RL HW 2

Surabhi S. Nath

__/__/____

Q1 States : Low, High

Actions : Search, Wait, Recharge

A(Low) = {search, wait} A(high) = {search, wait, recharge}.

$p(s', r | s, a) = \underbrace{p(s'|s,a)}_{\text{can be found from the table}} \times \underbrace{p(r|s',s,a)}_{\text{we know}}$ since independent

$r(s,a,s') = \sum_r r \underbrace{p(r|s,a,s')}_{\text{we can get this from the table.}}$

Therefore, calculating these equations we get

s	a	s'	r	$p(s', r \mid s, a)$.
high	search	high	0	$\alpha(1-r_{search})$
high	search	high	1	$\alpha(r_{search})$
high	search	low	0	$[1-\alpha][1-r_{search}]$
high	search	low	1	$[1-\alpha](r_{search})$
low	search	high	1	$\cancel{\alpha(r_{search})}$ $[1-\beta]$
low	search	low	0	$\beta(1-r_{search})$
low	search	low	1	βr_{search}
high	wait	high	0	r_{wait}
high	wait	~~high~~ low	1	r_{wait}
low	wait	low	1	r_{wait}
low	wait	low	0	$1 - r_{wait}$
low	recharge	high	0	1

Q3 a)

To show: Adding a constant c to all rewards adds a constant v_c to the values of all states and does not affect relative values of any states under any policy.

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

Now adding constant c to all rewards,

- $G_t' = (R_{t+1} + c) + \gamma(R_{t+2}+c) + \gamma^2 (R_{t+3}+c) + \cdots$
  $= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c).$

$$V_\pi'(s) = E_\pi[G_t' \mid S_t = s] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1}+c) \mid S_t = s\right] \quad \forall s.$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c \mid S_t = s\right]$$

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

$$= E_\pi[G_t \mid S_t = s] + \sum_{k=0}^{\infty} \gamma^k c$$

$$\text{const.} \leftarrow c \times \frac{1}{1-\gamma}$$

$\therefore$ value of all states is increased by $v_c = \frac{c}{1-\gamma}$ and relative values of $V(s)$ are not affected. i.e. signs of the rewards are not important as adding a constant to make them all positive doesnot affect the relative V values of the learning

b) Episodic task

In an episodic task, the number of steps are limited by T.

Thus we get:

$$V'_\pi(s) = E_\pi[G_t \mid S_t = s] + \sum_{k=0}^{T} \gamma^k c$$

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

v_c here depends on T

ie the number of steps.

~~in the simulation~~

However, T is fixed for an episode ~~task~~ and all $V_\pi(s)$ will hence be added by the same value. $\therefore$ v_c is constant within ~~[illegible]~~ episodes. Across episodes T may change $\therefore$ relative values across episodes can be different.

4) This can also be visualized as a special case of the previous formulation with ∞ steps. Here, reward after time instant T will repeatedly be 0. $\therefore$ for a given T, it is same as the previous case.

Q5 Equation for v_* in terms of q_*

$$v_*(s) = \max_{a \in A(s)} q_*(s, a) \quad (1)$$

$$v_*(s) = \max_{a \in A(s)} \sum p(s', r \mid s, a)\left(r + \gamma v_*(s')\right) \quad (2)$$

Using (1) and (2) we get:

$$v_*(s) = \max_{a \in A(s)} \sum p(s', r \mid s, a)\left(r + \gamma \max_{a' \in A(s')} q_*(s', a')\right)$$