8 Markov Decision Processes

Exercise 8.1

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 8.2

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 8.3

Consider the following modeling of a classification problem as sequential decision

making problem:

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

Does this correspondence makes sense? Comment adequately your answer.

Exercise 8.4

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 8.5

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;
- The probability that an MDP will be played in the next round is γ.

Provide adequate motivations for your answers.

Exercise 8.6

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:

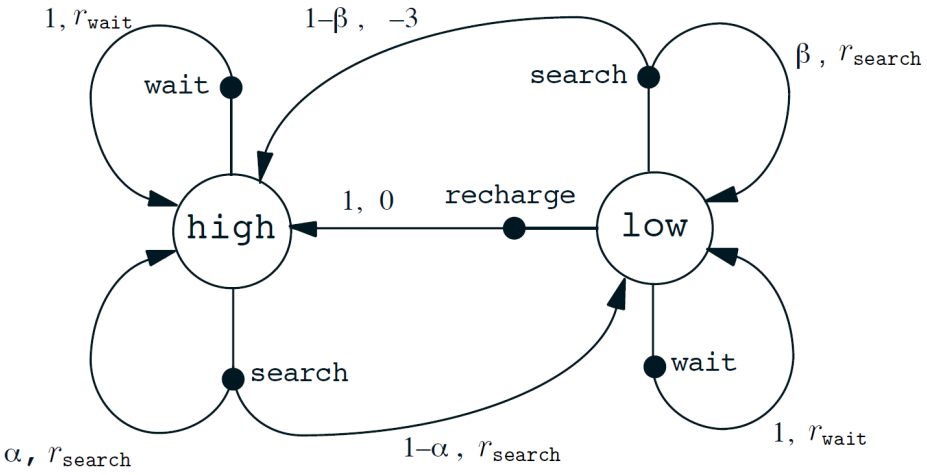

Figure 8.1: The MDP for the cleaning robot problem.

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

Exercise 8.7

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.

Exercise 8.8

Consider the MDP in Figure 8.1 with $\alpha = 0.3$, $\beta = 0.5$, $\gamma = 1$, $r_{search} = 2$, $r_{wait} = 0$ and the following policy:

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

1. Compute the Value function where the MDP stops after two steps.
2. Compute the Values function in the same setting assuming a discount factor of $\gamma = 0.5$.
3. Compute the action-value function for each action value pair in the case the MDP stops after a single step.

Exercise 8.9

Provide the formulation of the Bellman expectation for V equations for the MDP in Figure 8.1, 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 8.10

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 8.11

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 8.1
3. Maze escape
4. Tic-tac-toe

Provide adequate motivations for your answers.

Exercise 8.12

Consider the MDP in Figure 8.2:

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

Exercise 8.13

Consider the following snippet of code:

```
1 V1 = np.linalg.inv(np.eye(nS) - gamma * pi @ P_sas) @ (pi @ R_sa)
2
3 V_old = np.zeros(nS)
4 tol = 0.0001
5 V2 = pi @ R_sa
6 while np.any(np.abs(V_old - V2) > tol):
7   V_old = V2
8   V2 = pi @ (R_sa - gamma * P_sas @ V)
```

1. Describe the purpose of the procedure of line `1` and the purpose of the procedure of lines `3-8`. Are they correct? If not, propose a correction.

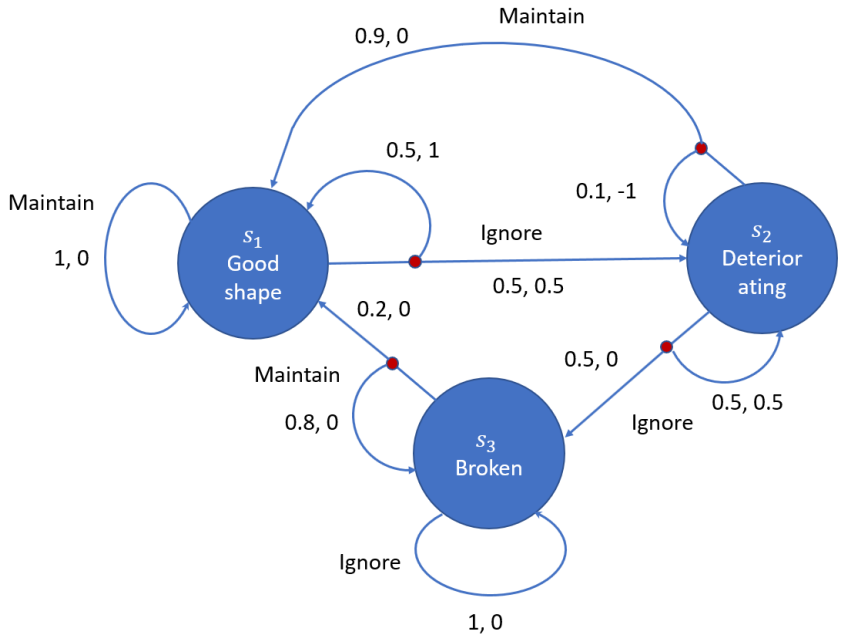

Figure 8.2: The MDP for machinery maintenance problem.

2. What is the main disadvantage of the procedure of line `1` compared to the one of lines `3-8`?

3. What happens to the two procedures when `gamma`$= 1$?

Exercise 8.14

We are given an Heating, Ventilation, and Air Conditioning (HVAC) system in which the states are cold (c), medium (m), warm (w) temperature. We can perform three actions: heat (h), refrigerate (r), and do nothing (d). Assume to have the following partial episodes for the HVAC functioning.

$$(c, d, 0) \to (c, h, 1) \to (m, h, 1) \to (m, h, -1) \to (w, r, 1) \to (m, \cdot, \cdot) \to \dots$$
$$(m, r, -2) \to (c, h, -2) \to (c, h, 1) \to (m, h, 1) \to (m, h, 1) \to (w, \cdot, \cdot) \to \dots$$

where a tuple (S, A, R) correspond to the State, Action, and Reward at a specific time.

1. Model it as an MDP and draw the corresponding graphical representation, specifying the transition probabilities and rewards (estimated from the episodes) for each transition.

2. Can you tell if the reward of this process is stochastic or deterministic? And what about the transitions?

3. Assuming we want to evaluate the performance of the HVAC, tell which kind of

problem we are in and suggest a technique to solve it.

Exercise 8.15

Tell whether the following statements about MDP and DP are true or false. Motivate your answers.

1. After the policy improvement step, the policy always changes.
2. Some sequential decision problems cannot be modeled as an MDPs.
3. It is always better to perform Policy Evaluation using Dynamic Programming instead of solving a linear system.
4. There are some tasks in which we may not need to discount rewards.
5. The best approach to solve a large MDP, exploiting the knowledge of the one-step dynamics, is Dynamic Programming.
6. The only optimal policy of an MDP can be non-Markovian.
7. We can solve Bellman Optimality Equation with a linear system.
8. With the optimal value function, I can compute the optimal policy, even without knowing one-step dynamics.
9. In Policy Evaluation, we can converge to the correct value function, only if we properly initialize its values.
10. In an MDP, the past actions you perform might influence the future rewards you will gain.
11. Given a value function, there is only one policy that corresponds to it.
12. Value Iteration may not converge to the optimal value function.

Exercise 8.16

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

1. In a finite state MDP there only exist optimal policies which are Markovian, stationary and deterministic;
2. If we observe a finite episode on an MDP, we should consider the undiscounted ($\gamma = 1$) cumulative reward as performance metric;
3. The Bellman Expectation Operator can be used to compute the optimal policy of

an MDP;

4. There is a unique optimal policy in an MDP (if true provide a theoretical result, if false a counterexample).

Exercise 8.17

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

1. Starting from any value function, we are not assured to converge to a solution when we apply repeatedly the Bellman optimality operator;
2. The closed form solution to the Bellman equation is always a good choice to compute the value function of an MDP;
3. The application of the Bellman optimality operator T for 5 times to a generic value function V_0 guarantees that $||V^* - T^5 V_0||_\infty \leq \gamma^3 ||V^* - V_0||_\infty$;
4. Starting from any value function, we are assured to converge to a solution when we apply repeatedly the Bellman expectation operator.

Exercise 8.18

Comment the following statements about Markov Decision Processes (MDPs).

1. In a finite state MDP there are only Markovian, stationary and deterministic optimal policies;
2. Given a policy π for an MDP, we can compute the corresponding policy transition matrix P^π and the policy expected rewards R^π in each state;
3. Given the value function $V^{\pi^*}(s)$ corresponding to the optimal policy π^* on a specific MDP, we are able to execute the optimal policy π^* on the MDP;
4. There is a unique optimal value function $V^{\pi^*}(s)$ in an MDP.

Exercise 8.19

Provide an MDP modeling, specifying all its defining elements, of the following processes:

1. Car manifacturing;
2. Chess.

Exercise 8.20

Consider an MDP with three states $\{s_1, s_2, s_3\}$ and actions $\{d, so, em\}$. Applying the policy π we have the following action-value function $Q(s, a)$ for the three states and the three actions when we consider different discount factors γ:

	$\gamma = 0.9$			$\gamma = 0.95$			$\gamma = 0.99$		
	d	so	cm	d	so	cm	d	so	cm
s_1	35	25	0	95	90	0	780	785	0
s_2	55	0	45	120	0	125	810	0	825
s_3	165	0	0	240	0	0	940	0	0

- Provide the optimal policy π^* for each discount factor γ.
- What is the expected reward for π^* if the initial state distribution is $(0.4, 0.4, 0.2)$.
- Which γ would you choose for this specific problem?

Exercise 8.21

Consider the MDP in Figure 8.1 with transition probabilities $\alpha = 0$ and $\beta = 0.5$, discount factor $\gamma = 1$, instantaneous rewards $r_{search} = 4$ and $r_{wait} = 1$ and the following policy:

$$\pi(high, search) = 0.5,$$
$$\pi(high, wait) = 0.5,$$
$$\pi(low, search) = 1.$$

1. Compute the Value function in the case the MDP stops after two steps.
2. Compute the action-value function for each action value pair in the case the MDP stops after a single step.