The University of Vermont

CS253A QR: Reinforcement Learning
Assignment №5

Ayat Ospanov

October 7, 2018

Contents

1 Exercise 4.1

Let's calculate $q_\pi(11, \text{down})$:

$$q_\pi(11, \text{down}) = \sum_{s',r} p(s', r|11, \text{down})[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s', a')] =$$
$$= -1 + \frac{1}{4} \sum_{a'} q_\pi(T, a') = -1$$

Now, let's calculate $q_\pi(7, \text{down})$:

$$q_\pi(7, \text{down}) = \sum_{s',r} p(s', r|7, \text{down})[r + \gamma v_\pi(s')] =$$
$$= -1 + v_\pi(11) = -1 - 14 = -15$$

2 Exercise 4.2

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

$$\begin{aligned} v_\pi(15) &= \sum_a \frac{1}{4} \sum_{s'(\text{given } a)} [-1 + v_\pi(s')] = \\ &= \frac{1}{4} \sum_{s' \in \{12,13,14,15\}} [-1 + v_\pi(s')] = \\ &= \frac{1}{4}[v_\pi(12) + v_\pi(13) + v_\pi(14) + v_\pi(15) - 4] \end{aligned}$$

Thus, when transitions for original states are unchanged, we have:

$$\begin{aligned} v_\pi(15) &= \frac{1}{3}[v_\pi(12) + v_\pi(13) + v_\pi(14) - 4] = \\ &= \frac{1}{3}[-22 - 20 - 14 - 4] = -20 \end{aligned}$$

If the action “down” on the state 13 transits the state to 15, we have the same situation. It happens, because the state 15 is the “copy” of state 13: all neighboring states are the same for both states, so going from 13 to 15 is the same as going from 13 to 13.

3 Exercise 4.3

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

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

4 Exercise 4.4

I can suggest two solutions:

1) As we know each iteration improves the policy, we can terminate when we reach maximum number of iterations we set in advance

2) If several policies have the same value, then they are all maximums. Thus, we can just check if our old policy is in the Argmax (not to be confused with argmax which is the value, while Argmax is the set)

5 Exercise 4.5

```
1. Initialization
    $q(s,a) \in \mathbb{R}$ arbitrarily for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

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

3. Policy Improvement
    policy-stable $\leftarrow$ true
    For each $s \in \mathcal{S}$:
        old-action $\leftarrow \pi(s)$
        $\pi \leftarrow Arg \max_a q(s,a)$
        $\pi(s) = random(\pi)$
        if old-action $\notin \pi$, then policy-stable $\leftarrow$ false
    if policy-stable, then stop and return $Q \approx q_*$ and $\pi \approx \pi_*$; else go to 2
```

6 Exercise 4.6

We will ignore termination conditions and describe only significant changes

Changes in **step 3**:

$$a_{opt} = arg \max_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V_\pi(s')]$$

Then we have to change the probabilities of not optimal actions to $\frac{\varepsilon}{|\mathcal{A}(s)|}$, and the probability of the optimal to the $1 - \frac{\varepsilon}{|\mathcal{A}(s)|}(|\mathcal{A}(s)| - 1) = 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|}$

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|}, & \text{if } a = a_{opt} \\ \frac{\varepsilon}{|\mathcal{A}(s)|}, & \text{if } a \neq a_{opt} \end{cases}$$

Changes in **step 2**:

$$V(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V_\pi(s')]$$

Changes in **step 1**: π becomes a list of lists of probabilities (not a matrix as each row has different lengths) and is initialized with $\frac{\varepsilon}{|\mathcal{A}(s)|}$, while a random action for each state has the probability of $1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|}$

7 Exercise 4.10

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