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

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

4 Exercise 4.4

The policy iteration algorithm on page 80 has a subtle bug in that it may never terminate if the policy continually switches between two or more policies that are equally good. This is ok for pedagogy, but not for actual use. Modify the pseudocode so that convergence is guaranteed.

In part 3 of the algorithm, instead of checking if the policy is stable, we should check if the policy has improved as follows:

policy-improved $\leftarrow false$

For each $s \in \mathcal{S}$

$v \leftarrow V(s)$

$\pi(s) \leftarrow \underset{a}{\text{argmax}} \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$

$v_{new} \leftarrow \sum_{s',r} p(s', r|s, \pi(s))[r + \gamma V(s')]$

If $v_{new} > v$, then *policy-improved* $\leftarrow true$

If *policy-improved*, then go to 2, else stop and return $V \approx v_*, \pi \approx \pi_*$.

5 Exercise 4.5

How would policy iteration be defined for action values? Give a complete algorithm for computing q_*, analogous to that on page 80 for computing v_*. Please pay special attention to this exercise, because the ideas involved will be used throughout the rest of the book.

The algorithm is

1. Initialisation
   $Q(s,a) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   Loop for each $a \in \mathcal{A}(s)$:
   $q \leftarrow Q(s,a)$
   $Q(s,a) \leftarrow \sum_{s',r} p(s',r|s,a)\Big[r + \gamma Q(s',\pi(s'))\Big]$
   $\Delta \leftarrow \max(\Delta, |q - Q(s,a)|)$
   until $\Delta < \theta$

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \underset{a}{\operatorname{argmax}} \sum_{s',r} p(s',r|s,a)\Big[r + \gamma Q(s',\pi(s'))\Big]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow false$
   If *policy-stable*, then stop and return $Q \approx q_*, \pi \approx \pi_*$, else go to 2.

6 Exercise 4.5

Suppose you are restricted to considering only policies that are ϵ-soft, meaning that the probability of selecting each action in each state, s, is at least $\epsilon/|\mathcal{A}(s)|$. Describe qualitatively the changes that would be required in each of the steps 3, 2, and 1, in that order, of the policy iteration algorithm for v_* on page 80.

Assume that ϵ-soft means that, for all $s \in \mathcal{S}$, $\pi(s) = a$ for some $a \in \mathcal{A}(s)$ with probability $1 - \epsilon$, else $\pi(s)$ is chosen uniformly at random from $\mathcal{A}(s)$.