Rienforcement learning: Assignment 2

Oluwatomilayo, Adegbite
500569283

Nikolas, Maier
500461990

February 21, 2019

1

1.1 Exercise 3.17

Must give action value q_π(s,a) in terms of q_π(s',a')

$$V_\pi(s) = \sum_a \pi(s,a) \sum P^a_{ss'}(R^a_{ss'} + \alpha V_\pi(s')) \tag{1}$$

The value function represented as a bellman equation taking state and summing over all possible actions in that state and value of successor states

$$P^a_{ss'} = p(s' \mid s,a) = Pr\left\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\right\} = \sum_{r \in R} p(s', r \mid s, a) \tag{2}$$

Equation to represent probability of state transition given action and previous state is the same as probability of moving to successor state and receiving reward from future state and action

$$r(s,a) = E[R_t \mid S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R} r \sum_{s' \in S} p(s', r \mid s, a) \tag{3}$$

state action reward function gives expected reward of state and action

$$G_t = p(s' \mid s,a) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{4}$$

G_t represents the discounted future reward

$$q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = s] \tag{5}$$

$$= E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a] \; expand \; G_t \tag{6}$$

$$= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \tag{7}$$

$$q_\pi(s,a) = E_\pi[\sum_{s'} P^a_{ss'}(R^a_{ss'} + \sum_{a'} \alpha q_\pi(s', a')) \mid S_t = s, A_t = a] \tag{8}$$

(7) Represent expected future reward as next reward plus discounted future reward (8) To get the value of all possible successor states sum over s', multiply the probability of moving from current state to next state given action a by the result of the reward of that action and state transition and the sum of all possible actions in s' times a discount times the value of q_π(s',a')

1.2 Exercise 3.19

$$P^a_{ss'} = p(s' \mid s, a) = Pr\left\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\right\} = \sum_{r \in R} p(s', r \mid s, a) \tag{9}$$

$$V_\pi(s) = \sum_a \pi(s,a) \sum P^a_{ss'}(R^a_{ss'} + \alpha V_\pi(s')) \tag{10}$$

The value function represented as a bellman equation taking state and summing over all possible actions in that state and value of successor states

$$\begin{align} G_t = p(s' \mid s, a) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + ... \tag{11} \\ = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + ...) \tag{12} \\ = R_{t+1} + \gamma G_{t+1} \tag{13} \end{align}$$

G_t can be represented as current reward + discounted future reward

$$q_\pi(s,a) = E[G_t \mid S_t = s, A_t = a] \tag{14}$$

$$q_\pi(s,a) = E[R_{t+1} + \alpha V_\pi(S_{t+1}) \mid S_t = s, A_t = a] \tag{15}$$

$q_\pi(s,a)$ Given in terms of future expected reward considering discount and expected value of future states

1.2.1 Second Equation

$$\begin{align} q_\pi(s,a) &= E_\pi[G_t \mid S_t = s, A_t = a] \tag{16} \\ &= E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a] \tag{17} \\ &= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \tag{18} \\ q_\pi(s,a) &= \sum_{s'} p(s', r \mid s, a)(R^a_{ss'} + \sum_{a'} \alpha q_\pi(s', a')) \tag{19} \end{align}$$