Date: ....

RL Homework 2

=> Given $p(s'|s,a)$, we have :-

(1)
$$p(r|s',s,a) = \frac{p(r,s'|s,a)}{p(s'|s,a)}$$

Also, $r(s,a,s') = \sum_r r\, p(r|s,a,s')$

[Expected Reward]

=> Reward can either be 0 [when no can is found] or 1 [when can empty can is found].

As from example 3.3; for $(s,a,s') = ($high, search, high$)$:-

$$r(s,a,s') = r_{search} = \sum_{r\in[0,1]} r\, p[r|s,a,s')$$

& $p(s'|s,a) = \alpha$ [Given]

$$\Rightarrow r_{search} = 0\, \frac{p(r=0,s'|s,a)}{p(s'|s,a)} + 1\, \frac{p(r=1,s'|s,a)}{p(s'|s,a)}$$

$$\boxed{\Rightarrow r_{search}.\alpha = p(r=1,s'|s,a)}$$

Also, $\sum_r p[r,s'|s,a] = p(s'|s,a)$

$$\therefore \alpha = \alpha r_{search} + p(r=0,s'|s,a)$$

$$\boxed{\therefore p(r=0,s'|s,a) = \alpha - \alpha r_{search}}$$

Similar calculations are done for other parts as well.

Final Table :-

s	a	s'	r	$p(s',r\|s,a)$
High	Search	High	0	$\alpha - \alpha r_{search}$
High	Search	High	1	αr_{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	$\beta . r_{search}$

s	a	s'	r	$p(s', r \mid s, a)$
Low	Search	High	−3	$1-\beta$
High	Wait	High	0	$1-\gamma$ wait
High	Wait	High	1	γ wait
Low	Wait	Low	0	$1-\gamma$ wait
Low	Wait	Low	1	γ wait
Low	Recharge	High	0	1

Ex 3·15

The sign does not matter. Only the relative value of one action as compared to the others matter, so it depends on the "Intervals between rewards" rather than the signs.

•> $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$.

adding a constant (c) to all the rewards :-

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

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

$$+ c + c\gamma + c\gamma^2 + \cdots$$

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

It being a constant remains as it is over expectation.

•> $\therefore V_c = \left(\frac{c}{1-\gamma}\right)$

It being a constant does not affect the relative values of any states under any policies.

Ex 3.16

In case of an episodic task,

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

Adding a constant to all rewards :-

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

$$= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{n-1} R_{t+n} + c\left[\frac{1-\gamma^n}{1-\gamma}\right]$$

$$V_c = c\left[\frac{1-\gamma^n}{1-\gamma}\right]$$

As V_c is not a constant, this would have an effect over the expectation & hence the value function of all states

Scanned by CamScanner

Date: ....

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

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

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

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

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

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

$\therefore$

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