Autour des problèmes de bandit combinatoires

Contexte. Les modèles de bandit à plusieurs à plusieurs bras [2] décrivent une situation où un agent choisit de manière répétée une action parmi un ensemble d'actions disponibles, et observe une récompense aléatoire liée à l'action choisie. Pour découvrir la meilleure action tout en maximisant ses récompenses, il doit réaliser un compromis entre exploration de son environement et exploitation de ses connaissances actuelles. Si leur nom fait référence à un casino où il s'agirait de découvrir la machine à sous (ou "bras") la plus performante, la popularité de ces modèles vient plutôt d'applications dans le domaine de l'optimisation de contenu web (systèmes de recommandation), l'allocation adaptative de spectre dans les radio cognitives, ou encore l'exploration optimiste d'un arbre minimax pour la résolution de jeux.

Certaines de ces applications ont motivé l'introduction de modèles de bandit plus complexes, où une structure de graphe est associée à l'ensemble des actions. Ainsi dans les modèles de bandits combinatoires, plutôt que de choisir une action (un nœud), l'agent est amené à choisir un groupe d'actions, et à observer une récompenses qui dépend du groupe choisi [5, 4]. Dans les modèles à information latente (*bandits with side information*), le graphe encode la structure de récompense : lorsqu'un noeud est choisi, la récompense de tous ses voisins et aussi observée, ce qui accélère l'apprentissage du meilleur nœud [3, 1].

Objectifs du stage. L'objectif du stage est d'acquérir une bonne compréhension de l'un de ces modèles de bandit structurés, en cherchant à répondre à la question suivante: comment la performance des algorithmes dépent-elle de la structure du graphe? Les algorithmes de l'état de l'art seront implentés, et l'une des pistes de recherche suivante pourra être explorée:

- dans les bandits combinatoires, on investigera l'utilisation de l'échantillonnage de Thompson [6], une stratégie dont l'optimalité a été prouvée dans les modèles simples
- dans les modèles à information latente, [1] met en évidence plusieurs régimes pour les performances des algorithmes en fonction de la structure du graphe, dans un contexte assez général où les récompenses peuvent être choisies par un adversaire. On cherchera à voir si cette propriété subsiste lorsque les récompenses sont générées aléatoirement.

L'implémentation des différents algorithmes pourra être effectuée en Python ou Julia.
(`http://julialang.org/`)

Informations pratiques. Le stage se déroulera au laboratoire CRIStAL à Lille, au sein de l'équipe Sequel (`https://sequel.lille.inria.fr/`), sous la direction d'Emilie Kaufmann (contact: emilie.kaufmann@inria.fr).

References

[1] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graph: Beyond bandits. In *Conference On Learning Theory (COLT)*, 2015.

[2] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Fondations and Trends in Machine Learning*, 5(1):1–122, 2012.

[3] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Levaraging side observations in stochastic bandits. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[4] N. Cesa-Bianchi and G. Lugosi. Combinatorial Bandits. *Journal of Computer and System Sciences*, 78:1404–1422, 2012.

[5] R. Combes, S. Talebi, A. Proutière, and M. Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[6] E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd conference on Algorithmic Learning Theory*, 2012.