Apprentissage par renforcement - Problèmes de bandits

Travaux Pratiques

L'apprentissage par renforcement en général, et les problèmes de bandits en particuliers, seront présentés dans un premier temps en cours (transparents disponibles sur le site pédagogique). Dans ce TP, on mettra en oeuvre les algorithmes présentés, afin d'expérimenter leurs qualités respectives et de jouer un peu avec leurs paramètres.

Il est extrêmement difficile de tester les algorithmes de bandits sur données réelles, puisqu'on ne peut pas s'appuyer sur un échantillon pré-défini (il se constuit au fur et à mesure en fonction des actions choisies). On se contentera donc de simuler des données : on considérera un scénario à 5 bras dont les distributions sont des lois de Bernoulli de paramètres respectifs $0.5, 0.7, 0.6, 0.4, 0.2$ (scénario 1) et $0.03, 0.01, 0.01, 0.01, 0.01$ (scénario 2).

- Borne de Chernoff / Hoeffding -

Soit $X_1, \dots, X_n$ des variables aléatoires iid à valeur dans l'intervalle $[0, 1]$. On note μ leur espérance commune, et $\bar{X}_n = (X_1 + \cdots + X_n)/n$ leur moyenne empirique. On définit la fonction $\phi : \mathbb{R} \to \mathbb{R}$ par la relation

$$\phi(\lambda) = \ln \mathbb{E}\left[\exp(\lambda X_1)\right] \ .$$

1. Calculer $\phi(\lambda)$ si $X_1 \sim \mathcal{B}(\mu)$.
2. Montrer que pour tout $\lambda \in \mathbb{R}$,

$$\phi(\lambda) \leq \ln\left(1 - \mu + \mu \exp(\lambda)\right) \ .$$

3. Déduire de l'inégalité de Markov, que pour tout $x \geq \mu$ et pour tout $\lambda \in \mathbb{R}$:

$$P\left(\bar{X}_n \geq x\right) \leq \exp\left(-n\left(\lambda x - \phi(\lambda)\right)\right) ; .$$

4. En déduire que

$$P\left(\bar{X}_n \geq x\right) \leq \exp(-n \,\mathrm{kl}(x, \mu)) \ ,$$

où $\mathrm{kl}(p, q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1-p}{1-q}$ pour $p, q \in]0, 1[$. Que dire des cas particuliers $p \in \{0, 1\}$ ou $q \in \{0, 1\}$?

5. Montrer l'inégalité dite de Pinsker :

$$\mathrm{kl}(p, q) \geq 2(p - q)^2 \ .$$

6. Pour quelles valeurs de μ et x l'inégalité dite de Hoeffding

$$P\left(\bar{X}_n \geq x\right) \leq \exp\left(-2n(x-\mu)^2\right)$$

est-elle bonne ? pour quelles valeurs est-elle sous-optimale ?

- Algorithme ε-greedy -

7. Programmer l'algorithme ε-greedy (avec ε constant).
8. Pour un horizon $n = 1000$, tracer en fonction de ε le regret moyen (estimé par Monte-Carlo avec 100 répétitions) pour chacun des deux scénarios.
9. Modifier votre fonction pour qu'elle puisse utiliser des valeurs de ε qui varient avec t. Essayer plusieurs façons de choisir ε_t : est-il possible de faire aussi bien qu'à la question 8 ?

- Algorithmes Upper-Confidence Bound -

10. Programmer l'algorithme UCB, puis l'algorithme kl-UCB.
11. Dans cette question, on ne considère que le scénario 1. Tracer une estimation du regret moyen sur un horizon $n = 1000$ en fonction de la constante multiplicative règlant la largeur de l'intervalle de confiance.
12. Dans chaque scénario, tracer une estimation (par Monte-Carlo) de la fonction qui à n associe le regret cumulé jusqu'à l'instant n, pour n compris entre 1 et 1000.
13. Comparer les algorithmes optimistes et les algorithmes ε-greedy

Références

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2) :235–256, 2002.

[2] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[3] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *23rd Conf. Learning Theory (COLT)*, Budapest, Hungary, 2011.

[4] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301) :13–30, 1963.