Q1

$S = \{ \text{high}, \text{low} \}$

$A(\text{high}) = \{ \text{search}, \text{wait} \}$

$A(\text{low}) = \{ \text{search}, \text{wait}, \text{recharge} \}$

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
high	wait	high	r_{wait}	1
low	search	high	-3	$1-\beta$
low	search	low	r_{search}	β
low	wait	low	r_{wait}	1
low	recharg	high	0	1

where $p(s', r \mid s, a) = Pr\{ S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a \}$

Teacher's Signature

05

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s, a)$$

$$v_*(s) = \max_{a \in A(s)} E[G_t \mid S_t = s, A_t = a]$$

$$v_*(s) = \max_{a \in A(s)} \sum_{s', r} p(s', r \mid s, a) [r + \gamma v_*(s')]$$

and,

$$v_*(s') = \max_{a' \in A(s')} q_*(s', a')$$

$$\therefore v_*(s) = \max_{a \in A(s)} \sum_{s', r} p(s', r \mid s, a) [r + \gamma \max_{a' \in A(s')} q_*(s', a')]$$

Q3 Exercise 3.15 & 3.16

→ Signs of these rewards do not affect the learning algorithm and hence, not matter b/c adding a large +ve constant to all the rewards s.t all rewards become positive would not affect the learning algorithm and only increase value function by a constant which we would see next. Similarly, adding a very small -ve constant to all the rewards s.t. all rewards become would not affect the learning algo.

- But inverting the signs of all the rewards would affect the learning algorithm. As, then, our reward function would become Cost function.

Claim: Adding a constant c to all rewards adds a constant, v_c, to the values of all states.

Proof: $v_\pi(s) = E_\pi[G_t | S_t = s]$

$$v_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

Now, add c to all the rewards

$$v'_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \mid S_t = s\right]$$

$$v'_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c \mid S_t = s\right]$$

$$v'_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = 0\right] + E\left[\sum_{k=0}^{\infty} \gamma^k c \mid S_t = s\right]$$

let $v_c = \sum_{k=0}^{\infty} \gamma^k c = c \sum_{k=0}^{\infty} \gamma^k$

as v_c is a constant

$$v'_\pi(s) = v_\pi(s) + v_c$$

As $0 \le \gamma < 1$, we have

$$v_c = \frac{c}{1-\gamma}$$

(using G.P sum for infinite series)

Teacher's Signature

Clearly, relative values of any state under any policy is unaffected as v_c only depends on c & γ.

In case of episodic task, we will have

$$v_c = c \sum_{k=0}^{T} \gamma^k$$

where T is a r.v which denotes termination time

$$v_c = c(\gamma^0 + \gamma + \gamma^2 + \ldots + \gamma^T)$$

$$v_c = c\left(\frac{\gamma^{T+1} - 1}{\gamma - 1}\right)$$

2) v_c is not a constant & is a r.v and function of T,

for different episodes, one may have different value-functions.

But v_c will be a constant for a single episode, hence won't affect the learning algorithm.