Prof. Joschka Boedecker, Gabriel Kalweit, Maria Huegle

Reinforcement Learning
Solution 2

1 Markov Decision Processes

Assume the following problem: there are 5 parking spaces and you start at parking space 5. In each step, you can either try to park or drive on. A parking space is free with probability ρ. If a parking space i was occupied or you drove on, you move to the next parking space $i-1$. You want to be as close to your home – which is at parking space 1 – as possible. However, you want to avoid to reach the end of parking spaces without parking successfully.

(a) Formalize the above problem as an MDP.

Solution. An MDP is a 4-tuple $\langle \mathcal{S}, \mathcal{A}, p, \mathcal{R} \rangle$.

The set of states consists of states for each parking space and two additional terminal states where we transition in for parking successfully or for not finding a parking space, i.e. $\mathcal{S} = \{5, 4, 3, 2, 1, S, F\}$.

The set of actions is $\mathcal{A} = \{\text{park}, \text{drive on}\}$.

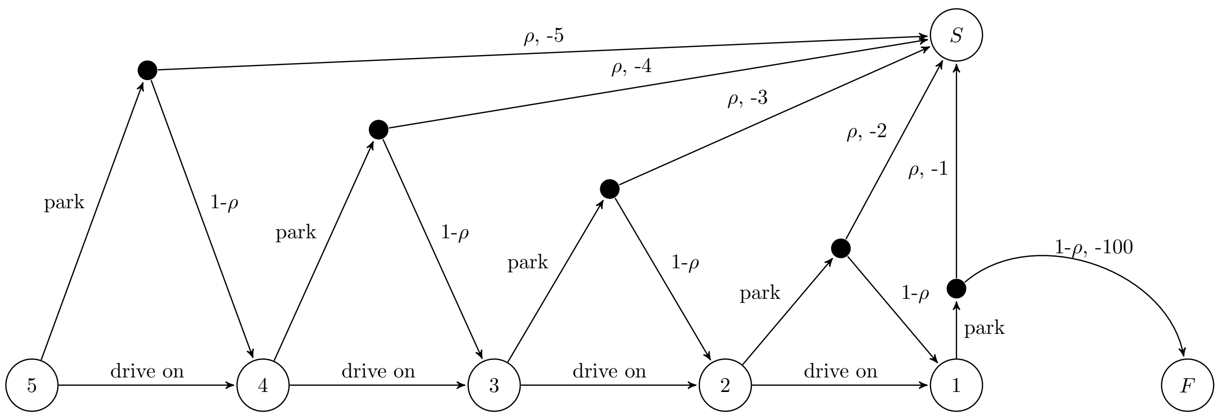
The set of rewards is $\mathcal{R} = \{0, -100\} \cup \{-i | 1 \leq i \leq 5\}$.

Hence, the transition probabilities are:

- $p(S, -i | i, \text{park}) = \rho$
- $p(F, -100 | 1, \text{park}) = (1 - \rho)$
- $p(i-1, 0 | i, \text{park})_{5 \geq i > 1} = (1 - \rho)$
- $p(i-1, 0 | i, \text{drive on})_{5 \geq i > 1} = 1$

(b) Draw the transition graph.

Solution. The transition graph is:



(c) Do we have to discount? Explain your answer.

Solution. We do not have to discount, since all paths in this MDP lead to a fixed terminal state.

2 Markov Property

Assume a biased slot machine in a casino. Each round, the player can win 1\$. However, whenever the outcome of the last two rounds is larger than 1\$, the machine lowers the probability of winning. Is the Markov property fulfilled?

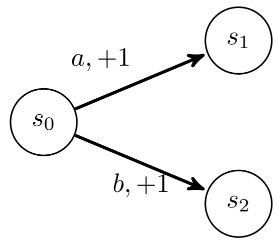
Solution. No, the Markov property is not fulfilled, since the transition probability of some state-reward pair depends on the last two former states. Therefore:

$$Pr\{S_{t+1}, R_{t+1}|S_t, A_t\} \neq Pr\{S_{t+1}, R_{t+1}|S_t, A_t, \ldots, S_0, A_0\}.$$

3 Optimal Value Function

(a) Disprove by counter example: for any MDP with optimal value function v_*, the optimal deterministic policy π_* is unique.

Solution. The statement is not true. Assume states s_0, s_1 and s_2. Further assume two actions a and b which can be applied in s_0 and which have a return of $+1$. Therefore both actions, a and b, are optimal. Hence, the optimal policy is not unique.



(b) Disprove by counter example: the optimal value function v_* for state s_t at time step t is always larger than for state s_{t+1} under the optimal deterministic policy π_*, i.e. $v_*(s_t) > v_*(s_{t+1})$.

Solution. The statement is not true. Assume state s_1 with $v_*(s_1) = 4$ and state s_2 with $v_*(s_2) = 2$ (we can easily define parts of an MDP s.t. these assumptions hold). Further assume the immediate reward function R in state s_0 to be $R(s_0, a) = -1$ and $R(s_0, b) = +0$. If we set the transition probabilities p for reaching s_1 and s_2 to $p(s_0, a, s_1) = 1$ and $p(s_0, b, s_2) = 1$, then the statement does not hold.