7 Markov Decision Processes

Exercise 7.1

Tell if the following statements about MDPs are true or false. Motivate your answers.

1. To solve an MDP we should take into account state-action pairs one by one;
2. An action you take on an MDP might influence the future rewards you gained;
3. A state considered by the agent acting on an MDP is always equal to the environment state;
4. Problems in which an agent knows the state of the environment and the MDP do not require the use of RL;
5. Policies applied to an MDP are influenced by other learning processes ongoing on the considered MDP.

Exercise 7.2

Tell whether the following statements about MDPs are true or false. Motivate your answers.

1. The Policy Evaluation procedure always outputs the optimal value function.
2. The value function may decrease on some steps of Policy Iteration, but in the end, the algorithm outputs the optimal one.
3. Employing a discount factor in the computation of the cumulative return in MDPs is only a mathematical trick to ensure the convergence of the return.
4. Given an MDP with a certain reward function, there is only a single policy that is optimal for it, and, for each optimal policy, there is only a single reward function for which it is optimal.
5. As many policies can be optimal, there can be multiple optimal value functions in an MDP.
6. All the sequential decision problems can be modeled as MDPs.

Exercise 7.3

State if the following applications may be modeled by means of an MDP:

1. Robotic navigation in a grid world;
2. Stock Investment;
3. Robotic soccer;
4. Playing Carcassonne (board game).

Define the possible actions and states of each MDP you considered.

Exercise 7.4

Consider the following modeling of a classification problem as sequential decision making problem:

$$
\begin{aligned}
o_i &\leftarrow x_i \\
a_i &\leftarrow \hat{y}_i \\
r_i &\leftarrow 1 - |t_i - \hat{y}_i|
\end{aligned}
$$

Does this correspondence makes sense? Comment adequately your answer.

Exercise 7.5

For each one of the following dichotomies in MDP modeling provide examples of problems with the listed characteristics:

1. Finite/infinite actions;
2. Deterministic/stochastic transitions;
3. Deterministic/stochastic rewards;
4. Finite/indefinite/infinite horizon;
5. Stationary/non–stationary environment.

Exercise 7.6

Are the following statements about the discount factor γ in a MDP correct?

- A myopic learner corresponds to have low γ values in the definition of the MDP;

- In an infinite horizon MDP we should avoid using $\gamma = 1$, while it is reasonable if the horizon is finite;
- γ is an hyper-parameter for the policy learning algorithm;
- Probability that an MDP will be played in the next round is γ.

Provide adequate motivations for your answers.

Exercise 7.7

The generic definition of policy is a stochastic function $\pi(h_i) = \mathbb{P}(a_i|h_i)$ which given a history $h_i = \{o_1, a_1, s_1, \ldots, o_i, a_i, s_i\}$ provides a distribution over the possible actions $\{a_i\}_i$.

Formulate the specific definition of a policy if the considered problem is:

1. Markovian, Stochastic, Non-stationary
2. History based, Stochastic, Stationary
3. Markovian, Deterministic, Stationary

Exercise 7.8

Comment the following statements about solving MDPs. Motivate your answers.

1. In a finite state MDP we just need to look for Markovian, stationary and deterministic optimal policies;
2. For finite time MDPs we should consider non-stationary optimal policies;
3. The results of coupling a specific policy and an MDP is a Markov process;
4. Given a policy we can compute P^π and R^π on an MDP;
5. The value function $V^{\pi^*}(s)$ contains all the information to execute the optimal policy π^* on a given MDP;
6. The action-value function $Q^{\pi^*}(s, a)$ contains all the information to execute the optimal policy π^* on a given MDP;
7. There is a unique optimal policy in an MDP;
8. There is a unique optimal value function in an MDP.

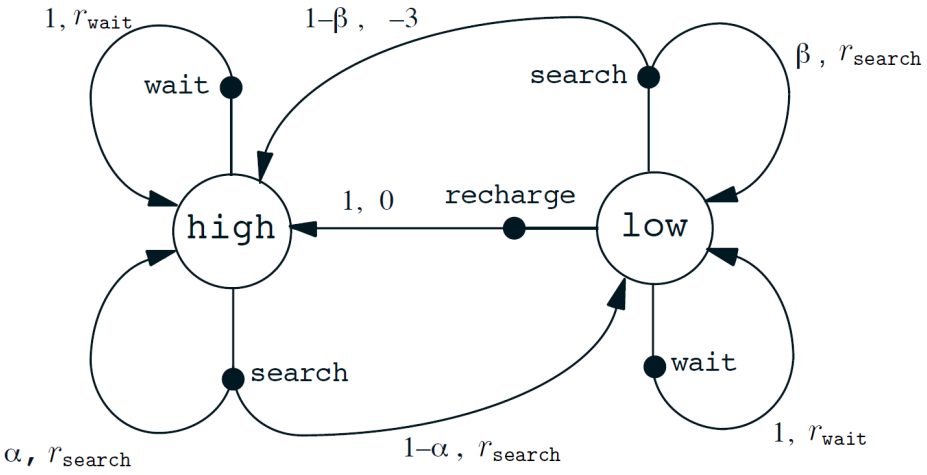

Figure 7.1: The MDP for the cleaning robot problem.

Exercise 7.9

Consider the MDP in Figure 7.1 with $\alpha = 0.3$, $\beta = 0.5$, $\gamma = 1$, $r_{search} = 2$, $r_{wait} = 0$ and the following policy:

$$\pi(s|H) = 1$$
$$\pi(s|L) = 0.5$$
$$\pi(r|L) = 0.5$$

Compute the Value function where the MDP stops after two steps. What happens if we consider a discount factor of $\gamma = 0.5$.

Compute the action-value function for each action value pair in the case the MDP stops after a single step.

Exercise 7.10

Provide the formulation of the Bellman expectation for V equations for the MDP in Figure 7.1, with $\alpha = 0.2$, $\beta = 0.1$, $r_{search} = 2$, $r_{wait} = 0$, $\gamma = 0.9$ and in the case we consider the policy:

$$\pi(H|s) = 1$$
$$\pi(L|r) = 1$$

Exercise 7.11

Tell if the following statements are TRUE or FALSE. Motivate your answers.

1. We are assured to converge to a solution when we apply repeatedly the Bellman expectation operator;
2. We are assured to converge to a solution when we apply repeatedly the Bellman optimality operator;
3. The Bellman solution to Bellman expectation operator is always a good choice to compute the value function for an MDP;
4. The solution provided by the iterative use of the Bellman expectation operator is always less expensive than computing the exact solution using the Bellman expectation equation;
5. The application of the Bellman optimality operator 10 times applied to a generic value function V_0 guarantees that $||V^* - T^{10}V_0||_\infty \leq \gamma^{10}||V^* - V_0||_\infty$

Exercise 7.12

Which one would you chose between the use of the Bellman recursive equation vs. Bellman exact solution in the case we are considering the following problems:

1. Chess
2. Cleaning robot problem in Figure 7.1
3. Maze escape
4. Tic-tac-toe

Provide adequate motivations for your answers.

Exercise 7.13

Consider the MDP in Figure 7.2:

1. Provide the transition matrix for the policy $\pi(I|s_1) = 1, \pi(M|s_2) = 1, \pi(M|s_3) = 1$;
2. Provide the expected instantaneous reward for the previous policy;
3. Compute the value function for the previous policy in the case the MDP stops after two steps;

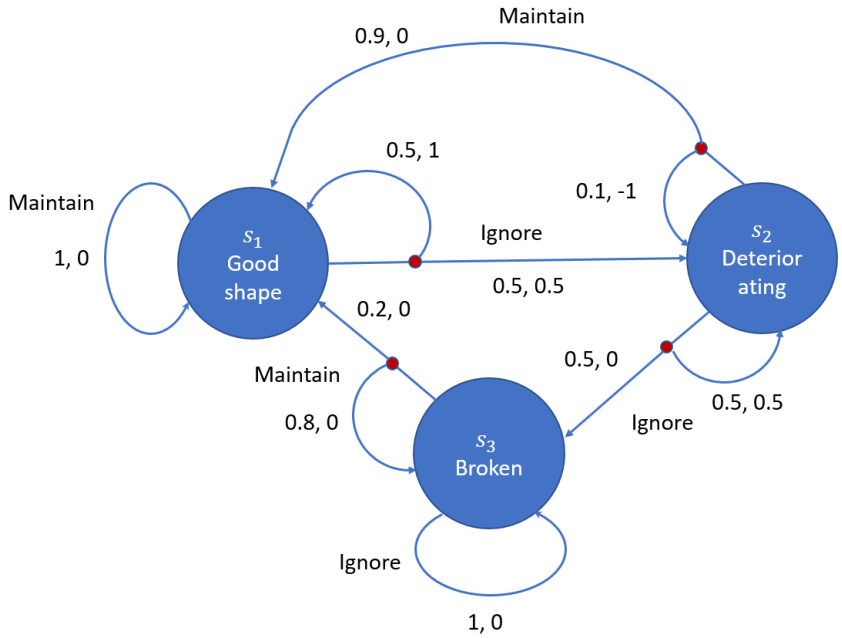

Figure 7.2: The MDP for machinery maintenance problem.

4. Compute the action-value function for each state-action pair in the case the MDP stops after a single step.