Charbel - Raphael Segerie DM1 RL

Ex 1 :

2)

1- We can take $\forall s \neq 1, s \neq 14 \quad r(s,a) = -\varepsilon$.
with $\varepsilon > 0$

ε should be positive in order to avoid the infinite cycle policy, but it should not be too big, because the agent would then suicide directly
But by taking $\varepsilon = 1$, because the distance between the cases 1 and 14 is only of 4, the agent has no need to suicide if he appears on the case 2

$$V(2) = \underset{2 \to 6}{-\varepsilon} \underset{6 \to 10}{-\varepsilon} \underset{10 \to 14}{-\varepsilon} \underset{14 \to \emptyset}{+10} = -3\varepsilon + 10$$

So $-3\varepsilon + 10 > -\varepsilon + V(1) = -10 \cdot \varepsilon \Rightarrow \varepsilon < \frac{20}{2} = 10$

It is the same reasoning for $V(5)$

So by taking $\varepsilon = 1$ there is no risk to suicide

The resulting value is :

5	-10	7	6
6	7	8	5
7	8	9	4
8	9	10	3

final state

2. We suppose a general MDP with reward $r(s,a)$.
If we take an affine transformation $r'(s,a) = Ar(s,a) + B$
The new value function is $V' = AV + B$ (because we keep the same policy). But the optimal policy is not the same.
Even if $A > 0$, by selecting the appropriate B we can change the policy. For example if $A = 1$, $B = -1000$ if we take the same environement as in Q1, the optimal policy is the quickest suicide.

3. Because $r'(s,a) = r(s,a) + 5 = 4 > \forall s \neq (1, 14)$
So the goal become equivalent to staying for eternity on the board by jinging the infinite cycle
$5 \to 6 \to 10 \to 9 \to 5$. And the value is $+\infty$ almost everywhere but in states 1 and 14.

DM1 RL

Ex2 $V^*(s) - V^{\pi_Q}(s) = \underbrace{Q^*(s,\pi^*(s)) - Q^*(s,\pi_Q(s))}_{A} + \underbrace{Q^*(s,\pi_Q(s)) - Q^{\pi_Q}(s,\pi_Q(s))}_{B}$ (le terme $Q^*(s,\pi^*(s)) - Q^*(s,\pi_Q(s))$ est ≤ 0)

$$A = Q^*(s,\pi^*(s)) - \underbrace{Q(s,\pi^*(s)) + Q(s,\pi^*(s))}_{0} - Q^*(s,\pi_Q(s))$$

$$\leq Q^*(s,\pi^*(s)) - Q(s,\pi^*(s)) + Q(s,\pi_Q(s)) - Q^*(s,\pi_Q(s))$$

car $\pi_Q(s) = \operatorname{argmax} Q(s,a)$

$$\leq \max_s \left| Q^*(s,\pi^*(s)) - Q(s,\pi^*(s)) \right| + \max_s \left| Q^*(s,\pi_Q(s)) - Q(s,\pi_Q(s)) \right|$$

$$\leq 2 \max_{s,a} \left| Q^*(s,a) - Q(s,a) \right| \leq 2 \|Q^* - Q\|_\infty$$

$$B = Q^*(s,\pi_Q(s)) - Q^{\pi_Q}(s,\pi_Q(s))$$

$$= \cancel{r(s,\pi_Q(s))} + \gamma \sum_{s'} P(s'|s,\pi_Q(s)) V^*(s') - \cancel{r(s,\pi_Q(s))} - \gamma \sum_{s'} P(s'|s,\pi_Q(s)) V^{\pi_Q}(s')$$

$$\leq \gamma \sum_{s'} P(s'|s,\pi_Q(s)) \|V^* - V^{\pi_Q}\|_\infty$$

$$\leq \gamma \|V^* - V^{\pi_Q}\|_\infty$$

Donc $\forall s$, $V^*(s) - V^{\pi_Q}(s) \leq 2\|Q^* - Q\|_\infty + \gamma \|V^* - V^{\pi_Q}\|_\infty$

$$\Rightarrow \|V^* - V^{\pi_Q}\| \leq 2\|Q^* - Q\|_\infty + \gamma \|V^* - V^{\pi_Q}\|_\infty$$

$$\Rightarrow \|V^* - V^{\pi_Q}\| \leq \frac{2}{1-\gamma} \|Q^* - Q\|_\infty \quad (\gamma < 1)$$

$$\Rightarrow \forall s, \quad V^*(s) - V^{\pi_Q}(s) \leq \frac{2}{1-\gamma} \|Q^* - Q\|_\infty$$

et $\boxed{V^{\pi_Q}(s) \geq V^*(s) - \frac{2}{1-\gamma} \|Q^* - Q\|_\infty}$

Chorbel - Rannal Sraoui

Q3 We develop the wanted expression

$$\sum_s \mu^{\pi'}(s) \sum_a (\pi'(a|s) - \pi(a|s)) Q^{\pi}(s,a)$$

$$= \sum_s \mu^{\pi'}(s) \sum_a (\pi'(a|s) - \pi(a|s)) \Big(\underbrace{r(s,a) - g^{\pi}}_{A} + \underbrace{\sum_{s'a'} p(s'|s,a)\pi(a'|s') Q^{\pi}(s',a')}_{B} \Big)$$

$$= A + B$$

$$A = \sum_{s,a} \mu^{\pi'}(s) (\pi'(a|s) - \pi(a|s)) (r(s,a) - g^{\pi})$$

$$= \sum_{s,a} \mu^{\pi'}(s) \Big(\underbrace{\pi'(a|s) r(s,a)}_{\text{We remark that } \sum_{s,a} \mu^{\pi'}(s)\pi'(a|s) r(s,a) = g^{\pi'}} - \underbrace{\pi'(a|s) g^{\pi}}_{\text{We sum first on } a \text{ and then on } s \rightarrow g^{\pi}} - \pi(a|s) r(s,a) + \pi(a|s) g^{\pi} \Big)$$

$$= g^{\pi'} - \sum_{s,a} \mu^{\pi'}(s) \pi(a|s) r(s,a)$$

$$B = \sum_{\substack{s,a \\ s',a'}} \mu^{\pi'}(s) (\pi'(a|s) - \pi(a|s)) p(s'|s,a) \underbrace{\pi(a'|s') Q^{\pi}(s',a')}_{\text{cste of } s,a}$$

$$= \sum_{s',a'} \pi(a'|s') Q^{\pi}(s',a') \underbrace{\sum_{s,a} \mu^{\pi'}(s) (\pi'(a|s) - \pi(a|s)) p(s'|s,a)}$$

$$= \mu^{\pi'}(s') - \sum_{s,a} \mu^{\pi'}(s) \pi(a|s) p(s'|s,a)$$

inversion of Sum

$$= - \sum_{s,a} \mu^{\pi'}(s) \pi(a|s) \underbrace{\sum_{s'a'} \pi(a'|s') Q^{\pi}(s',a') p(s'|s,a)}_{\text{Bellman } = Q^{\pi}(s,a) - r(s,a) + g^{\pi}}$$

$$+ \sum_{s',a'} \pi(a'|s') Q^{\pi}(s',a') \mu^{\pi'}(s')$$

$$B = -\sum_{s,a} \mu^{\pi'}(s)\,\pi(a|s)\,Q^{\pi}(s,a) \longrightarrow$$

$$-\sum_{s,a} \mu^{\pi'}(s)\,\pi(a|s)\,\left(-r(s,a) + g^{\pi}\right) \longrightarrow 0$$

$$+\sum_{s',a'} Q^{\pi}(s',a')\,\pi(a'|s')\,\mu^{\pi'}(s') \longrightarrow$$

$$= -\sum_{s,a} \mu^{\pi'}(s)\,\pi(a|s)\,\underbrace{\left(g^{\pi} - r(s,a)\right)}_{\text{sum on } a \text{ then on } s}$$

$$= -g^{\pi} + \sum_{s,a} \mu^{\pi'}(s)\,\pi(a|s)\,r(s,a)$$

$$\boxed{A + B = g^{\pi'} - g^{\pi}}$$

Chorbel - Raphael Sigerie

Ex 4 We have a 6 story building with 2 elevators.

1) We have to define the tuple $\Pi = (S, A, p, r)$.

1. $S = (z_t^1, z_t^2)$ with no history.
$(z^1_t, z^2_t) \in [\![0, 6]\!]^2$.
We neu do not modelise the time the elevators takes to go from one story to another one. And we use a discrete time. One unit of time is the time corresponding to opening the doors, entering and exiting the elevator, pushing the button, waiting to close the door and going to another story, not necessary the one just above or just below.
This modulisation is valid if the elevator is fast (to go from a story to another one) but the doors and the people are slow.

But we have also to modulise the different people.
$S = ((z_t^1, z_t^2), W, E)$. W_{ij}^t is the number of people currenty waiting on the i-th story willing to go to the j-th story. $E_{i,t}^e$ is the number of people currenty in the elevator e (which is currently in the story z_t^e) willing to go to the story $i \neq z_t^e$.

The set of possible actions is to go to another story
So $A = ((z_{t+1}, z_{t+1}) \in [\![0, 6]\!]^2)$
If the elevator was slow or the number of story was much bigger, we would choose
$A = ((dz_1, dz_2) \in [-1, 1]^2)$, but that is not the case in our assumptions.

$p(s'/s,a)$? We consider the elevator is responding pefectly to the orders of the agent. So $z^e_{t+1} = a^e_t$ with no stochasticity. We make the hypothesis the elevators can contain any number of people. And people in the story just want to go on the floor, and they enter the first elevator they see. They do not exit an elevator unless they are at the right story.

So $\left| E^e_{z,t+1} = E^e_{z,t} + W^t_{z^e_t, z} \quad \text{if } z \neq z^e_{t+1} \right.$

" the number of people in elevator e, willing to go to story z equals The previous number of people willing to go to z plus the people willing to go to z who were on the prescedent story. *

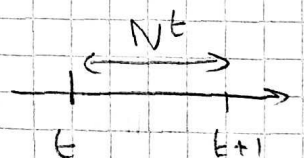
$$E^e_{z,t+1} = 0 \quad \text{if } z = z^e_{t+1}$$

" They exited the elevator "

$$W^{t+1}_{z_1,z_2} = W^t_{z_1,z_2} + N^t_{z_1,z_2} \quad \text{if } z^e_t \neq z_1$$

" People who were waiting are still waiting and did not took the stairs if the elevator did not come "

$$W^{t+1}_{z_1,z_2} = N^t_{z_1,z_2} \quad \text{if } z^e_t = z_1$$



" People who were waiting took the elevator and a bunch of new people arrived "

* NB = if elevator 1 and 2 simultenceusly come to a story, the waiting people all choose the elevator with the smallest number of people within.

Charbel - Raphael Segerie RL1 exo 4

$N^t_{i,j}$ is the number of people coming between time t and $t+1$.

So they always wait until $t+1$, and if the elevators come, they enter in.

$N^t_{i,j}$ follows a Poisson law $\mathcal{P}(\lambda^t_{i,j})$

If we want the process to be Markovian, we should not consider the time dependance of $\lambda^t_{ij} \rightarrow \lambda_{ij} = cte$.

So in our modelisation, there is no Night and Day cycle.

Finally, the reward function is $r(s,a,s')$: We can take a penalty function of the number of people currently waiting.

$$r(s_t, a, s_{t+1}) = -\sum_{i,j} w^t_{ij} - \sum_{\substack{e=1,2 \\ i}} E^e_{i,t} \quad \forall a$$

The penalty is linear in the total waiting time.

If we choose instead ~~the~~ to modelise the penalty as the sum of the squares of the individuals waiting time, we could not do it unless we modelise each individual.

Cherbel - Raphael RL1

Ex 5 1 : See Code

2. See Code

3. For proba_succ = 0,8 : (stochastic)

The value iteration terminated with 1194 iterations.

The policy iteration terminated in 9 iterations.

And the value funtion converges to 691.

For proba_succ = 0,99 (almost non stochastic)

The value iteration terminated with 1175 iterations

The policy iteration terminated with 12 iterations

And the value function converges to 545.

545 is smaller than 691 which is normal, because the non stochastic environement is less risky.

The number of iterations in the value iteration is almost the same

But the number of iterations decreases in the policy iteration if there is stochasticity.

If we are in the non-stochastic environment, there is multiplicity of the optimal policy, because it is equivalent to go first on the right and then in the bottom. or to go on the bottom and then on the right.

But in the stochastic environment, → ↓ or ↓→ are not the same because cases contening a trap. So the optimal policy in the stochastic environment tries to reduce the time on the frontier of the cases with a trap.

4. Value Iteration needs more iterations but each iteration is quick to run. Policy iteration needs less iterations but each iteration takes more time

There is no unicity in the policy iteration, but there is unicity in the value iteration.

The time needed to compute the value iteration is longer (22 s) than the time needed to compute policy iteration (0,3 s).

The stopping criterion in the policy iteration is straightforward (equality of the policies) whereas the stopping criterion of the value iteration is a little bit more technical, because only asymptotic.