Reinforcement Learning: An Introduction
Attempted Solutions
Chapter 4

Scott Brownlie & Rafael Rui

1 Exercise 4.1

In Example 4.1, if π is the equiprobable random policy, what is $q_\pi(11, \texttt{down})$. What is $q_\pi(7, \texttt{down})$?

As moving downwards from 11 results in the terminal state, $q_\pi(11, \texttt{down}) = -1$. Moving right from 7 leaves the state unchanged, so

$$q_\pi(7, \texttt{down}) = -1 + v_\pi(7) = -1 + -20 = -21.$$

2 Exercise 4.2

In Example 4.1, suppose a new state 15 is added to the gridworld just below state 13, and its actions, `left`, `up`, `right`, and `down`, take the agent to states 12, 13, 14, and 15, respectively. Assume that the transitions from the original states are unchanged. What, then, is $v_\pi(15)$ for the equiprobable random policy? Now suppose the dynamics of state 13 are also changed, such that action `down` from state 13 takes the agent to the new state 15. What is $v_\pi(15)$ for the equiprobable random policy in this case?

When the transitions from the original states are unchanged we have

$$\begin{aligned} v_\pi(15) &= -1 + 0.25(v_\pi(12) + v_\pi(13) + v_\pi(14) + v_\pi(15)) \\ &= -1 + 0.25(-22 - 20 - 14) + 0.25 \cdot v_\pi(15) \\ \iff 0.75 \cdot v_\pi(15) &= -15 \\ \iff v_\pi(15) &= -20. \end{aligned}$$

Now suppose that the dynamics of state 13 are changed such that action `down` from state 13 takes the agent to the new state 15. Since $v_\pi(15) = -20 = v_\pi(13)$ in the case of unchanged dynamics, $v_\pi(15)$ should remain -20. We can formally prove this:

$$\begin{aligned} v_\pi(15) &= -1 + 0.25(v_\pi(12) + v_\pi(13) + v_\pi(14) + v_\pi(15)) \\ &= -1 + 0.25(-22 - 14) + 0.25 \cdot v_\pi(13) + 0.25 \cdot v_\pi(15) \\ &= -10 + 0.25 \cdot v_\pi(13) + 0.25 \cdot v_\pi(15), \end{aligned}$$

where

$$\begin{aligned} v_\pi(13) &= -1 + 0.25(v_\pi(9) + v_\pi(12) + v_\pi(14) + v_\pi(15)) \\ &= -1 + 0.25(-20 - 22 - 14) + 0.25 \cdot v_\pi(15) \\ &= -15 + 0.25 \cdot v_\pi(15). \end{aligned}$$

Hence,

$$\begin{aligned} v_\pi(15) &= -10 + 0.25(-15 + 0.25 \cdot v_\pi(15)) + 0.25 \cdot v_\pi(15) \\ &= -13.75 + 0.3125 \cdot v_\pi(15) \\ \iff 0.6875 \cdot v_\pi(15) &= -13.75 \\ \iff v_\pi(15) &= -20. \end{aligned}$$

3 Exercise 4.3

What are the equations analogous to (4.3), (4.4), and (4.5) for the action-value function q_π and its successive approximation by a sequence of functions q_0, q_1, q_2,...?

We have

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

and

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