Reinforcement Learning

MDP

Christian Sallinger

12.4.2021

Aufgabe 7. Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions?

Lösung. In the MDP framework the probability of each possible value for S_t and R_t depend only on the immediately preeceding state and action and given them not at all on earlier states and action. We can still obtain information from earlier states if we define our states accordingly but here problems might arise. If we were to include some information form earlier states into every preeceding state the amount of memory nedded for just one state would probably soon be too high. In this case it would probably be useful to use some alternatives.
Another example would be when we try to have an agent learn poker. To get good at poker and yield high wins, the agent would have to try to find out with what probability the opponents will bluff. If it is playing against humans, there is no clear way to find this probability. The human component of bluffing would probably make it very hard to fit this problem into the MDP framework. □

Aufgabe 8. Give a table analogous to that in Example 3.3 (texbook p. 52), but for $p(s', r \mid s, a)$. It should have columns for s, a, s', r and $p(s', r \mid s, a)$, and a row for every 4-tuple for which $p(s', r \mid s, a) > 0$.

Lösung. We assume that the number of cans the robot collects while performing the respective action are distributed according to the pmf's $r_{\text{search}}(k), r_{\text{wait}}(k)$ for $k \in \mathbb{N}$. With this we get the table:

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	k	$\alpha \cdot r_{\text{search}}(k)$
high	search	low	k	$(1-\alpha) \cdot r_{\text{search}}(k)$
low	search	low	k	$\beta \cdot r_{\text{search}}(k)$
low	wait	low	k	$r_{\text{wait}}(k)$
high	wait	high$t$	k	$r_{\text{wait}}(k)$
low	search	high	-3	$1-\beta$
low	recharge	high	0	1

□

Aufgabe 9. Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for -1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task?

Lösung. The terminal state of the episode would always be the state where we fail to balance the pole, so we would always get reward 0 an get exactly one reward of -1. All in all we would get

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k = -\gamma^{T-t-1}$$

In the discounted, continuing formulation where we fail at steps $T_1, T_2, \cdots > t$ we would get

$$G_t = -\sum_{j=1}^{\infty} \gamma^{T_j - t - 1}$$

□

Aufgabe 10. Imagine that you are designing a robot to run a maze. You decide to give it a reward of +1 for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes - the successive runs through the maze - so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (1). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \tag{1}$$

Lösung. As we can see in (1), it does not matter at all at which step we get our reward of 1. So we show no improvement because to the robot there is no difference as to when it leaves the maze. We have not effectively communicated to the agent what we want it to achieve, our goal is not to have it just leave the maze, but to leave the maze as fast as possible. One possible solution would be to give the agent a reward of -1 for each step that it does not leave the maze, to incentivice it to leave the maze as soon as possible. Another possible way of doing it would be to use discounting, so it gets less reward the longer it takes. □

Aufgabe 11. Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are $G_0, G_1, \ldots, G_5$? Hint: Work backwards.

Lösung. We follow the hint and note that for the terminal state $T = 5$ we defined $G_T = 0$. Now we use the recursive formula

$$G_t = R_{t+1} + \gamma G_{t+1}.$$

So we end up with

$$\begin{aligned}
G_5 &= 0 \\
G_4 &= 2 + 0 = 2 \\
G_3 &= 3 + \frac{1}{2} \cdot 2 = 4 \\
G_2 &= 6 + \frac{1}{2} \cdot 4 = 8 \\
G_1 &= 2 + \frac{1}{2} \cdot 8 = 6 \\
G_0 &= -1 + \frac{1}{2} \cdot 6 = 2
\end{aligned}$$

□

Aufgabe 12. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0?

Lösung. We use the formula (4) and the geometric series to compute the values:

$$\begin{aligned}
G_1 &= \sum_{k=0}^{\infty} \gamma^k R_{k+2} = 7 \cdot \sum_{k=0}^{\infty} \gamma^k = 7 \cdot \frac{1}{1/10} = 70 \\
G_0 &= 2 + 7 \cdot \sum_{k=1}^{\infty} \gamma^k = 2 + 7 \cdot \left(\sum_{k=0}^{\infty} \gamma^k - 1 \right) = 2 + 7 \cdot 9 = 65
\end{aligned}$$

□

Aufgabe 13. If the current state is S_t, and actions are selected according to stochastic policy π, then what is the expectation of R_{t+1} in terms of π and the four-argument function p (2)?

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \tag{2}$$

Lösung. We first remind of the definition of the expected rewards for state-action pairs:

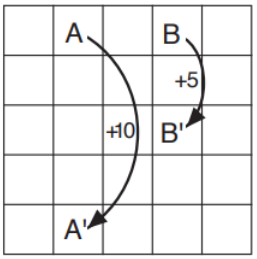
$$r(s \mid a) = \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$$

With this and the total law of expectation we show

$$\mathbb{E}_\pi[R_{t+1} \mid S_t] = \sum_{a \in \mathcal{A}} \pi(a \mid S_t) \mathbb{E}[R_{t+1} \mid S_t, A_t = a] = \sum_{a \in \mathcal{A}} \pi(a \mid S_t) r(S_t \mid a) = \sum_{a \in \mathcal{A}} \sum_{r \in \mathcal{R}} r\pi(a \mid S_t) \sum_{s' \in \mathcal{S}} p(s', r \mid S_t, a)$$

□

Aufgabe 14. The Bellman equation (3) must hold for each state for the value function v_π shown in Figure 1 (right). Show numerically that this equation holds for the center state, valued at $+0.7$, with respect to its four neighboring states, valued at $+2.3$, $+0.4$, -0.4, and $+0.7$. (These numbers are accurate only to one decimal place.)



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Abbildung 1: Gridworld example: exceptional reward dynamics (left) and state-value function for the equiprobable random policy (right).

$$v_\pi(s) \doteq \sum_a \pi(a \mid s) \sum_{s', r} p(s', r \mid s, a)[r + \gamma v_\pi(s')], \qquad \text{for all } s \in \mathcal{S} \tag{3}$$

Lösung. We give the the states the names m, u, d, l, r (for mid, up, down, left, right with respect to the state valued at 0.7) and the actions $north, east, south, west$. According to the policy $\pi(a \mid m) = 0.25$ for all the actions a. Once we take an action the four argument function $p(s', 0 \mid m, a) = 1$ for exactly one pair of (s', a) and 0 otherwise. If we omit all the pairs with probability 0 we get:

$$\begin{aligned}
&\pi(north \mid m)p(u, 0 \mid m, north)[0 + \gamma v_\pi(u)] + \pi(east \mid m)p(l, 0 \mid m, east)[0 + \gamma v_\pi(l)] \\
&+\pi(south \mid m)p(d, 0 \mid m, south)[0 + \gamma v_\pi(d)] + \pi(west \mid m)p(r, 0 \mid m, west)[0 + \gamma v_\pi(r)] \\
&= 0.25 \cdot \big([0.9 \cdot 2.3] + [0.9 \cdot 0.4] - [0.9 \cdot 0.4] + [0.9 \cdot 0.7]\big) = 0.675 \approx 0.7 = v_\pi(m)
\end{aligned}$$

□

Aufgabe 15. In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals

between them? Prove, using (4), that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies. What is v_c in terms of c and γ?

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{4}$$

Lösung. In the following we write v_π^+ for the state value where we added a constant and v_π for the state value function where we have not added the constant. By use the definition of v_π as well as basic properties of the expectation we get

$$\begin{aligned} v_\pi^+(s) &= \mathbb{E}_\pi\Big[G_t^+ \big| S_t = s\Big] = \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \big| S_t = s\Big] \\ &= \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c \big| S_t = s\Big] = \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \big| S_t = s\Big] + \mathbb{E}_\pi\Big[\frac{c}{1-\gamma} \big| S_t = s\Big] \\ &= v_\pi(s) + \frac{c}{1-\gamma} \qquad \forall s \in \mathcal{S}. \end{aligned}$$

So we get

$$v_\pi^+(s) = v_\pi(s) + v_c, \qquad \forall s \in \mathcal{S}$$

with $v_c = \frac{c}{1-\gamma}$. □

Aufgabe 16. Now consider adding a constant c to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.

Lösung. This would change the task at hand. We write G_t^+ for the expected return where we add the constant to all the rewards and G_t for the expected return without the constants. Without discounts we would get:

$$G_t^+ = \sum_{k=t+1}^{T} (R_k + c) = G_t + (T - t - 1) \cdot c$$

With discounts:

$$G_t^+ = \sum_{k=t+1}^{T} \gamma^{k-t-1}(R_k + c) = G_t + c \sum_{k=t+1}^{T} \gamma^{k-t-1} = G_t + c\frac{1-\gamma^{T-t}}{1-\gamma}$$

In both cases we see that the added constant to the reward is not constant when considering the expected return (and with that also v_π) but instead is depenend both on the terminal state of the episode as well as the current timestep.
An example where this would have a heavy impact would be the maze example: If we would give a reward of -1 in every timestep the agent does not leave the maze and a reward of 0 in the terminal state where we leave the state and then add 1 to the rewards, we would end up with the same problem as in exercise 10. □

Aufgabe 17. What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s_0, a_0)$, of possible successors to the state-action pair (s, a). Hint: the backup diagram to the right corresponds to this equation. Show the sequence of equations analogous to (4), but for action values.

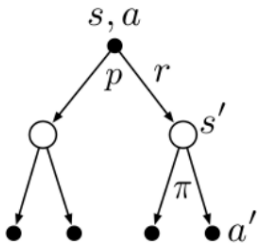


Abbildung 2: q_π backup diagram

Lösung. We use the law of total expectation and get:

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

□

Aufgabe 18. The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action:

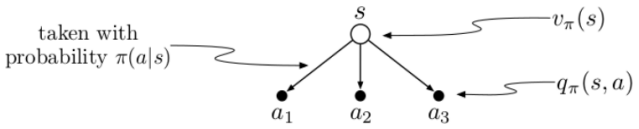


Abbildung 3

Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s,a)$, given $S_t = s$. This equation should include an expectation conditioned on following the policy, π. Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a \mid s)$ such that no expected value notation appears in the equation.

Lösung. We use the law of total expectation and write

$$v_\pi(s) = \mathbb{E}_\pi\big[G_t \mid S_t = s\big] = \sum_{a\in\mathcal{A}}\pi(a \mid s)\mathbb{E}_\pi\big[G_t \mid S_t = s, A_t = a\big] = \sum_{a\in\mathcal{A}}\pi(a \mid s)q_\pi(s,a)$$

□

Aufgabe 19. The value of an action, $q_\pi(s,a)$, depends on the expected next reward and the expected sum of the remaining rewards. Again we can think of this in terms of a small backup diagram, this one rooted at an action (state-action pair) and branching to the possible next states:

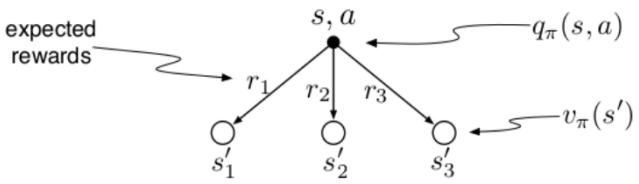


Abbildung 4

Give the equation corresponding to this intuition and diagram for the action value, $q_\pi(s, a)$, in terms of the expected next reward, R_{t+1} , and the expected next state value, $v_\pi(S_{t+1})$, given that $S_t = s$ and $A_t = a$. This equation should include an expectation but not one conditioned on following the policy. Then give a second equation, writing out the expected value explicitly in terms of $p(s_0, r \mid s, a)$ defined by (2), such that no expected value notation appears in the equation.

Lösung. With the law of total expectation we get:

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

□

Aufgabe 20. Consider the continuing MDP shown below. The only decision to be made is that in the top state, where two actions are available, **left** and **right**. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, $\pi_{\textbf{left}}$ and $\pi_{\textbf{right}}$. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

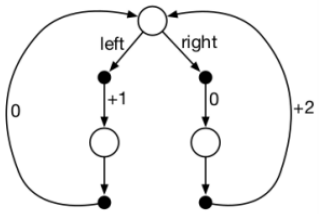


Abbildung 5

Lösung. 1. $\gamma = 0$: Since we discard all future rewards and only consider the immediate reward so we get

$$v_{\pi_{\textbf{right}}}(\text{top}) = 0 \qquad v_{\pi_{\textbf{left}}}(\text{top}) = 1$$

In both bottom states we get the same reward regardless of the policy we follow, so in this case the optimal policy would be $\pi_{\textbf{left}}$.
2. $\gamma = 0.9$: We call the bottom states left and right and with the Bellman equation (3) we get

$$\begin{aligned}
v_{\pi_{\textbf{right}}}(\text{top}) = 0.9 \cdot v_{\pi_{\textbf{right}}}(\text{right}) = 0.9 \cdot (2 + 0.9 \cdot v_{\pi_{\textbf{right}}}(\text{top})) \implies v_{\pi_{\textbf{right}}}(\text{top}) = \frac{180}{19} \\
v_{\pi_{\textbf{left}}}(\text{top}) = 1 + (0.9 \cdot v_{\pi_{\textbf{left}}}(\text{left})) = 1 + 0.81 \cdot v_{\pi_{\textbf{left}}}(\text{top}) \implies v_{\pi_{\textbf{left}}}(\text{top}) = \frac{100}{19}
\end{aligned}$$

For the bottom states we get:

$$
\begin{aligned}
v_{\pi_{\mathbf{right}}}(\text{right}) &= 2 + 0.9 \cdot v_{\pi_{\mathbf{right}}}(\text{top}) = 10 \\
v_{\pi_{\mathbf{right}}}(\text{left}) &= 0.9 \cdot v_{\pi_{\mathbf{right}}}(\text{top}) = \frac{162}{19} \\
v_{\pi_{\mathbf{left}}}(\text{right}) &= 2 + 0.9 \cdot v_{\pi_{\mathbf{left}}}(\text{top}) = \frac{118}{19} \\
v_{\pi_{\mathbf{left}}}(\text{left}) &= 0.9 \cdot v_{\pi_{\mathbf{left}}}(\text{top}) = \frac{90}{19}
\end{aligned}
$$

So in this case the optimal policy would be $\pi_{\mathbf{right}}$.
3. $\gamma = 0.5$: With the Bellman equation (3) we get:

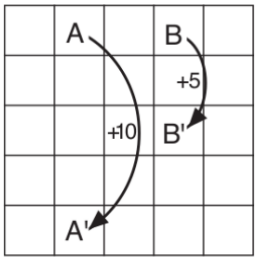
$$
\begin{aligned}
v_{\pi_{\mathbf{right}}}(\text{top}) = 0.5 \cdot v_{\pi_{\mathbf{right}}}(\text{bottom}) = 0.5 \cdot (2 + 0.5 \cdot v_{\pi_{\mathbf{right}}}(\text{top})) &\implies v_{\pi_{\mathbf{right}}}(\text{top}) = \frac{4}{3} \\
v_{\pi_{\mathbf{left}}}(\text{top}) = 1 + (0.5 \cdot v_{\pi_{\mathbf{left}}}(\text{bottom})) = 1 + 0.25 \cdot v_{\pi_{\mathbf{left}}}(\text{top}) &\implies v_{\pi_{\mathbf{left}}}(\text{top}) = \frac{4}{3}
\end{aligned}
$$

For the bottom states we get:

$$
\begin{aligned}
v_{\pi_{\mathbf{right}}}(\text{right}) &= 2 + 0.5 \cdot v_{\pi_{\mathbf{right}}}(\text{top}) = \frac{8}{3} \\
v_{\pi_{\mathbf{right}}}(\text{left}) &= 0.5 \cdot v_{\pi_{\mathbf{right}}}(\text{top}) = \frac{2}{3} \\
v_{\pi_{\mathbf{left}}}(\text{right}) &= 2 + 0.5 \cdot v_{\pi_{\mathbf{left}}}(\text{top}) = \frac{8}{3} \\
v_{\pi_{\mathbf{left}}}(\text{left}) &= 0.5 \cdot v_{\pi_{\mathbf{left}}}(\text{top}) = \frac{2}{3}
\end{aligned}
$$

So in this case both policies are optimal. □

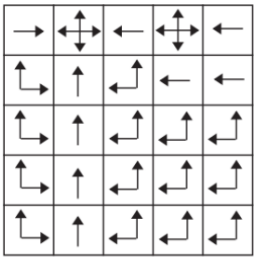
Aufgabe 21. Figure 6 (below) gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (4) to express this value symbolically, and then to compute it to three decimal places.



(a) Gridworld

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

(b) v_*



(c) π_*

Abbildung 6: Optimal solutions to the gridworld example.

Lösung. Once we are in the best state A and follow the optimal policy, we get a reward of 10 for moving out of the state and then move up again, getting no reward for 4 steps until we are in state A again. With this we compute

$$v_*(A) = \mathbb{E}_{\pi_*}\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \mid S_t = A\right] = \sum_{k=0}^{\infty}\gamma^k \mathbb{E}_{\pi_*}\left[R_{t+k+1} \mid S_t = A\right]$$
$$= \sum_{k=0}^{\infty}\gamma^k \mathbb{1}_{\{k=0 \mod 5\}} \cdot 10 = 10 \cdot \sum_{k=0}^{\infty}\left(\gamma^5\right)^k = \frac{10}{1-(0.9)^5} \approx 24,4194$$

□