Reinforcement Learning (SS18) - Exercise 2

Daniel Hennes

02.05.2018 (due 09.05.2018)

1. For k-armed bandits, we defined the value as:

$$q(a) = \mathbb{E}[R_t \mid A_t = t]$$

For MDPs, the state-action value is defined as follow:

$$q(s, a) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \dots \mid S_t = s, A_t = a]$$

Argue why we do not need to consider future rewards in the bandit setting.

2. Show that $v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a)$.

3. We introduced the *Bellman equation* for v_π in terms of the four-argument function p. Express the recursive relationship of v_π in terms of $p(s' \mid s, a)$ and $r(s, a, s')$.

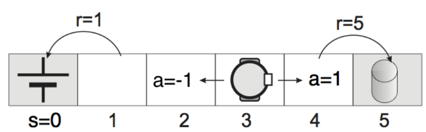

Figure 1: Cleaning robot

Consider the following problem: A cleaning robot has to collect cans and also recharge its batteries. The robot can move left ($a = -1$) or right ($a = 1$) and is in one of 6 distinct states at all times. State transitions are deterministic. Non–zero rewards are only received for transitions into the far–left or far–right states as indicated in the figure above.

4. Formulate the problem as a MDP.

5. Calculate the optimal value function v_* for the cleaning robot problem.