CRIStAL
Centre de Recherche en Informatique,
Signal et Automatique de Lille

Structured Bandits: Should Optimism strike back?

Emilie Kaufmann (CNRS & CRIStAL, Inria Lille (SequeL team))

A Multi-Armed Bandit (MAB) model is a simple framework that can describe several sequential resource allocation tasks, that range from the design of adaptive clinical trials to that of recommenders systems or cognitive radios. In this model, an agent is interacting with arms, that are (unknown) probability distribution, by sequentially querying samples from these distributions, in order to optimize some criterion, e.g. collect as much rewards as possible if the samples collected can be interpreted as rewards (like clicks or saved patients in the aforementioned examples) or identifying quickly the arm with largest mean (the so-called "best arm identification" problem), see [Kaufmann and Garivier, 2017] for a survey of the two problems.

In the simplest setup in which arms are distributions parameterized by their means with no constraints on these possible means, good strategies for maximizing rewards in MABs are well understood, in that there is an asymptotic lower bound on the number of samples needed from all sub-optimal arms, and algorithms that attain this lower bound. Most of them are based on the celebrated *optimism-in-face-of-uncertainty principle*. Given an Upper Confidence Bound (UCB) on the value of each mean, they select at each round the arm with highest UCB, which amounts to choosing the arm that is best in an "optimistic" model in which all arms have the highest value that is statistically plausible. Among such strategies, the kl-UCB algorithm proposed by [Cappé et al., 2013] has been proved to be asymptotically optimal.

Many more structured variants of this bandit problem have been studied, in which there is extra knowledge on the possible values for the means that has to be taken into account (for example some linear or unimodal structure). Lower bounds on the algorithms' performance can also be obtained in this setting, and although they are less explicit they provide an "oracle" minimal number of samples from the sub-optimal arms. Building on the ability to compute this oracle, [Combes et al., 2017] and [Lattimore and Szepesvári, 2017] have recently proposed asymptotically optimal algorithms for some instances of structured bandits. Both algorithms rely on two ingredients: forced exploration to make sure that the means of all arms are well-estimated and a "tracking" mechanism to ensure that the number of draws of the sub-optimal arms follow the oracle (evaluated on the empirical means). Similar ideas (forced exploration + tracking) have also recently been used for the best arm identification problem [Garivier and Kaufmann, 2016]. However, the need for forced exploration usually makes the analysis of such algorithms very asymptotic, an can furthermore be very armful in practice.

The goal of this internship is to understand when forced exploration is actually needed in the oracle-based algorithms described above and whether the optimism principle (or at least the ability to construct confidence intervals on the arms' mean) can be combined with the access to the oracle. The ultimate goal is to propose a new, simple algorithm combining these tools that is efficient in practice and for which a finite-time analysis is possible.

Practical informations. The candidate should have a strong background in mathematics (especially some knowledge about statistics), and should also be able to perform simple numerical experiments. The internship will take place at Inria Lille, in the SequeL team, under the supervision of Emilie Kaufmann. It will be part of the ANR project BADASS.

Contact: `emilie.kaufmann@univ-lille1.fr`

CRIStAL - UMR 9189
Cité Scientifique - Bât. M3 extension
59655 Villeneuve d'Ascq Cedex - France
www.cristal.univ-lille.fr

References

[Cappé et al., 2013] Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541.

[Combes et al., 2017] Combes, R., Magureanu, S., and Proutière, A. (2017). Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*.

[Garivier and Kaufmann, 2016] Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference On Learning Theory (to appear)*.

[Kaufmann and Garivier, 2017] Kaufmann, E. and Garivier, A. (2017). Learning the distribution with largest mean: two bandit frameworks. *ESAIM: Proceedings and Surveys*.

[Lattimore and Szepesvári, 2017] Lattimore, T. and Szepesvári, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *AISTATS*.