Research internship proposal
Centre Inria de l'Université de Lille
Team project Scool – Spring 2024

"Bandits exploiting contextual structure in the non-parametric, Huber-outlier setup "

Keywords: Multi-armed bandits, Sequential statistics, Societal challenge.
Supervision: The intern will be advised by Odalric-Ambrym Maillard and Timothée Mathieu from Inria team-project Scool.
Place: This internship will be primarily held at the research center Inria Lille – Nord Europe, 40 avenue Halley, 59650 Villeneuve d'Ascq, France, in the Inria team-project Scool (previously known as SequeL).

Context Multi-armed bandit theory has witnessed tremendous progress over the last decade, yielding algorithms achieving strong learning guarantees (regret minimization, best-arm identification) in increasingly challenging context involving sequential decision making in uncertain environment. In particular, provably optimal strategies such as KL-UCB [Cappé et al., 2013], TS [Korda et al., 2013] or IMED [Honda and Takemura, 2015] have been shown to achieve strong optimality in parametric context, while other strategies such as NPTS from [Riou and Honda, 2020] or SDA [Baudry et al., 2020, Baudry et al., 2021b] obtained non-parametric optimality, enabling application of multi-armed bandit to a large range of applications when reward distributions are not easily modeled with classical families. Recently, a complementary bandit model considering Huber-outlier distributions (mixture between an parametric distribution of interest and an arbitrary one) has been studied, offering an interesting complementary perspective compared to non-parametric assumptions, see [Basu et al., 2022] and [Agrawal et al., 2023]. Further, progress has been made to handle risk-averse objective rather than simply obtaining guarantees in expectation [Baudry et al., 2021a], or when reward feedback is available in infrequent batches rather than immediately [Gautron et al., 2022]. This setup is related to group-sequential clinical trials and hypothesis testing, but with the bandit, that is adaptive sampling approach. Last, the extension of classical bandits to structured bandits [Magureanu, 2016, Saber, 2022], including e.g. contextual bandits enable novel algorithms in recommender systems that can learn to provide decisions (actions) personalized to a given context in an efficient way [Abbasi-Yadkori et al., 2011], see also [Kirschner and Krause, 2019], and even to Markov decision processes [Pesquerel and Maillard, 2022]. Despite this, applying multi-armed bandit in a real-life application comes with many additional challenges.

In this internship, we consider two related questions around the notion of model misspecification and context. In linear contextual multi-armed bandits, strategies have been introduced to obtain cumulative regret bounds exploiting structure on the context. One natural extension of the work on non-parametric bandits is to design and study algorithms for non-parametric linear bandits, combining non-parametric assumption on the distributions and exploitation of the contextual structure. A variant of this problem is to study misspecified assumption, say under an Huber-corruption of the reward distributions. A second related to topic is to study a different different objective than regret minimization, and related to best-arm identification, which can be seen related to hypothesis testing. Here, we would like to consider contextual hypothesis testing from a non-parametric or Huber-outlier perspective. To summarize, we want to study either Regret minimization of Best-arm identification objectives, in the contextual setup assuming either Huber-outlier or Non-parametric distributions.

CENTRE DE RECHERCHE
LILLE - NORD EUROPE
Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq - France
Tél. : +33 (0)3 59 57 78 00
Fax : +33 (0)3 59 57 78 51

www.inria.fr

Proposal This internship focuses on basic research, theory of multi-armeb bandits, and for this rasonf requires a candidate with a strong background in mathematical statistics. The objective of this internship is to study the bibliography related to these related problems, especially the proof techniques and design of existing algorithms. Then, the candidate will develop novel algorithm, combining and extending ideas from linear contextual, non-parametric, and misspecified bandit literature, as well as regret minimization and best-arm identification literature, together with an analysis controlling the performance of the provided algorithm. The study will be completed with illustrative numerical experiments comparing the proposed method to alternative work and baselines. Depending on the output of the study, a publication in an international conference or journal will be considered.

All the work done during the internship will be made reproducible and open-source, following an open-science and open knowledge philosophy.

Host institution and supervision The student will be hosted at Centre Inria de l'Université de Lille, in the Scool team. Scool (Sequential COntinual and Online Learning) is an Inria team-project. It was created on November 1st, 2020 as the follow-up of the team SequeL. In a nutshell, the research topic of Scool is the study of the sequential decision making problem under uncertainty. Most of our activities are related to either bandit problems, or reinforcement learning problems. Through collaborations, we are working on their application in various fields including health, agriculture and ecology, sustainable development. More information, please visit `https://team.inria.fr/scool/projects`.

Odalric-Ambrym Maillard is a permanent 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 provide decision making in uncertain environments. He was PI of several projects, including ANR-JCJC project BADASS (BAnDits Against non-Stationarity and Structure), Inria Action Exploratoire SR4SG (Sequential Recommendation for Sustainable Gardening) and Inria-Japan Associate team RELIANT (Reliable Bandit strategies). His goal is to push forward key fundamental and applied questions related to the grand-challenge of making reinforcement learning applicable in real-life societal applications.

References

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

[Agrawal et al., 2023] Agrawal, S., Mathieu, T., Basu, D., and Maillard, O.-A. (2023). Crimed: Lower and upper bounds on regret for bandits with unbounded stochastic corruption.

[Basu et al., 2022] Basu, D., Maillard, O.-A., and Mathieu, T. (2022). Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithm. *arXiv preprint arXiv:2203.03186*.

[Baudry et al., 2021a] Baudry, D., Gautron, R., Kaufmann, E., and Maillard, O. (2021a). Optimal thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, pages 716–726. PMLR.

[Baudry et al., 2020] Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020). Sub-sampling for efficient non-parametric bandit exploration. *Advances in Neural Information Processing Systems*, 33:5468–5478.

[Baudry et al., 2021b] Baudry, D., Saux, P., and Maillard, O.-A. (2021b). From optimality to robustness: Adaptive re-sampling strategies in stochastic bandits. *Advances in Neural Information Processing Systems*, 34:14029–14041.

[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. *The Annals of Statistics*, pages 1516–1541.

[Gautron et al., 2022] Gautron, R., Baudry, D., Adam, M., Falconnier, G. N., and Corbeels, M. (2022). Towards an efficient and risk aware strategy for guiding farmers in identifying best crop management. *arXiv preprint arXiv:2210.04537*.

[Honda and Takemura, 2015] Honda, J. and Takemura, A. (2015). Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756.

[Kirschner and Krause, 2019] Kirschner, J. and Krause, A. (2019). Stochastic bandits with context distributions. *Advances in Neural Information Processing Systems*, 32.

[Korda et al., 2013] Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26.

[Magureanu, 2016] Magureanu, S. (2016). *Structured Stochastic Bandits*. PhD thesis, KTH Royal Institute of Technology.

[Pesquerel and Maillard, 2022] Pesquerel, F. and Maillard, O.-A. (2022). Imed-rl: Regret optimal learning of ergodic markov decision processes. *Advances in Neural Information Processing Systems*, 35:26363–26374.

[Riou and Honda, 2020] Riou, C. and Honda, J. (2020). Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pages 777–826. PMLR.

[Saber, 2022] Saber, H. (2022). *Structure Adaptation in Bandit Theory*. PhD thesis, Université de Lille.

CENTRE DE RECHERCHE
LILLE - NORD EUROPE
Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq - France
Tél. : +33 (0)3 59 57 78 00
Fax : +33 (0)3 59 57 78 51

www.inria.fr