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

We know, $r(s, a, s') = \sum_{r \in R} r\, p(r, s' \mid s, ($

Consider, we only have binary reward (

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

$p(\,, s' \mid s, a) = \sum_{r \in R} p(r, s' \mid s, a)$

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

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

— 2

Hence, we can find all values in the table using 1 and 2 to obtain the final values as :

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

Q.2) a) Yes, the rewards are not important but the intervals b/w them are. Policy depends on optimal action which remains same given their diff. is preserved.

We know,

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

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

$$\text{New } G_t = R_{t+1} + c + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \cdots$$

$$= (R_{t+1} + \gamma R_{t+2} + \cdots) + (c + \gamma c + \gamma^2 c + \cdots)$$

Scanned with CamScanner

$= \text{Original } G_t + c(\gamma + \gamma^2 + \gamma^3 + \dots)$

$= \text{Original } G_t + \dfrac{c}{1-\gamma}$

Hence, even if we take Expectation of this for $v_\pi(s)$ this additional term will come out of Expectation.

So, $v_c = c/(1-\gamma)$

b) For ~~epsi~~odic task, consider $n$ steps after time $t$.

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

New $G_t = (R_{t+1}+c) + \gamma(R_{t+2}+c) + \dots \gamma^{n-1}(R_{t+n}+c)$

$= (R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots) + c(1+\gamma+\gamma^2+\dots \gamma^{n-1})$

$= \text{Original } G_t + c\left(\dfrac{1-\gamma^n}{1-\gamma}\right)$

Here, additional term $= c\dfrac{(1-\gamma^n)}{(1-\gamma)}$

So, this would change the task as the additional term in $G_t$ and hence in $V_\pi(s)$ [as Expectation $(G_t|s) = V_\pi(s)$] depends on $n$ which is no. of episodes after $t$.

Example of an episodic task is playing single game of Go. In episodic task, we will have one reward at the end of the game only and will not get reward at each time step or at time when action is taken.
Here, also we can see that if we ~~incea~~ change reward of all outcomes by some constant, we will get a different policy based on the length of episode.

Q.3) $v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$

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

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

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

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

$$q_*(s,a) = E[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a]$$

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