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

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

G_t represents the discounted future reward

$$
\begin{aligned}
q_\pi(s,a) &= E_\pi[G_t \mid S_t = s, A_t = s] \\
&= E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = s] \\
&= E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = s]
\end{aligned} \tag{4}
$$

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

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

$$
\begin{aligned}
G_t = p(s' \mid s,a) &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + ... && (5) \\
&= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + ...) && (6) \\
&= R_{t+1} + \gamma G_{t+1} && (7)
\end{aligned}
$$

G_t can be represented as current reward + discounted future reward

$$
q_\pi(s,a) = E[G_t \mid S_t = s, A_t = s] \tag{8}
$$

$$
q_\pi(s,a) = E[R_{t+1} + \alpha V_\pi(s'_{t+1}) \mid S_t = s, A_t = s] \tag{9}
$$

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