Q1 ~~Running average return~~

s	a	s'	r	$p(s, a \mid s' r)$
high	search	big	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
low	search	high	r_{search}	$1-\beta$
low	search	low	r_{search}	β
high	wait	high	r_{wait}	1
low	wait	low	r_{wait}	1
low	recharge	high	0	1

Solved using transition graph. As r_{search}, r_{wait} are guaranteed rewards fixed for given (s, s', a) triple, $\therefore p(s', r(s,a,s') \mid s, a) = p(s' \mid s, a)$

Q3 ~~$G_t = R_{t+1}$~~ let R' be new rewards

$$G'_t = R'_t + \gamma R'_{t+1} \cdots = \sum_{k=0}^{\infty} \gamma^k \left(R'_{t+k+1}\right)$$

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

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

$$= G_t + \frac{c}{1-\gamma} \quad \therefore \boxed{v_c = c/(1-\gamma)}$$

$$\therefore v'_\pi(s) = E\left[G'_t \mid S_t = s\right] = E\left[G_t + v_c \mid S_t = s\right]$$

$= E[G_t | S_t = s] + v_c$

$= v_\pi(s) + v_c$

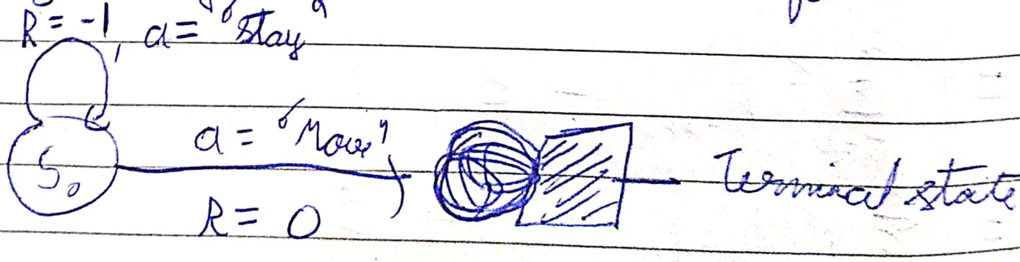
∴ adding constant to all rewards has no effect on relative values of states under any policy.

Q3b) $G'_t = \sum_{k=0}^{T} \gamma^k R'_{t+k+1} = \sum_{k=0}^{T} \gamma^k (R_t + c)$

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

As there is still a factor of T which can depend on policy, there is an effect of the constant.

example: ~~consider a maze-solving~~, ~~if normally reward is -1~~ for



In this case the best policy goes straight to terminal state.

However if we add c = 2, then the best policy 'stay' is forever looping.

Date
DELTA Pg No.

~~Date Page~~

Q5) $v_*(s) = \max_{a \in A(s)} q_*(s, a)$

Optimal policy makes best decision greedily