

# Rienforcement learning: Assignment 2

Oluwatomilayo, Adegbite  
500569283

Nikolas, Maier  
500461990

February 21, 2019

## 1

### 1.1 Exercise 3.17

Must give action value  $q_\pi(s,a)$  in terms of  $q_\pi(s',a')$

$$V_\pi(s) = \sum_a \pi(s,a) \sum_{s'} P_{ss'}^a (R_{ss'}^a + \alpha V_\pi(s')) \quad (1)$$

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

$$P_{ss'}^a = p(s' | s, a) = \Pr \{S_t = s' | S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in R} p(s', r | s, a) \quad (2)$$

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

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

$G_t$  represents the discounted future reward

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

$$q_\pi(s, a) = E_\pi\left[\sum_{s'} P_{ss'}^a (R_{ss'}^a + \sum_{a'} \alpha q_\pi(s', a')) \mid S_t = s, A_t = a\right]$$

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  $q_\pi(s', a')$

## 1.2 Exercise 3.19

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

$G_t$  can be represented as current reward + discounted future reward

$$q_\pi(s, a) = E[G_t \mid S_t = s, A_t = a] \tag{6}$$

$$q_\pi(s, a) = E[R_{t+1} + \alpha V_\pi(s'_{t+1}) \mid S_t = s, A_t = a] \tag{7}$$

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