Q1

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r	$(\alpha)\, P_{search}(r)$
high	search	low	r	$(1-\alpha)\, P_{search}(r)$
low	search	high	-3	$(1-\beta)\, P_{search}(r)$
low	search	low	r	$(\beta)\, P_{search}(r)$
high	wait	high	r	$(1)\, P_{wait}(r)$
low	wait	low	r	$(1)\, P_{wait}(r)$
low	recharge	high	0	$(1)(1)$

Here $P_{search}(r)$ is a probability distribution with mean 'r_{search}' and $P_{wait}(r)$ is a probability distribution with mean 'r_{wait}'

$$\Rightarrow \quad E[P_{search}(r)] = \text{'}r_{search}\text{'}$$

$$\text{and} \quad E[P_{wait}(r)] = \text{'}r_{wait}\text{'}$$

Q3 Ex 3.15

Signs of rewards are NOT important and only the intervals between them are. This is because if a large constant is added or subtracted, then all rewards can be made of the same sign. Since the relative order is preserved, the algorithm is not affected. This can be seen below:

we know that

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

where

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots$$

$$= \sum_{k=0} \gamma^k R_{t+k+1}$$

Now, if a constant 'c' is added to all rewards

$\hat{G_t}$ becomes $\sum_{k=0}^{\infty} \gamma^k \left(R_{t+k+1} + c\right)$

$$\Rightarrow \quad \hat{G_t} = G_t + \sum_{k=0}^{\infty} c\gamma^k$$

$$\therefore \quad \hat{G_t} = G_t + \frac{c}{1-\gamma} \qquad (\text{since } \gamma < 1)$$

Now, $\hat{V_\pi}(s) = E\left[\hat{G_t} \mid S_t = s\right]$

$$= E\left[G_t + \frac{c}{1-\gamma} \,\middle|\, S_t = s\right]$$

$$= E\left[G_t \mid S_t = s\right] + \frac{c}{1-\gamma}$$

or

$$\boxed{\hat{V_\pi}(s) = V_\pi(s) + \frac{c}{1-\gamma}}$$

Scanned with CamScanner

Ex 3.16

For an episodic task, let the terminal timestep be $t = T$

Then,

$$\hat{G_t} = \sum_{k=0}^{T} \gamma^k \hat{R}_{t+k+1}$$

$$\hat{G_t} = \sum_{k=0}^{T} \gamma^k (R_{t+k+1} + c)$$

$$\hat{G_t} = \sum_{k=0}^{T} \gamma^k R_{t+k+1} + \sum_{k=0}^{T} c\gamma^k$$

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

or $$\hat{G_t} = G_t + c\left(\frac{1-\gamma^{T+1}}{1-\gamma}\right) \quad [\text{since } \gamma < 1]$$

Following a similar approach for $\hat{V_\pi}(s)$

we get

$$\hat{V_\pi}(s) = E\left[G_t + c\left(\frac{1-\gamma^{T+1}}{1-\gamma}\right) \middle| S_t = s\right]$$

Now, T is a random variable that depends on S_t

Thus, $c\left(\frac{1-\gamma^{T+1}}{1-\gamma}\right)$ cannot come out of the expectation

$\therefore$ a simple linear mapping doesn't exist

between the two.

Q5. By Eqn 3.17

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

Also, by eqn 3.18

$$v_*(s) = \max_a E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$\therefore \boxed{\underbrace{v_*(s)}_{\text{optimal value at state 's'}} = \max_a \underbrace{q_*(s,a)}_{\text{optimal value at state 's' after taking action 'a'}}}$$