Examen M2 IAAA option AR

Durée 2h - documents autorisés - connexion internet interdite

31 janvier 2019

Exercice 1 (Bandits avec stratégie ϵ-greedy, 7 à 9 points)**.** On considère un problème de bandits à K bras. On note $q^*(k)$ la valeur des bras pour $k \in \{1, \dots, K\}$. On définit la stratégie greedy ainsi :

- initialisation de la valeur estimée de chaque bras : $\forall k \in \{1, \dots, K\}, \widehat{q}(k) \leftarrow 0$
- initialisation du nombre de tirages de chaque bras : $\forall k \in \{1, \dots, K\}, n(k) \leftarrow 0$.
- choix du bras : $\arg\max_k \widehat{q}(k)$
- mise à jour du bras k avec la récompense r :
  $n(k) \leftarrow n(k) + 1$ puis $\widehat{q}(k) \leftarrow \widehat{q}(k) + \frac{1}{n(k)} (r - \widehat{q}(k))$ (ce qui revient à faire la moyenne des récompenses obtenues pour le bras k).

La stratégie ϵ-greedy est similaire à la stratégie greedy, sauf pour le choix du bras : on garde le choix greedy avec probabilité $1 - \epsilon$ et on choisit un bras au hasard avec probabilité ϵ.

1. Quel est l'intérêt de choisir ϵ petit ?
2. Quel est l'inconvénient de choisir ϵ petit ?
   Première variante : on modifie l'initialisation des $\widehat{q}(k)$ en affectant une valeur q_0 à la place de la valeur 0 : $\forall k, \widehat{q}(k) \leftarrow q_0$. On prendra une valeur élevée $q_0 \gg \max_k q^*(k)$ en supposant qu'elle est supérieure à toutes les récompenses reçues.
3. Que se passe-t-il à la première itération de la nouvelle stratégie greedy ou ϵ-greedy ? À la deuxième ? Dans les K premières itérations ? En quoi est-ce une façon de contourner l'inconvénient évoqué ci-dessus lorsque ϵ est petit ?
4. Montrez que dès que le bras k a été choisi une fois, sa valeur estimée $\widehat{q}(k)$ est la même avec la nouvelle stratégie qu'avec la stratégie initiale $q_0 = 0$.
   Deuxième variante : pour que q_0 ait une influence au-delà du premier tirage de chaque bras, on modifie également l'initialisation $n(k) \leftarrow n_0$ de $n(k)$ en lui affectant une valeur $n_0 = 1$ au lieu de $n_0 = 0$. Cela revient à considérer qu'un tirage initial a déjà eu lieu pour chaque bras avec une récompense q_0, et à prendre celle-ci en compte dans la moyenne des récompenses pour l'estimation $\widehat{q}(k)$.
5. Une simulation de chaque stratégie est représentée sur la figure 1.
   (a) Quelle est la limite en $+\infty$ de chaque courbe ?
   (b) Pourquoi la courbe de la stratégie initiale ($q_0 = 0, n_0 = 0$) est-elle au-dessus des deux autres dans les 100 premières itérations ?
   (c) Pourquoi la courbe de la première variante ($q_0 = 10, n_0 = 0$) passe-t-elle au-dessus ensuite ?
   (d) Pourquoi la courbe de la deuxième variante ($q_0 = 10, n_0 = 1$) reste-t-elle alors sous la courbe de la première variante puis passe au-dessus ?

Exercice 2 (Un peu de modélisation, 2 à 3 points)**.** Décrivez un sçénario applicatif pouvant être modélisé par un processus de décision de Markov (MDP), et définissez aussi précisément

que possible le MDP. Plus la réponse sera originale par rapport aux exemples déjà vus ensemble, plus elle rapportera de points.

Exercice 3 (Un peu de calcul : équations de Bellman, environ 3 points)**.** On rappelle la définition des fonctions d'évaluation pour les MDP :

$$\forall s \in \mathcal{S}, v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

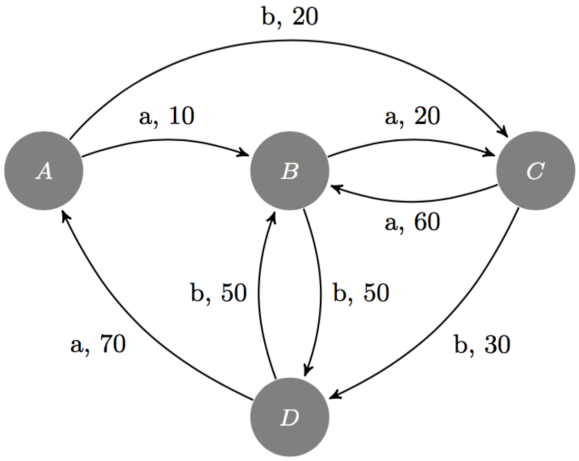
avec $G_t = \sum_{k=0}^{+\infty} \gamma^k R_{t+k+1} = R_{t+1} + \gamma G_{t+1}$ et les équations de Bellman

$$\forall s \in \mathcal{S}, v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_\pi(s') \right]$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, q_\pi(s,a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \sum_{a'} q_\pi(s', a') \pi(a'|s') \right]$$

Avec un raisonnement de probabilités similaire à celui vu en cours (loi de bayes, marginalisation), établir une équation récursive qui exprime $q_\pi(s,a)$ en fonction des $v_\pi(s')$ pour $s' \in \mathcal{S}$.

Exercice 4 (Q-learning, adapté de JB Alonso, Universitat Politècnica de Catalunya)**.** Cet exercice a pour objectif de mettre en œuvre l'algorithme de Q-learning vu en cours, sur le graphe suivant. Chaque nœud y représente un état et chaque arc une action, ainsi que la récompense associée au couple état/action. Par exemple, effectuer l'action a à partir du nœud A amène au nœud B, et la récompense associée à ce couple état/action (A, a) a la valeur 10.



Nous allons procéder à la mise à jour de la fonction $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ où :
— $\mathcal{S} = \{A, B, C, D\}$ est l'ensemble des états ;
— $\mathcal{A} = \{a, b\}$ l'ensemble des actions possibles ;
en utilisant l'algorithme de Q-learning.

Pour rappel, l'équation de mise à jour du Q-learning est :

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha \left(r(s_t, a_t) + \gamma \max_{c \in \mathcal{A}} q(s_{t+1}, c) - q(s_t, a_t) \right) \tag{1}$$

où il est entendu que la séquence d'états/actions visitée est $(s_1, a_1), \ldots, (s_T, a_T)$ pour un épisode de taille T. Notez que, par rapport à l'équation vue en cours, la récompense dépend ici d'un couple état/action et non pas seulement de l'état — ce qui ne change rien aux propriétés de l'algorithme.

Répondre aux questions suivantes :

1. A quoi sert l'algorithme Q-learning ? Pourquoi est-il qualifié d'algorithme *off-policy* ?
2. Faire un tableau (4 lignes, 2 colonnes) qui répertorie les récompenses associées à chaque couple état/action.
3. En supposant que : i) la fonction q (qui est simplement un tableau ici) est initialisée à 0, ii) que $\alpha = 1$ et $\gamma = 0.9$ et que iii) une séquence de visite observée est initiée au nœud A et poursuivie par la séquence d'actions $\{a, a, b, a, b, a\}$
   (a) donnez la séquence des états visités ;
   (b) montrez, en donnant le détail des calculs, que la fonction q obtenue en utilisant (1) peut se résumer dans le tableau suivant :

	a	b
A	10	47
B	20	0
C	78	30
D	79	0

4. Comment gérer la situation où l'espace produit état×action est gigantesque, voire infini, de sorte que toutes les valeurs de q ne peuvent être stockées dans un tableau ?

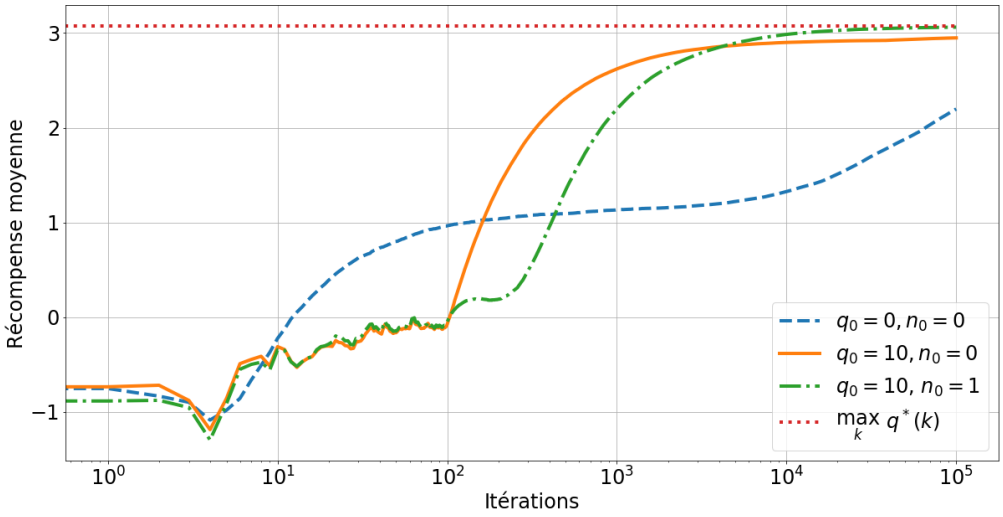


FIGURE 1 – Récompense moyenne pour la stratégie ϵ-greedy avec diverses initialisations, dans le cas de bandits à $K = 100$ bras, $\epsilon = 0.001$, avec des distributions gaussiennes (résultats moyennés sur 10 runs).