TD 1 Apprentissage par Renforcement - Master AIC

Laurent Cetinsoy, Diviyan Kalainathan, Michèle Sebag

20 novembre 2017

1 Exercice de cours

Ces questions ont pour but de vous aider à clarifier les différentes notions. Réponses courtes attendues.

Q. 1.1 *Quel est le but d'un algorithme d'un apprentissage par renforcement ?* what is the goal of Reinforcement Learning

Q. 1.2 *Dans ce cadre quel est le but d'un agent ?* In this framework what is the goal of an agent

Q. 1.3 *Comment le traduire formellement ?* How can you formalize it mathematically ?

Q. 1.4 *Qu'est ce qui différencie d'après vous un problème de RL d'un problème supervisé ? Qu'est ce qui rend un problème de RL Difficile ?* What make reinforcement learning different from supervised learning ? What makes reinforcement learning hard ?

Q. 1.5 *Formellement, qu'est ce qu'un MDP ? Donnez la signification de ses composants.* Give the formal definition of a Markov Decision Process (MDP). Give the meaning of it components

Q. 1.6 *Que represente le taux d'actualisation (discount), souvent noté γ ?* What does the discount factor γ mean or represent ?

Q. 1.7 *Pouvez vous donner des exemples de problèmes où les transitions ne sont pas déterministes ?* Can you give exemples of setup With non deterministic transitions ?

Q. 1.8 *De quoi dépend la récompense R ?* On what depends the reward function R

Q. 1.9 *Pour les problèmes suivants indiquer si l'espace des actions et des états est continu ou discret :* For each setups, tells if the space of action and state are continuous or discrete

- Cartpole. *Cartpole*
- La voiture autonome *Autonomous car*
- Le poker. *Poker*

Q. 1.10 *Quelle est la différence entre un problème épisodique et continue ? Donner un exemple de chaque.* What is the difference of an episodic setup and a continuous one ? Give an example of each

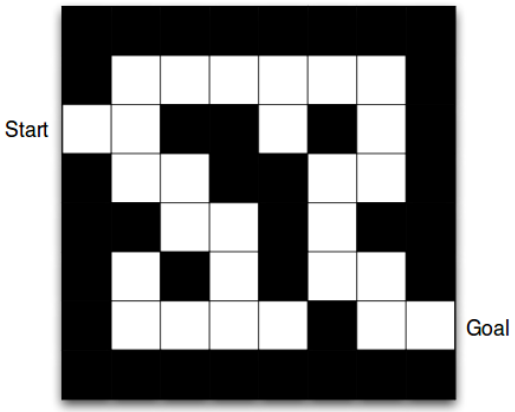
Q. 1.11 *Que réprésentent / signifient les fonctions de valeur V et Q ?* What does the V and Q functions means or represents ?

2 Exercice du labyrinthe

On considère un agent dans le labyrinthe ci dessous qui contient 27 cases. *Let us consider the following labyrinth made of 27 white boxes*

Afin de lui apprendre à chercher la sortie, on lui donne une récompense de $+r$ quand il est dans la case de sortie (Goal) et une punition de moins $-\alpha r$ à chaque fois qu'il se trouve dans une autre case. On considère cet l'environement déterministe : si l'agent décide d'aller à une case blanche adjacente, il se retrouve bien à cette case.

In order to teach him to find to reach the exit, we give it a positive reward +r if is reach the goal and and punition of $-\alpha r$ each time he is on another white square. We consider the environment as deterministic : if the agent decides to go to an adjacent white box, it indeed goes there.



Q. 2.1 *Quel est l'espace des états ?* What is the space of states ?

Q. 2.2 *Quel est l'espace des actions ?* What is the space of actions ?

Q. 2.3 *Pourquoi met-on on un reward négative à chaque itération qui n'atteint pas l'objectif ?* Why do we give a negative reward when the agent does not reach the goal ?

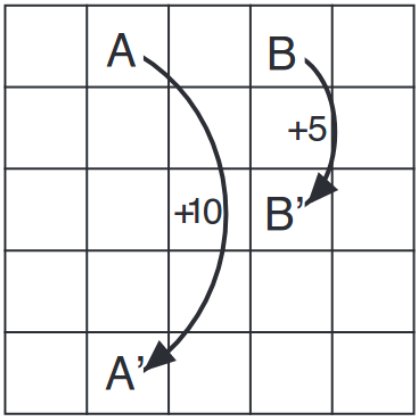
Q. 2.4 *Donner l'équation de bellman vérifiée par V* Give the bellman equation folloWed by V

Q. 2.5 *En déduire V pour chacun des états pour la politique optimale (c'est à dire si à chaque état, l'agent prend la meilleur décision possible).* Deduce what is V for each state for the optimal policy (the one where the agent takes the best decision at each step)

3 Grille

On considère l'environement suivant : un agent peut se déplacer sur une grille dans les quatre direction de l'espace (sauf quand il se retrouve au bord). Quand il se retrouve en A, l'agent reçoit une récompense de +10 et se retrouve téléporté en A'. Quand l'agent se trouve en B, il reçoit une récompense de +5 et se retrouve téléporté en B.

Let us consider the following environment : an agent can move a 2D grid in the four dimensions (except when it reaches the border). when it is on A, it receives a +10 reward and is teleported to A'. when it is in B, it receives a +5 reward and is teleported to B'



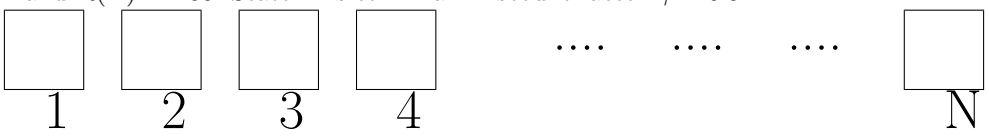
Q. *3.1* *On considère la politique explicite suivante : Toujours aller vers le haut. Estimer la fonction de valeur pour chacun des états. On considère un horizon infini et on prendra pour facteur d'actualisation* $\gamma = 0.9$. We consider the following policy : always going upside. Give the value function for all states. We consider an infinite horizon and we will take $\gamma = 0.9$ as discount factor.

Q. *3.2* *S'agit-il de la meilleur politique possible ? Dans le cas contraire proposer la meilleur.* Is it the best policy one can find ? If not give the best one

4 Rivière

On souhaite faire traverser une rivière à un agent que l'on représente par N cases. La case une est le versant de départ, la case N le versant d'arrivé. L'espace des état est donc 1..N et l'espace des action est gauche, droite. La fonction de récompense est la suivante : R(s) = 0 pour $s < N$ *et R(N) = 100. L'état N est terminal. Le facteur d'escompte est* $\gamma = .9$.

we want to make an agent cross a river modelised by N boxes. The box number 1 is the starting point and the box N is the goal. The state space is thus 1..N and the action space is left, right the reward function if R(s) = 0 for $s < N$ and R(N) = 100. State N is terminal. Discount factor $\gamma = 0.9$



Q. 4.1 *Let* π *be a random policy (*$\pi(s) = Left$ *or Right with probability 1/2). Let the transition function be defined as :*

$$p(i, Right, i+1) = 1 \quad if \;\; i < N$$
$$p(i, Left, i-1) = 1 \quad if\; i > 0 \qquad else\; p(0, Left, 0) = 1$$

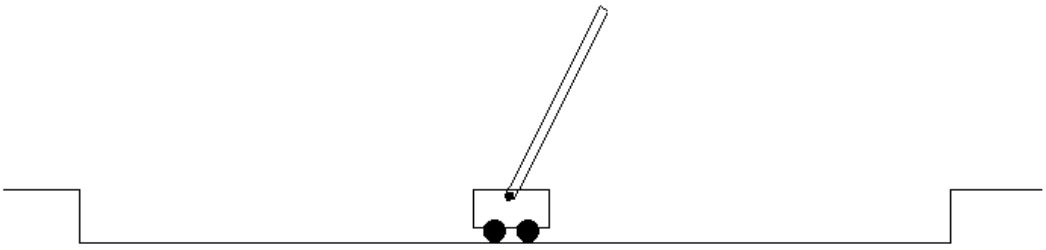
What is the probability of arriving in state N after 100 time steps ? After 1,000 time steps ?
Soit π une politique aléatoire ($\pi(s) = Left$ ou *Right* avec probabilité 1/2). Avec la fonction de transition ci-dessus, quelle est la probabilité d'arriver à l'état N en 100 pas de temps ? en 1,000 pas de temps ?

Q. 4.2 *Let* π *be the constant policy,* $\pi(s) = Right$ *for all s. With same function transition as above, compute the value function (function of N).*
Soit π la politique constante qui va toujours à droite. Avec la meme fonction de transition que pour la question précédente, calculez la fonction de valeur associée à π (fonction de N).

Q. 4.3 *Same question with* $p(i, Right, i+1) = .9;\; p(i, Right, i) = .1$.
Même question avec une fonction de transition probabiliste, $p(i, Right, i+1) = .9;\; p(i, Right, i) = .1$.

5 Pole balancing

On considère le problème de maintenir en équilibre un baton tenu par sa base à chariot pouvant se déplacer. Let us consider the CartPole problem where on wants to maintain a stick vertically. The stick is fixed at the bottom to a cartpole that can move.



Q. 5.1 *Proposez une structure de reward qui permettrait aux algorithmes de RL à apprendre à garder en équilibre la barre.* Propose a reward structure which would likely induce the desired behavior

Q. 5.2 *Étant donné une Q function définie, comment peut-on choisir la prochaine action ; cad comment à partir de Q obtient on la policy ? Voyez-vous un problème à cette approche ?* Given a Q function, how one could use it to choose the next action to perform ? Can you see a problem to this approach ?