Q1

State set = {high, low}

A(high) = {search, wait}

A(low) = {search, wait, 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
high	wait	low	~~$r_{recharge}$~~ —	—
low	search	high	-3	$1-\beta$
low	search	low	r_{search}	β
low	wait	low	r_{wait}	1
low	wait	high	—	—
low	recharge	high	0	1
low	recharge	low	—	—

Q3

Rewards are +ve for goals, −ve for running into the edge of the world and zero the rest of the time

Only the internals between the rewards are important and not their signs.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \qquad \gamma = \text{discount rate}$$

$$0 \leq \gamma \leq 1$$

Using the above equation, one can show that adding a constant c to all the rewards adds a constant V_c to the values of all the states and hence does not affect the relative values of any states.

Here rewards can be made the same sign by adding/subtracting a large positive quantity to/from all the rewards. Which leads to an increase or a decrease in the value function by a constant. This whole thing has no affect on the algorithm.

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

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

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

$$= E\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right] + E\left[\sum_{k=0}^{\infty} \gamma^k c \mid S_t = s\right]$$

$$= V_\pi(s) + \sum_{k=0}^{\infty} \gamma^k c$$

$$= V_\pi(s) + \underbrace{\frac{c}{1-\gamma}}_{\text{constant term } V_c}$$

Q5

Equation for V_* in terms of q_*

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma V_*(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right]$$

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right]$$

From the Bellman Optimality equation

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

$$= \max_a \sum_{s',r} p(s',r|s,a)\cdot\left[r + \gamma V_*(s')\right]$$

$$= \max_a \sum_{s'r} \cdot p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right]$$

Q3.16(b)

In case of episodic task

Let terminal time be T

solving the equation

$$V_\pi'(s) = E\left[G_t' \mid S_t = s\right]$$

$$= E\left[\sum_{k=0}^{T} \gamma^k (R_{t+k+1} + c) \mid S_t = s\right]$$

$$= E\left[\sum_{k=0}^{T} \gamma^k R_{t+k+1} \mid S_t = s\right] +$$

$$E\left[\sum_{k=0}^{T} \gamma^k c_t \mid S_t = s\right]$$

$$= V_\pi(s) + E\left[\sum_{k=0}^{T} \gamma^k c \mid S_t = s\right]$$

$V_c = E\left[\sum_{k=0}^{T} \gamma^k c \mid S_t = s\right]$ is a function of T

$\therefore T$ is a random variable, which normally varies from episode to episode.

$\Rightarrow$ diff episodes will have diff value functions

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

Thus it will increase V_π when T will increase