APS 1080 REINFORCEMENT LEARNING - ASSIGNMENT 2

ARMANDO ORDORICA
STUDENT ID: 1005592164

JANUARY 2023

Exercise 3.7

Imagine that you're 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 for 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. 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?

Escaping the Maze $\rightarrow R = +1$

All other times $\rightarrow R = \varnothing$

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

Discounted:

$$G_T = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$ where $0 \leq \gamma \leq 1$ and γ is called the discount rate

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 (-1) every time step before the escape.

Exercise 3.8

Suppose that $\gamma = 0.5$ and the following sequence of rewards is received:

$R_1 = -1$
$R_2 = 2$
$R_3 = 6$
$R_4 = 3$
$R_5 = 2$

with $T = 5$

What are $G_0, G_1, \ldots G_5$

We define $G_T = 0$

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

$$G_4 = R_5 + \gamma G_5$$
$$= 2 + (0.5)(0) = 2$$

$$G_3 = R_4 + \gamma G_4 = 3 + (0.5)(2) = 3 + 1 = 4$$

$$G_2 = R_3 + \gamma G_3 = 6 + 0.5(4) = 8$$

$$G_1 = R_2 + \gamma G_2 = 2 + (0.5)(8) = 2 + 4 = 6$$

$$G_0 = R_1 + \gamma G_1 = -1 + (0.5)(6) = -1 + 3 = 2$$

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?

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

$$G_0 = R_1 + \gamma G_1$$

$= 2 + 0.9\,(G_1) = 2 + 0.9\left(\frac{7}{1-0.9}\right) = 2 + \frac{6.3}{0.1} = 65$

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

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

3.12 Give an equation for v_π in terms of q_π and π.

v_π is a state value function,

$$v_\pi(s) = \mathbb{E}\left[G_t \mid S_t = s\right] \quad \forall s$$

$$q_\pi(s,a) = \mathbb{E}\left[G_t \mid S_t = s,\ A_t = a\right]$$

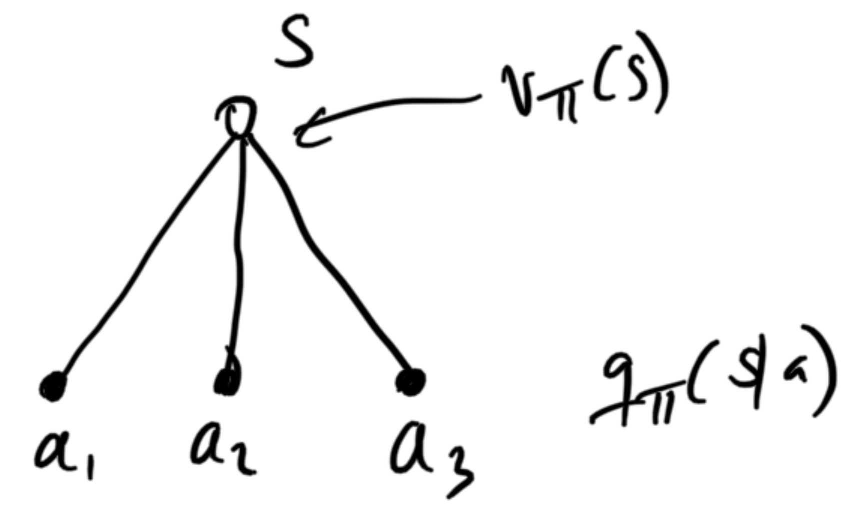
↓

Helps to solve the action selection problem because the maximum value of the expectation can be computed directly.

$$v_\pi(s) \doteq \sum_a \pi(a \mid s)\, q_\pi(s,a)$$

3.18 The value of a state depends on the values of the actions possible in that state and 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:

taken with probability $\pi(a|s)$



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 condition on following the policy, π. Then, 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.

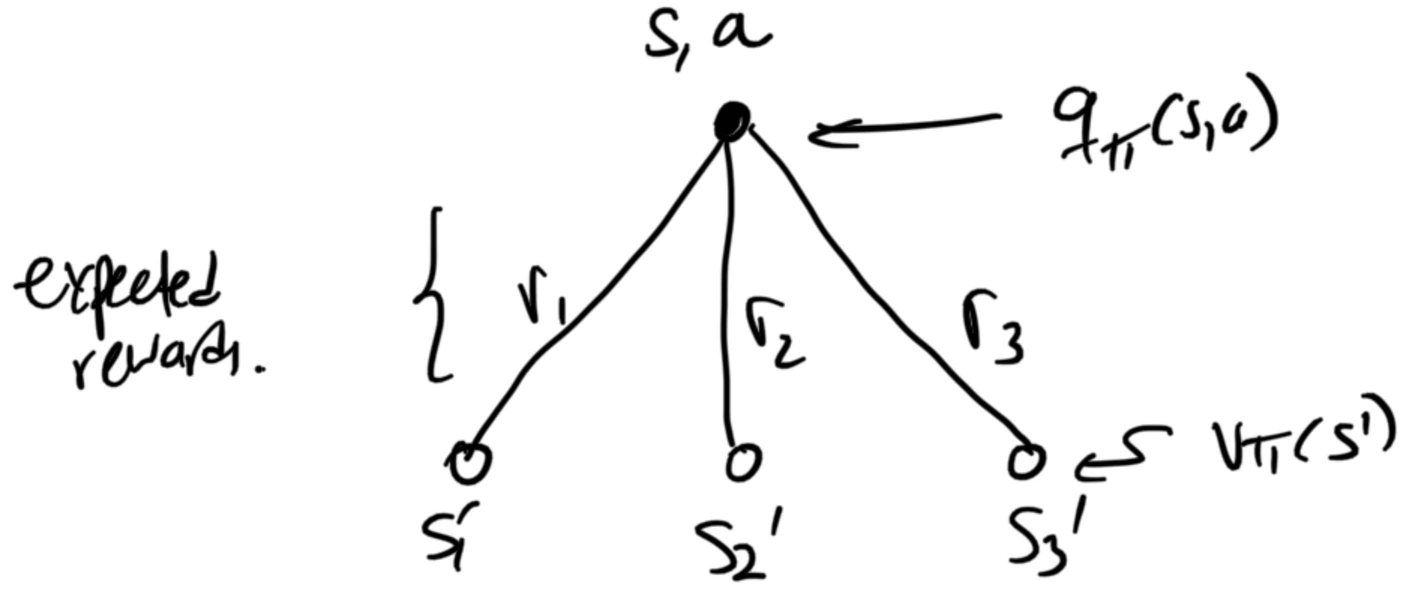
$V_\pi \propto$ Actions possible in that state
$\times$
prob of each action given policy.

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

$$v_\pi(s) = \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \mid S_t = s, A_t = a \right]$$

$$= \sum_a \pi(a|s)\, q_\pi(s,a)$$

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 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 for this intuition and diagram for the action value, $q_\pi(s,a)$, in terms of the expected 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 $p(s',r \mid s,a)$ defined by (3.2) such that no expected value notation appears in the equation.

$q_\pi(s,a) \rightarrow$ Action value function for policy π.

$\hookrightarrow R_{t+1} \rightarrow$ Expected Reward

$\hookrightarrow v_\pi(S_{t+1}) \rightarrow$ Expected Next State value.

(3.2)
$$\underbrace{p(s',r \mid s,a)}_{\text{Dynamics of the MDP}} \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

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

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

$$q_\pi(s,a) = \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a \right]$$

$$= \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]$$