Reinforcement Learning HW 2

Mert Bilgin (7034879)
mert.bilgin@student.uni-tuebingen.de
Lalitha Sivakumar (6300674)
lalitha.sivakumar@student.uni-tuebingen.de
Kevin Van Le (7314700)
kevin-van.le@student.uni-tuebingen.de

October 27, 2025

1 State-Action Value Function and Policy Iteration

a)

The state-action value function is given by:

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

For $q_\pi(11, \text{down})$:

$$q_\pi(11,\text{down}) = \sum_{s'} p(s' \mid 11,\text{down})\big[-1 + v_\pi(s')\big]$$

Since only $p(\text{Terminal} \mid 11,\text{down}) = 1$, we have:

$$q_\pi(11,\text{down}) = -1 + v_\pi(\text{Terminal}) = -1 + 0 = -1$$

Similarly, for $q_\pi(7, \text{down})$:

$$q_\pi(7,\text{down}) = \sum_{s'} p(s' \mid 7,\text{down})\big[-1 + v_\pi(s')\big]$$

Since only $p(11 \mid 7,\text{down}) = 1$, we have:

$$q_\pi(7,\text{down}) = -1 + (-14) = -15$$

Finally, for $q_\pi(9, \text{left})$:

$$q_\pi(9,\text{left}) = \sum_{s'} p(s' \mid 9,\text{left})\big[-1 + v_\pi(s')\big]$$

Since only $p(8 \mid 9,\text{left}) = 1$, we have:

$$q_\pi(9,\text{left}) = -1 + (-20) = -21$$

b)

We know that the state-value function under a policy π is given by:

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

The optimal value can be written in a similar fashion. That is,

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

For the optimal policy π_*, we have:

$$v_*(s) = \sum_{a \in \mathcal{A}} \pi_*(a \mid s)\, q_*(s, a)$$

Since the optimal policy selects the action that maximizes $q_*(s, a)$,

$$\pi_*(a \mid s) = \begin{cases} 1, & \text{if } a = \arg\max_{a \in \mathcal{A}} q_*(s, a), \\ 0, & \text{otherwise.} \end{cases}$$

Substituting this back, we obtain:

$$v_*(s) = \sum_{a \in \mathcal{A}} \left[\mathbb{I}\left(a = \arg\max_{a' \in \mathcal{A}} q_*(s, a') \right)\, q_*(s, a) \right] = \max_a q_*(s, a)$$

where we use $\mathbb{I}$ for the indicator operator.

c)

We know that for any policy π, the following holds for the action-value function:

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a\, v_\pi(s')$$

Taking the maximum over all policies, we obtain:

$$\max_\pi q_\pi(s, a) = q_*(s, a) = \max_\pi \left(R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a\, v_\pi(s') \right)$$

Since the reward R_s^a and transition probabilities $P_{ss'}^a$ do not depend on π, we can move the maximization inside the sum:

$$q_*(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a\, \left(\max_\pi v_\pi(s') \right)$$

By definition of the optimal value function $v_*(s') = \max_\pi v_\pi(s')$, we have:

$$\boxed{q_*(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a\, v_*(s')}$$

d)

An optimal policy can be found by maximizing over the optimal action-value function $q_*(s, a)$:

$$\pi_*(a \mid s) = \begin{cases} 1, & \text{if } a = \arg\max_{a' \in \mathcal{A}} q_*(s, a'), \\ 0, & \text{otherwise.} \end{cases}$$

It greedily selects the action with the highest estimated return according to q_*.

e)

We know that the state-value function is defined as:

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

For any policy π, the action-value function satisfies:

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a\, v_\pi(s')$$

Substituting the definition of $v_\pi(s')$ into the above equation gives:

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a \left(\sum_{a' \in \mathcal{A}} \pi(a' \mid s')\, q_\pi(s', a') \right)$$

Simplifying, we obtain the Bellman expectation equation for the action-value function:

$$\boxed{q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P_{ss'}^a\, \pi(a' \mid s')\, q_\pi(s', a')}$$

2 Value Iteration