# Offer #2024-07416

# PhD Position F/M Bandits-Inspired Reinforcement Learning to Explore Large Stochastic Environments

**Contract type :** Fixed-term contract

**Level of qualifications required :** Graduate degree or equivalent

**Fonction :** PhD Position

## About the research centre or Inria department

The Inria University of Lille centre, created in 2008, employs 360 people including 305 scientists in 15 research teams. Recognised for its strong involvement in the socio-economic development of the Hauts-De-France region, the Inria University of Lille centre pursues a close relationship with large companies and SMEs. By promoting synergies between researchers and industrialists, Inria participates in the transfer of skills and expertise in digital technologies and provides access to the best European and international research for the benefit of innovation and companies, particularly in the region.For more than 10 years, the Inria University of Lille centre has been located at the heart of Lille's university and scientific ecosystem, as well as at the heart of Frenchtech, with a technology showroom based on Avenue de Bretagne in Lille, on the EuraTechnologies site of economic excellence dedicated to information and communication technologies (ICT).

## Context

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 grandchallenge of reinforcement learning that is to make it applicable in real-life situations.

The student will be hosted at Inria, 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. For more information, please visit https://team.inria.fr/scool/projects/

## Assignment

Having emerged in the last decade, Deep Reinforcement Learning (DRL), resulting from the combination of Deep Learning (DL) and Reinforcement Learning (RL) techniques has predominantly been explored in environments characterized by determinism or low stochasticity, such as games and robotic tasks. However, the applicability of RL extends far beyond these controlled settings to encompass realworld
scenarios with substantial stochastic elements. This includes for instance, autonomous driving, where stochasticity can be inherently induced by many factors like the unpredictable behavior of other vehicles, pedestrians, and varying weather conditions. Another typical example is that of supporting decision making in agrosystems subject to stochastic weather or pests conditions and inherent variability
due to latent variables, or in healthcare where similar treatments may affect individuals differently. In these complex and dynamic real-world environments, deep approaches make sense in approximating complex functions in a non-parametric manner, serving as effective feature extractors, especially when agents contend with vast amounts of information. However, traditional deep RL approaches [14, 2] face considerable challenges, primarily stemming from their inherent sample inefficiency. Existing promising algorithms often adopt a model-based approach, explicitly attempting to learn the underlying distribution of dynamics, then sample from it and derive an optimal policy using traditional RL methods. However, such methods appear to be very brittle in the context of stochastic environments.

In stark contrast, in simpler scenarios, without dynamics but characterized by high stochasticity, bandit algorithms efficiently manage the exploration-exploitation tradeoff. Over the past few years, significant progress has been made in the field, in addressing more structured and complex bandit problems

[15, 17, 18, 20, 7, 8], including facing non-parametric [4, 6] or corrupted [1] reward distributions. Rcvently, bandit strategies have been shown to achieve significant progress in handling Markov Decision Processes, see [16, 19], though restricted to small, discrete environments, paving the path to more challenging environments.

This Ph.D. aims to draw inspiration from these promising strategies to develop bandit-inspired deep reinforcement
learning approaches, specifically designed to effectively navigate stochastic environments with substantial information complexity. Indeed, as a special and simplified subset of Markov Decision Processes, bandits algorithm offer a profound theoretical understanding. The notion of regret comparing the disparity between the potential optimal cumulated reward and the actual collected reward by an agent following a specific learning strategy motivates creating sample-efficient strategies. Drawing inspiration from bandits, we aspire to enhance the theoretical guarantees of Deep RL algorithms, traditionally
viewed as black boxes, to become more reliable including in stochastic environments.

The developed method aims to be more adapted to address real-world problems, a research topic that is gaining increasing popularity, as evidenced by events like the Real Life Workshop at NeurIPS 2022. In real-world scenarios, addressing challenges such as high randomness, risk, information abundance, reward sparsity, and sample efficiency is crucial. The resulting algorithms will be instrumental for the Inria team-project Scool and its collaborators, who are actively engaged in real-world applications across diverse fields, with a primary focus on health, agriculture, ecology, and sustainable development.


[1] Shubhada Agrawal, Timoth´ee Mathieu, Debabrota Basu, and Odalric-Ambrym Maillard. Crimed: Lower and upper bounds on regret for bandits with unbounded stochastic corruption, 2023.
[2] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement
learning: A brief survey. IEEE Signal Processing Magazine, 34(6):26–38, 2017.
[3] 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.
[4] Dorian Baudry, Emilie Kaufmann, and Odalric-Ambrym Maillard. Sub-sampling for Efficient Non-Parametric Bandit Exploration. In NeurIPS 2020, Vancouver, Canada, December 2020.
[5] Dorian Baudry, Patrick Saux, and Odalric-Ambrym Maillard. From optimality to robustness: Adaptive re-sampling strategies in stochastic bandits. Advances in Neural Information Processing Systems, 34:14029–
14041, 2021.
[6] 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.
[7] Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. Advances in neural information processing systems, 28, 2015.
[8] Thibaut Cuvelier, Richard Combes, and Eric Gourdin. Statistically efficient, polynomial-time algorithms for
combinatorial semi-bandits. Proceedings of the ACM on Measurement and Analysis of Computing Systems,
5(1):1–31, 2021.
[9] Awi Federgruen and Paul J Schweitzer. Successive approximation methods for solving nested functional
equations in markov decision problems. Mathematics of operations research, 9(3):319–344, 1984.
[10] Romain Gautron, Dorian Baudry, Myriam Adam, Gatien N Falconnier, and Marc Corbeels. Towards an efficient and risk aware strategy for guiding farmers in identifying best crop management. arXiv preprint arXiv:2210.04537, 2022.
[11] Yu-Chi Larry Ho and Xi-Ren Cao. Perturbation analysis of discrete event dynamic systems, volume 145.
Springer Science & Business Media, 2012.
[12] Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support
models. In COLT, pages 67–79. Citeseer, 2010.
[13] Johannes Kirschner and Andreas Krause. Stochastic bandits with context distributions. Advances in Neural
Information Processing Systems, 32, 2019.
[14] Maxim Lapan. Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks,
value iteration, policy gradients, TRPO, AlphaGo Zero and more. Packt Publishing Ltd, 2018.
[15] Stefan Magureanu. Structured Stochastic Bandits. PhD thesis, KTH Royal Institute of Technology, 2016.
[16] Fabien Pesquerel and Odalric-Ambrym Maillard. Imed-rl: Regret optimal learning of ergodic markov decision
processes. Advances in Neural Information Processing Systems, 35:26363–26374, 2022.
[17] Hassan Saber, Pierre M´enard, and Odalric-Ambrym Maillard. Optimal Strategies for Graph-Structured Bandits.
working paper or preprint, July 2020.
[18] Hassan Saber, Pierre M´enard, 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.

[19] Hassan Saber, Fabien Pesquerel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Logarithmic
regret in communicating mdps: Leveraging known dynamics with bandits. In Asian Conference on Machine
Learning, 2023.
[20] Hassan Saber, L´eo Saci, Odalric-Ambrym Maillard, and Audrey Durand. Routine Bandits: Minimizing Regret on Recurring Problems. In ECML-PKDD 2021, Bilbao, Spain, September 2021.

## Main activities

More particularly we propose to define three research axis, each of them mixing both theoretical and practical advances:

2.1 Axis 1 - Adapting RL to stochastic environments:
In this axis, we aim to leverage insights from bandits, particularly in minimizing regret within highly stochastic problems, to propose extensions or variants that can advance the capabilities of RL in large stochastic environments. We intend to strengthen the development of RL methods for stochastic environments,
extending [16] from ergodic to generic MDPs and leveraging results from perturbation
analysis [11] and higher-order optimal control [9] to better handle the bias function hence build more effective
model-based RL methods. We will continue the efforts in injecting promising bandit strategies,
such as [12], [5] both in tabular and approximate algorithms, to develop hybrid deep-bandit strategies for stochastic environments. We will also investigate how to extend the literature on RL in deterministic environments, such as complex games and deterministic control tasks, to stochastic environments incorporating
popular Bayesian approaches. This research will contribute to the growing body of literature
on RL in stochastic environments and help practitioners develop methods for RL in these challenging environments. To test these new methods, we aim to develop new benchmark/environments that emphasis the manage of stochasticity.

2.2 Axis 2 - Structured interactions and auxiliary information:
To efficiently handle the rich diversity of interactions, we propose to investigate the problem of leveraging
auxiliary information in RL for systems with many state variables. Our focus will first be on
identifying the most influential variables, uncovering underlying structures, and developing methods to forecast the evolution of these variables inspired from related questions in bandits [13]. We will then investigate the structure that may be known to the domain expert, such as factored, causality or combinatorial
structures, to reduce the complexity of the learning task. More generally, this research will
contribute to the growing body of literature on leveraging auxiliary information in RL and help practitioners
improve the performance of their RL algorithms by leveraging the available expert knowledge.

2.3 Axis 3 - Context-goal exploration-exploitation trade-off
In this axis, We aim to extend the literature on contextual bandits and contextual MDPs by exploring the concept of context-goal, seen as context variable shaping the objectives of the agent. Our goal is to design an agent that is both general and efficient across diverse objectives, adapting to the changing nature of the context to achieve optimal performance. Specifically, we propose to investigate the problem of jointly addressing diverse objectives, such as different risk-aversion levels, across multiple environments under the linear-context structural assumption. To address this problem, we will revisit the exploration-exploitation trade-off to understand how to explore in one environment to gather information
that is relevant to another practitioner's objective, in a collectively optimal way. This can be
done extending lower-bound techniques to this problem. Besides, we will also contribute to the literature
on risk-aversion in MDPs, extending our work [3, 10] from bandits to MDPs and to the contextual case. Overall, this task will fill a gap in the literature by providing a framework for efficiently learning from many environments with diverse objectives tailored to the preferences of the decision-maker, and will contribute to the growing body of research on risk-aversion in RL.

During the initial six months, the student will delve into the literature to identify potential bandit and deep RL algorithms for integration and extension. Simultaneously, benchmarking environments will be developed. The subsequent three semesters will be dedicated to the development of new algorithms, with a focus on achieving tangible results on real-world problems by the end of the fifth semester. The sixth semester will be dedicated to thesis writing and job applications.

## Skills

Technical skills and level required : Master in AI and RL related domains.

Languages : English (scientific writing)

Relational skills : Ability to interact with supervisor and other team members, present work in seminar or conference.

## Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

## Remuneration

1st and 2nd year : 2100 € (grossly salary)

3st year: 2190 € (grossly salaray)

## General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
- **Town/city :** Villeneuve d'Ascq
- **Inria Center :** Centre Inria de l'Université de Lille
- **Starting date :** 2024-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-06-30

## Contacts

- **Inria Team :** SCOOL
- **PhD Supervisor :**
  Maillard Odalric-ambrym / Odalric.Maillard@inria.fr

## About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

## The keys to success

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.

**Warning** : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

## Instruction to apply

**Defence Security :**
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

**Recruitment Policy :**
As part of its diversity policy, all Inria positions are accessible to people with disabilities.