life expectancy in the current year does not necessarily increase the chance of a child living until 2100. For example, a policy might be introduced which favours the elderly but has a negative impact on the younger generation.

3 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?

As stated in this chapter, "the agent-environment boundary represents the limit of the agent's absolute control". In this example it would appear that the limit of the agent's absolute control is where the body meets the machine. Therefore, we would define the actions in terms of the accelerator, steering wheel and brake. When driving a car the driver has full control over these apparatus, assuming that nothing jams.

Farther out, where the rubber meets the road, we probably do not have absolute control. For example, the torque applied to the tyres may depend on the condition of the car or even the weather. It is unlikely that pressing the accelerator to a given angle always results in exactly the same torque. At an even higher level, such as the specification of *where* to drive, we certainly do not have absolute control.

Farther in, we do have absolute control over our own limbs, but this boundary is not on the limit of absolute control and defining actions as muscle twitches would appear to be overkill.

4 Exercise 3.4

5 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).

We need to add the terminal state to the possible next states:

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

6 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?

Suppose that failure occurs on time step T. Then the reward at time step t is

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T = -\gamma^{T-t-1}.$$

In the discounted, continuing formulation of this task the reward received at each time step not only depends on the rewards of subsequent time steps in the same episode, but also time steps in all subsequent episodes. Suppose that failure occurs on time step T_i during episode i for $i = 1, 2, 3, \ldots$. Then the reward at time step t of episode j is

$$G_{j,t} = -\gamma^{T_j - t - 1} - \gamma^{T_{j+1} + T_j - t - 1} - \gamma^{T_{j+2} + T_{j+1} + T_j - t - 1} - \ldots$$

7 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?

The reward on each time step will be 1, regardless of how long it takes the agent to escape. As we only give the robot a reward of $+1$ on the final time step, it has no incentive to escape on the 10th time step as opposed to the 1000th time step. Assuming that we want the robot to escape quickly, a much better idea would be to give it a reward of -1 on each time step that is it inside the maze.

8 Exercise 3.8

Suppose $\gamma = 0.5$ 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.

We have

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

9 Exercise 3.9

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?

We have

$$\begin{aligned} G_1 &= \sum_{k=0}^{\infty} \gamma^k R_{k+2} \\ &= \sum_{k=0}^{\infty} 0.9^k \cdot 7 \\ &= 7 \cdot \sum_{k=0}^{\infty} 0.9^k \\ &= \frac{7}{1-0.9} \\ &= 70 \end{aligned}$$

and

$$G_0 = R_1 + \gamma G_1 = 2 + 0.9 \cdot 70 = 65.$$

10 Exercise 3.10

Prove the second equality in (3.10).

We have

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma^k &= 1 + \sum_{k=1}^{\infty} \gamma^k \\ &= 1 + \gamma \sum_{k=0}^{\infty} \gamma^k \\ \iff (1-\gamma) \sum_{k=0}^{\infty} \gamma^k &= 1 \\ \iff \sum_{k=0}^{\infty} \gamma^k &= \frac{1}{1-\gamma}. \end{aligned}$$

11 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)?

The expectation is

$$\mathbb{E}[R_{t+1}|S_t = s] = \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sum_{s' \in \mathcal{S}^+} \sum_{r \in \mathcal{R}} p(s', r|s, a) r, \text{ for all } s \in \mathcal{S}.$$

12 Exercise 3.12

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

We have

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) q_\pi(s, a), \text{ for all } s \in \mathcal{S}.$$

13 Exercise 3.13

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

We have

$$q_\pi(s,a) = \sum_{s' \in \mathcal{S}^+} p(s', r|s, a)[r + \gamma v_\pi(s')], \text{ for all } s \in \mathcal{S}.$$

14 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.)

As π is the equiprobable random policy, we have $\pi(a|s) = 0.25$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, and given the action the next state is deterministic. Thus, the right hand side of the Bellman equation is

$$\sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] = 0.25 \cdot 0.9(2.3 + 0.4 - 0.4 + 0.7) = 0.675,$$

which equals the value of the centre state, $+0.7$, when rounded to one decimal place.

15 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?

As the agent tries to maximise the long term reward, the signs of the rewards are clearly important. If we switched the signs then the agent would learn to run of the edge of the world.

Suppose we add a constant c to all the rewards. Then for all $s \in \mathcal{S}$ we have

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \middle| S_t = s\right] \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] + \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k c \middle| S_t = s\right] \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] + c\sum_{k=0}^{\infty} \mathbb{E}_\pi[\gamma^k|S_t = s] \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] + c\sum_{k=0}^{\infty} \gamma^k \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] + \frac{c}{1-\gamma}.
\end{aligned}$$

Therefore, adding a constant c to all the rewards adds $v_c = \frac{c}{1-\gamma}$ to the values of all states.

16 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.

It depends on the value of c. Suppose that initially the agent receives a reward of $-r$ on each time step that it is inside the maze. In this case the agent will try to escape from the maze as quickly as possible in order to maximize the total reward. Suppose now that we add a constant c to all the rewards. If $c < r$ then the rewards are still negative and the task is unchanged. If $c = r$ then all rewards are 0 and the agent will learn nothing. If $c > r$ then all rewards are positive and the agent will stay in the maze for as long as possible in order to maximise the total reward.

17 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.

We have

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

18 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.

We have

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[q_\pi(s, a) | S_t = s] \\ &= \sum_a \pi(a|s) q_\pi(s, a). \end{aligned}$$

19 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.

We have

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s, A_t = a] \\ &= \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]. \end{aligned}$$

20 Exercise 3.20

21 Exercise 3.21

22 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 $\gamma = 0$? If $\gamma = 0.9$? If = $\gamma = 0.5$?

Suppose that we start in the top state. Then

$$v_{\pi_{\text{left}}}(\text{top}) = 1 + 0 + \gamma^2 + 0 + \gamma^4 + \cdots = \sum_{k=0,2,4,\ldots} \gamma^k$$

and

$$v_{\pi_{\text{right}}}(\text{top}) = 0 + 2\gamma + 0 + 2\gamma^3 + 0 + \cdots = \sum_{k=1,3,5,\ldots} 2\gamma^k.$$

If $\gamma = 0$ then $v_{\pi_{\text{left}}}(\text{top}) = 1$ and $v_{\pi_{\text{right}}}(\text{top}) = 0$, so $v_{\pi_{\text{left}}}$ is optimal.

If $\gamma = 0.9$ then

$$\begin{aligned} v_{\pi_{\text{right}}}(\text{top}) &= 2 \cdot 0.9 + 2 \cdot 0.9^3 + 2 \cdot 0.9^5 + \ldots \\ &= 1.8(1 + 0.9^2 + 0.9^4 + \ldots) \\ &= 1.8 v_{\pi_{\text{left}}}(\text{top}), \end{aligned}$$

so $v_{\pi_{\text{right}}}$ is optimal.

If $\gamma = 0.5$ then

$$\begin{aligned} v_{\pi_{\text{right}}}(\text{top}) &= 2 \cdot 0.5 + 2 \cdot 0.5^3 + 2 \cdot 0.5^5 + \ldots \\ &= 1 + 0.5^2 + 0.5^4 + \ldots \\ &= v_{\pi_{\text{left}}}(\text{top}), \end{aligned}$$

so $v_{\pi_{\text{left}}}$ and $v_{\pi_{\text{right}}}$ are both optimal.

In fact, $v_{\pi_{\text{left}}}$ is optimal for $\gamma \leq 0.5$ and $v_{\pi_{\text{right}}}$ is optimal for $\gamma \geq 0.5$.

23 Exercise 3.23

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

When the battery is high the robot can either search or wait, and when the battery is low the robot can either search, wait or recharge. Thus there are five equations in total:

$$\begin{aligned}
q_*(\mathtt{h},\mathtt{s}) &= p(\mathtt{h}|\mathtt{h},\mathtt{s})\big[r(\mathtt{h},\mathtt{s},\mathtt{h}) + \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\}\big] \\
&\quad + p(\mathtt{l}|\mathtt{h},\mathtt{s})\big[r(\mathtt{h},\mathtt{s},\mathtt{l}) + \gamma \max\{q_*(\mathtt{l},\mathtt{s}), q_*(\mathtt{l},\mathtt{w}), q_*(\mathtt{l},\mathtt{re})\}\big] \\
&= \alpha\big[r_{\text{search}} + \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\}\big] + (1-\alpha)\big[r_{\text{search}} + \gamma \max\{q_*(\mathtt{l},\mathtt{s}), q_*(\mathtt{l},\mathtt{w}), q_*(\mathtt{l},\mathtt{re})\}\big] \\
&= \alpha\gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\} + r_{\text{search}} + (1-\alpha)\gamma \max\{q_*(\mathtt{l},\mathtt{s}), q_*(\mathtt{l},\mathtt{w}), q_*(\mathtt{l},\mathtt{re})\},
\end{aligned}$$

$$\begin{aligned}
q_*(\mathtt{h},\mathtt{w}) &= r(\mathtt{h},\mathtt{w},\mathtt{h}) + \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\} \\
&= r_{\text{wait}} + \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\},
\end{aligned}$$

$$\begin{aligned}
q_*(\mathtt{l},\mathtt{s}) &= p(\mathtt{h}|\mathtt{l},\mathtt{s})\big[r(\mathtt{l},\mathtt{s},\mathtt{h}) + \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\}\big] \\
&\quad + p(\mathtt{l}|\mathtt{l},\mathtt{s})\big[r(\mathtt{l},\mathtt{s},\mathtt{l}) + \gamma \max\{q_*(\mathtt{l},\mathtt{s}), q_*(\mathtt{l},\mathtt{w}), q_*(\mathtt{l},\mathtt{re})\}\big] \\
&= (1-\beta)\big[-3 + \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\}\big] + \beta\big[r_{\text{search}} + \gamma \max\{q_*(\mathtt{l},\mathtt{s}), q_*(\mathtt{l},\mathtt{w}), q_*(\mathtt{l},\mathtt{re})\}\big],
\end{aligned}$$

$$\begin{aligned}
q_*(\mathtt{l},\mathtt{w}) &= r(\mathtt{l},\mathtt{w},\mathtt{l}) + \gamma \max\{q_*(\mathtt{l},\mathtt{s}), q_*(\mathtt{l},\mathtt{w}), q_*(\mathtt{l},\mathtt{re})\} \\
&= r_{\text{wait}} + \gamma \max\{q_*(\mathtt{l},\mathtt{s}), q_*(\mathtt{l},\mathtt{w}), q_*(\mathtt{l},\mathtt{re})\},
\end{aligned}$$

$$\begin{aligned}
q_*(\mathtt{l},\mathtt{re}) &= r(\mathtt{l},\mathtt{re},\mathtt{h}) + \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\} \\
&= \gamma \max\{q_*(\mathtt{h},\mathtt{s}), q_*(\mathtt{h},\mathtt{w})\}.
\end{aligned}$$

24 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.

Suppose that the agent is at position $\mathtt{A}$ on time step t. Then it will move to position $\mathtt{A}'$ and receive an undiscounted reward of 10 on time step $t+1$. Next it will move north back towards position $\mathtt{A}$, arriving there on time step $t+5$. On each of time steps $t+2$ to $t+5$ it will receive a reward of 0. Then on time step $t+6$ it will move back to $\mathtt{A}'$ and receive a reward of $10 \cdot 0.9^5$, and so on.

Thus we have

$$v_*(\mathtt{A}) = 10 \sum_{k \in \{0,5,10,15,\dots\}} 0.9^k = 24.419$$

to three decimal places.

25 Exercise 3.25

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

We have

$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_*(s,a).$$

26 Exercise 3.26

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

We have

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

27 Exercise 3.27

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

We have

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \text{argmax}_x\, q_*(s,x) \\ 0 & \text{otherwise} \end{cases}$$

28 Exercise 3.28

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

We have

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \text{argmax}_x \sum_{s',r} p(s',r|s,x)\big[r + \gamma v_*(s')\big] \\ 0 & \text{otherwise} \end{cases}$$

29 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).

We have

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

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

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

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