## Exercise 1 part A


- Armando Ordorica
- APS 1080
- 1005592164
- Jan 2023


**Exercise 3.7**

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 e↵ectively communicated to the agent what you want it to achieve?

- Escaping the maze $R = +1$.
- All other times $R = 0$

$$G_T = \text{Expected Return } = R_{t+1} + R_{t+2} + \ldots + R_T$$

Discounted:

- Cumulative sum of all future rewards

$$G_T = \sum_{k=t+1}^{T} \gamma^{k-(t+1)} R_k$$

where $\gamma$ is the discount factor and is bounded $0 \leq \gamma \leq 1$.

If the goal is to maximize the expected total reward $G_T$, this number will always have a maximum value of 1, regardless of how long it takes for the agent to escape. In order to ensure that the agent learns that speed is important, we can penalize with a -1 every time step before the escape.

**Exercise 3.8**

*Exercise 3.8* 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 $G_0$, $G_1$, ..., $G_5$? Hint: Work backwards. □

The return at the terminal state will be 0, i.e. when $t = T$.

We define $G_T = 0$.

In this case $T = 5$, so $G_5 = 0$.

$$G_4 = R_5 + \gamma G_5 = 2 + (0.5)2 = 2$$
$$G_3 = R_4 + \gamma G_4 = 3 + (0.5)2 = 3 + 1 = 4$$
$$G_2 = R_3 + \gamma G_3 = 6 + (0.5)4 = 8$$
$$G_1 = R_2 + \gamma G_2 = 2 + (0.5)8 = 6$$
$$G_0 = R_1 + \gamma G_1 = -1 + (0.5)6 = 2$$

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

$$G_t = R_{t+1} + \gamma G_{t+1}$$
$$G_0 = R_1 + \gamma G_1$$
$$= 2 + 0.9(G_1)$$
$$= 2 + (0.9)(7/(1 - 0.9)) = 2 + 6.3/0.1$$
$$G_0 = 65$$

The expected discounted return is given by

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Therefore:

$$G_1 = \sum_{k=0}^{\infty} (0.9)^k 7 = \frac{1}{1-\gamma}(7) = \frac{7}{1 - 0.9}$$

**Exercise 3.12** Give an equation for $v_\pi$ in terms of $q_\pi$ and $\pi$.

$v_\pi$ is the state value function

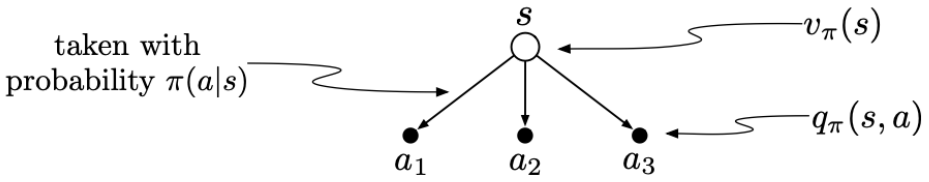
$$v_\pi(s) := \mathbb{E}_\pi [G_t | S_t = s]$$

for all $s \in \mathbb{S}$

and $q_\pi$ is given by:

$$q_\pi(s, a) = \mathbb{E} [G_t | S_t = s, A_t = a]$$
$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s, a)$$

*Exercise 3.18* 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:



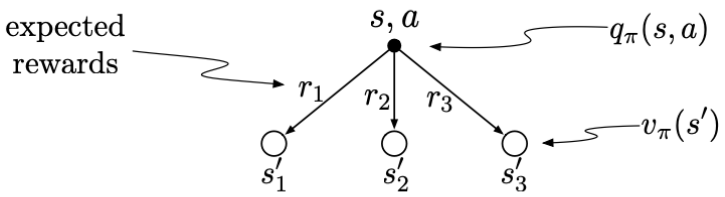

Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s,a)$, given $S_t = s$. This equation should include an expectation conditioned on following the policy, $\pi$. Then give a second equation 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. □

$v_\pi \propto$ Actions possible in that state × probability of each action given the policy

$$v_\pi(s) = \sum_{a\in\mathbb{A}} \pi(a|s) \sum_{s'\in\mathbb{S}} \sum_{r\in\mathbb{R}} [p'(s', r|s, a)[r + \gamma v_\pi(s')]]$$

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[q_\pi(S_t, A_t)|S_t = s, A_t = a] \\ &= \sum_a \pi(a|s) q_\pi(s|a) \end{aligned}$$

*Exercise 3.19* 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:




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. □

$q_\pi(s,a)$ is the Action value function for policy $\pi$

$R_{t+1}$ is the expected reward

$v_\pi(s_{t+1})$ is the expected value of the next state

We know that the dynamics of the MDP are given by

$$p(s', r|s, a) = Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\}$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$q_\pi(s, a) = \mathbb{E}_\pi [G_t|S_t = s, A_t = a]$$

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi [R_{t+1} + \gamma v_\pi(s_{t+1})|S_t = s, A_t = a] \\ &= \sum_{s', r} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right] \end{aligned}$$