Prof. Joschka Boedecker, Gabriel Kalweit, Maria Huegle, Andreas Saelinger

Reinforcement Learning
Solution 4

1 Monte Carlo Prediction

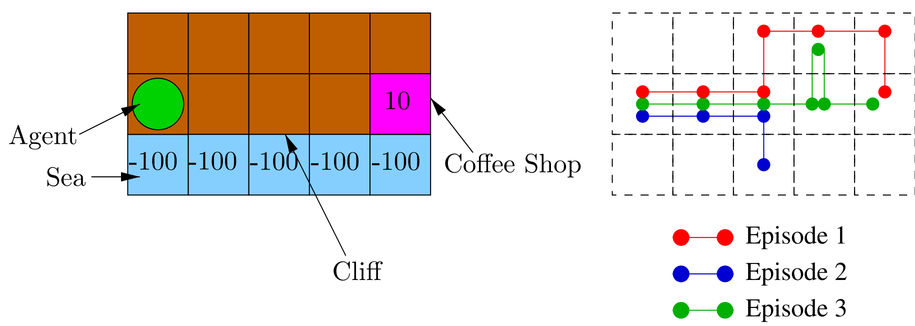

Figure 1: Cliff MDP

Consider the MDP in Figure 1, where all actions (an action moves the agent in a desired direction: up, down, left or right) succeed with a probability of 0.8. With a probability of 0.2 the agent moves randomly in another direction. All transitions result in a reward of -1, except when the coffee shop is reached (terminal state $s_{2,5}$: reward of 10) or if the agent falls of the cliff (terminal states $s_{3,1} \dots s_{3,5}$: reward of -100). The agent always starts in state $s_{2,1}$ as indicated in Figure 1.

Using Monte-Carlo policy evaluation, calculate $V(i)$ for all states i based on the illustrated episodes 1 to 3 (right part of Figure 1). Use the first-visit-method, i.e. every state is updated only once – on the first-visit – per episode, even if the state is visited again during the episode. In this task, we estimate the value by a running mean with $\alpha_t = \frac{1}{t}$ for episode t and initialize $V(i) = 0$ for all i. We do not discount, i.e. $\gamma = 1$.

Solution. We have to iteratively calculate the different returns for the states of a trajectory and then update our estimation. Let G_s^t denote the return in episode t starting from state s.

For trajectory 1:

- $G_0^1 = -1 - 1 - 1 - 1 - 1 + 10 = 5$
- $G_1^1 = -1 - 1 - 1 - 1 + 10 = 6$
- $G_2^1 = -1 - 1 - 1 + 10 = 7$

- $G_3^1 = -1 - 1 + 10 = 8$
- $G_4^1 = -1 + 10 = 9$
- $G_5^1 = 10$

Following MC-policy evaluation, we update by $V_{t+1}(s) = V_t(s) + \alpha_t(G_s^{t+1} - V_t(s))$ and $\alpha_1 = \frac{1}{1} = 1$, we get for V_1:

- $V_1(s) = V_0(s)$ for all states s that are not visited on this trajectory, i.e. for $s \in \{s_{1,1}, s_{1,2}, s_{2,4}, s_{3,k}\}$ with $1 \leq k \leq 5$
- $V_1(s_{2,1}) = 0 + 1(5 - 0) = 5$
- $V_1(s_{2,2}) = 0 + 1(6 - 0) = 6$
- $V_1(s_{2,3}) = 0 + 1(7 - 0) = 7$
- $V_1(s_{1,3}) = 0 + 1(8 - 0) = 8$
- $V_1(s_{1,4}) = 0 + 1(9 - 0) = 9$
- $V_1(s_{1,5}) = 0 + 1(10 - 0) = 10$

For trajectory 2:

- $G_0^2 = -1 - 1 - 100 = -102$
- $G_1^2 = -1 - 100 = -101$
- $G_2^2 = -100$

With $\alpha_2 = \frac{1}{2}$, we get for V_2:

- $V_2(s) = V_1(s)$ for all states that are not visited on this trajectory, i.e. for:

$$s \in \{s_{1,1}, s_{1,2}, s_{1,3}, s_{1,4}, s_{1,5}, s_{2,4}, s_{2,5}, s_{3,1}, s_{3,2}, s_{3,4}, s_{3,5}\}$$

- $V_2(s_{2,1}) = 5 + \frac{1}{2}(-102 - 5) = 5 - 53\frac{1}{2} = -48\frac{1}{2}$
- $V_2(s_{2,2}) = 6 + \frac{1}{2}(-101 - 6) = 6 - 53\frac{1}{2} = -47\frac{1}{2}$
- $V_2(s_{2,3}) = 7 + \frac{1}{2}(-100 - 7) = 7 - 53\frac{1}{2} = -46\frac{1}{2}$

For trajectory 3:

- $G_0^3 = -1 - 1 - 1 - 1 - 1 + 10 = 5$
- $G_1^3 = -1 - 1 - 1 - 1 + 10 = 6$
- $G_2^3 = -1 - 1 - 1 + 10 = 7$
- $G_3^3 = -1 - 1 + 10 = 8$
- $G_4^3 = -1 + 10 = 9$

- $G_5^3 = G_3^3$ **calculated on first-visit (see above)**

With $\alpha_3 = \frac{1}{3}$, we get for V_3:

- $V_3(s_{2,1}) = -48\frac{1}{2} + \frac{1}{3}(5 - (-48\frac{1}{2})) = -48\frac{1}{2} + 17\frac{5}{6} = -30\frac{2}{3}$
- $V_3(s_{2,2}) = -47\frac{1}{2} + \frac{1}{3}(6 - (-47\frac{1}{2})) = -47\frac{1}{2} + 17\frac{5}{6} = -29\frac{2}{3}$
- $V_3(s_{2,3}) = -46\frac{1}{2} + \frac{1}{3}(7 - (-46\frac{1}{2})) = -46\frac{1}{2} + 17\frac{5}{6} = -28\frac{2}{3}$
- $V_3(s_{2,4}) = 0 + \frac{1}{3}(8 - 0) = 0 + 2\frac{2}{3} = 2\frac{2}{3}$
- $V_3(s_{1,4}) = 9 + \frac{1}{3}(9 - 9) = 9$
- $V_3(s_{2,4})$ **updated on first-visit (see above)**