Q1) r_{search} and r_{wait} are expectations of the reward while searching and waiting.
Therefore consider a probability distribution P_{search} that has expectation r_{search} and P_{wait} that has expectation r_{wait}

	s	a	s'	r	$P(s', r \mid s, a)$
1)	high	search	high	r_1	$\alpha \cdot P_{search}(r_1)$
2)	high	search	low	r_2	$(1-\alpha)\ P_{search}(r_2)$
3)	low	search	high	-3	$(1-\beta)$. 1
4)	low	search	low	r_3	$\beta\ P_{search}(r_3)$
5)	high	wait	high	r_4	$1 \cdot P_{wait}(r_4)$
6)	low	wait	low	r_5	$1 \cdot P_{wait}(r_5)$

~~low~~ ~~recharge~~

For the rest of the rows such as, high wait low gives $P(s', r \mid s, a)$ as 0 because if you wait in high state, you cannot go to low state.
Similarly for,

low wait high → 0 as transition probability is 0.

low recharge high → 0 because reward is 0.

Explanation of table

Row 1 → At state high, you take an action search but return to the same state. The probability of this is α. The reward has expectation r_{search} hence we have r_1 that belongs to the distribution P_{search} with expectation r_{search}. ∴ $P(s', r|s, a)$ becomes transition probability times reward probability.

Row 2 → r_2 belongs to the distribution of P_{search} and now from high we move to low with probability $1-\alpha$. Hence $(1-\alpha).P_{search}(r_2)$

To check our answer for row 1 and row 2, we can sum $P(s', r|s, a)$ and it should be equal to 1.

$\Sigma \alpha P_{search}(r_1) + \Sigma (1-\alpha) P_{search}(r_2)$

$$= \alpha \sum P_{search}(r_1) + (1-\alpha)\sum P_{search}(r_2)$$

$$= \alpha + (1-\alpha)$$

$$= 1$$

Row 3 → At state low, action search and reaches high, gives a reward of -3 with probability 1, ~~here~~ and $1-\beta$ is the probability of going to state high, Hence $(1-\beta).1$

Row 4 → $\beta . P_{search}(r_3)$. stay at state low and get reward r_3 from Psearch distribution.

Summation of $(1-\beta).1 + \beta . \sum P_{search}(r_3) = 1$

Row 5 → state is high, action is wait, state is high, reward is from distribution with expectation r_{wait}. Hence $1 . P_{wait}(r_4)$

Row 6 → Similarly as above but with starting state low, wait then come back to low with probability 1 and receive

reward r_5 with probability $P_{wait}(r_5)$.
Hence 1. $P_{wait}(r_5)$.

Q3) Adding a constant c to all rewards, in case of continuing task

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

$$= E_\pi [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s]$$

Adding c to all rewards.

$$= E_\pi [(R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \dots | s]$$

$$= E_\pi [(c + \gamma c + \gamma^2 c \dots) + (R_{t+1} + \gamma R_{t+2} + \dots) | S_t = s]$$

$$= E_\pi [\frac{c}{1-\gamma} + G_t | S_t = s].$$

Expectation of constant is constant

$$V'_\pi(s) = \frac{c}{1-\gamma} + E_\pi [G_t | S_t = s].$$

$$V'_\pi(s) = \frac{c}{1-\gamma} + V_\pi(s).$$

Therefore all the values are just incremented by a factor of $\frac{c}{1-\gamma}$ and hence ~~choosing the actions~~ relatively it is the same thing.

In case of episodic task,

$$V_\pi(s) = E_\pi\left[\underbrace{R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{k-1} R_{t+k}}_{G_t} \;\middle|\; S_t = s\right]$$

Adding c

$$V'_\pi(s) = E_\pi\left[(R_{t+1}+c) + \gamma(R_{t+2}+c) + \cdots + \gamma^{k-1}(R_{t+k}+c) \mid S_t = s\right]$$

$$V'_\pi(s) = E_\pi\left[(c + \gamma c + \cdots + \gamma^{k-1} c) + G_t \mid S_t = s\right]$$

$$V'_\pi(s) = E_\pi\left[\frac{c(1-\gamma^k)}{1-\gamma} + G_t \mid S_t = s\right]$$

We get this additional term of $\frac{c(1-\gamma^k)}{1-\gamma}$

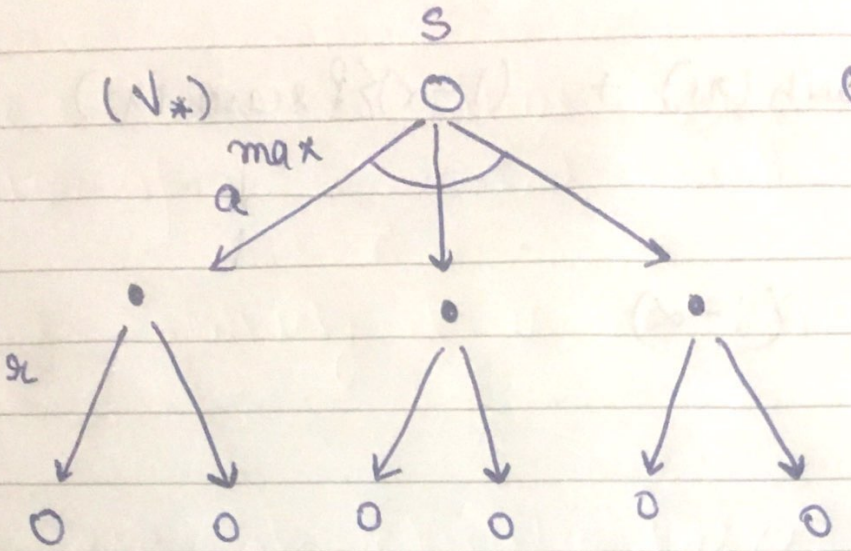
that is a random variable because k is random k is the time step at which terminating state is reached. Hence this depends on

policy we choose and the state we begin in. If k is more then this means we started from a state that is far from the terminating state and hence $1-\gamma^k$ is ~~this~~ more ($0<\gamma<1$) and therefore $V_\pi(s)$ gets a ~~small~~ larger constant.

If k is less, $1-\gamma^k$ is less and hence $V_\pi(s)$ gets a smaller constant.

So states that are closer will relatively have smaller value.

Q5).



when you are at state s, choose an action a, you reach the • position. From there on choose the optimal ~~stat~~ action value function i.e. $q_*(s,a)$. for each action you took from state s, you will have an optimal further direction given by $q^*(s,a)$ $\forall\ a \in A(s)$.

Out of these, choose the best action and this gives the value V_*

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