Ques 1. We know r_{search}

- For $s = high$, $a = search$, $s' = high$

$$r_{search} = \frac{(1)\, p(s', r=1 \mid s, a)}{p(s' \mid s, a)}$$

$p(s', r=1 \mid s,a) = \alpha\, r_{search}$

$p(s' \mid s,a) = p(s', r=1 \mid s,a) + p(s', r=0 \mid s,a)$

$\Rightarrow p(s', r=0 \mid s,a) = \alpha - \alpha\, r_{search}$

- Similarly for $s = high$, $a = search$, $s' = low$

$p(s', r=1 \mid s,a) = (1-\alpha)\, r_{search}$

$p(s', r=0 \mid s,a) = (1-\alpha) - (1-\alpha) r_{search}$

- Similarly for $s = low$, $a = search$, $s' = low$

$p(s', r=1 \mid s,a) = \beta\, r_{search}$

$p(s', r=0 \mid s,a) = \beta - \beta\, r_{search}$

- For $s = high$, $a = wait$, $s' = high$

$r_{wait} = p(s', r=1 \mid s,a)$

$\Rightarrow p(s', r=0 \mid s,a) = 1 - r_{wait}$

- Similarly for $s = low$, $a = wait$, $s' = low$

$p(s', r=1 \mid s,a) = r_{wait}$

$p(s', r=0 \mid s,a) = 1 - r_{wait}$

Table:-

s	a	s'	r	$p(s', r \mid s, a)$
High	Search	High	0	$\alpha - \alpha\, r_{search}$
High	Search	High	1	$\alpha\, r_{search}$
High	Search	Low	0	$(1-\alpha) - (1-\alpha) r_{search}$
High	Search	Low	1	$(1-\alpha) r_{search}$
Low	Search	Low	0	$\beta - \beta\, r_{search}$
Low	Search	Low	1	$\beta\, r_{search}$
Low	Search	High	-3	$1-\beta$
Low	Wait	Low	0	$1 - r_{wait}$
Low	Wait	Low	1	r_{wait}

P.T.O.

s	a	s'	r	$p(s', r \mid s, a)$
High	Wait	High	0	$1 - r_{wait}$
High	Wait	High	1	r_{wait}
Low	Recharge	High	0	1

Ques 3.

(a) Exercise 3.15

The signs of the rewards are not important but intervals between them are.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

If we add constant to all rewards then,

$$G_t' = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) = G_t + \sum_{k=0}^{\infty} \gamma^k c$$

$$\Rightarrow G_t' = G_t + \frac{c}{1-\gamma} = G_t + v_c$$

$$\begin{aligned} V_\pi(s)' &= E_\pi[G_t' \mid S_t = s] \\ &= E_\pi[G_t + v_c \mid S_t = s] \\ &= E_\pi[G_t \mid S_t = s] + E_\pi[v_c \mid S_t = s] \\ &= V_\pi(s) + v_c \end{aligned}$$

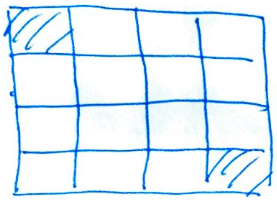
$$\Rightarrow v_c = \frac{c}{1-\gamma}$$

{Thus only constant v_c is added to v_π on increasing rewards to v_π}

(b) Exercise 3.16

Adding a constant (c) to all rewards in an episodic task will effect the tasks.

Consider the following episodic task:-



(0,0) and (3,3) are terminal states

reward for each action (-1)

Optimal Policy would ensure shortest path to either of the terminal states.

Now if the rewards are $-1+c$ $(c>1)$ then optimal policy changes, also value functions will for any policy will be different as it would never try to go to terminal states for maximizing reward

Ques 5. $V_*(s) = \max_{a \in A(s)} q_{\pi^*}(a, s)$

$$= \max_a E_{\pi^*}[G_t \mid A_t = a, S_t = s]$$

$$= \max_a E_{\pi^*}[R_t + \gamma G_{t+1} \mid A_t = a, S_t = s]$$

$$= \max_a E_{\pi^*}[R_t + \gamma V_*(S_{t+1}) \mid A_t = a, S_t = s]$$

$$= \max_a \sum_{s'} \sum_{r} p(s', r \mid s, a) \{ r + \gamma V_*(s') \}$$

$$= \max_a \sum_{s'} \sum_{r} p(s', r \mid s, a) \{ r + \gamma \max_{a} q_*(s', a) \}$$