Reinforcement Learning
Prof. Milica Gašić

Release: Thursday, 17 October 2024
Discussion: Wednesday, 23 October 2024

Exercise set #2

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. **Discounted returns**

(a) Assume you observe a sequence of five rewards

$$R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2$$

until you reach a terminal state, i.e. a state that always transitions back to itself with a reward of 0. Calculate the returns $G_0, \ldots, G_5$ for a discount factor of $\gamma = 0.5$.

Answer: From the lecture:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = R_{t+1} + \gamma G_{t+1}$$

The rewards after the terminal state are all zero, i.e. $R_6 = 0, R_7 = 0, \ldots$

$$\begin{aligned} \Rightarrow G_5 &= R_6 + \gamma R_7 + \gamma^2 R_8 + \ldots = 0 \\ G_4 &= R_5 + 0.5 \cdot G_5 = 2 + 0.5 \cdot 0 = 2 \\ G_3 &= R_4 + 0.5 \cdot G_4 = 3 + 0.5 \cdot 2 = 4 \\ G_2 &= R_3 + 0.5 \cdot G_3 = 6 + 0.5 \cdot 4 = 8 \\ G_1 &= R_2 + 0.5 \cdot G_2 = 2 + 0.5 \cdot 8 = 6 \\ G_0 &= R_1 + 0.5 \cdot G_1 = -1 + 0.5 \cdot 6 = 2 \end{aligned}$$

(b) Assume an MDP produces an infinite sequence of rewards of 5, i.e.

$$R_1 = 5, R_2 = 5, R_3 = 5, \ldots$$

Calculate the return G_0 for the discount factors $\gamma \in \{0, 0.2, 0.5, 0.9, 0.95, 0.99, 0.999\}$. What would happen if the discount factor was $\gamma = 1$?

Hint: You can use the closed form of a special case of the power series.

Exercises by Stefan Harmeling, used with permission

Answer:

$$G_0 = \sum_{k=0}^{\infty} \gamma^k R_{0+k+1} = \sum_{k=0}^{\infty} \gamma^k \cdot 5$$

This is a geometric series with ratio γ and coefficient 5.
Since $|\gamma| < 1$, there is a closed form of this series:

$$G_0 = \frac{5}{1-\gamma}$$

Therefore:

$$\begin{aligned}
\gamma = 0 : &\quad G_0 = \frac{5}{1} = 5 & \gamma = 0.95 : &\quad G_0 = \frac{5}{0.05} = 100 \\
\gamma = 0.2 : &\quad G_0 = \frac{5}{0.8} = 6.25 & \gamma = 0.99 : &\quad G_0 = \frac{5}{0.01} = 500 \\
\gamma = 0.5 : &\quad G_0 = \frac{5}{0.5} = 10 & \gamma = 0.999 : &\quad G_0 = \frac{5}{0.001} = 5000 \\
\gamma = 0.9 : &\quad G_0 = \frac{5}{0.1} = 50 & &
\end{aligned}$$

For $\gamma = 1$ the return would diverge to infinity.

(c) Assume you observe a sequence of $T > 1$ rewards

$$R_1 = 0, R_2 = 0, \ \dots, \ R_{T-1} = 0, R_T$$

until you reach a terminal state. Note that all rewards except R_T are zero. How can you choose γ such that the initial return G_0 is equal to $\epsilon > 0$? Calculate these γ for the following situations:

i. $\epsilon = 0.1,\ R_T = 1,\ T = 10$
ii. $\epsilon = 0.1,\ R_T = 1,\ T = 50$
iii. $\epsilon = 0.01,\ R_T = 1,\ T = 50$

Answer:

$$G_0 = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = 0 + \ldots + 0 + \gamma^{T-1} R_T + 0 + \ldots = \gamma^{T-1} R_T$$

$$\begin{aligned}
G_0 &= \epsilon \\
\Leftrightarrow \gamma^{T-1} R_T &= \epsilon \\
\Leftrightarrow \gamma^{T-1} &= \frac{\epsilon}{R_T} \\
\Leftrightarrow \gamma &= \left(\frac{\epsilon}{R_T}\right)^{\frac{1}{T-1}} \quad \text{since } T > 1
\end{aligned}$$

i. $\gamma = \left(\frac{0.1}{1}\right)^{\frac{1}{10-1}} \approx 0.7743$
ii. $\gamma = \left(\frac{0.1}{1}\right)^{\frac{1}{50-1}} \approx 0.9541$
iii. $\gamma = \left(\frac{0.01}{1}\right)^{\frac{1}{50-1}} \approx 0.9103$

2. **Value functions**

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

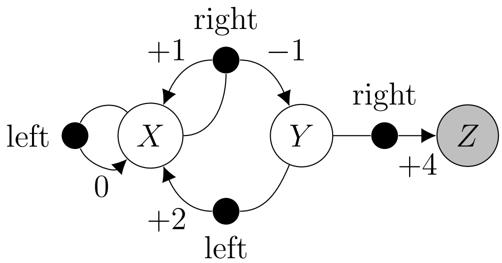
Answer:

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

$$\begin{aligned} v_\pi(E) &= \sum_a \pi(a|E) \left(\underbrace{\mathcal{R}(E,a)}_{=0} + \gamma \underbrace{\mathcal{P}(E|E,a)}_{=1} v_\pi(E) \right) \\ &= \underbrace{\sum_a \pi(a|E)}_{=1} \cdot \gamma v_\pi(E) = \gamma v_\pi(E) \end{aligned}$$

$$\begin{aligned} \Leftrightarrow \qquad && v_\pi(E) &\overset{!}{=} \gamma v_\pi(E) \\ \Leftrightarrow \qquad && (1-\gamma) v_\pi(E) &\overset{!}{=} 0 \\ \Rightarrow \qquad && v_\pi(E) &= 0 \text{ since } \gamma < 1 \end{aligned}$$

(b) Consider the MDP and policy π_1 from the previous exercise set. 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 $\pi_1(X) = \text{right}$, $\pi_1(Y) = \text{right}$.



Calculate the values of states X and Y under policy π_1, i.e. $v_{\pi_1}(X)$ and $v_{\pi_1}(Y)$, using a discount factor of $\gamma = 0.9$.

Hint: Start with the value of Y.

Answer: We convert π_1 to a stochastic policy:

$$\begin{aligned} \pi_1(\text{left}|X) &= 0 & \qquad \pi_1(\text{left}|Y) &= 0 \\ \pi_1(\text{right}|X) &= 1 & \qquad \pi_1(\text{right}|Y) &= 1 \end{aligned}$$

Now we compute the values:

$$\begin{aligned} v_{\pi_1}(Y) &= \sum_a \pi_1(a|Y) \left(\mathcal{R}(Y,a) + 0.9 \cdot \sum_{s'} \mathcal{P}(s'|Y,a)\, v_{\pi_1}(s') \right) \\ &= \underbrace{\pi_1(\text{left}|Y) \cdot \ldots}_{=0} + \underbrace{\pi_1(\text{right}|Y)}_{=1} \left(\underbrace{\mathcal{R}(Y,\text{right})}_{=4} + 0.9 \cdot \sum_{s'} \mathcal{P}(s'|Y,\text{right}) v_{\pi_1}(s') \right) \\ &= 4 + 0.9\,(0 \cdot v_{\pi_1}(X) + 0 \cdot v_{\pi_1}(Y) + 1 \cdot v_{\pi_1}(Z)) \\ &= 4 + 0.9 \cdot \underbrace{v_{\pi_1}(Z)}_{=0} = 4 \end{aligned}$$

$$
\begin{aligned}
v_{\pi_1}(X) &= \sum_a \pi_1(a|X) \left(\mathcal{R}(X,a) + 0.9 \cdot \sum_{s'} \mathcal{P}(s'|X,a)\, v_{\pi_1}(s') \right) \\
&= \underbrace{\pi_1(\text{left}|X) \cdot \ldots}_{=0} + \underbrace{\pi_1(\text{right}|X)}_{=1} \left(\underbrace{\mathcal{R}(X,\text{right})}_{=0.5} + 0.9 \cdot \sum_{s'} \mathcal{P}(s'|X,\text{right}) v_{\pi_1}(s') \right) \\
&= 0.5 + 0.9 \left(0.75 \cdot v_{\pi_1}(X) + 0.25 \cdot \underbrace{v_{\pi_1}(Y)}_{=4} + 0 \cdot v_{\pi_1}(Z) \right) \\
&= 0.5 + 0.9\,(0.75 \cdot v_{\pi_1}(X) + 1) \\
&= 1.4 + 0.675 \cdot v_{\pi_1}(X)
\end{aligned}
$$

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

3. **Policy iteration**

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