Reinforcement Learning
Prof. Milica Gašić

Release: Thursday, 24 October 2024
Discussion: Wednesday, 30 October 2024

Exercise set #3

You do not have to hand in your solutions to the exercises and they will **not** be graded. However, there will be four short tests during the semester. You need to reach ≥50% of the total points in order to be admitted to the final exam (Klausur). The tests are held at the start of a lecture (room 2522.U1.74) at the following dates:

Test 1: Thursday, 31 October 2024, 10:30 - 10:45
Test 2: Thursday, 21 November 2024, 10:30 - 10:45
Test 3: Thursday, 5 December 2024, 10:30 - 10:45
Test 4: Thursday, 9 January 2025, 10:30 - 10:45

Please ask questions in the RocketChat
The exercises are discussed every Wednesday, 14:30 - 16:00 in room 2512.02.33.

1. **Recursive Bellman equations**

   Prove the recursive Bellman expectation equations for the value function v_π and the action value function q_π using the state transition function $\mathcal{P}$ and the reward function $\mathcal{R}$. You are allowed to use the equations from Theorem 1 in Section 4.

   (a) $v_\pi(s) = \sum_a \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s') \right)$

   (b) $q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a')$

   Answer:

   (a)
   $$\begin{aligned} v_\pi(s) &= \sum_a \pi(a|s) q_\pi(s,a) && \text{Theorem 6 (iii)} \\ &= \sum_a \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s') \right) && \text{Theorem 6 (iv)} \end{aligned}$$

   (b)
   $$\begin{aligned} q_\pi(s,a) &= \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s') && \text{Theorem 6 (iv)} \\ &= \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a') && \text{Theorem 6 (iii)} \end{aligned}$$

2. **Action value functions**

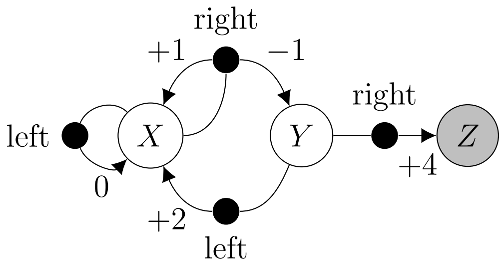
   (a) For any given MDP, policy π, *terminal state* E and action a, what is $q_\pi(E,a)$? All transitions from a terminal state are back to itself with a reward of 0.

   Answer:

   $$\begin{aligned} q_\pi(E,a) &= \underbrace{\mathcal{R}(E,a)}_{=0} + \gamma \sum_{s'} \underbrace{\mathcal{P}(s'|E,a)}_{=1 \text{ for } s'=E} v_\pi(s') \\ &= \gamma v_\pi(E) = 0 \quad \text{(from previous exercise set)} \end{aligned}$$

Exercises by Stefan Harmeling, used with permission

(b) Consider the MDP and policy π_1 from the previous exercise sets. Note that if action *right* is taken in state X, then the transitions to X and Y occur with probabilities 0.75 and 0.25, respectively. The deterministic policy π_1 is defined as $\pi_1(X) = \text{right}$, $\pi_1(Y) = \text{right}$.



Compute the action value of state X and action *left* under policy π_1, i.e. $q_{\pi_1}(X, \text{left})$, using *only* the action value function (don't use the values from the last exercise set). The discount factor is $\gamma = 0.9$.

Answer:

$$\begin{aligned}
q_{\pi_1}(Y, \text{right}) &= \underbrace{\mathcal{R}(Y, \text{right})}_{=4} + \gamma \sum_{s'} \mathcal{P}(s'|Y, \text{right}) \sum_{a'} \pi_1(a'|s') q_{\pi_1}(s', a') \\
&= 4 + 0.9 \sum_{a'} \pi_1(a'|Z) \underbrace{q_{\pi_1}(Z, a')}_{\overset{(a)}{=} 0} = 4
\end{aligned}$$

$$\begin{aligned}
q_{\pi_1}(X, \text{right}) &= \underbrace{\mathcal{R}(X, \text{right})}_{=0.5} + \gamma \sum_{s'} \mathcal{P}(s'|X, \text{right}) \sum_{a'} \pi_1(a'|s') q_{\pi_1}(s', a') \\
&= 0.5 + 0.9 \Big(\mathcal{P}(X|X, \text{right}) \sum_{a'} \pi_1(a'|X) q_{\pi_1}(X, a') + \\
&\qquad \mathcal{P}(Y|X, \text{right}) \sum_{a'} \pi_1(a'|Y) q_{\pi_1}(Y, a') \Big) \\
&= 0.5 + 0.9 (0.75 \cdot q_{\pi_1}(X, \text{right}) + 0.25 \cdot \underbrace{q_{\pi_1}(Y, \text{right})}_{=4}) \\
&= 1.4 + 0.675 \cdot q_{\pi_1}(X, \text{right})
\end{aligned}$$

$$\begin{aligned}
\Leftrightarrow (1 - 0.675) q_{\pi_1}(X, \text{right}) &= 1.4 \\
\Leftrightarrow 0.325 \cdot q_{\pi_1}(X, \text{right}) &= 1.4 \\
\Rightarrow q_{\pi_1}(X, \text{right}) &\approx 4.308
\end{aligned}$$

$$\begin{aligned}
q_{\pi_1}(X, \text{left}) &= \underbrace{\mathcal{R}(X, \text{left})}_{=0} + \gamma \sum_{s'} \mathcal{P}(s'|X, \text{left}) \sum_{a'} \pi_1(a'|s') q_{\pi_1}(s', a') \\
&= 0.9 \sum_{a'} \pi_1(a'|X) q_{\pi_1}(X, a') \\
&= 0.9 \cdot q_{\pi_1}(X, \text{right}) \approx 3.8769
\end{aligned}$$

(c) In the lecture we defined the policy iteration algorithm to find the optimal policy using value functions. Write down a modified version of policy iteration that finds the optimal policy using action value functions (known as Q-Policy iteration).

Answer:

1. Policy evaluation:
   $q_{k+1}(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi_k(a'|s') q_k(s',a')$
   for all $s \in \mathcal{S}, a \in \mathcal{A}$
2. Policy improvement:
   $\pi_{k+1}(s) = \arg\max_a q_{k+1}(s,a)$
   for all $s \in \mathcal{S}$

3. **Value iteration**

   (a) Perform two steps of value iteration for the MDP from exercise 2 (b), i.e. calculate $v_1(s)$ and $v_2(s)$ for $s \in \{X, Y\}$. Initialize the values with $v_0(X) = 0$ and $v_0(Y) = 0$. You can assume that the value of the terminal state Z is zero in each step.

   Answer:

$$v_{k+1}(s) = \max_a \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_k(s') \right)$$

$$\begin{aligned}
k = 0\!: \quad & v_0(X) = 0 \\
& v_0(Y) = 0 \\
k = 1\!: \quad & v_1(X) = \max_a \Big\{ \underbrace{\mathcal{R}(X, \text{left})}_{=0} + 0.9 \sum_{s'} \mathcal{P}(s'|X, \text{left}) \underbrace{v_0(s')}_{=0}, \\
& \qquad \underbrace{\mathcal{R}(X, \text{right})}_{=0.5} + 0.9 \sum_{s'} \mathcal{P}(s'|X, \text{right}) \underbrace{v_0(s')}_{=0} \Big\} \\
& \quad = \max_a \{0, 0.5\} = 0.5 \\
& v_1(Y) = \max_a \Big\{ \underbrace{\mathcal{R}(Y, \text{left})}_{=2} + 0.9 \sum_{s'} \mathcal{P}(s'|Y, \text{left}) \underbrace{v_0(s')}_{=0}, \\
& \qquad \underbrace{\mathcal{R}(Y, \text{right})}_{=4} + 0.9 \sum_{s'} \mathcal{P}(s'|Y, \text{right}) \underbrace{v_0(s')}_{=0} \Big\} \\
& \quad = \max_a \{2, 4\} = 4 \\
k = 2\!: \quad & v_2(X) = \max_a \Big\{ 0 + 0.9 \cdot 1 \cdot v_1(X),\ 0.5 + 0.9\,(0.75 \cdot v_1(X) + 0.25 \cdot v_1(Y)) \Big\} \\
& \quad = \max_a \{0.45, 1.7375\} = 1.7375 \\
& v_2(Y) = \max_a \Big\{ 2 + 0.9 \cdot 1 \cdot v_1(X),\ 4 + 0.9 \cdot 1 \cdot v_1(Z) \Big\} \\
& \quad = \max_a \{2.45, 4\} = 4
\end{aligned}$$

   (b) Implement value iteration and apply it to the Maze environment from the lecture. Follow the instructions in the Jupyter notebook `value-iteration.ipynb`.