ML-A4

Theory Que 1)

Initial values given :-

$V(S_0) = 0 \; ; \; V(S_1) = 0 \; ; \; V(S_2) = 0 \; ; \; V(S_3) = 0$

(a) We know that

$$V^{k+1}(s) = \max_a \left(R(s,a) + \gamma \sum P(s'|s,a) V^k(s') \right)$$

R - Reward
P - Prob
a - actions
γ - Discount

Iteration 1

$$V^1(S_0) = \max_a \left(1 + 0.9 \left(0.5 * 0 + 0.5 * 0 \right) \right)$$

$$= 1$$

$V^1(S_1) = 2$

$V^1(S_2) = 3$

$V^1(S_3) = 10$

In first iteration, values turns out to be same as reward initial values are for all states

Iteration 2

$$V^2(S_0) = \max \left(1 + 0.9 \left(0.5 * V^1(S_1) + 0.5 * V^1(S_2) \right) \right)$$

$$= 1 + 0.9 \left(0.5 * 2 + 0.5 * 3 \right)$$

$$= 1 + 0.9 (1 + 1.5)$$

$$= 1 + 0.9 (2.5)$$

$$= 3.25$$

$$V^2(S_1) = \max \left(2 + 0.9 (0.5 * V^1(S_1) + 0.5 * V^1(S_2)) \right.$$
$$\left. 2 + 0.9 (0.1 * V^1(S_2)) \right)$$

$$= \max \left(2 + 0.9 (0.5 * 2 + 0.5 * 10) \right.$$
$$\left. 2 + 0.9 (3) \right)$$

$$= \max \left(2 + 0.9 (6), 2 + 0.9 (3) \right)$$

$$= 7.4$$

$V^2(S_2) = \max\left(3 + 0.9(1 * V'(S_0))\right)$

$= 3 + 0.9 * 1 \quad = 3.9$

$V^2(S_3) = \max\left(10 + 0.9(1 * V'(S_3))\right)$

$= 10 + 0.9(1 * 10)$

$= 10 + 9 \quad = 19$

Iteration 3

$V^2(S_0) = \max\left(1 + 0.9(0.5 * 7.4 + 0.5 * 3.9)\right)$

$= 1 + 0.9(3.7 + 1.95)$

$= 6.085$

$V^2(S_1) = \max\left(2 + 0.9(0.5 * 7.4 + 0.5 * 19),\right.$

$\left.2 + 0.9(3.9)\right)$

$= \max\left(2 + 0.9(13.2), 2 + 0.9(3.9)\right)$

$= \max(13.88, 5.51)$

$= 13.88$

$V^2(S_2) = \max\left(3 + 0.9(1 * 3.25)\right)$

$= 3 + 0.9(3.25) \quad = 5.925$

$V^2(S_3) = \max\left(10 + 0.9(1 * 19)\right)$

$= 10 + 0.9(19) \quad = 27.1$

(b) It can be noticed that the optimal policy for state S_1 is :- the movement from S_1 to S_3. This is because S_2 has a large reward.

$$V'(S_3) = R(S_3) + \gamma * V'(S_2).$$
$$= 10 + 0.9 * V'(S_2)$$
$$= 100$$

Now, optimal value for state S_1

$$V'(S_1) = 2 + 0.9(0.5 * V'(S_1) + 0.5 * V'(S_3))$$
$$= 85.454545\ldots$$

Thus, it is clear that best policy for state S_1 is moving from S_1 to S_3.

(c) (i) False

Since the value will get updated in every iteration, it won't converge.

$$V(S_2)^{t} = \max_{a} \left(R(S,a) + \gamma \sum P(S^{t+1} | S, a) V(S^{t}) \right)$$

(ii) False
MDP does not converge. The value keeps getting updated in every iteration.

(iii) True
If $\gamma = 0$, then $V(s_0)^t = \max(R(s,a))$
So, value will be same as the reward for that state.

(iv) True,
Acyclic MDP ⇒ having no cycles.
After every iteration in an acyclic MDP, at least one state gets to optimal value.
Thus after N iterations, acyclic MDP converges.

(v) False
Since it is given that there are no absorbing goal states, so, MDP won't converge.