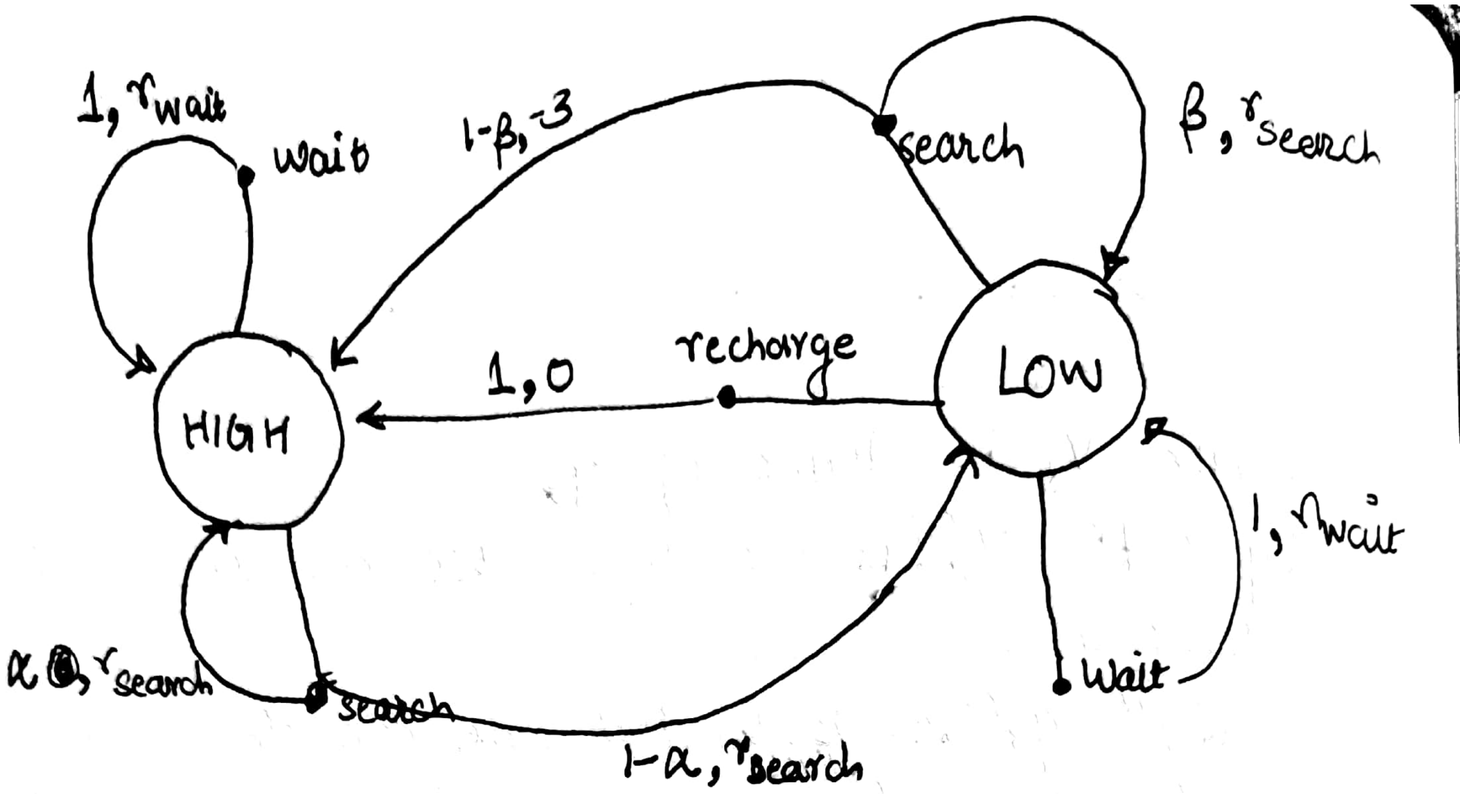
Q1



let $P_{wait}(r)$ be a gausian distribution with mean r_{wait}, standard deviation σ_{wait}.
similarly $P_{search}(r) \sim N(r_{search}, \sigma_{search})$

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

Q3(a) $v_\pi(s) = E[G_t \mid S_t = s]$

$= $ [illegible]

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

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

let $\hat{r}_t$ be the new reward with ~~cost~~ constant c added to it.

$$\Rightarrow \hat{r}_t = r_t + c$$

$$\Rightarrow \hat{G}_t = \sum_{k=0}^{\infty} \gamma^k \hat{R}_{t+k+1}$$

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

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

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

$$\Rightarrow \hat{v}_\pi(s) = E[\hat{G}_t \mid S_t = s] = E\left[G_t + \frac{c}{1-\gamma} \mid S_t = s\right]$$

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

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

Scanned by CamScanner

Q(3)(b) If the task is episodic, let the task finish after $t = N$

i.e. S_N is terminal state.

following from the above question

$$\hat{r}_t = r_t + c$$

$$\hat{G}_t = \sum_{k=0}^{N} \gamma^k \hat{R}_{t+k+1} = \sum_{k=0}^{N} \gamma^k (R_{t+k+1} + c)$$

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

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

$$\Rightarrow \hat{V}_\pi(s) = E[\hat{G}_t \mid S_t = s]$$

$$= \cancel{V_\pi(s) + c\left(\frac{\gamma^N - 1}{\gamma - 1}\right)}$$

$$= E\left[G_t + c\left(\frac{\gamma^N - 1}{\gamma - 1}\right) \Big| S_t = s\right]$$

since N is a random variable that depends on S_t, therefore it can't come out of expectation.

Hence, there is no simple mapping b/w $\hat{V}_\pi(s)$ and $V_\pi(s)$

Q5 write V_* in terms of q_*

by defn $V_*(s)$ is the best we can do starting at state s

by defn $q_*(s,a)$ is the best we can do starting at state s and performing action a.

therefore,

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