Exercise 4.1
In Example 4.1, if π is the equiprobable random policy, what is qπ(11, down)? What is qπ(7, down)?

Answer

$$q_\pi(s,a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p\left(s', r | s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right]$$

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

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

Q_{pi} (11, down) = r + 1*0 = -1.
Q_{pi} (7, down) = r + 1*v_{pi}(11) = -1 - 1*14 = -15.

$$\begin{aligned} q_\pi(7, \texttt{down}) &= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \\ &= R_t + \underbrace{\gamma}_{1} \underbrace{\mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']}_{v(11)} \\ &= -1 - 14 \\ &= -15. \end{aligned}$$

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 vπ(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 vπ(15) for the equiprobable random policy in this case?

Answer
unchanged policy
$v_\pi(15)$ = = 0.25 (-1 - 22 - 1 - 20 - 1 - 14 - 1 + $v_\pi(15)$)
$v_\pi(15)$ = -20.

Changed policy

Suppose the dynamics of 13 also changed. We do the iterative policy evaluation. The initialization is natural: we let V0(s) = vπ(s), where vπ is the value when the dynamics are unchanged, as shown in the k = ∞ case in Figure 4.1.
V0(15) = −20, where −20 is what we just derived for vπ(15).
Particularly, we do the “immediate overwriting”. First we update the value of 13.
v(13) = 0.25 * (-1 - 20 - 1 - 22 - 1 - 14 - 1 - 20) = 20.
We then proceed to update V(15) with V(13):
V(13) = 0.25 * (-1 - 22 - 1 - 20 - 1 - 14 - 1 - 20) = 20.

As we proceed, we observe no change – the initialization is good enough. Hence the iterative policy evaluation ends with 1 iteration. V1(15) = −20 for this case.

Exercise 4.3

What are the equations analogous to (4.3), (4.4), and (4.5) for the action- value function qπ and its successive approximation by a sequence of functions q0, q1, q2,...?

$$q_\pi(s,a) = r(s,a) + \sum_{a'} \mathbb{E}_\pi \left[\gamma q_\pi(S_{t+1}, a') | S_t = s, A_t = a\right]$$

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

$$q_{k+1}(s,a) = \sum_{s',a} p\left(s', r | s, a\right) \left[r + \sum_{a'} \pi_k\left(a' | s'\right) \gamma q_k(s', a')\right]$$