Imperial College
London

October 20, 2022

Reinforcement Learning – Prof Aldo Faisal Dr Paul Bilokon
Assignment design: Manon Flageat
Special thank: Borja G. Leon for proofreading

Extended solution of lab Assignment 1: Understanding of MRPs and MDPs

Question 1: Simple Markov Reward Process

1. The only terminal state is s_7 (Sleep), as it is the only state you cannot exit to go to another state. We also refer to such state as "absorbing" state.

2. Multiple examples of such traces are given in the other questions.
For example:

$$\tau = s_{class_1}\ s_{facebook}\ s_{facebook}\ s_{facebook}\ s_{class_1}\ s_{class_2}\ s_{sleep} = s_1\ s_4\ s_4\ s_4\ s_1\ s_2\ s_7$$

3. The state transition probability matrix gives the probability to transition from state s (line) to state s' (column). For example the coefficient on line 2, column 3 gives the probability to transition from state s_2 to state s_3. Each line of this matrix should sum to 1, as you necessarily go somewhere at the next step.
We give the complete matrix:

$$P_{ss'} = \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0 & 0.4 & 0.6 & 0 \\ 0.1 & 0 & 0 & 0.9 & 0 & 0 & 0 \\ 0.2 & 0.4 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

As the last state is terminal, it would also be correct to write:

$$P_{ss'} = \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0 & 0.4 & 0.6 & 0 \\ 0.1 & 0 & 0 & 0.9 & 0 & 0 & 0 \\ 0.2 & 0.4 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

4. According to the definition of MRP, rewards are associated to the state. Thus we can give $R_{ss'}$ as a vector:

$$R_{ss'} = \begin{pmatrix} -2 \\ -2 \\ -2 \\ -1 \\ 1 \\ 10 \\ 0 \end{pmatrix}$$

However, it is also correct to write $R_{ss'}$ as a matrix with the reward associated to the transition from s (line) to state s' (column).

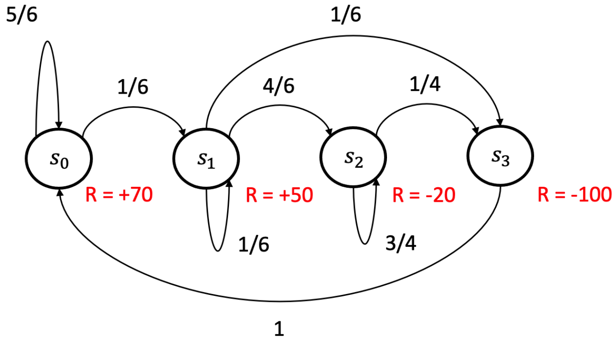
5. The return associated with a trace of length T is defined as $R = \sum_{t=0}^{T} \gamma^t r_{t+1}$.
   Here we accept two possible answers. Either you consider that the reward is associated with the departing state, in this case you get for the returns **-3.0625** and **-3.1875**. Either you consider that the reward is associated with the arriving state and you get **-2.125** and **-2.375**.

Question 2: Understanding Markov Decision Process

1. You can fill the transition matrix using the fact that the machine necessarily needs to transition to a new state when performing an action at each timestep, so all the probabilities associated with a state given an action should sum to 1.
   For example, imagine you take action a_0 in state s_0. You know you can only transition to state s_1 or s_3 (given by the lines in the table), and you know you have probability 5/6 to transition to s_1. Thus you necessarily have probability 1/6 to transition to s_3.
   The full filled table is given below.

2. As indicated in the description of the problem, the reward associated with a transition depends on the starting state and the action taken, for example if you perform action a_2 in state s_1, your reward would be $r(s_1, a_2) = p(s_1) - c(a_2) = 50 - 100 = -50$.
   The full filled table is given below.

3. Below is the MDP graph corresponding to policy π. As explained in the question, you know the action chosen in each state, so you only need to represent the states, the transitions and the rewards.



4. $\tau = s_0\ s_0\ s_1\ s_2\ s_3\ s_0$ Note that any trace is accepted as long as it starts at s_0, has length 6 and the transitions are "possible" according to the transition matrix.

5. The transition probability matrix of this MDP given the policy corresponds to the matrix of transitions knowing the action you are taking in each state (given by the policy). Thus, as for the graph in the previous question, it only needs to give the probability to transition from state s (line) to s' (column).
   We fill it using the table, and we get:

$$P^\pi = \begin{pmatrix} 5/6 & 1/6 & 0 & 0 \\ 0 & 1/6 & 4/6 & 1/6 \\ 0 & 0 & 3/4 & 1/4 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

   If you were to give the transition matrix associated with the full MDP, it would be of dimension $4 * 4 * 4$ to take into account the 4 possible starting states, 4 possible actions and 4 possible arriving states. Here we only need a $4 * 4$ matrix because we already know which action is taken.

6. Similarly, as the policy is given and the rewards are associated associated with the departing state and the action, we can give the reward matrix as a vector. Each line gives the reward associated with being in state s as the action taken in this state is already known (given by the policy).
   We get the following matrix using the table:

$$R_\pi = \begin{pmatrix} 70 \\ 50 \\ -20 \\ -100 \end{pmatrix}$$

   If you were to give the reward matrix associated with the full MDP, it would be of dimension $4 * 4 * 4$ to take into account the 4 possible starting states, 4 possible actions and 4 possible arriving states. Here we only need a $4 * 1$ matrix because we already know which action is taken and the reward only depends on the starting state.

7. The Bellman equation for the MDP is given as follow:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma V^\pi(s') \right)$$

   Where $\mathcal{P}^a_{ss'}$ is the probability to transition from s to s' taking action a, and $\mathcal{R}^a_{ss'}$ is the reward associated with the transition from s to s' taking action s.
   So we get:

$$V^\pi(s) = \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi(s,a) \mathcal{P}^a_{ss'} \mathcal{R}^a_{ss'} + \gamma \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi(s,a) \mathcal{P}^a_{ss'} V^\pi(s')$$

   In this case, the policy is deterministic and given, thus this expression can easily be simplified. For example in the case of s_0, you would always take action a_1, thus:

$$\begin{aligned} V^\pi(s_0) &= \sum_{s' \in \mathcal{S}} \mathcal{P}^{a_1}_{s_0 s'} \mathcal{R}^{a_1}_{s_0 s'} + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{a_1}_{s_0 s'} V^\pi(s') \\ &= R^{a_1}_{s_0} \sum_{s' \in \mathcal{S}} \mathcal{P}^{a_1}_{s_0 s'} + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{a_1}_{s_0 s'} V^\pi(s') \\ &= R^{a_1}_{s_0} + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{a_1}_{s_0 s'} V^\pi(s') \ \ \texttt{(as} \sum_{s' \in \mathcal{S}} \mathcal{P}^{a_1}_{s_0 s'} = 1\texttt{)} \end{aligned}$$

   In the general case:

$$V^\pi(s) = R^\pi_s + \gamma \mathcal{P}^\pi_{ss'} V^\pi(s')$$

   We can express this results using vectors:

$$V^\pi = R^\pi + \gamma * P^\pi * V^\pi$$

$$(I - \gamma * P^\pi) * V^\pi = R^\pi$$

   You do not necessarily need to use this vector expression, you can also compute the value for each state using the equation directly, you would get the same results.
   We get:

$$(I - \gamma * P^\pi) = \begin{pmatrix} 0.5 & -0.1 & 0 & 0 \\ 0 & 0.9 & -0.4 & -0.1 \\ 0 & 0 & 0.55 & -0.15 \\ -0.6 & 0 & 0 & 1 \end{pmatrix}$$

   So finally: $V^\pi \approx \begin{pmatrix} 147 \\ 36.7 \\ -39.5 \\ -11.6 \end{pmatrix}$

8. To improve the action taken in state s_1 by policy π, we can choose the action a_k that maximised the expected value of the state when taking that action: $\delta_{s_1}^{a_k} = R_{s_1}^{a_k} + \gamma * \sum_{j=0}^{3} P_{s_1 s_j}^{a_k} * V_{s_j}^{\pi}$. We get:

$$\delta_{s_1}^{a_0} = 50 + 0.6 * (1/6 * 36.7 + 4/6 * -39.5 + 1/6 * -11.6) = 36.7$$

$$\delta_{s_1}^{a_1} = 20 + 0.6 * (4/5 * 36.7 + 1/5 * -39.5) = 32.9$$

$$\delta_{s_1}^{a_2} = -50 + 0.6 * 1 * 147 = 38.2$$

According to these values, a one-step improvement of the action taken in state s_1 would be to now systematically take action a_2 (renovate).
To continue improving this deterministic policy, we would need to improve similarly the actions for all the other states (s_0, s_2 and s_3), and then re-evaluate the value function of this new improved policy. We would then continue improving π iteratively this way until convergence to the optimal policy π^*.

Here is the transition table with all data:

Initial state	Action	Final state	Transition probability	Reward
s_0 (new)	a_0 (do nothing)	s_1 (good condition)	5/6	100
		s_3 (broken)	1/6	
	a_1 (maintain)	s_0 (new)	5/6	70
		s_1 (good condition)	1/6	
	a_2 (renovate)	s_0 (new)	1	0
s_1 (good condition)	a_0 (do nothing)	s_1 (good condition)	1/6	50
		s_2 (poor condition)	4/6	
		s_3 (broken)	1/6	
	a_1 (maintain)	s_1 (good condition)	4/5	20
		s_2 (poor condition)	1/5	
	a_2 (renovate)	s_0 (new)	1	-50
s_2 (poor condition)	a_0 (do nothing)	s_2 (poor condition)	2/3	10
		s_3 (broken)	1/3	
	a_1 (maintain)	s_2 (poor condition)	3/4	-20
		s_3 (broken)	1/4	
	a_2 (renovate)	s_0 (new)	1	-90
s_3 (broken)	a_0 (do nothing)	s_3 (broken)	1	0
	a_1 (maintain)	s_3 (broken)	1	-30
	a_2 (renovate)	s_0 (new)	1	-100