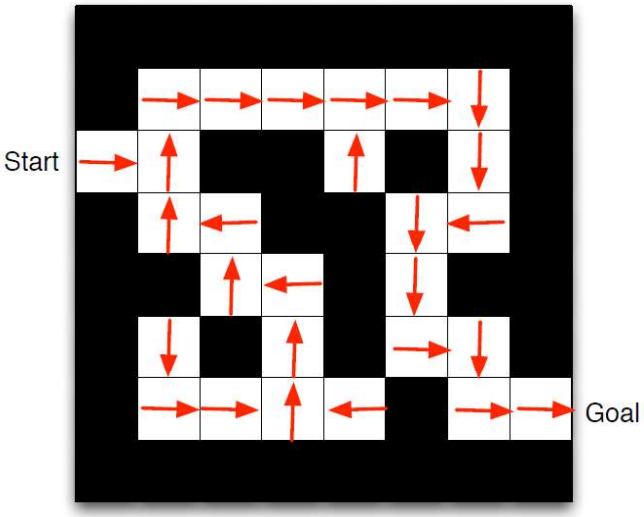
MI0B903T - Exercices Processus de décisions markoviens

Exercice 1
Dans l'exemple du cours du labyrinthe, nous considérons la politique suivante :



Déterminez la fonction valeur associée à cette politique en supposant que un pas coûte -1.

Exercice 2
Considérons la grille suivante :

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Il y a deux états terminaux : les cases grisées 1 et 16. Ce sont les cases cibles à atteindere par l'agent. Elles auront toujours une fonction valeur égale à 0.

Pour chaque transition, la récompense (coût) est de -1. On choisit la politique uniforme. Plus précisément, les actions possibles sont

— les quatre déplacements pour les quatre points centraux : haut, bas, droite et gauche. L'agent suit la politique uniforme : il choisit l'une des quatre directions avec probabilité 1/4.

— seulement trois déplacements pour les points du bord. L'agent suit la politique uniforme : il choisit l'une des trois directions avec probabilité 1/3.

— seulement deux déplacements pour les points en coin. L'agent suit la politique uniforme : il choisit l'une des deux directions avec probabilité 1/2.

Ici la politique π est fixée et on souhaite
— d'abord déterminer la fonction valeur V^π associée en utilisant l'algorithme de value-iteration
— puis en déduire une amélioration de cette politique.

1. Première étape : value-iteration pour déterminer la fonction valeur.
   Donnez les valeurs des états après trois itérations de l'algorithme value-iteration en partant de la fonction valeur nulle en chaque état :

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

   La formule de mise à jour de la fonction valeur est la suivante :

$$V_0(s) = 0$$
$$V_{k+1}(s) = r(s,a) + \sum_{s' \in S} p(s'|s,a) V_k(s').$$

   En itérant cette procédure, on converge vers la vraie fonction valeur associée à la politique :

$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

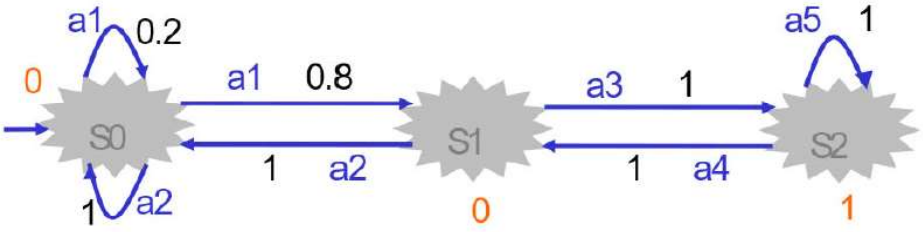
$\leftarrow v_\pi$.

2. Deuxième étape : amélioration de la politique.
   Proposer une nouvelle politique π' qui sera meilleure que la politique initiale π. Calculer la fonction valeur correspondante.

Exercice 3
Soit le processus de décision markovien suivant :
dont les transitions sont étiquetées par les noms des actions a_i et les probabilités de transitions ; les états s_j sont étiquetés par les récompenses correspondantes :

$$R(s_0) = 0, \quad R(s_1) = 0, \quad R(s_2) = 1.$$

1. Value-iteration pour calculer la fonction valeur optimale. Donnez les valeurs des états à la fin de trois itérations de l'algorithme value-iteration, en supposant qu'on utilise le critère infini



pondéré avec un facteur d'atténuation $\gamma = 0.5$ et en partant initialement avec des valeurs des états toutes égales à 0 :

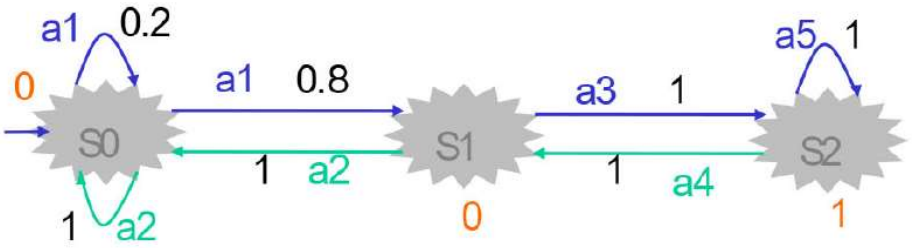
$$V(s_0) = 0, \quad V(s_1) = 0, \quad V(s_2) = 0.$$

Donnez le plan d'actions (politique) correspondant à ces valeurs.

2. Policy-iteration pour déterminer la politique optimale. Déterminer la politique optimale en prenant comme politique initiale :

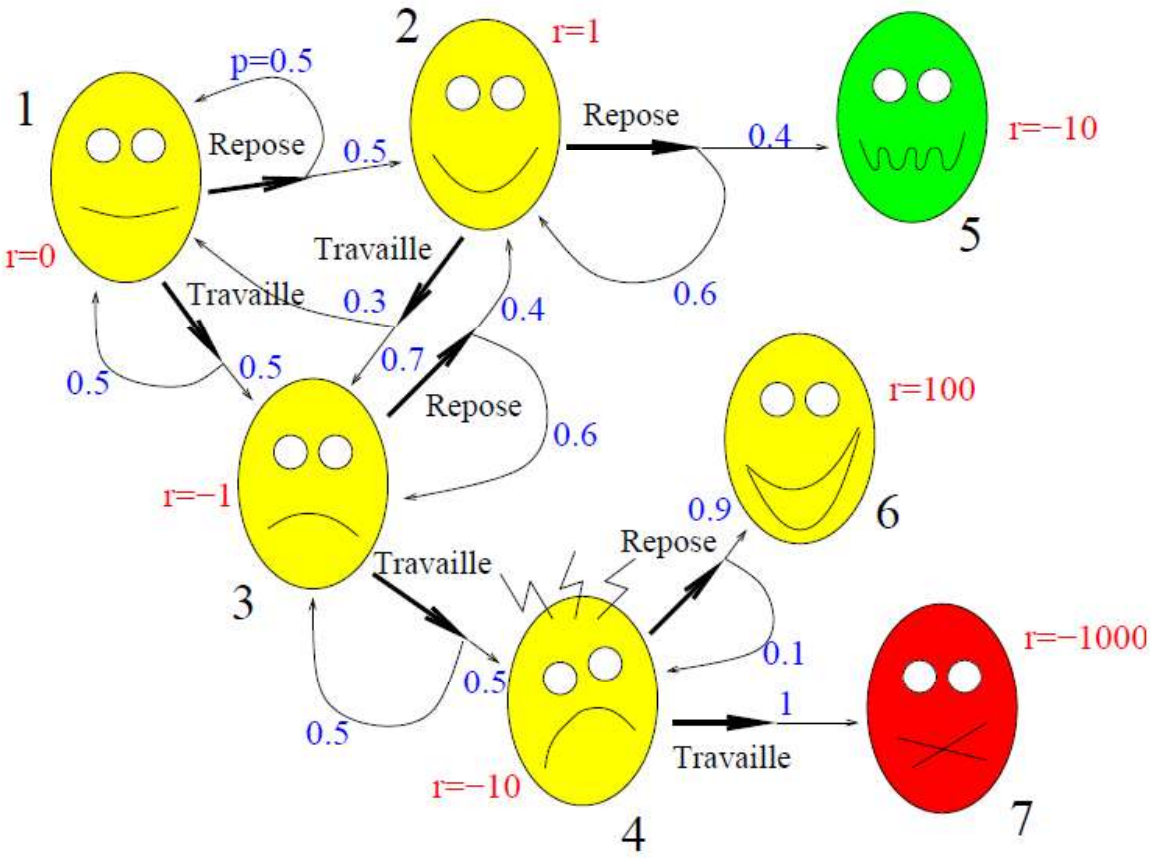
$$\pi_0(s_0) = a_2, \quad \pi_0(s_1) = a_2, \quad \pi_0(s_2) = a_4$$

indiquée en vert dans le graphique ci-dessous et le critère infini pondéré avec facteur d'atténuation $\gamma = 0.5$.



Exercice 4

Déterminer la politique optimale pour l'exemple du cours Travail ou repos ?



Exercice 5

On souhaite maximiser le revenu provenant de la production d'une machine. Or, le niveau de production dépend de l'état de la machine, et l'état de la machine dépend de son entretien. Les données du problème sont les suivantes.

— Les états sont : 0 (neuf), 1 (bon état), 2 (mauvais état) et 3 (en panne).
— Les rendements respectifs (par période) sont de 30, 15, 5 et 0.
— Les actions possibles sont : entretenir (1), ne rien faire (2), rénover (3).
— Le revenu par objet produit est de 100$.
— Le taux d'actualisation est $\gamma = 0.8$.

1. Compléter le tableau suivant qui fournit les informations sur les revenus associés aux différentes combinaisons d'états et de décision, ainsi que les probabilités de transition.

état	action	coût	probabilité de transition	nouvel état	revenu (gain - coût)
0	entretenir	500$	3/4	0	
			1/4	1	
	ne rien faire	0$	4/5	1	
			1/5	3	
1	entretenir	1000$	4/7	1	
			2/7	2	
			1/7	3	
	ne rien faire	0$	4/5	2	
			1/5	3	
	rénover	3000$	1	0	
2	entretenir	1000$	3/4	2	
			1/4	3	
	ne rien faire	0$	1/2	2	
			1/2	3	
	rénover	3000$	1	0	
3	rénover	3000$	1	0	

2. On considère la politique π :
   — entretenir une machine neuve ($\pi(0) = 1$) ;
   — ne rien faire si la machine est en bon état ($\pi(1) = 2$) ;
   — entretenir une machine en mauvais état ($\pi(2) = 1$) ;
   — réparer une machine en panne ($\pi(3) = 3$).

   Déterminer la matrice de transition P^π et le coût R^π de cette politique.
3. Déterminer la fonction valeur V^π associée à la politique π en résolvant le système linéaire

$$V^\pi = R^\pi + \gamma V^\pi \quad \text{soit} \quad (I - \gamma P^\pi)V^\pi = R^\pi.$$

4. Améliorer la politique en améliorant π sur une seule étape.