

Spring 2020

Exercises (Solutions)

An Introduction to Reinforcement Learning

Author
Kurath Samuel

Contents

1	Reinforcement Learning	2
1.1	Übung 1	2
1.1.1	Episode	2
1.1.2	Reward	3
1.1.3	Discounted Reward	3
1.1.4	State-Value-Function	3

1 Reinforcement Learning

Ausgangslage Die Übungen zu Reinforcement Learning basieren auf dem Markov Decision Process zu **Kurs absolvieren**, welcher in Abbildung 1.1) illustriert ist. Gestartet wird im Zustand *Attend Cours*. *Fail Cours* und *Success Cours* sind beides Terminalzustände.

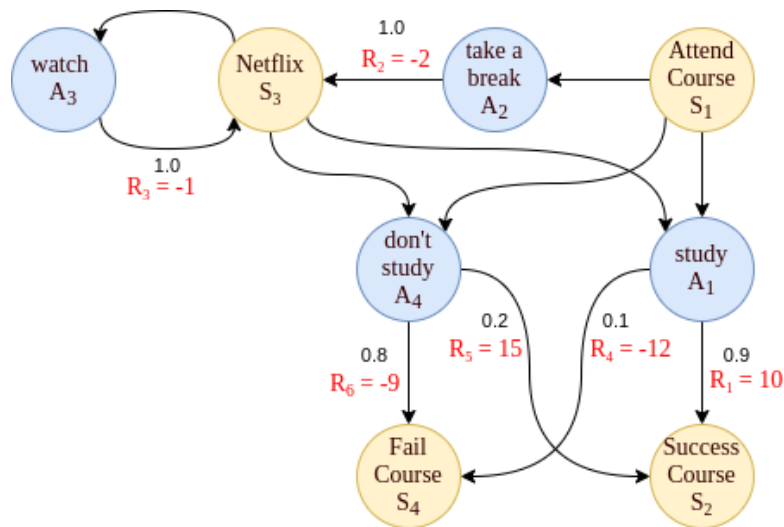


Figure 1.1: Markov Decision Process - Kurs absolvieren

1.1 Übung 1

1.1.1 Episode

Welcher der folgenden **Episoden** sind valide in Bezug auf **Kurs absolvieren**?

- a) S_1, A_1, R_1, S_2 (richtig)
- b) $S_1, A_2, R_2, S_3, A_3, R_3, S_3, A_4, R_6, S_4$ (richtig)
- c) $S_1, A_2, R_2, S_3, A_3, R_3, S_3, R_{15}, S_2$ (falsch)
- d) S_1, A_4, R_4, S_4 (falsch)
- e) $S_1, A_2, R_2, S_3, A_3, R_3, S_3, A_3, R_3, S_3$ (falsch)

1.1.2 Reward

Welchen **Reward** erhalten wir in der Episode (1.1)?

$$S_1, A_2, R_2, S_3, A_3, R_3, S_3, A_3, R_3, S_3, A_4, R_5, S_2 \quad (1.1)$$

(Lösung)

$$R_{total} = R_2 + R_3 + R_3 + R_5 = -2 + -1 + -1 + 15 = 11 \quad (1.2)$$

1.1.3 Discounted Reward

Angenommen wir gehen von einem Discount Factor von $\gamma = 0.8$ aus und befinden uns im Startzustand S_1 ($t = 0$). Was für ein **Discounted Reward** resultiert in Bezug der Episode (1.1)?

(Lösung)

$$R_{discount} = \gamma^0 R_2 + \gamma^1 R_3 + \gamma^2 R_3 + \gamma^3 R_5 = 0.8^0 \cdot (-2) + 0.8^1 \cdot (-1) + 0.8^2 \cdot (-1) + 0.8^3 \cdot 15 = 4.24 \quad (1.3)$$

1.1.4 State-Value-Function

a) Bestimme den Erwartungswert im Zustand S_1 unter Bezug der Policy (1.4) aka. "Musterstudent" und $\gamma = 0.8$.

$$\pi : \left\{ S_1 \rightarrow A_1 \right. \quad (1.4)$$

(Lösung)

$$\begin{aligned} E^\pi(S_1) &= 0.1 \cdot (-12) + 0.9 \cdot 10 = 7.8 \\ E^\pi(S_{terminal}) &= 0 \\ V^\pi(S_1) &= \gamma^0 \cdot E^\pi(S_1) + \gamma^1 \cdot E^\pi(S_2) = 0.8^0 \cdot 7.8 + 0.8^1 \cdot 0 = 7.8 \end{aligned} \quad (1.5)$$

b) Nun ändern wir unser Verhalten vom "Musterstudent" zum "Lazy guy". Dies bedeutet, dass wir die Policy (1.6) wählen, die restlichen Parameter bleiben gleich. (Zustand S_1 , $\gamma = 0.8$) Wie verändert sich der Erwartungswert?

$$\pi : \left\{ \begin{array}{l} S_1 \rightarrow A_2 \\ S_3 \rightarrow A_4 \end{array} \right. \quad (1.6)$$

(Lösung)

$$\begin{aligned} E^\pi(S_1) &= 1.0 \cdot (-2) = -2 \\ E^\pi(S_3) &= 0.8 \cdot (-9) + 0.2 \cdot 15 = -4.2 \\ E^\pi(S_{terminal}) &= 0 \\ V^\pi(S_1) &= \gamma^0 \cdot E^\pi(S_1) + \gamma^1 \cdot E^\pi(S_3) = 0.8^0 \cdot -2 + 0.8^1 \cdot -4.2 + 0.8^2 \cdot 0 = -5.36 \end{aligned} \quad (1.7)$$