RL - Theory - 2

Vishal Samal
2017209.

1.

s	a	s'	r	p(s', r | s, a)
h	s	h	r_s	α
h	s	l	r_s	$1-\alpha$
h	w	h	r_w	1
l	s	h	-3	$1-\beta$
l	s	l	r_s	β
l	w	l	r_w	1
l	r	h	0	1

For s, h → high
l → low

For a, s → search
w → wait
r → recharge

For s', h → high
l → low

To obtain the table it took all the possible s, a, s', r' values and then found the prob. with which they will occur.

3. 3.15

Signs of the rewards are important since it would affect overall expected sum reward.

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

Adding constant 'c' to each reward,

$$G'_t = (R_{t+1} + C) + \gamma(R_{t+2} + C) + \cdots$$

$$\Rightarrow G'_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$
$$+ c + c\gamma + c\gamma^2 + \cdots$$

$$\Rightarrow G'_t = G_t + \frac{c}{1-\gamma} \quad \text{assuming } 0 \le \lambda < 1$$

$$\Rightarrow V_c = \frac{c}{1-\gamma}$$

Hence, adding constant to each term leads to addition of a constant to the overall sum.

3.11

For episodic task,

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T$$

$$= \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

Adding a constant 'c' to all rewards, would again similar to 3.1 would add a constant to all the returns

Since, we are adding a constant to all rewards it would be similar to assigning new rewards thus, task remains unchanged.

Eg:- In a grid world, from (0,1) to (2,3) reward is 3, and 0 for all actions. and then we add +2 to it making the reward +5 from (0,1) to (2,3) and +2 otherwise.

But all func^n would be same
reltaviley ~~action values~~ remain similar.
Thus, tags want change

5. $V_*(s)$ is the optimal state-value function
$Q_*(s,a)$ is the optimal action-value fn

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

Since, we ~~need~~ would greedily chose optimal action for optimal state value function.

3. 3.16 *

The constant here varys with the time step $c\left[\frac{1-\gamma^M}{1-\gamma}\right]$ M is the no. of step left

Thus, as we go forward this toward 0 would decrease returns.

eg:- For a maze solving if we reward 1 for each step then earlier steps would have larger return as M is larger. and G_t will also be larger.

3.15 *

Negative rewards can be offset by a constant their effect will be nullified.