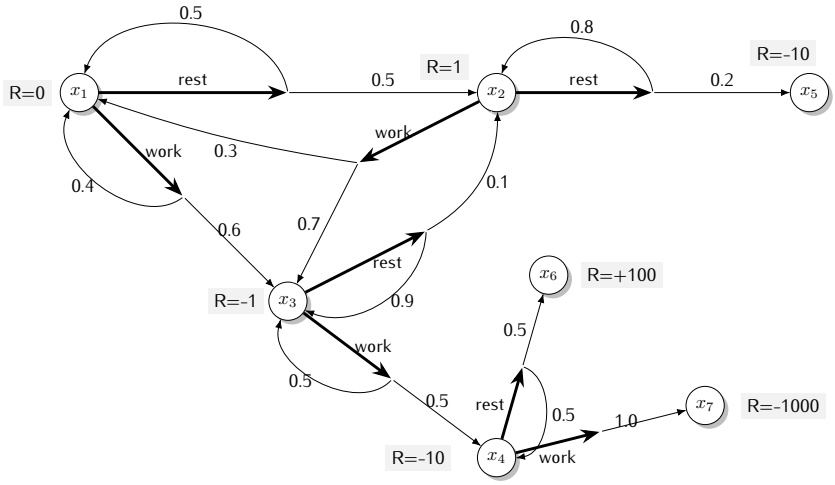
Examen de première session

La durée de l'examen est de deux heures. Vous n'avez pas droit à des documents ni à la calculatrice. Le sujet comporte trois pages.

Exercice 1 – MDP



Les actions sont indiquées avec des flêches en gras, la fonction de transition avec des flêches plus fines et l'indication des probabilités. La fonction de récompenses est deterministe est est indiquée par l'étiquette $R = value$.

1. Si un agent effectue la séquence d'action ⟨work,work,rest⟩ à partir de l'état x_1, dans quels états peut il se retrouver et avec quelle probabilité ?
2. Définissez *une* politique π et écrivez le système linéaire qui définit v_π pour cette politique. Votre réponse finale devrait être prête pour faire l'application numérique si on avait accès à un solveur de système linéaire.

Exercice 2 - Modélisation

Au début de chaque nuit, un cambrioleur doit décider s'il prend sa retraite de sa vie de cambrioleur. Si il décide de continuer, il doit aussi décider s'il fait un ou deux cambriolage pendant la nuit. Si le cambriolage est un succès, il gagne en moyenne 500€. Chaque cambriolage a une probabilité de 75% de réussir. Si un cambriolage échoue, le cambrioleur perd tous les gains de la nuit (et ne peut pas faire un autre cambriolage cette nuit là !) et le cambrioleur est condamné avec sursis. Si le cambrioleur est pris une seconde fois, c'est fini ! Le cambrioleur est mis en prison et il est forcé de prendre sa retraite. Dans ce cas, le cambrioleur évalue cette fin comme une perte de 5.000 €.

1. Formulez ce problème comme un PDM : quels sont les états, les actions, le modèle de transition et le modèle de récompenses. Vous pouvez faire un schéma et l'étiquetter de manière claire.
2. Donnez explicitement (avec des nombres ou des fractions) l'équation de Bellman optimale pour ce PDM.

Exercice 3 - Q-learning et SARSA

On a un PDM avec 6 états (nommés 1, 2, 3, 4, 5 et 6) et 4 actions (nommées $\uparrow$, $\downarrow$, $\leftarrow$, et $\rightarrow$) qui se passe dans la grille ci-dessous. L'état 1 est l'état initial, l'état 6 est l'état final. On suppose qu'on ne peut pas prendre une action qui nous mène en dehors de la grille. Quand l'état 6 est atteint, on reçoit une récompense de 10 et l'épisode recommence à l'état 1. Pour toute autre action qui ne mène pas à l'état 6, la récompense est de -1.

4	5	6
1	2	3

Supposons qu'on ait une estimation des valeurs de q ci-dessous :

actions	$\uparrow$	$\downarrow$	$\leftarrow$	$\rightarrow$
1	4			3
2	6		3	8
3	9		7	
4		2		5
5		6	5	8

1. Serait-il une bonne idée d'utiliser une politique gloutonne maintenant ? Justifiez votre réponse.
2. Pourquoi travailler avec des valeurs de q plutôt qu'avec des valeurs de v ?
3. Supposons que l'on continue l'apprentissage en utilisant Q-learning et que l'on obtienne la trace suivante (succession d'états et d'actions) :
   $$2 - \uparrow - 5 - \rightarrow - 6$$
   Mettez à jour la table des valeurs de q.
4. Nous revenons maintenant aux valeurs de q de la table ci-dessus et supposons maintenant que la trace soit le résultat de l'exécution de SARSA. Mettez à jour les valeurs de q.

Exercice 4 - Monte Carlo

Supposons qu'on ait une chaîne de Markov avec récompenses avec deux états A et B. On utilise comme paramètres $\gamma = 1$ et $\alpha = 0.1$. On ne connait pas ni le modèle de transitions ni celui de récompenses mais on observe les deux trajectoires suivantes :

$A \rightarrow (+3)A \rightarrow (+2)B \rightarrow (-4)A \rightarrow (+4)B \rightarrow (-3)$

$B \rightarrow (-2)A \rightarrow (+3)B \rightarrow (-3)$

où $X \rightarrow (R)Y$ signifie qu'on est parti de l'état X, on a fait une transition vers l'état Y et on a obtenu la récompense R.

1. Estimez $v(A)$ et $v(B)$ avec l'algorithme de Monte Carlo en version première visite
2. Estimez $v(A)$ et $v(B)$ avec l'algorithme de Monte Carlo en version chaque visite

Exercice 5 - questions diverses

1. Quels sont les limitations des méthodes de Monte Carlo ?
2. Pour l'algorithme SARSA, on a parlé de méthode "on policy" alors que pour l'algorithme Q-learning, on a parlé de méthode "off policy". Expliquez pourquoi en quoi on a "on" ou "off" policy. Quels sont les différences clés entre les méthodes "on policy" et les méthodes "off policy" ?
3. Supposons que l'on fasse une descente de gradient avec une fonction qui approxime la fonction de valeurs. Expliquez ce qui ne va pas bien avec la mise à jour suivante :
   $$w_{t+1} \leftarrow w_t + \alpha \left[v_\pi(s) - \hat{v}(s_t, w_t) \right] \nabla \hat{v}(s_t, w_t)$$
4. Que devient l'expression précédente si on utilise une approximation linéaire ?