Sujet de Thèse

par **Odalric-ambrym Maillard**

dans l'équipe Scool, Inria Lille – Nord Europe:
"Real-life Challenges for Reinforcement Learning Theory"

Résumé : Rendre applicable les techniques d'apprentissage par renforcement dans des applications réelles (telles que la recommandation de pratiques agroécologiques en agriculture) nécessite de surmonter plusieurs verrous scientifiques. Etendre les stratégies existantes à des distributions non-paramétriques, prendre en compte l'aversion au risque, être robuste à des changements de dynamique ou à des données corrompues le tout dans le cadre d'un système dynamique partiellement connu sont autant de défis qu'il est devenu envisageable de lever suites aux avancées récentes de l'état-de-l'art. Ils nécessitent une attention particulière tant sur le plan mathématique qu'informatique pour être effectivement levés.

Mots-clés : Apprentissage par Renforcement, Bandits, Processus Decisionnels de Markov, Non-paramétrique, Risque

Directeur de thèse : Odalric-Ambrym Maillard (Inria Scool, CRIStAL)

This PhD proposal is part of the Chaire IA "AppRenf" project (R-PILOTE-19-004-APPRENF), funded by the Fondation I-SITE ULNE within the project PILOTE from cluster HumAIn@Lille. This is a PhD in Machine Learning, more specifically in Reinforcement Learning.

Odalric-Ambrym Maillard is a researcher at Inria. He has worked for over a decade on advancing the theoretical foundations of reinforcement learning, using a combination of tools from statistics, optimization and control, in order to build more efficient algorithms able to better estimate uncertainty, exploit structures, or adapt to some non-stationary context. He was the PI of the ANR-JCJC project BADASS (BAnDits Against non-Stationarity and Structure) until Oct. 2021. He is also leading the Inria Action Exploratoire SR4SG (Sequential Recommendation for Sustainable Gardening) and is involved in a series of other projects, from more applied to more theoretical ones all related to the grand-challenge of reinforcement learning that is to make it applicable in real-life situations, see `http://odalricambrymmaillard.neowordpress.fr` for further details.

AppRenf is a research project in Artificial Intelligence more precisely in the domain of machine learning known as *Reinforcement Learning*. AppRenf is co-headed by Philippe Preux, Professor in Computer Science at the Université de Lille and Odalric-Ambrym Maillard, Researcher at Inria. Both belong to the research group named Scool at Inria-Lille, and UMR CNRS CRIStAL. Considering the current state of Artificial Intelligence, AppRenf proposes to develop a set of fundamental research questions from their theoretical investigation, down to their concrete application. The PhD proposal fits into this project and is naturally aligned with the objectives of the program `AI_PhD@Lille`.

Motivated by a real-life use-case application to sharing of good practice in an agricultural context, recent developments of the fields of sequential decision making under uncertainty has focused on developing strategies able to handle first *non-parametric* families of distributions, and then various *risk-averse* criteria, in a provably optimal way, see [2], [3], [1]. While most of the existing work in bandit has focused on risk-neutral criterion and distributions having a simple parametric form (Bernolli, Gaussian, Exponential, etc.), moving wy from these assumptions is justified by the fact that the reward distributions in such a real-life context usually do not display any nice parametric form, and that a farmer may not only seek best recommendations in expectation but rather recommendations incorporating her own aversion to risk. These promising developments have been explored in the simplified multi-armed bandit framework (removing the dynamical aspect of the decisions). In order to foster the development of more elaborate application, we would like to explore the adaptation of such techniques to the full-blown Markov-decision process (MDP) framework. Such an extension is non trivial, first due to the fact MDP exhibit a rich structure that is not easy to exploit from a bandit perspective. Fir this reason, we may first explore the intermediate models of structured and contextual multi-armed bandits from this non-parametric, risk-averse perspective, building on some alternative approaches for the parametric risk-neutral case [7], [6]. In particular, it is currently unknown how to extend the promising sub-sampling bandit strategies to exploit a given structure in a non-parametric context. Risk-aversion is a complementary challenge due to the estimation that must be done differently and the non-linearity of the considered operators, when compared to a simple expectation. Furthermore, solving an MDP is considerably more theoretically and computationally demanding than solving a bandit (see [4, 5]). Hence care should be given to the built solution, that should be both theoretically appealing, and computationally reasonable in practice. Now, beyond these specific topics, the PhD proposal is motivated by addressing a series of challenges bridging the nice but over-simplified MDP model to more realistic, but more intricate real-life systems. We expect the PhD candidate to explore such complementary challenges as robustness to possibly changing dynamics, outliers in the data, and trade-offs between fully-sequential vs fully-batch learning that arise when attempting to apply sequential decision making strategies in real-life application. Besides addressing the fundamental challenges, the PhD candidate will in particular be able to work with and get hands on a set of agriculture simulators studied or developed in the team, due to collaboration with CIRAD and CGIAR that have been initiated in relation with complementary research projects of the Scool team.

The PhD requires a solid background in statistics, probability, Markov chains, concentration of measure and confidence regions, a good knowledge of multi-armed bandit, active sampling and Markov decision processes methods, strong analytical skills, as well as the capacity to code, conduct relevant numerical experiments and prove theoretical guarantees of the considered strategies. The successful candidate is expected to learn quickly, have a solid mathematical background, and have good to excellent programming skills. After getting familiar with the relevant literature, and through numerous discussions with the PhD advisor, the candidate will investigate such questions and is expected to publish its outcome in the top-tier conferences and journals of the field.

References

[1] Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric-Ambrym Maillard. Optimal Thompson Sampling strategies for support-aware CVaR bandits. In *ICML 2021 - International Conference on Machine Learning*, Virtual Conference, United States, July 2021.

[2] Dorian Baudry, Emilie Kaufmann, and Odalric-Ambrym Maillard. Sub-sampling for Efficient Non-Parametric Bandit Exploration. In *NeurIPS 2020*, Vancouver, Canada, December 2020.

[3] Dorian Baudry, Patrick Saux, and Odalric-Ambrym Maillard. From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits. In *Neurips 2021*, Sydney, Australia, December 2021.

[4] Hippolyte Bourel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Tightening Exploration in Upper Confidence Reinforcement Learning. In *International Conference on Machine Learning*, Vienna, Austria, July 2020.

[5] Sayak Ray Chowdhury, Aditya Gopalan, and Odalric-Ambrym Maillard. Reinforcement Learning in Parametric MDPs with Exponential Families. In PLMR, editor, *International Conference on Artificial Intelligence and Statistics*, volume 130, San diego, United States, 2021.

[6] Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard. Optimal Strategies for Graph-Structured Bandits. working paper or preprint, July 2020.

[7] Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard. Indexed Minimum Empirical Divergence for Unimodal Bandits. In *NeurIPS 2021 - International Conference on Neural Information Processing Systems*, Virtual-only Conference, United States, December 2021.