Chapter 3 - Finite Markov Decision Processes

Stéphane Liem NGUYEN

August 16, 2021

Exercises with (***corrected***) were corrected based on the Errata

Exercise 3.1 **(p. 51)** Devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. Make the three examples as *different* from each other as possible. The framework is abstract and flexible and can be applied in many different ways. Stretch its limits in some way in at least one of your examples.

Exercise 3.2 **(p. 51)** Is the MDP framework adequate to usefully represent *all* goal-directed learning tasks? Can you think of any clear exceptions ?

Exercise 3.3 **(p. 51)** Consider the problem of driving. You could define the actions in terms of the accelerator, steering wheel, and brake, that is, where your body meets the machine. Or you could define them farther out—say, where the rubber meets the road, considering your actions to be tire torques. Or you could define them farther in—say, where your brain meets your body, the actions being muscle twitches to control your limbs. Or you could go to a really high level and say that your actions are your choices of *where* to drive. What is the right level, the right place to draw the line between agent and environment? On what basis is one location of the line to be preferred over another? Is there any fundamental reason for preferring one location over another, or is it a free choice?

Exercise 3.4 **(p. 53)** Give a table analogous to that in Example 3.3 (recycling robot), but for $p(s', r|s, a)$. It should have columns for s, a, s', r, and $p(s', r|s, a)$, and a row for every 4-tuple for which $p(s', r|s, a) > 0$.

Exercise 3.5 **(p. 55)** The equations in Section 3.1 (Agent-Environment Interface) are for the continuing case and need to be modified (very slightly) to apply to episodic tasks. Show that you know the modifications needed by giving the modified version of (3.3).

Exercise 3.6 **(p. 56)** Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for -1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task?

***Exercise 3.7* (p. 56)** Imagine that you are designing a robot to run a maze. You decide to give it a reward of $+1$ for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes—the successive runs through the maze—so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

***Exercise 3.8* (p. 56)** Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are $G_0, G_1, \ldots, G_5$? Hint: Work backwards.

***Exercise 3.9* (p. 56)** Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0?

***Exercise 3.10* (p. 56)** Prove the second equality in (3.10).

***Exercise 3.11* (p. 58)** If the current state is S_t, and actions are selected according to a stochastic policy π, then what is the expectation of R_{t+1} in terms of π and the four-argument dynamics function p (3.2)?

***Exercise 3.12* (p. 58)** Give an equation for v_π in terms of q_π and π.

***Exercise 3.13* (p. 58)** Give an equation for q_π in terms of v_π and the four-argument p.

***Exercise 3.14* (p. 60)** The Bellman equation (3.14) must hold for each state for the value function v_π shown in Figure 3.2 (right) of Example 3.5 (Gridworld). Show numerically that this equation holds for the center state, valued at $+0.7$, with respect to its four neighboring states, valued at $+2.3, +0.4, -0.4$, and $+0.7$. (These numbers are accurate only to one decimal place.)

***Exercise 3.15* (p. 61)** In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, using (3.8), that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies. What is v_c in terms of c and γ?

***Exercise 3.16* (p. 61)** Now consider adding a constant c to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.

Exercise 3.17 (p. 61) What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state–action pair (s, a). Hint: The backup diagram in Figure 1 corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values.
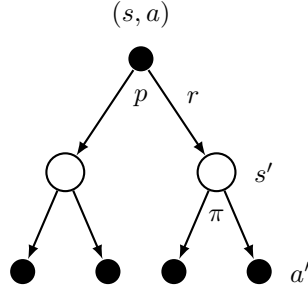

Figure 1: q_π backup diagram

Exercise 3.18 (p. 62) The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action:
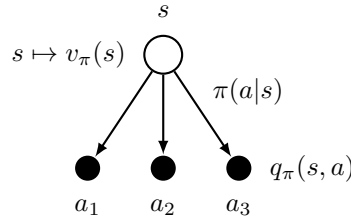

Figure 2: Small backup diagram showing v_π in terms of q_π and the policy π

Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s, a)$, given $S_t = s$. This equation should include an expectation conditioned on following the policy, π. Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ such that no expected value notation appears in the equation.

Exercise 3.19 **(p. 62)** The value of an action, $q_\pi(s, a)$, depends on the expected next reward and the expected sum of the remaining rewards. Again we can think of this in terms of a small backup diagram, this one rooted at an action (state-action pair) and branching to the possible next states:
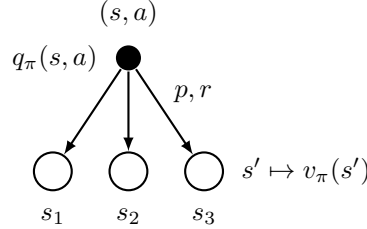

Figure 3: Small backup diagram showing q_π in terms of v_π and the dynamics function p

Give the equation corresponding to this intuition and diagram for the action value, $q_\pi(s, a)$, in terms of the expected next reward, R_{t+1}, and the expected next state value, $v_\pi(S_{t+1})$, given that $S_t = s$ and $A_t = a$. This equation should include an expectation but *not* one conditioned on following the policy. Then give a second equation, writing out the expected value explicitly in terms of $p(s', r | s, a)$ defined by (3.2), such that no expected value notation appears in the equation.

Exercise 3.20 **(p. 66)** Draw or describe the optimal state-value function for the golf example.

Exercise 3.21 **(p. 66)** Draw or describe the contours of the optimal action-value function for putting, $q_*(s, \textbf{putter})$, for the golf example.

Exercise 3.22 **(p. 66)** Consider the continuing MDP shown in Figure 4. The only decision to be made is that in the top state, where two actions are available, **left** and **right**. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

Exercise 3.23 **(p. 67)** Give the Bellman equation for q_* for the recycling robot.

Exercise 3.24 **(p. 67)** Figure 3.5 (Gridworld) gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places.

Exercise 3.25 **(p. 67)** Give an equation for v_* in terms of q_*.

Exercise 3.26 **(p. 67)** Give an equation for q_* in terms of v_* and the four-argument p.
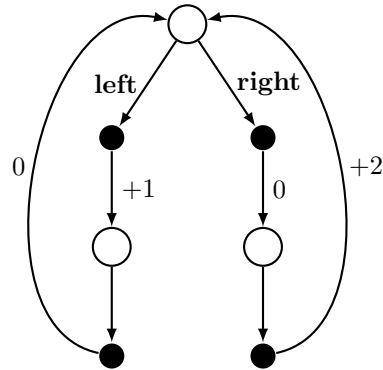
Figure 4: MDP transition graph

***Exercise 3.27* (p. 67)**   Give an equation for π_* in terms of q_*.

***Exercise 3.28* (p. 67)**   Give an equation for π_* in terms of v_* and the four-argument p.

***Exercise 3.29* (p. 67)**   Rewrite the four Bellman equations for the four value functions (v_π, v_*, q_π, and q_*) in terms of the three argument function p (3.4) and the two-argument function r (3.5).