Aix Marseille Université — 2019-2020
Master IAAA — 2ème année

UE apprentissage par renforcement

Séance 2 - Thompson Sampling

On considère ici des bandits dont les K bras sont des variables de Bernoulli :

- les récompenses immédiates r valent 0 ou 1 ;
- pour le bras k, la probabilité d'obtenir une récompense 1 vaut $p_k \in [0,1]$ et la probabilité d'obtenir une récompense 0 vaut $1 - p_k$; p_k est le paramètre de la loi de Bernoulli et correspond aussi à l'espérance du gain $q^*(k) = p_k$ du bras k.

L'algorithme 1 décrit d'échantillonnage de Thompson pour le cas de variables de Bernoulli.

Algorithme 1 Thompson Sampling pour le cas Bernoulli

Entrées: K (nombre de bras)
  $\boldsymbol{\alpha} \leftarrow \mathbf{1}_K$ {Initialisation tableau de K uns.}
  $\boldsymbol{\beta} \leftarrow \mathbf{1}_K$ {Initialisation tableau de K uns.}
  répéter
    pour chaque bras k **faire**
      tirer $\theta_k \sim \text{Beta}(\alpha_k, \beta_k)$
    fin pour
    $a \leftarrow \text{argmax}_k\, \theta_k$ {choisir l'action}
    $r \leftarrow r(a)$ {obtenir la récompense}
    $\boldsymbol{\alpha}[a] \leftarrow \boldsymbol{\alpha}[a] + r$ {Mise à jour des paramètres}
    $\boldsymbol{\beta}[a] \leftarrow \boldsymbol{\beta}[a] + 1 - r$ {Mise à jour des paramètres}
  jusqu'à la fin

Exercice 1 (*Mise en œuvre*)

Dans la continuité du TP précédent sur les bandits,

1. créez une nouvelle classe `ThompsonSamplingAlgorithm` pour mettre en œuvre l'échantillonnage de Thompson ; la fonction `np.random.beta` permet de faire les tirages selon la distribution Beta ;
2. utilisez l'environnement `BernoulliMultiArmedBandits` pour visualiser le gain moyen obtenu en fonction des itérations ; affichez également le gain optimal ; vous pouvez prendre par exemple 5 bras avec des paramètres tirés au hasard et 10^4 itérations
3. affichez également, sur une autre figure, les valeurs des paramètres α et β pour chaque bras en fonction des itérations, avec une couleur pour chaque bras et un style de ligne différent (continu, pointillé) pour chaque paramètre.

Exercice 2 (*Comparaison*)

En utilisant l'environnement `BernoulliMultiArmedBandits`, comparez les stratégies Thompson Sampling, UCB ($c = 1$) et ϵ-greedy en moyennant les résultats sur plusieurs réalisations. Vous pouvez faire cet exercice dans le notebook du TP1 à la suite du travail déjà réalisé.